Performance Is Not Delegation: The Human–AI Division of Labor*

Tsung-Hsien Li

This version: October 2026

Abstract

Conditional on using generative AI on a task, does the human or the AI execute it? I examine this division of labor using task-level aggregates from two weekly samples of roughly one million Claude.ai conversations each, in November 2025 and February 2026. For each O*NET task, I measure *delegation* as the share of conversations in which execution is handed to the AI, and *realized performance* as the share of conversations that Anthropic's classifier judges successful. I ask whether execution is handed to the AI more often in tasks with higher performance. Across tasks, higher delegation is associated with lower performance in both waves, with correlations ranging from -0.455 to -0.640. This relationship persists after accounting for the purpose of use, the type of work, and Claude's concentration in programming tasks. Between the two waves, average performance in the same tasks rose by roughly four percentage points, while delegation remained essentially unchanged. Among tasks with similar performance, delegation is lower where users are more often judged able to complete the task without AI, and delegation also varies systematically across types of work. These findings suggest that performance and delegation are distinct margins of AI use.

Keywords: Generative AI, Delegation, AI adoption, Human–AI collaboration, Division of labor, Task allocation

JEL Classification: D23, D91, J24, O33

*Assistant Research Fellow, Institute of Economics, Academia Sinica (email: thli@econ.sinica.edu.tw, website: https://tsunghsien1124.github.io/). I gratefully acknowledge financial support from the Academia Sinica Grand Challenge Program Seed Grant (AS-GCS-115-H02).

1 Introduction

Much of the economics of generative AI (GenAI) asks two questions: what AI can do, and where it is used. Measures of occupational exposure to AI and benchmark evaluations of AI models address the first question (Webb, 2020; Felten et al., 2021; Eloundou et al., 2024; Kwa et al., 2025; Patwardhan et al., 2026), and surveys of workers address the second (Bick et al., 2026a; Humlum and Vestergaard, 2025). Conditional on using AI on a task, a third question is whether the human or the AI executes it. The related literature examines in theory when AI should execute tasks rather than assist humans (Athey et al., 2020; Fügener et al., 2026), investigates how people decide what to hand over to AI in experiments and field studies (Fügener et al., 2022; Randazzo et al., 2025), and documents patterns of human–AI interaction in actual use (Handa et al., 2025; Chatterji et al., 2025).

Consider a programmer who asks an AI to fix a bug. The programmer has two broad options. First, the programmer can let the AI write the fix, at most running it and reporting any error message back to the AI. Second, the programmer can write the fix and ask the AI to check it, refine a draft fix together with the AI over several rounds, or ask the AI to explain what went wrong. In the first case, the AI executes the task, and I call this type of interaction *delegation*. In the second case, the human keeps control of producing the work and uses the AI to iterate, to learn, or to check, and I call this type of interaction *augmentation*. A task can be handled in either way, and different users handle the same task differently. The choice between these two modes of interaction determines who executes the task, a margin close to the allocation of tasks between workers and machines that task-based models of technology emphasize (Autor et al., 2003; Acemoglu and Autor, 2011; Autor, 2015; Acemoglu and Restrepo, 2018, 2019).

A natural question is whether users delegate execution more often in tasks with higher realized performance. This paper examines this relationship across tasks, how it changed over time, and which other characteristics are associated with delegation.

To this end, I use task-level aggregates from the Anthropic Economic Index (AEI). In each of two weekly waves, one in November 2025 and one in February 2026, Anthropic samples about one million conversations on its consumer product, Claude.ai. Anthropic's automated classifiers assign each conversation a task statement from the Occupational Information Network (O*NET),[1] one of five interaction patterns or none if no pattern can be identified, a judgment of whether the AI successfully completed the task, and other labels, including whether the conversation is for work and whether

[1] O*NET, sponsored by the U.S. Department of Labor, describes each occupation by a list of task statements, short descriptions of the work done in that occupation (National Center for O*NET Development, 2015).

the user could have completed the task without AI. Anthropic does not release individual conversations, but it reports the number of conversations and aggregate statistics for each O*NET task, including the share of conversations assigned to each interaction pattern and the share of conversations judged successful. The O*NET task is therefore my unit of observation. For each wave, the main sample consists of tasks that have at least 100 conversations. The main samples contain 947 O*NET tasks in November and 1,023 in February, with 883 appearing in both.

Of the five interaction patterns, I classify directive use and feedback loops as delegation patterns, in which the AI executes the task either directly or by revising its work in response to user feedback. I classify task iteration, learning, and validation as augmentation patterns, in which the user retains control while refining outputs, seeking explanations, or checking their own work.[2] For each task, I measure *delegation* as the share of conversations with an identified interaction pattern that fall into the two delegation patterns. Delegation therefore measures how often execution is handed to the AI across conversations, not what fraction of the work within a conversation is delegated. I measure *realized performance* as the share of a task's conversations that Anthropic's success classifier judges successful. This measure captures success in actual human–AI conversations rather than the AI model's standalone capability. I refer to it simply as performance below.

The paper has three main findings, which together suggest that performance and delegation are distinct margins of AI use. First, tasks with higher delegation tend to have lower performance. With equal weights, the correlation between delegation and performance across tasks is -0.455 in November and -0.558 in February. When tasks are weighted by their conversation counts, the correlations are -0.586 and -0.640, respectively. Software tasks such as debugging and writing programs tend to have high delegation and low performance, whereas tutoring, coursework help, advising, and copy editing tend to have low delegation and high performance. The negative relationship persists among work-focused tasks and after accounting for the composition of each task's conversations across work, coursework, and personal use.[3] It also holds within O*NET generalized work activities, after excluding computer-occupation tasks, and after reweighting Claude's intermediate work activities toward the survey-measured distribution of GenAI use among U.S. workers (Bick et al., 2026b).

[2] Anthropic groups the patterns in the same way but uses the term automation for the delegation patterns. I use the term delegation because the patterns record who executes the task in a conversation, not whether AI replaces workers.

[3] I define work-focused tasks as those with at least 50 percent of conversations classified as work and repeat the analysis using a 75 percent threshold. I also check whether delegation and performance remain negatively correlated after adjusting both measures for the shares of conversations used for work and for coursework. Because the shares of work, coursework, and personal use sum to about one, this adjusts for the whole composition of uses.

Second, among the 883 tasks observed in both waves, average performance increased by roughly four percentage points, while average delegation remained essentially unchanged. The performance gain was similarly large among tasks whose delegation changed by no more than two percentage points and remained evident after controlling for observed changes in conversation composition or excluding computer-occupation tasks. Task-fixed-effects estimates further show an upward shift in the fitted delegation–performance profiles of 4.4 and 3.7 percentage points at a delegation level of 0.5, with equal and conversation weights, respectively. The shift was approximately parallel across delegation levels, with no clear evidence of nonlinear curvature.

An illustrative framework offers a plausible interpretation of why these two margins need not move together. In this framework, users choose between the delegation mode and the augmentation mode based on their relative payoffs. If the probability of success for a given task rises by similar amounts under both modes, overall performance can increase with little change in the relative payoffs of the two modes and hence little change in how often users delegate. If this common rise is also similar across tasks, the delegation–performance relationship shifts upward approximately in parallel. The two waves span the release of two new models in Anthropic's most capable class, Claude Opus 4.5 and Claude Opus 4.6, each of which scored higher than its predecessor on most of the benchmarks.[4]

Third, among tasks with similar performance, delegation varies systematically with three characteristics: (1) *human-only ability*, defined as the share of conversations in which the user is judged able to complete the task without AI, (2) *human-only time*, defined as the estimated time a competent professional would need, without AI, to complete the work that the AI performed in a conversation, and (3) the type of work. I find that delegation is lower for tasks with higher human-only ability and longer human-only time. Even at similar performance, computing and information-technology work exhibits higher delegation than communication, advice, and information-gathering work. The human-only time result may seem counterintuitive, since work that would take a human longer offers greater potential time savings from delegation. One plausible explanation is that longer work is also costlier to delegate in full. The user may need to supply context, guide intermediate steps, and verify the output. AI success also tends to decline with the time a task takes humans (Kwa et al., 2025; Appel et al., 2026). As a result, step-by-step augmentation may be more attractive than delegating all of the work at once.

The paper relates to three strands of literature. The first studies delegation, both in organizations and to AI. In organizations, giving authority to a better-informed agent

[4]The benchmark comparisons are in their release announcements (Anthropic, 2025, 2026d). The data do not break down the rise in performance by its source.

with different objectives costs the principal control (Aghion and Tirole, 1997; Dessein, 2002). When the agent is an AI, recent theories study whether it works autonomously and the cost of verifying its output (Ide and Talamàs, 2025; Huang et al., 2026). Research on AI systems also treats an AI's capability and the amount of work handed to it as separate questions. Whether a case should be handed to an algorithm depends on more than the algorithm's accuracy (Raghu et al., 2019), and an AI system's autonomy is a dimension separate from its capability (Morris et al., 2024). In experiments, people give up earnings to keep the right to make a decision themselves (Bartling et al., 2014), stop relying on an algorithm after seeing it err, even when it outperforms a human forecaster (Dietvorst et al., 2015), and delegate cases to an AI poorly because they misjudge their own ability (Fügener et al., 2022). These studies examine delegation in theory or in controlled settings. By contrast, I measure it in actual AI use, using task-level aggregates of about two million Claude.ai conversations across O*NET tasks. I find that, among tasks with similar performance, delegation is lower for tasks with higher human-only ability and longer human-only time.

The second strand measures where and how GenAI is used. Handa et al. (2025) map Claude.ai conversations to O*NET tasks and introduce the interaction patterns behind my delegation measure, and Anthropic's later reports add labels for success, human-only ability, and human-only time and track the aggregate shares of the interaction patterns over time (Appel et al., 2026; Massenkoff et al., 2026). Anthropic also finds that users who have used Claude longer rely less on directive use and succeed more often (Massenkoff et al., 2026). These reports do not relate delegation to success across tasks or report how success changed between the two waves. Other studies use Anthropic's labels to examine usage, workflows, and employment (Lynch, 2026; Demirer et al., 2026; Brynjolfsson et al., 2025a), or measure how much autonomy users grant AI agents, finding evidence that autonomy in practice depends on user behavior as well as model capabilities (McCain et al., 2026). Tomlinson et al. (2025) separate users' goals from the AI's actions in Bing Copilot conversations, and Shao et al. (2025) compare workers' desire for automation of O*NET tasks with experts' ratings of AI capability. Bick et al. (2026b) measure GenAI *adoption* among U.S. workers across O*NET detailed work activities. Chatterji et al. (2025) study *request types*, distinguishing requests for information or advice, which they call Asking, from requests for task execution or output production, which they call Doing. Neither measure captures how work is divided between humans and AI. In particular, adoption records whether AI is used, whereas a Doing label is based on a single sampled message and does not rule out repeated human guidance and revision. I instead use Anthropic's labels to relate delegation, which is conditional on AI use, to performance across approximately 1,000 O*NET tasks, and compare delegation with the adoption rates of Bick et al. (2026b).

The third strand of literature studies human–AI production and the division of work. GenAI improves productivity in writing and customer support (Noy and Zhang, 2023; Brynjolfsson et al., 2025b), but can reduce accuracy when tasks exceed its capabilities (Dell'Acqua et al., 2026), and consultants who hand entire workflows to it are less accurate than those who use it selectively (Randazzo et al., 2025). Agarwal et al. (2023) find that AI assistance does not improve radiologists' average diagnostic accuracy because radiologists do not properly incorporate AI predictions. Fügener et al. (2026) show that the optimal allocation of tasks among humans, AI, and human–AI teams depends on their relative accuracy. These studies examine the effects of AI assistance or how work should be allocated in specific settings. I instead relate how users actually divide task execution with AI to measured performance across a large number of O*NET tasks. I find that delegation is more common in tasks with lower measured performance and remained nearly unchanged between the two waves, which span the release of new models, even as performance rose across the same tasks.

The rest of the paper is organized as follows. Section 2 introduces the data and measures. Section 3 presents an illustrative framework for the choice between delegation and augmentation. Section 4 examines the relationship between delegation and performance across tasks, while Section 5 compares the same tasks across two waves. Section 6 studies other correlates of delegation, including human-only ability, human-only time, and the type of work, and compares delegation with GenAI adoption. Section 7 concludes and discusses directions for future research.

2 Data and Measurement

2.1 Unit of observation and samples

My data come from the AEI, which describes the use of Claude, Anthropic's AI assistant. I use two weekly waves of the AEI, one in November 2025 and one in February 2026. In each week, Anthropic draws a sample of about one million Claude.ai conversations, and its automated classifiers assign labels to each conversation: (1) an O*NET task statement, (2) one of five interaction patterns or none, (3) a yes-or-no success judgment, and (4) other labels (Tamkin et al., 2024; Handa et al., 2025; Appel et al., 2026; Massenkoff et al., 2026). Anthropic does not release the conversations or any conversation-level records, but it publishes task-level aggregates of these labels. For each O*NET task, these aggregates give the number of conversations, the count for each interaction pattern, the count for each success label, and statistics for other la-

bels, such as the share of conversations that are for work and the share in which the user could have completed the task alone. My unit of observation is thus the O*NET task. Every measure from the AEI is a share or a mean computed from a task's conversations. Appendix A gives the full details of the data, and its Table A.1 defines the variables and specifies their sources.

Anthropic assigns each conversation to an O*NET task statement, drawn from approximately 19,500 statements in O*NET version 20.1 covering 974 occupations. These statements describe the activity discussed in a conversation, not necessarily the user's occupation. For example, a conversation in which a student seeks help with a programming assignment may be classified under a software-development task, even though the student is not a software developer. I therefore study how task execution is divided between humans and AI across conversations within each task, regardless of the user's occupation.

The November sample consists of 999,875 conversations from November 13 to 20, 2025, and the February sample of 1,000,000 conversations from February 5 to 12, 2026 (Appel et al., 2026; Massenkoff et al., 2026; Anthropic, 2026a).[5] The *main sample* in each week consists of the tasks that have at least 100 conversations and satisfy the data-availability criteria in Appendix A. It contains 947 tasks in November, holding 84.7 percent of the week's conversations, and 1,023 tasks in February, holding 83.9 percent. For the cross-wave analysis, I use the 883 tasks that belong to both main samples and call them the *balanced panel*. The panel follows tasks, not users or conversations, since each week's measures come from a different sample of conversations. The 883 tasks hold 99.1 and 97.9 percent of the main samples' conversations and 83.9 and 82.1 percent of the two weeks' conversations.

Unless noted otherwise, I report each statistic under two weightings. With *equal weights*, each task counts once, and a mean is an average over tasks. With *conversation weights*, each task receives weight proportional to its number of conversations n_j in the week. For performance, the resulting mean equals the share of all main-sample conversations judged successful. The two weightings can differ because a few programming tasks account for a large share of conversations. The ten largest tasks, mostly software work, hold 24 percent of the November week's conversations and 19 percent of the February week's. In the balanced panel, each task receives the same conversation weight in both waves, defined as its average conversation count,

[5] Both samples cover the consumer product Claude.ai. I do not use the other releases of the AEI. The three releases before the two I use, published in February, March, and September 2025, do not report task success, and the June 2026 release changes the sampling design and the task classifier and does not report success by task (Anthropic, 2026a). I also exclude data from Anthropic's first-party application programming interface (API), which are classified at the level of individual request–response exchanges rather than complete conversations. Consequently, their interaction-pattern and success measures are not directly comparable to those used here.

$\bar{n}_j = (n_{j,\text{Nov}} + n_{j,\text{Feb}})/2$. This holds the relative weights of tasks fixed, so changes in their conversation shares do not mechanically contribute to the cross-wave comparison.

2.2 Delegation

Anthropic's interaction-pattern classifier assigns each conversation to one of five patterns, or to none if it cannot identify a pattern.[6] The five patterns are directive use, feedback loops, task iteration, learning, and validation. Table 1 describes them, based on Anthropic's definitions. Following Handa et al. (2025), I divide the five interaction patterns into two groups. In the first group, I classify directive use and feedback loops as *delegation patterns*, in which the AI executes the task either directly or by revising its work in response to user feedback. In directive use, the human hands the task over with little further interaction. In a feedback loop, the human keeps the role of tester, relaying error messages or test results back to the AI. In the second group, I classify task iteration, learning, and validation as *augmentation patterns*, in which the user retains control while refining outputs, seeking explanations, or checking their own work. Anthropic calls the first group automation, and I use the term delegation because the measure records who executes the task in a conversation with the AI, not whether the AI replaces workers.

Let $\pi_{j,p}$ be the share of pattern p among task j's conversations with an identified pattern.[7] I measure *delegation* D_j, defined as the share of a task's conversations with an identified interaction pattern that fall into the two delegation patterns,

$$D_j = \pi_{j,\text{directive use}} + \pi_{j,\text{feedback loop}}. \tag{1}$$

Delegation D_j therefore measures how often execution is handed to the AI across a task's conversations, not what fraction of the work within a conversation is delegated.

[6] The label none indicates that the classifier could not identify an interaction pattern, rather than a distinct mode of use. These conversations account for approximately 2 percent of main-sample conversations and are excluded from the denominator of delegation. Appendix Table A.2 reports the results when they are included in the denominator.

[7] For privacy, Anthropic withholds any task-by-label count below 15 conversations. When an interaction-pattern count is withheld, I compute the pattern shares from the published counts and leave the withheld conversations out. These are 1.4 percent of main-sample conversations in November and 1.6 percent in February. Appendix A gives the details.

Table 1: The Five Interaction Patterns

Pattern	What the human does	What the AI does
Delegation patterns		
Directive use	Hands over the whole task with minimal further interaction	Executes the task
Feedback loop	Relays feedback from outside the conversation, such as error messages or test results, and so remains involved as a tester	Executes the task and revises it in response to the feedback
Augmentation patterns		
Task iteration	Refines the AI's output	Produces the output jointly with the human
Learning	Asks for explanations to understand the subject rather than to get the task done	Explains or instructs
Validation	Produces the work and asks the AI to check it	Checks the human's work

Notes: The middle column paraphrases the definitions of the AEI's interaction-pattern classifier, whose labels Anthropic calls collaboration patterns, and the last column summarizes what they imply for the AI (Handa et al., 2025; Anthropic, 2026e).

2.3 Performance

Anthropic's success classifier judges whether the AI successfully completed the task that the user requested, based on the conversation as a whole, and answers yes or no. From the published counts, I construct each task's *performance* S_j, defined as the share of the task's conversations that the success classifier judges successful.

Three features of this measure matter for reading the results. First, performance is measured in actual use and is not a standardized measure of capability. Benchmarks score what a model does alone on fixed problems (Kwa et al., 2025; Patwardhan et al., 2026), whereas performance is a joint outcome of a user and the AI, judged by a model, and it averages over the people who bring the task to Claude, what they ask for, and how they divide the work with the AI.[8]

Second, the same classifier model, Claude Sonnet 4.5, labeled both waves, and the success classifier used the same prompt in both, so a cross-wave change in the level of performance does not reflect a change in the classifier model or in the success prompt.[9] Because the same model also assigned the interaction patterns of both samples, a cross-wave change in delegation does not reflect a change in the classifier model either.

[8] Even benchmarks can misstate the value of a model in deployment (Gans, 2026). Performance also depends on the difficulty of what was asked, on how the user behaved, on the trajectory of the interaction, on the model that served the conversation, and on the classifier that judged it. Anthropic compared the success classifier with a human researcher's ratings of a small set of Claude.ai conversations and found it "directionally accurate" (Appel et al., 2026).

[9] Anthropic's published documentation names the classifier model only for the November sample, in its January 2026 report (Appel et al., 2026). Anthropic confirmed the February classifier model and the unchanged success prompt. Appendix Table A.3 documents the two samples.

Third, Anthropic reports success and interaction patterns separately for each task, without reporting how many conversations fall into each combination of the two. Consequently, I cannot compare the success rates of delegation and augmentation within a task. Instead, I examine how delegation and performance are related across tasks and how they change for the same tasks between the two waves.

2.4 Other task characteristics

Besides performance, I relate delegation to two groups of task characteristics. The first group consists of four other classifier labels: (1) whether the user could have completed the task alone, (2) how long a professional would need, without AI, for the work that the AI performed in the conversation, (3) whether the user was multitasking, and (4) the purpose of use. These four labels come from the same conversations and the same family of classifiers as D_j and S_j. The second group comes from O*NET, independently of Anthropic's classifiers, and describes the type of work and whether O*NET lists the task for a computer occupation.

Human-only ability is defined as the share of conversations in which the user is judged able to complete the task without AI. The classifier counts the user as able even if completing the task alone would have taken more time. *Human-only time* is defined as the estimated time a competent professional would need, without AI, to complete the work that the AI performed in a conversation. For each conversation, the classifier estimates the hours that a professional with the necessary skills, context, and tools, but without AI, would need for that work.[10] Anthropic reports the average estimated time across each task's conversations. Unlike the other labels, human-only time is measured in hours rather than as a share. Because human-only time is right-skewed across tasks, I use its logarithm in the following analysis.

The *multitasking share* is defined as the share of conversations in which the user is judged to work on more than one task over the course of the conversation, although each conversation is assigned a single O*NET task statement. The use-case classifier records the purpose of use, labeling each conversation as work, coursework, or personal use according to whether it mainly serves the user's job, the user's coursework, or neither. The *work share*, *coursework share*, and *personal share* are defined as the shares of a task's conversations with each label.

For the second group, I use O*NET to characterize the type of work associated with each task. O*NET lists task statements for each occupation and, where available, links them to *detailed work activities* (DWAs). Each DWA belongs to an *intermediate work activ-*

[10] This work may cover only part of the O*NET task to which the conversation is assigned.

Table 2: O*NET Work-Activity Levels

Level	Number in O*NET 20.1	Coverage, Nov. / Feb.	Linkage
Task	19,530	947 / 1,023	One or more DWAs, or none
DWA	2,070	496 / 535	One IWA
IWA	332	188 / 191	One GWA
GWA	41	35 / 35	One subcategory
Subcategory	9	9 / 9	One category
Category	4	4 / 4	–

Notes: Coverage counts the main-sample tasks and the DWAs, IWAs, GWAs, subcategories, and categories to which they link. Of the O*NET 20.1 tasks, 1,159 link to no DWA, as do 54 main-sample tasks in November and 55 in February.

ity (IWA), and each IWA belongs to a *generalized work activity* (GWA). A task statement, my unit of observation, describes one specific task, whereas a GWA is a broad type of work that groups many task statements across occupations. For example, the GWA Getting Information includes the librarian task "Search standard reference materials, including online sources and the Internet, to answer patrons' reference questions" and the computer systems analyst task "Read manuals, periodicals, and technical reports to learn how to develop programs that meet staff and user requirements." The GWAs are further grouped into nine subcategories, which belong to four broader categories. Table 2 summarizes these levels and their linkages.

A task's *type of work* is defined as the GWAs to which it links. Because a task may link to multiple GWAs, I define its *GWA shares* $\omega_{j,g}$ as the shares of its DWA links associated with each GWA g.[11] Separately, I identify *computer-occupation tasks* as those listed under at least one occupation in minor group 15-11, Computer Occupations, of the 2010 Standard Occupational Classification (SOC).

2.5 Summary statistics

Table 3 summarizes the two main samples in Panel A and the balanced panel in Panel B. With conversation weights, which give more weight to the few large programming tasks, mean delegation is higher, mean human-only ability is lower, and mean human-only time is slightly longer than with equal weights. In Panel B, each task's conversation weight is its two-week average $\bar{n}_j$ in both weeks.

[11] For a task without a DWA link, all GWA shares are zero and an indicator $\omega_{j,0}$ equals one, so that the GWA shares and this indicator sum to one for every task.

Table 3: Samples and Task-Level Measures

	November 2025		February 2026	
Panel A: main samples				
Tasks, number	947		1,023	
Conversations, number	846,472		838,965	
Share of the week's conversations, percent	84.7		83.9	
Computer-occupation tasks, percent of tasks	15.3		15.0	
Computer-occupation tasks, percent of conversations	37.6		33.4	
	Equal	Conversation	Equal	Conversation
Performance S_j, mean	0.678	0.671	0.712	0.699
Delegation D_j, mean	0.414	0.474	0.418	0.460
Human-only ability, mean	0.918	0.869	0.910	0.869
Human-only time, hours, mean	3.084	3.192	3.061	3.174
Multitasking share, mean	0.091	0.095	0.094	0.101
Work share, mean	0.426	0.472	0.449	0.462
Coursework share, mean	0.187	0.203	0.118	0.132
Personal share, mean	0.387	0.325	0.433	0.405
Panel B: balanced panel of tasks in both main samples				
Tasks, number	883		883	
Conversations, number	838,690		821,305	
Share of the main sample's conversations, percent	99.1		97.9	
Share of the week's conversations, percent	83.9		82.1	
Computer-occupation tasks, percent of tasks	16.1		16.1	
Computer-occupation tasks, percent of conversations	37.9		33.9	
	Equal	Conversation	Equal	Conversation
Performance S_j, mean	0.678	0.669	0.719	0.703
Delegation D_j, mean	0.406	0.464	0.410	0.466
Human-only ability, mean	0.918	0.876	0.904	0.857
Human-only time, hours, mean	3.097	3.141	3.130	3.270
Multitasking share, mean	0.091	0.095	0.094	0.101
Work share, mean	0.430	0.469	0.443	0.466
Coursework share, mean	0.177	0.189	0.128	0.147
Personal share, mean	0.393	0.343	0.429	0.387

Notes: Main samples are the O*NET tasks that have at least 100 conversations in the weekly sample and satisfy the data-availability criteria in Appendix A. "Equal" gives every task the same weight. "Conversation" weights each task by its number of conversations n_j in Panel A and by its two-week average $\bar{n}_j = (n_{j,\text{Nov}} + n_{j,\text{Feb}})/2$ in Panel B.

3 An Illustrative Framework

To guide the interpretation of the empirical results in Sections 4 to 6, I present an illustrative framework in which users choose between delegation and augmentation based on their relative payoffs. Huang et al. (2026) model a related choice, in which a worker decides whether to delegate a task to AI and how much to verify its output. Consider a user i who uses AI for task j in a conversation. Under delegation, the AI executes the work, with the user at most providing feedback. Under augmentation, the user retains production or control of the work and uses AI to iterate, learn, or check. These modes

correspond conceptually to the classification of interaction patterns into delegation and augmentation in Table 1.

Let $s_j^D(M)$ and $s_j^A(M)$ denote the probabilities of successfully completing task j under delegation and augmentation, where the model environment M summarizes the available models and their quality. Let V_j be the value of success. Each mode $m \in \{D, A\}$ also provides benefits B_j^m and entails costs C_j^m beyond success, such as time saved through delegation, learning or control retained through augmentation, and the effort required to specify, monitor, verify, or correct the AI's output.[12] The terms $\varepsilon_{i,j}^m$ capture user- and conversation-specific variation in mode payoffs, which allows different choices within the same task. For tractability, the mode-specific success probabilities are common across a task's conversations, and users choose based on these probabilities, which the framework equates with the success they anticipate. The expected utilities are:

$$U_{i,j}^D = V_j\, s_j^D(M) + B_j^D - C_j^D + \varepsilon_{i,j}^D, \tag{2}$$

$$U_{i,j}^A = V_j\, s_j^A(M) + B_j^A - C_j^A + \varepsilon_{i,j}^A. \tag{3}$$

The user delegates when $U_{i,j}^D \geq U_{i,j}^A$ and chooses augmentation otherwise. Define the systematic payoff advantage of delegation as follows:

$$\Lambda_j(M) = V_j \left[s_j^D(M) - s_j^A(M) \right] + B_j - C_j. \tag{4}$$

Here, $B_j = B_j^D - B_j^A$ and $C_j = C_j^D - C_j^A$ denote the benefits and costs of delegation relative to augmentation. Let $\eta_{i,j} = \varepsilon_{i,j}^A - \varepsilon_{i,j}^D$ denote the idiosyncratic payoff advantage of augmentation. Since the user delegates when $\eta_{i,j} \leq \Lambda_j(M)$, the delegation probability for task j is:

$$D_j^* = \Pr\left(U_{i,j}^D \geq U_{i,j}^A\right) = F_j\big(\Lambda_j(M)\big)\,. \tag{5}$$

Here, F_j is the distribution of $\eta_{i,j}$ across the task's users and conversations. I impose no parametric form on F_j and assume only that it has a density f_j. The empirical counterpart of D_j^* is delegation D_j, defined in Section 2.2.

Let $g_j^* = s_j^D(M) - s_j^A(M)$ denote the mode-specific success gap. A task's expected performance is the average success probability across the two modes, weighted by their shares of conversations:

$$S_j^* = D_j^*\, s_j^D(M) + (1 - D_j^*)\, s_j^A(M) = s_j^A(M) + D_j^*\, g_j^*. \tag{6}$$

[12] Experimental evidence supports the value of such control. Bartling et al. (2014) document that people value decision rights for their own sake, and Dietvorst et al. (2018) show that people are more willing to use an imperfect forecasting algorithm when they can modify its forecasts, even slightly.

The empirical performance measure S_j in Section 2.3 is its counterpart. Delegation depends on the success gap g_j^*, the value of success, the relative benefits and costs, and the distribution of idiosyncratic payoffs, as shown in equations (4) and (5). Performance, by contrast, also depends on the levels of success under the two modes, as shown in equation (6). Even if $g_j^* = 0$ for every task, equation (6) gives $S_j^* = s_j^A(M)$, whereas equations (4) and (5) give $D_j^* = F_j(B_j - C_j)$. Performance then differs across tasks with the success probability $s_j^A(M)$, and delegation with the relative benefits and costs and with the distribution of idiosyncratic payoffs. Since the framework imposes no restriction on how these terms vary together across tasks, the relationship between delegation and performance can have either sign.

Equations (2) to (6) describe a single environment. To interpret Section 5, which compares the same tasks in two weeks, consider two environments $t \in \{0, 1\}$ with model environments M_0 and M_1, and write $s_{j,t}^m \equiv s_j^m(M_t)$ for the success probability of mode m in environment t. The success probabilities change only through M, but the other payoff terms may also differ between the environments, because of changes in AI technology, costs, users, requests, or other factors. I therefore index them as $B_{j,t}$, $C_{j,t}$, and $F_{j,t}$, and for simplicity I hold V_j fixed. In environment t, equations (4) to (6) become:

$$\Lambda_{j,t} = V_j \left[s_{j,t}^D - s_{j,t}^A \right] + B_{j,t} - C_{j,t}, \tag{7}$$

$$D_{j,t}^* = F_{j,t}(\Lambda_{j,t}), \tag{8}$$

$$S_{j,t}^* = D_{j,t}^* \, s_{j,t}^D + (1 - D_{j,t}^*) \, s_{j,t}^A. \tag{9}$$

Let $\delta_j^m = s_{j,1}^m - s_{j,0}^m$ be the change in the success probability of mode $m \in \{D, A\}$, and let $\Delta B_j = B_{j,1} - B_{j,0}$ and $\Delta C_j = C_{j,1} - C_{j,0}$. The change in the systematic payoff advantage of delegation is then:

$$\Lambda_{j,1} - \Lambda_{j,0} = V_j(\delta_j^D - \delta_j^A) + \Delta B_j - \Delta C_j. \tag{10}$$

Delegation therefore changes only if success changes differently under the two modes, if the relative benefits or costs change, or if the distribution of idiosyncratic payoffs changes from $F_{j,0}$ to $F_{j,1}$, so a rise in overall performance alone need not necessarily raise delegation.

Proposition 1 (Common Success Improvements and the Delegation Choice)**.** *Consider task j in two environments, $t = 0, 1$. Suppose that both modes experience the same feasible improvement in success probability, $\delta_j^D = \delta_j^A = \delta_j > 0$, with $s_{j,0}^m + \delta_j \leq 1$ for $m \in \{D, A\}$. Suppose further that V_j, $B_{j,t}$, $C_{j,t}$, and $F_{j,t}$ remain unchanged across environments. Then:*

(a) *The systematic payoff advantage of delegation and the delegation probability remain unchanged:* $\Lambda_{j,1} = \Lambda_{j,0}$ *and* $D^*_{j,1} = D^*_{j,0}$.

(b) *Expected performance increases by* δ_j*:* $S^*_{j,1} - S^*_{j,0} = \delta_j$.

(c) *If these conditions hold for every task in a fixed set* J*, with* $\delta_j = \delta$ *for all* $j \in J$*, the weighted least-squares line of performance on delegation across* J *shifts upward by* δ *without changing its slope, provided the task weights are fixed and positive, and delegation varies across the tasks in* J.

Proof. See Appendix B.1.

The conditions of Proposition 1 are sufficient, not necessary. Delegation may also remain unchanged when success gains differ between the two modes, for example, if changes in the relative benefits or costs offset the change in the success gap.

To examine two other determinants of delegation, write $\eta_{i,j} = a(h_{i,j}) + \xi_{i,j}$, where $h_{i,j}$ is the user's skill at task j without AI, $a(h)$ captures its contribution to the idiosyncratic payoff advantage of augmentation, and $\xi_{i,j}$ captures the remaining idiosyncratic factors. Let ℓ_j denote the time a human would need to complete all the work.

Proposition 2 (Human Alternatives and the Costs of Delegation)**.** *Consider task* j *with* $\eta_{i,j} = a(h_{i,j}) + \xi_{i,j}$ *and length of work* ℓ_j*. Then:*

(a) *If* $a(h)$ *is nondecreasing, an increase in users' human-only skills* $h_{i,j}$*, holding* Λ_j *and each* $\xi_{i,j}$ *fixed, weakly reduces delegation* D^*_j.

(b) *Suppose that* F_j *does not vary with* ℓ_j *and has a density that is continuous and positive at* Λ_j*, and that the payoff terms are differentiable in* ℓ_j*. Then:*

$$\frac{dD^*_j}{d\ell_j} = f_j(\Lambda_j)\left[\frac{d(V_j g^*_j)}{d\ell_j} + B'_j(\ell_j) - C'_j(\ell_j)\right]. \tag{11}$$

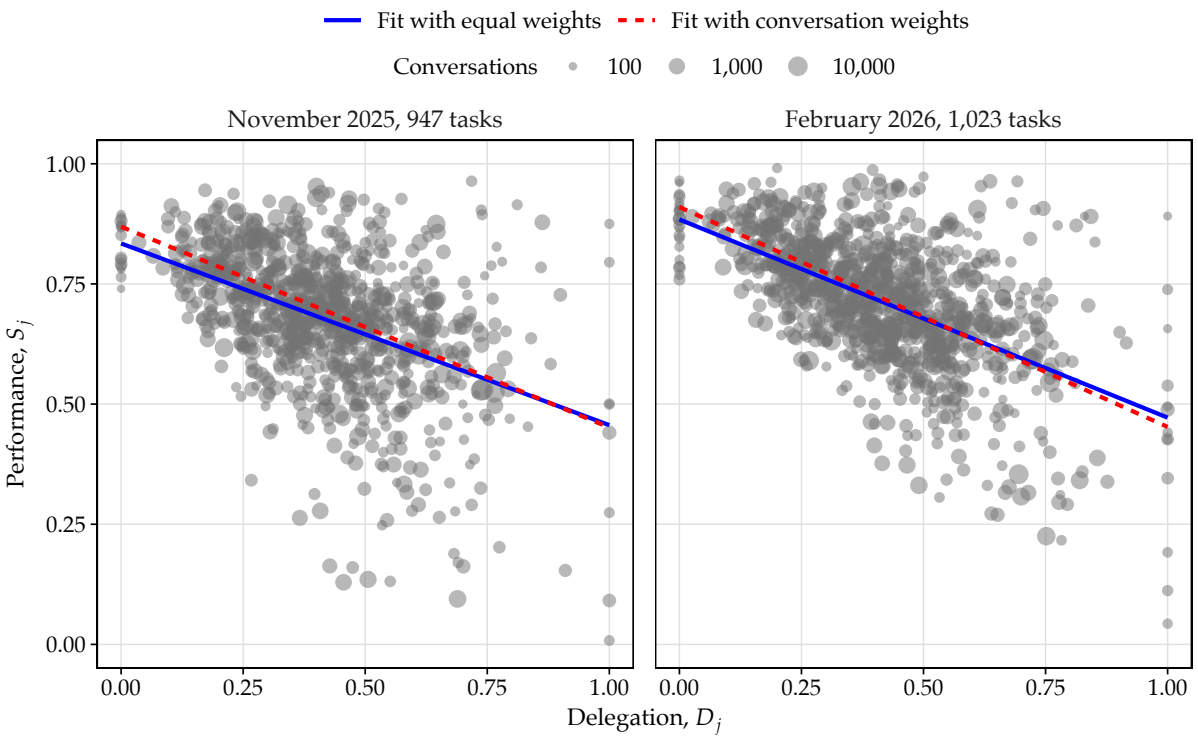
Delegation increases with work length if the expression in brackets is positive and decreases if it is negative.

Proof. See Appendix B.2.

Proposition 2(a) assumes that greater human-only skill favors augmentation over delegation. The prediction could reverse if skill instead facilitates delegation, for example, by improving the specification or verification of AI output. Part (b) has no general sign because longer work may increase both the benefits and costs of delegation.

Together, the propositions provide benchmarks for interpreting the empirical relationships among delegation, performance, and task characteristics. Appendix B pro-

Figure 1: Delegation and Performance across Tasks in the Two Weeks



Notes: Each point is an O*NET task in the main sample of the week, with at least 100 conversations. Delegation D_j is the share of the task's conversations with an identified interaction pattern that are directive use or feedback loops, and performance S_j is the share of its conversations that Anthropic's success classifier judges successful. Larger points are tasks with more conversations, on a logarithmic scale. The lines are least-squares fits of S_j on D_j.

vides their proofs and examines the identification of mode-specific success probabilities from the published data.

4 Delegation and Performance across Tasks

This section examines whether users delegate execution more often in tasks with higher realized performance. Instead, tasks with higher performance tend to have lower delegation. This negative relationship persists after accounting for differences in the purpose of use, broad types of work, and Claude's activity composition.

4.1 The baseline relationship

Figure 1 plots each task's performance against its delegation in the two weeks, on the same axes, together with the least-squares lines of performance on delegation with equal weights and with conversation weights. Table 4 reports the delegation–performance relationships, where Panel A gives the correlation and the least-squares

Table 4: Negative Delegation–Performance Relationship and Composition Checks

	November 2025		February 2026	
	Equal	Conversation	Equal	Conversation
Panel A: baseline				
Correlation between D_j and S_j	−0.455	−0.586	−0.558	−0.640
95% confidence interval	[−0.514, −0.397]	[−0.680, −0.451]	[−0.605, −0.508]	[−0.701, −0.556]
Slope of S_j on D_j	−0.378	−0.418	−0.412	−0.458
	(0.030)	(0.031)	(0.024)	(0.038)
Opposite-median share, percent	67.2	77.4	70.0	76.9
Number of tasks	947	947	1,023	1,023
Panel B: use composition, correlation between D_j and S_j				
Tasks with work share ≥ 0.50	−0.475	−0.642	−0.476	−0.559
Number of tasks	406	406	475	475
Tasks with work share ≥ 0.75	−0.548	−0.583	−0.436	−0.525
Number of tasks	169	169	232	232
Adjusted for work/coursework	−0.445	−0.559	−0.542	−0.634
Number of tasks	947	947	1,023	1,023
Panel C: work-type composition, tasks linked to a GWA				
Slope of S_j on D_j				
without GWA fixed effects	−0.398	−0.414	−0.415	−0.451
	(0.030)	(0.031)	(0.024)	(0.039)
with GWA fixed effects	−0.429	−0.449	−0.434	−0.497
	(0.036)	(0.046)	(0.028)	(0.042)
Within-GWA correlation	−0.487	−0.547	−0.566	−0.629
Within-GWA share of covariance, percent	77	47	77	57
Number of tasks	893	893	968	968

Notes: Brackets are 95 percent intervals from 2,000 bootstrap draws of tasks. Parentheses hold heteroskedasticity-robust (HC1) standard errors, clustered by task in Panel C. The opposite-median share is the share of tasks, with equal weights, or of conversations, with conversation weights, in the two groups with delegation at or above the week's unweighted median and performance below its median, or the reverse. The adjusted correlation in Panel B is between the residuals of D_j and S_j from regressions on the work and coursework shares, which leave out the personal share because the three shares sum to about one. Panel C uses one row per task and linked GWA, with each task's weight split across its GWAs in proportion to its GWA shares, and the within-GWA share of covariance is the share of the covariance between D_j and S_j that lies within GWAs. Every correlation in Panels B and C differs from zero at the 1 percent level by the same bootstrap, and Appendix Table C.1 reports the intervals.

slope of performance on delegation in each week, and Panels B and C give the composition checks on the purpose of use and the type of work.

All four lines slope downward: tasks in which users delegate execution more often have lower performance. In Panel A of Table 4, the correlation is −0.455 in November and −0.558 in February with equal weights, and −0.586 and −0.640 with conversation weights. The slope is −0.378 and −0.412 with equal weights and −0.418 and −0.458 with conversation weights, so under either weighting, tasks whose delegation is 10 percentage points higher have, on average, about 4 percentage points lower performance. Every correlation and every slope is negative and statistically different from

zero. The 95 percent intervals of the correlations lie between -0.701 and -0.397, and each slope is more than ten standard errors from zero.

Using each week's medians to classify tasks, I find in Panel A of Table 4 that 67 percent of tasks in November and 70 percent in February fall into either the high-delegation, low-performance group or the low-delegation, high-performance group, against about one half if delegation and performance were unrelated. These groups account for 77 percent of conversations in both waves. The high-delegation, low-performance group includes software debugging, programming, and network troubleshooting, while the low-delegation, high-performance group includes tutoring, coursework help, advising, and copy editing. Appendix Table C.2 provides the detailed breakdown. The negative relationship is robust to alternative delegation measures, sample thresholds, and Anthropic's request groupings, and to measuring delegation in one week and performance in the other, as shown in Appendices A and C.1.

4.2 Use and work-type composition

The negative relationship may reflect differences in the purpose of use, with work-related tasks exhibiting higher delegation and lower performance. If so, the relationship should weaken when restricting the sample to predominantly work-related tasks or adjusting for work and coursework shares. Panel B of Table 4 shows that the correlation remains negative among tasks with at least 50 or 75 percent of conversations classified as work, where it is similar to or stronger than the baseline in November and weaker in February, down to -0.436 against -0.558 with equal weights. In the full sample, adjusting delegation and performance for the work and coursework shares moves each correlation only slightly toward zero, at most from -0.586 to -0.559 with conversation weights in November. The observed purposes of use therefore explain little of the relationship.

The negative relationship may also reflect differences across broad types of work. Programming tasks tend to have high delegation and low performance, while teaching and advice tasks show the opposite pattern. To examine whether the relationship persists within work types, Panel C of Table 4 allocates each task's weight across its linked O*NET GWAs according to the shares of its DWA links. Including GWA fixed effects leaves the estimated slopes negative in both waves and under both weightings, ranging from -0.429 to -0.497. The pooled within-GWA correlations are also negative, ranging from -0.487 to -0.629.

Table 5: Platform Composition: Computer-Occupation Tasks and RPS Reweighting

	November 2025			February 2026			Sample size	
	Equal	Conversation	RPS	Equal	Conversation	RPS	Nov.	Feb.
Panel A: task-level correlations								
Full main sample	−0.455	−0.586		−0.558	−0.640		947	1,023
Excluding computer-occupation tasks	−0.460	−0.475		−0.565	−0.657		802	870
Tasks on RPS-covered IWAs		−0.590	−0.531		−0.639	−0.478	882	958
Same, excluding computer-occupation tasks			−0.517			−0.450	739	808
Panel B: IWA-level correlations								
RPS-covered IWAs	−0.451	−0.615	−0.533	−0.456	−0.640	−0.441	181	184

Notes: Entries are correlations between D_j and S_j across tasks in Panel A and across IWAs in Panel B, where each task's conversations are split across its linked IWAs and an IWA's delegation and performance are conversation-weighted means. Equal weights count each task or IWA once, and conversation weights use Claude's conversation counts. RPS weights give each IWA covered by both sources its share of U.S. workers' GenAI use in the Real-Time Population Survey (RPS) of Bick et al. (2026b), as Appendix C.3 describes. Within a week, each row uses the same tasks or IWAs under every weighting. Every correlation differs from zero at the 1 percent level by a bootstrap of 2,000 draws of tasks, or of IWAs in Panel B. Table 4 reports the 95 percent intervals of the first row, and Appendix Table C.1 those of the other rows.

To assess the contribution of work-type composition, I further decompose the total delegation–performance covariance into between- and within-GWA components. The within-GWA component accounts for 77 percent of the negative covariance with equal weights in both waves, compared with 47 percent in November and 57 percent in February with conversation weights. Thus, differences across work types matter, particularly with conversation weights, but the negative relationship also persists within them. Appendix C.2 provides the decomposition, and Appendix Table C.6 shows that the relationship remains negative after controlling for additional O*NET task characteristics measured independently of Anthropic's classifiers.

4.3 Platform-specific task composition

Claude's use is concentrated in programming, with computer-occupation tasks accounting for 40 percent of conversations on IWA-linked tasks in November and 36 percent in February. Table 5 shows that excluding these tasks leaves the negative correlation largely intact, ranging from −0.460 to −0.657 across waves and weightings. Only the November conversation-weighted correlation weakens, from −0.586 to −0.475, but remains substantially negative.

To examine the broader role of platform-specific task composition, I reweight IWAs covered by both sources using their shares of U.S. workers' GenAI use in the Real-Time Population Survey (RPS) of Bick et al. (2026b), reducing the computer-occupation

share to 19 and 18 percent. Panel A of Table 5 shows that RPS reweighting weakens the correlation from -0.590 to -0.531 in November and from -0.639 to -0.478 in February. Panel B shows that the correlation remains significantly negative at the IWA level under all three weighting schemes in both waves. Thus, platform-specific task composition affects the magnitude of the relationship but does not account for its negative sign among RPS-covered IWAs. Appendix C.3 describes the reweighting and its limits.

5 Performance Rose While Delegation Barely Changed

The negative cross-task relationship does not reveal whether delegation increases alongside performance within the same task. I therefore examine the 883 tasks observed in both waves, accounting for persistent differences across tasks. Section 5.1 examines average changes, Section 5.2 estimates within-task profiles, and Section 5.3 discusses their interpretation.

5.1 Average changes in delegation and performance

For the 883 tasks, let ΔD_j and ΔS_j denote the February-minus-November changes in delegation and performance. Panel A of Table 6 shows that average delegation barely changed, increasing by 0.004 with equal weights and 0.002 with conversation weights, neither statistically significant.[13] The 95 percent intervals of these changes, $[-0.001, 0.009]$ with equal weights and $[-0.006, 0.010]$ with conversation weights, exclude average changes of more than one percentage point in either direction. Average performance, by contrast, rose substantially, by 0.041 and 0.035, respectively. Delegation increased in 49 percent of tasks, accounting for 44 percent of conversations, whereas performance increased in 76 percent of tasks, accounting for 81 percent of conversations. Under either weighting, the average change in delegation was at most one-tenth of the performance gain.

Figure 2 plots the task-level changes in delegation and performance. To examine whether the average performance gain is concentrated among tasks with changing delegation, Panel B of Table 6 restricts the sample to tasks whose delegation changed by at most two or five percentage points. Among the 282 tasks within two percentage points, accounting for 52 percent of conversations, performance increased by 0.040

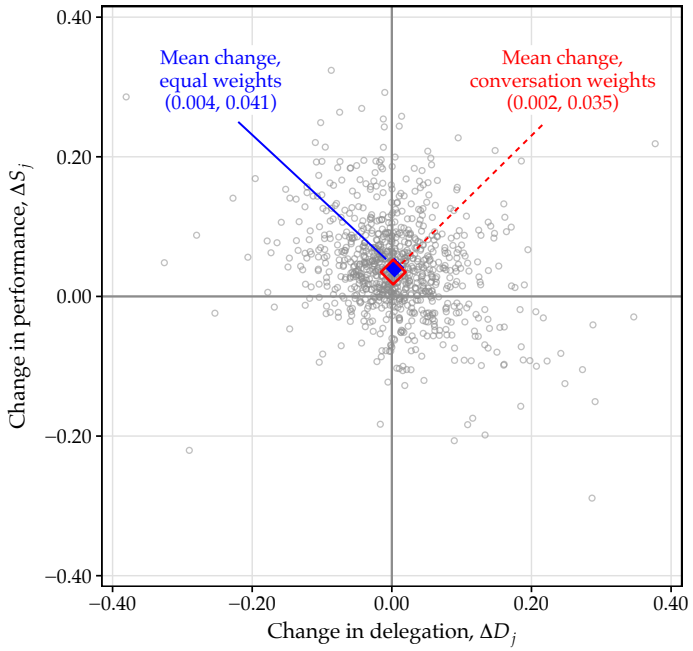
[13] Measured over all conversations, by any of the three measures of Appendix A, delegation rose by 0.005 to 0.007 with equal weights, which is statistically significant, and by about 0.003 with conversation weights, which is not. Part of the equal-weight increase reflects fewer conversations labeled none and fewer withheld cells rather than a shift among the interaction patterns. Anthropic's automation share, pooled over all Claude.ai conversations and tasks, was 45 percent in November and 44 percent in February (Massenkoff et al., 2026).

Table 6: Cross-Wave Results on the Balanced Panel of 883 Tasks

	Weights	
	Equal	Conversation
Panel A: changes, February 2026 minus November 2025		
Mean ΔD_j	0.004	0.002
	(0.003)	(0.004)
$\Delta D_j > 0$		
Number of tasks	437	437
Percent of tasks or conversations	49	44
Mean ΔS_j	0.041	0.035
	(0.002)	(0.004)
$\Delta S_j > 0$		
Number of tasks	674	674
Percent of tasks or conversations	76	81
Panel B: tasks with nearly unchanged delegation		
$\lvert\Delta D_j\rvert \leq 0.02$		
Number of tasks	282	282
Percent of tasks or conversations	32	52
Mean ΔS_j	0.040	0.040
	[0.033, 0.048]	[0.031, 0.048]
$\lvert\Delta D_j\rvert \leq 0.05$		
Number of tasks	543	543
Percent of tasks or conversations	61	76
Mean ΔS_j	0.042	0.040
	[0.037, 0.048]	[0.034, 0.047]
Panel C: predicted performance changes from first-difference regressions		
All 883 tasks	0.042	0.036
	(0.002)	(0.003)
All 883 tasks, with composition controls	0.052	0.041
	(0.004)	(0.004)
741 tasks, after excluding the 142 computer-occupation tasks	0.042	0.032
	(0.003)	(0.003)

Notes: The 883 O*NET tasks in the main samples of both weeks. Changes and means are in proportions, so 0.041 is 4.1 percentage points. Percentages are of the panel's tasks with equal weights and of its conversations with conversation weights, the two-week average conversation counts $\bar{n}_j$. Parentheses hold HC1 standard errors, and brackets 95 percent percentile intervals from 9,999 bootstrap draws of tasks. Panel C reports regression intercepts. Its first and third rows use $\Delta S_j = \gamma + b\,\Delta D_j + u_j$, whose intercept is the predicted change in performance when delegation is unchanged. Its second row adds the within-task changes in the work and coursework shares, human-only ability, the multitasking share, and log human-only time. In that specification, the intercept sets the changes in delegation and in all these characteristics to zero.

Figure 2: Changes in Delegation and Performance between the Two Waves



Notes: The 883 O*NET tasks in the main samples of both weeks. Each point is a task's change in performance against its change in delegation, February 2026 minus November 2025, in proportions. The filled blue diamond marks the mean change $(\Delta D_j, \Delta S_j)$ with equal weights and the hollow red diamond the mean change with conversation weights, the two-week average conversation counts $\bar{n}_j$.

under both weightings. Among the 543 tasks within five percentage points, accounting for 76 percent of conversations, the increases were 0.042 with equal weights and 0.040 with conversation weights. These gains are comparable to those in the full panel, indicating that performance rose even where delegation barely changed.

Panel C of Table 6 regresses changes in performance on changes in delegation, with the intercept representing the predicted performance change when $\Delta D_j = 0$. The intercept is 0.042 with equal weights and 0.036 with conversation weights, close to the unconditional gains in Panel A because average delegation barely changed. With composition controls, the predicted performance gain when delegation and the observed conversation characteristics are unchanged is 0.052 with equal weights and 0.041 with conversation weights. Without controls, excluding computer-occupation tasks yields intercepts of 0.042 and 0.032. Thus, neither observed changes in conversation characteristics nor computer-occupation tasks account for the performance gain. Tables D.1 and D.2 in Appendix D.1 report the full regression results and additional robustness checks.

Table 7: Changes in Within-Task Delegation–Performance Profiles

	Weights	
	Equal	Conversation
Panel A: linear profiles with task fixed effects		
February shift at $D = 0.5$, $\phi + 0.5\,\tau$	0.044	0.037
	(0.003)	(0.003)
Change in slope, February minus November, τ	0.020	0.030
	(0.015)	(0.014)
November slope, β	−0.274	−0.374
	(0.050)	(0.080)
February slope, $\beta + \tau$	−0.254	−0.344
	(0.047)	(0.079)
Panel B: nonlinear profile checks with task fixed effects		
Joint p-value, both weeks' profiles linear		
quadratic	0.64	0.76
linear spline, knot at $D = 0.5$	0.71	0.74
February shift at $D = 0.8$ minus at $D = 0.2$, quadratic	0.007	0.020
	(0.012)	(0.010)

Notes: Balanced panel of 883 tasks. Panel A reports equation (12), and Panel B uses quadratic and linear-spline alternatives. The February shift is the fitted difference between the weeks at a common level of delegation. The joint F tests assess linearity in both weeks. All regressions are estimated in first differences, with HC1 standard errors in parentheses. Conversation weights are the two-week average conversation counts $\bar{n}_j$. Appendix D.2 reports the nonlinear specifications.

5.2 An approximately parallel shift

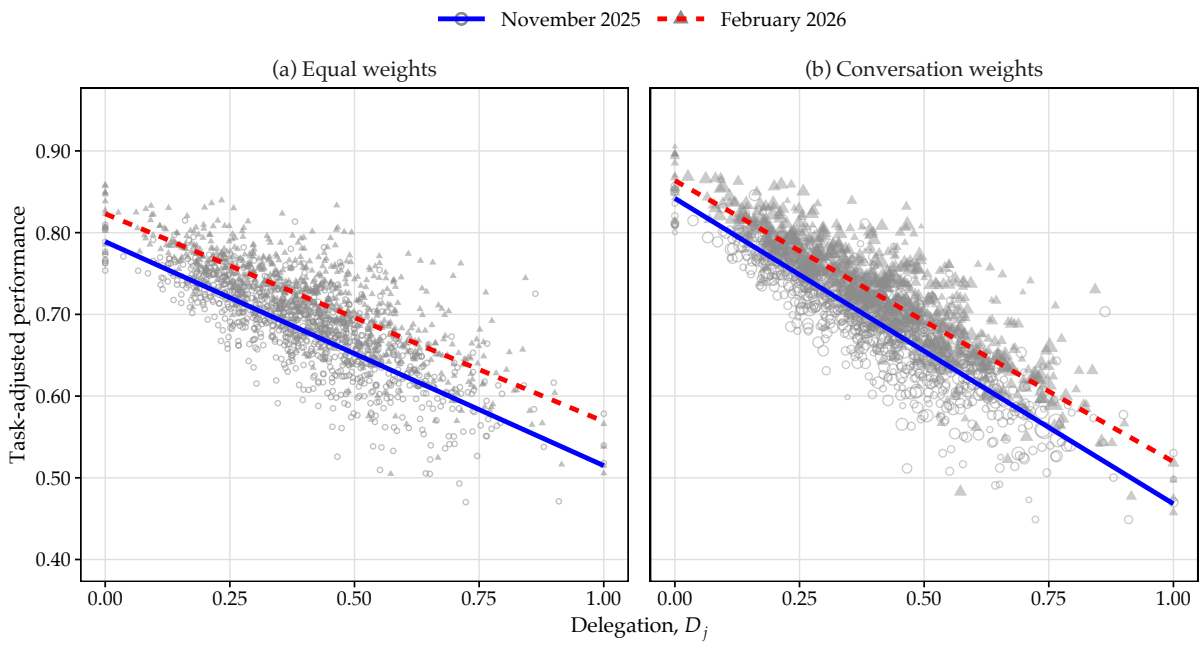
The average changes do not reveal whether performance improved similarly across delegation levels. To examine this, I estimate the following task-fixed-effects regression for task j in week w:

$$S_{j,w} = \alpha_j + \beta D_{j,w} + \phi\,\mathrm{Feb}_w + \tau\,\mathrm{Feb}_w D_{j,w} + \varepsilon_{j,w}. \tag{12}$$

Here, Feb_w indicates February, and the task fixed effects α_j absorb persistent differences across tasks. The February shift at a common delegation level d is $\phi + \tau d$, while τ measures the change in slope. Unlike the intercept in Panel C of Table 6, this shift compares the fitted profiles at the same delegation level.

Panel A of Table 7 shows that the February profile lies above the November profile across delegation levels. At $D = 0.5$, the upward shift is 0.044 with equal weights and 0.037 with conversation weights. The change in slope is 0.020 and 0.030, respectively, indicating a modest flattening that is statistically significant only with conversation weights. Thus, the profiles shifted upward approximately in parallel, as illustrated in Figure 3.

Figure 3: Within-Task Profiles with Equal and Conversation Weights



Notes: The 883 O*NET tasks in the main samples of both weeks, with equal weights in Panel (a) and conversation weights, the two-week average conversation counts $\bar{n}_j$, in Panel (b). The lines are the fitted profiles of the two weeks from equation (12), drawn through the mean task effect. Task-adjusted performance is a task's performance in one week with its estimated fixed effect replaced by the mean task effect under the panel's weighting. Each point is one task in one week, a circle for November and a triangle for February, and in Panel (b) its area grows with $\bar{n}_j$. Both panels use the same axes.

Panel B of Table 7 examines quadratic and linear-spline alternatives. The joint tests provide no clear evidence against linearity, with all p-values exceeding 0.60. In the quadratic specification, the February shift at $D = 0.8$ exceeds that at $D = 0.2$ by 0.007 with equal weights and 0.020 with conversation weights, significant only under the latter. Overall, performance rose at low, middle, and high levels of delegation, with little evidence of nonlinear changes in the profiles. Appendix D.2 reports the full specifications.

5.3 Interpretation and limitations

Sections 5.1 and 5.2 document a robust pattern: average performance rose by roughly four percentage points while average delegation barely changed, and the within-task profiles shifted upward approximately in parallel.[14] The performance gain persists among tasks with nearly unchanged delegation, after controlling for observed changes in conversation characteristics, and after excluding computer-occupation tasks. Parts (a) and (b) of Proposition 1 provide a benchmark for interpreting these

[14] Anthropic's earlier reports show larger movements in the pooled share of directive use among Claude.ai conversations, which rose from 27 percent in December 2024 and January 2025 to 39 percent in August 2025 and fell to 32 percent in November 2025, in samples labeled with different classifier models (Appel et al., 2025, 2026).

findings. Equal improvements in success under delegation and augmentation, holding other payoff determinants fixed, raise performance without changing delegation. The evidence is consistent with this benchmark, although the published data do not separately identify success improvements under the two modes, as Appendix B.4 explains. Part (c) concerns the least-squares line across tasks rather than the within-task profiles. On the balanced panel, that line also shifted up between the waves with little change in its slope, as Appendix D.1 reports.

The increase in measured performance reflects classifier-assessed conversation success rather than a direct measure of model capability. Anthropic released Claude Opus 4.5 and Claude Opus 4.6 around the two waves, with reported benchmark improvements over their predecessors.[15] However, the model used in each conversation is unobserved, and changes in users and requests may also have contributed. The same classifier model, Claude Sonnet 4.5, labeled both waves with the same success prompt, so the rise does not reflect a change in the classifier model or its prompt.

6 What Organizes Delegation beyond Performance?

Sections 4 and 5 show that performance alone does not fully characterize how users divide work with AI. Across tasks, higher performance is associated with lower delegation, and delegation changed little even as performance rose between the two waves. Section 6.1 examines how human and task characteristics are associated with delegation after controlling for performance. Section 6.2 then distinguishes delegation, which concerns how work is divided when GenAI is used, from adoption, which concerns whether GenAI is used at all.

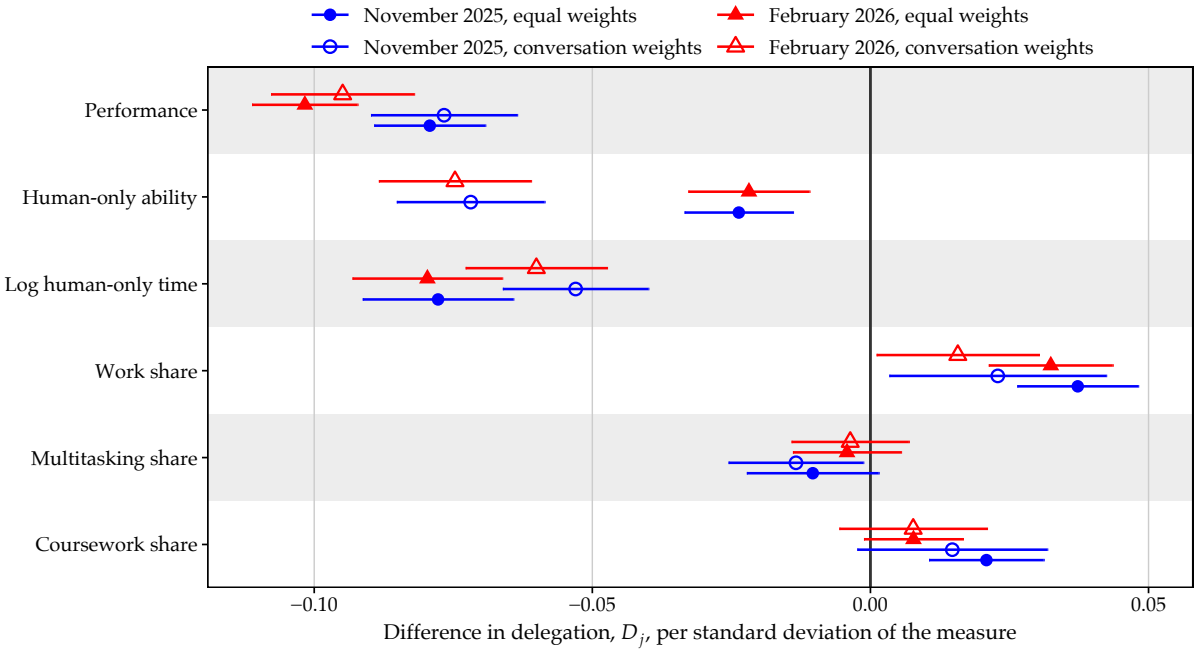
6.1 Human and task characteristics

To examine how human and task characteristics are associated with delegation beyond performance, I estimate the following regression separately for each week:

$$D_{j,w} = \alpha_w + \beta_w S_{j,w} + X'_{j,w}\Gamma_w + \sum_{g \neq g_0} \lambda_{g,w}\,\omega_{j,g} + \varepsilon_{j,w}. \tag{13}$$

[15] Claude Opus 4.5 was released on November 24, 2025, four days after the November sample ended, and Claude Opus 4.6 on February 5, 2026, the first day of the February sample. Anthropic's release comparisons report higher scores for Opus 4.5 than for Opus 4.1 on all nine scores reported for both models, and for Opus 4.6 than for Opus 4.5 on 14 of the 16 scores reported for both models (Anthropic, 2025, 2026d).

Figure 4: What Organizes Delegation across Tasks



Notes: Standardized coefficients from equation (13), estimated separately by week and weighting on the main samples of 947 and 1,023 tasks, with controls for the task's GWA shares and for tasks without a DWA link. Each coefficient is the difference in delegation associated with a one-standard-deviation difference in the measure, holding the other measures and the GWA shares fixed. Bars are 95 percent intervals from HC1 standard errors.

Here, j indexes tasks and w the two weeks. The vector $X_{j,w}$ contains five measures: (1) human-only ability, (2) log human-only time, (3) work share, (4) multitasking share, and (5) coursework share.[16] The GWA shares $\omega_{j,g}$ of Section 2.4, with the indicator $\omega_{j,0}$ for tasks without a DWA link, control for broad types of work, and one GWA, g_0, is omitted. Performance and the five measures are standardized within each week and weighting, and the regressions use equal and conversation weights.

Figure 4 plots the estimated coefficients from equation (13), and Appendix Table E.1 reports the full results. Four measures have coefficients with the same sign and are statistically significant at the 5 percent level in both weeks and under both weightings: performance, human-only ability, log human-only time, and work share. Controlling for the other measures and GWA shares, performance remains negatively associated with delegation, with coefficients ranging from -0.077 to -0.102 per standard deviation. Compared with regressions using performance alone, the coefficients remain nearly unchanged under equal weights and become less negative under conversation weights. Appendix E.1 reports additional robustness checks.

Proposition 2(a) predicts lower delegation when greater human-only skill makes augmentation more attractive relative to delegation. Consistent with this prediction, delegation is lower in tasks whose users are more often judged able to complete the

[16] The personal share is left out because the work, coursework, and personal shares sum to one, apart from the few conversations that carry none of the three labels.

work without AI, controlling for performance and other task characteristics. A 0.1 increase in this share is associated with delegation lower by 2.7 to 3.0 percentage points under equal weights and 6.3 to 6.5 percentage points under conversation weights.

Human-only time measures how long a competent professional would need to complete, without AI, the work performed by AI in a conversation. A doubling of this estimated time is associated with delegation lower by 7.8 to 8.0 percentage points under equal weights and 5.8 to 6.0 percentage points under conversation weights, controlling for performance and other task characteristics. This negative association contrasts with a simple time-saving argument, under which longer work favors delegation. One possible explanation is that longer work also raises the costs of specifying requests (Tirole, 2009), providing context, guiding intermediate steps, and verifying AI output (Subramonyam et al., 2024; Huang et al., 2026). Anthropic engineers report that they tend to delegate tasks that are well defined or easy to verify (Huang et al., 2025), while AI success tends to decline with the time tasks take humans (Kwa et al., 2025; Appel et al., 2026). In the model of Demirer et al. (2026), handing a longer chain of steps to AI saves human checks but lowers the probability that the whole chain succeeds. As Proposition 2(b) suggests, if the costs and risks of delegation rise with the length of the work faster than the time it saves, augmentation becomes relatively more attractive. Human-only time, however, measures only the work performed by AI rather than the full task.

Among the remaining measures, the work share is positively associated with delegation in both waves and under both weightings, although this association is less robust in additional checks. The multitasking and coursework shares show no consistent statistically significant associations.

Consistent with the findings in Section 4.2, the type of work also matters for delegation. Adding GWA shares to the regression with performance and the five other measures increases R^2 by 0.10 to 0.14 across waves and weighting schemes, reaching 0.53 to 0.58 under equal weights and 0.77 under conversation weights. Thus, differences across work types help explain delegation even after accounting for performance and other measured characteristics. With conversation weights, computing and information-technology GWAs have higher delegation than their performance predicts, and communication, advice, and information-gathering GWAs have lower delegation, as Appendix C.2 shows. Appendix E.2 relates delegation to task characteristics that O*NET measures independently of Anthropic's classifiers.

Table 8: RPS Adoption and Claude's Delegation across DWAs

	Weights		
	Equal	RPS observations	Conversation
Panel A: November 2025, 214 DWAs			
Correlation of the RPS adoption rate with			
Claude's delegation	0.004	−0.067	0.448
	[−0.132, 0.143]	[−0.246, 0.130]	[0.141, 0.644]
Claude's log usage share	0.344	0.250	0.353
	[0.209, 0.460]	[0.045, 0.436]	[−0.050, 0.613]
Panel B: February 2026, 243 DWAs			
Correlation of the RPS adoption rate with			
Claude's delegation	0.092	−0.008	0.335
	[−0.035, 0.217]	[−0.165, 0.158]	[0.014, 0.562]
Claude's log usage share	0.328	0.169	0.334
	[0.206, 0.440]	[−0.032, 0.373]	[−0.021, 0.590]

Notes: Correlations across O*NET DWAs between the RPS adoption rate, the survey-weighted share of U.S. workers performing the DWA who regularly use GenAI for it (Bick et al., 2026b), and two Claude measures of the DWA, computed by allocating each task's conversations equally across its DWA links: its delegation and the log of its share of Claude's main-sample conversations on tasks linked to a DWA. The sample consists of the DWAs with a published adoption rate and at least 100 attributed main-sample conversations. "Equal" counts each DWA once, "RPS observations" weights DWAs by their number of RPS observations, and "Conversation" weights them by their attributed Claude conversations. Brackets are 95 percent percentile intervals from 9,999 bootstrap draws of DWAs, which treat DWAs that share tasks as independent.

6.2 Adoption is not delegation

Adoption and delegation capture different margins of GenAI use, but whether they are related across activities is an empirical question. I examine whether activities with higher GenAI adoption rates among U.S. workers also exhibit higher delegation in Claude conversations, comparing the two measures across O*NET DWAs.

I measure adoption using Bick et al. (2026b), who report, for O*NET DWAs, the share of U.S. workers performing a DWA who regularly use GenAI for it. To construct the corresponding delegation measure, I allocate each O*NET task's Claude conversations equally across its DWA links and calculate delegation for each DWA using the allocated conversation counts as weights. The matched samples contain 214 DWAs in November and 243 in February with published adoption rates and at least 100 attributed Claude conversations, accounting for 67.6 and 65.6 percent of main-sample conversations, respectively. Table 8 reports correlations using equal DWA weights, weights proportional to the number of RPS observations, and weights proportional to Claude conversations. It also compares adoption with the log of each DWA's share of Claude conversations. Since the RPS measures regular GenAI use at work among U.S. workers, whereas Claude delegation covers conversations across countries and purposes, the comparison is across activities rather than individual users.

Table 8 shows that the relationship between adoption and delegation depends on weighting. With equal DWA weights, their correlations are 0.004 in November and 0.092 in February. With RPS observation weights, they are -0.067 and -0.008. None of these correlations is statistically significant. With Claude conversation weights, however, the correlations increase to 0.448 and 0.335. By comparison, adoption is positively correlated with the log share of Claude conversations under equal weights, with correlations of 0.344 and 0.328 in the two waves. Thus, adoption is related to where Claude is used, but its relationship with delegation varies substantially with the weighting of activities.

Claude conversations are concentrated in programming DWAs, which combine high delegation with above-median adoption. These activities account for only 17 percent of matched DWAs in November and 15 percent in February, but 47 and 41 percent of matched conversations, respectively. Teaching and advice activities, such as tutoring, instead tend to have both lower delegation and below-median adoption. Excluding the single largest programming DWA reduces the conversation-weighted correlations to 0.326 and 0.215, respectively, but both remain positive. Thus, the positive conversation-weighted relationship reflects these two groups of activities, at opposite ends of both measures, rather than a general correspondence between adoption and delegation across activities. Appendix C.3 defines the two groups, gives the correlation without both of them, and reports additional checks at the IWA level.

7 Conclusion

This paper studies how task execution is divided between humans and AI, conditional on GenAI use. Using task-level aggregates of about two million Claude.ai conversations from two weekly waves, I examine how delegation relates to realized performance across tasks, how both changed over time, and which human and task characteristics are associated with delegation at similar levels of performance. I also compare Claude delegation with survey-measured GenAI adoption across DWAs.

Three findings emerge. First, higher delegation is associated with lower realized performance across tasks, and this relationship persists across checks on the purpose and type of work and the composition of AI use. Second, among the same tasks observed 12 weeks apart, performance rose by roughly four percentage points while delegation remained essentially unchanged, and the fitted delegation–performance profiles shifted upward approximately in parallel. Third, at a given level of performance, delegation is lower in tasks with higher human-only ability and longer human-only time, and it differs systematically across types of work. Delegation shows little associ-

ation with GenAI adoption, except under conversation weights, where the association is positive and reflects programming activities and teaching and advice activities at opposite ends of both measures. Together, these results suggest that realized performance alone does not characterize how task execution is divided between humans and AI.

These findings have implications for measuring AI use and modeling its effects on work. First, Tamkin and McCrory (2025) estimate productivity gains from task-level time savings, and Appel et al. (2026) adjust such estimates by observed success rates. These adjusted estimates rely on performance measured under existing patterns of human involvement. The performance gains with nearly unchanged delegation documented here show that improved success need not reflect a greater transfer of execution to AI. Adoption, performance, and delegation therefore need to be measured separately. Related measurement work distinguishes delegation from augmentation by assigning greater weight to automated than augmentative use in occupational exposure (Massenkoff and McCrory, 2026). Second, for models of technology and work, the consequences of automation depend on which tasks within a job are affected (Autor and Thompson, 2025). This paper adds a further consideration, the division of execution within tasks already performed with AI. Improvements in AI can support both delegation and augmentation, and their implications for the human role depend on how they change the relative payoffs of these modes. Models linking AI progress to labor demand therefore need to account for how users organize task execution, alongside which tasks they adopt AI for.

Future research can build on these findings in several directions. Following the same users across tasks and over time would help distinguish changes in individual delegation choices from changes in the composition of AI users. Linking each conversation to the model used and measuring success separately under delegation and augmentation would clarify how improvements reach the two modes. Randomized access to new models or plausibly exogenous variation in their rollout could then identify how model improvements affect both performance and delegation. Combining these data with measures of human time, effort, and output quality would connect changes in task execution to productivity and show how the benefits of AI depend on the way humans work with it.

References

ACEMOGLU, D., AND D. H. AUTOR (2011): "Skills, Tasks and Technologies: Implications for Employment and Earnings," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card. Elsevier, vol. 4B, 1043–1171.

ACEMOGLU, D., AND P. RESTREPO (2018): "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment," *American Economic Review*, 108, 1488–1542.

ACEMOGLU, D., AND P. RESTREPO (2019): "Automation and New Tasks: How Technology Displaces and Reinstates Labor," *Journal of Economic Perspectives*, 33, 3–30.

AGARWAL, N., A. MOEHRING, P. RAJPURKAR, AND T. SALZ (2023): "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," Working Paper 31422, National Bureau of Economic Research.

AGHION, P., AND J. TIROLE (1997): "Formal and Real Authority in Organizations," *Journal of Political Economy*, 105, 1–29.

ANTHROPIC (2025): "Introducing Claude Opus 4.5," November 24, 2025, https://www.anthropic.com/news/claude-opus-4-5 (accessed October 30, 2026).

ANTHROPIC (2026a): "The Anthropic Economic Index," Hugging Face dataset Anthropic/EconomicIndex, releases of January 15, 2026, March 24, 2026, and June 26, 2026, https://huggingface.co/datasets/Anthropic/EconomicIndex (accessed October 30, 2026).

ANTHROPIC (2026b): "Anthropic Economic Index: New Building Blocks for Understanding AI Use," January 15, 2026, https://www.anthropic.com/research/economic-index-primitives (accessed October 30, 2026).

ANTHROPIC (2026c): "Appendix to 'Anthropic Economic Index Report: Learning Curves,'" March 2026, https://cdn.sanity.io/files/4zrzovbb/website/a3cdcd9e67c3c4c51440429dd016cacba514b35b.pdf (accessed October 30, 2026).

ANTHROPIC (2026d): "Introducing Claude Opus 4.6," February 5, 2026, https://www.anthropic.com/news/claude-opus-4-6 (accessed October 30, 2026).

ANTHROPIC (2026e): "Online Appendix: Economic Primitives and Classification Prompts," January 15, 2026, https://huggingface.co/datasets/Anthropic/EconomicIndex/blob/main/release_2026_01_15/aei_v4_appendix.pdf (accessed October 30, 2026).

APPEL, R., M. MASSENKOFF, P. MCCRORY, M. MCCAIN, R. HELLER, T. NEYLON, AND A. TAMKIN (2026): "Anthropic Economic Index Report: Economic Primitives," Anthropic, January 15, 2026, https://www.anthropic.com/research/anthropic-economic-index-january-2026-report (accessed October 30, 2026).

APPEL, R., P. MCCRORY, A. TAMKIN, M. MCCAIN, T. NEYLON, AND M. STERN (2025): "Anthropic Economic Index Report: Uneven Geographic and Enterprise AI Adoption," Anthropic, September 15, 2025, https://www.anthropic.com/research/anthropic-economic-index-september-2025-report (accessed October 30, 2026).

ATHEY, S. C., K. A. BRYAN, AND J. S. GANS (2020): "The Allocation of Decision Authority to Human and Artificial Intelligence," *AEA Papers and Proceedings*, 110, 80–84.

AUTOR, D. H. (2015): "Why Are There Still So Many Jobs? The History and Future of Workplace Automation," *Journal of Economic Perspectives*, 29, 3–30.

AUTOR, D. H., F. LEVY, AND R. J. MURNANE (2003): "The Skill Content of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, 118, 1279–1333.

AUTOR, D. H., AND N. THOMPSON (2025): "Expertise," *Journal of the European Economic Association*, 23, 1203–1271.

BARTLING, B., E. FEHR, AND H. HERZ (2014): "The Intrinsic Value of Decision Rights," *Econometrica*, 82, 2005–2039.

BICK, A., A. BLANDIN, AND D. J. DEMING (2026a): "The Rapid Adoption of Generative AI," *Management Science*, 72, 8440–8456.

BICK, A., A. BLANDIN, D. J. DEMING, AND T. R. SCHUMACHER (2026b): "What Work Does Generative AI Do?" Working Paper 35677, National Bureau of Economic Research.

BRYNJOLFSSON, E., B. CHANDAR, AND R. CHEN (2025a): "Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence," Working Paper, Stanford Digital Economy Lab.

BRYNJOLFSSON, E., D. LI, AND L. RAYMOND (2025b): "Generative AI at Work," *The Quarterly Journal of Economics*, 140, 889–942.

CHATTERJI, A., T. CUNNINGHAM, D. J. DEMING, Z. HITZIG, C. ONG, C. Y. SHAN, AND K. WADMAN (2025): "How People Use ChatGPT," Working Paper 34255, National Bureau of Economic Research.

CROSS, P. J., AND C. F. MANSKI (2002): "Regressions, Short and Long," *Econometrica*, 70, 357–368.

DELL'ACQUA, F., E. MCFOWLAND, III, E. MOLLICK, H. LIFSHITZ, K. C. KELLOGG, S. RAJENDRAN, L. KRAYER, F. CANDELON, AND K. R. LAKHANI (2026): "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality," *Organization Science*, 37, 403–423.

DEMIRER, M., J. J. HORTON, N. IMMORLICA, B. LUCIER, AND P. SHAHIDI (2026): "Chaining Tasks, Redefining Work: A Theory of AI Automation," Working Paper 34859, National Bureau of Economic Research.

DESSEIN, W. (2002): "Authority and Communication in Organizations," *The Review of Economic Studies*, 69, 811–838.

DIETVORST, B. J., J. P. SIMMONS, AND C. MASSEY (2015): "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err," *Journal of Experimental Psychology: General*, 144, 114–126.

DIETVORST, B. J., J. P. SIMMONS, AND C. MASSEY (2018): "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them," *Management Science*, 64, 1155–1170.

ELOUNDOU, T., S. MANNING, P. MISHKIN, AND D. ROCK (2024): "GPTs Are GPTs: Labor Market Impact Potential of LLMs," *Science*, 384, 1306–1308.

FELTEN, E., M. RAJ, AND R. SEAMANS (2021): "Occupational, Industry, and Geographic Exposure to Artificial Intelligence: A Novel Dataset and Its Potential Uses," *Strategic Management Journal*, 42, 2195–2217.

FÜGENER, A., J. GRAHL, A. GUPTA, AND W. KETTER (2022): "Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path toward Productive Delegation," *Information Systems Research*, 33, 678–696.

FÜGENER, A., D. D. WALZNER, AND A. GUPTA (2026): "Roles of Artificial Intelligence in Collaboration with Humans: Automation, Augmentation, and the Future of Work," *Management Science*, 72, 538–557.

GANS, J. S. (2026): "Artificial Jagged Intelligence: When AI Benchmarks Misstate Deployment Value," Working Paper 34712, National Bureau of Economic Research.

HANDA, K., A. TAMKIN, M. MCCAIN, S. HUANG, E. DURMUS, S. HECK, J. MUELLER, J. HONG, S. RITCHIE, T. BELONAX, K. K. TROY, D. AMODEI, J. KAPLAN, J. CLARK, AND D. GANGULI (2025): "Which Economic Tasks Are Performed with AI? Evidence from Millions of Claude Conversations," arXiv:2503.04761.

HANSEN, M. C., J. J. NORTON, C. M. GREGORY, A. W. MEADE, L. F. THOMPSON, D. RIVKIN, P. LEWIS, AND J. NOTTINGHAM (2014): "O*NET Work Activities Project Technical Report," Report prepared for the U.S. Department of Labor, Employment and Training Administration, National Center for O*NET Development, February 2014, https://www.onetcenter.org/reports/DWA_2014.html (accessed October 30, 2026).

HUANG, L., W. XIAO, AND N. K. VISHNOI (2026): "Delegation and Verification under AI," in *Proceedings of the 43rd International Conference on Machine Learning*, PMLR, vol.

306, 48039–48071.

HUANG, S., B. SEETHOR, E. DURMUS, K. HANDA, M. MCCAIN, M. STERN, AND D. GANGULI (2025): "How AI Is Transforming Work at Anthropic," Anthropic, December 2, 2025, https://www.anthropic.com/research/how-ai-is-transforming-work-at-anthropic (accessed October 30, 2026).

HUMLUM, A., AND E. VESTERGAARD (2025): "The Unequal Adoption of ChatGPT Exacerbates Existing Inequalities among Workers," *Proceedings of the National Academy of Sciences*, 122, e2414972121.

IDE, E., AND E. TALAMÀS (2025): "Artificial Intelligence in the Knowledge Economy," *Journal of Political Economy*, 133, 3762–3800.

KISH, L. (1992): "Weighting for Unequal P_i," *Journal of Official Statistics*, 8, 183–200.

KWA, T., B. WEST, J. BECKER, A. DENG, K. GARCIA, M. HASIN, S. JAWHAR, M. KINNIMENT, N. RUSH, S. VON ARX, R. BLOOM, T. BROADLEY, H. DU, B. GOODRICH, N. JURKOVIC, L. H. MILES, S. NIX, T. LIN, C. PAINTER, N. PARIKH, D. REIN, L. J. KOBA SATO, H. WIJK, D. M. ZIEGLER, E. BARNES, AND L. CHAN (2025): "Measuring AI Ability to Complete Long Software Tasks," in *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, Neural Information Processing Systems Foundation, Inc., 102382–102435.

LYNCH, P. D. (2026): "The Collaboration Code: How Humans and AI Work Together across Millions of Conversations on Job Tasks," *Frontiers in Artificial Intelligence*, 9, 1801235.

MACKINNON, J. G., AND H. WHITE (1985): "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305–325.

MANSKI, C. F. (2003): *Partial Identification of Probability Distributions*, Springer Series in Statistics, New York: Springer.

MASSENKOFF, M., E. LYUBICH, P. MCCRORY, R. APPEL, AND R. HELLER (2026): "Anthropic Economic Index Report: Learning Curves," Anthropic, March 24, 2026, https://www.anthropic.com/research/economic-index-march-2026-report (accessed October 30, 2026).

MASSENKOFF, M., AND P. MCCRORY (2026): "Labor Market Impacts of AI: A New Measure and Early Evidence," Anthropic, March 5, 2026, https://www.anthropic.com/research/labor-market-impacts (accessed October 30, 2026).

MCCAIN, M., T. MILLAR, S. HUANG, J. EATON, K. HANDA, M. STERN, A. TAMKIN, M. KEARNEY, E. DURMUS, J. SHEN, J. HONG, B. CALVERT, J. S. CHAN, F. MOSCONI, D. SAUNDERS, T. NEYLON, G. NICHOLAS, S. POLLACK, J. CLARK, AND D. GANGULI (2026): "Measuring AI Agent Autonomy in Practice,"

Anthropic, February 18, 2026, https://www.anthropic.com/research/measuring-agent-autonomy (accessed October 30, 2026).

MORRIS, M. R., J. SOHL-DICKSTEIN, N. FIEDEL, T. WARKENTIN, A. DAFOE, A. FAUST, C. FARABET, AND S. LEGG (2024): "Position: Levels of AGI for Operationalizing Progress on the Path to AGI," in *Proceedings of the 41st International Conference on Machine Learning*, PMLR, vol. 235, 36308–36321.

NATIONAL CENTER FOR O*NET DEVELOPMENT (2015): "O*NET 20.1 Database," U.S. Department of Labor, Employment and Training Administration, released October 2015, https://www.onetcenter.org/db_releases.html (accessed October 30, 2026).

NOY, S., AND W. ZHANG (2023): "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, 381, 187–192.

PATWARDHAN, T., R. DIAS, E. PROEHL, G. KIM, M. WANG, O. WATKINS, S. POSADA FISHMAN, M. ALJUBEH, P. THACKER, L. FAUCONNET, N. S. KIM, P. CHAO, S. MISERENDINO, G. CHABOT, D. LI, M. SHARMAN, A. BARR, A. GLAESE, AND J. TWOREK (2026): "GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks," in *International Conference on Learning Representations (ICLR 2026)*.

PETERSON, N. G., M. D. MUMFORD, W. C. BORMAN, P. R. JEANNERET, E. A. FLEISHMAN, K. Y. LEVIN, M. A. CAMPION, M. S. MAYFIELD, F. P. MORGESON, K. PEARLMAN, M. K. GOWING, A. R. LANCASTER, M. B. SILVER, AND D. M. DYE (2001): "Understanding Work Using the Occupational Information Network (O*NET): Implications for Practice and Research," *Personnel Psychology*, 54, 451–492.

RAGHU, M., K. BLUMER, G. CORRADO, J. KLEINBERG, Z. OBERMEYER, AND S. MULLAINATHAN (2019): "The Algorithmic Automation Problem: Prediction, Triage, and Human Effort," arXiv:1903.12220.

RANDAZZO, S., H. LIFSHITZ, K. C. KELLOGG, F. DELL'ACQUA, E. MOLLICK, F. CANDELON, AND K. R. LAKHANI (2025): "Cyborgs, Centaurs and Self-Automators: The Three Modes of Human-GenAI Knowledge Work and Their Implications for Skilling and the Future of Expertise," Working Paper 26-036, Harvard Business School.

SHAO, Y., H. ZOPE, Y. JIANG, J. PEI, D. NGUYEN, E. BRYNJOLFSSON, AND D. YANG (2025): "Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the U.S. Workforce," arXiv:2506.06576.

SUBRAMONYAM, H., R. PEA, C. PONDOC, M. AGRAWALA, AND C. SEIFERT (2024): "Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ACM, 1–19.

TAMKIN, A., M. MCCAIN, K. HANDA, E. DURMUS, L. LOVITT, A. RATHI, S. HUANG, A. MOUNTFIELD, J. HONG, S. RITCHIE, M. STERN, B. CLARKE, L. GOLDBERG,

T. R. SUMERS, J. MUELLER, W. MCEACHEN, W. MITCHELL, S. CARTER, J. CLARK, J. KAPLAN, AND D. GANGULI (2024): "Clio: Privacy-Preserving Insights into Real-World AI Use," arXiv:2412.13678.

TAMKIN, A., AND P. MCCRORY (2025): "Estimating AI Productivity Gains from Claude Conversations," Anthropic, November 2025, https://www.anthropic.com/research/estimating-productivity-gains (accessed October 30, 2026).

TIROLE, J. (2009): "Cognition and Incomplete Contracts," *American Economic Review*, 99, 265–294.

TOMLINSON, K., S. JAFFE, W. WANG, S. COUNTS, AND S. SURI (2025): "Working with AI: Measuring the Applicability of Generative AI to Occupations," arXiv:2507.07935, forthcoming in *Advances in Neural Information Processing Systems (NeurIPS 2026)*.

WEBB, M. (2020): "The Impact of Artificial Intelligence on the Labor Market," Working Paper, Stanford University.

Appendices

A Data and Measurement

This appendix documents the data and measures of Section 2. It describes the source files, defines the variables, reports the questions that the classifiers answer, explains how withheld cells are handled, repeats the main comparisons with delegation measured over all conversations, and collects what is known about the two waves.

Data files. I use the aggregation over all countries in the Claude.ai files of the AEI repository (Anthropic, 2026a). The November file was released on January 15, 2026, and the February file on March 24, 2026. The replication code labels the two files R4 and R5, following their order among the AEI's data releases.

Variables. Table A.1 defines the variables used in the paper and reports their construction and sources. Anthropic refers to the characteristics assigned to conversations, such as purpose of use and task success, as facets. Most of these characteristics are reported separately for each O*NET task. For example, `onet_task::task_success` reports task success by O*NET task. Categorical characteristics are summarized by conversation counts for each category, while numeric characteristics are reported as task-level means. Request groups, used in Appendix C.1, have their own facets, which report the same labels by request group instead of by task.

Classifier prompts. The online appendix to the January 2026 report archives the prompts of Anthropic's success, interaction-pattern, and other conversation-label classifiers, though not of its task classifier (Anthropic, 2026e). The success classifier asks "Did the Assistant complete the task provided by the User successfully?" and answers yes or no. The interaction-pattern classifier uses the five definitions that Table 1 paraphrases, asks for the pattern most representative of the user's response, and answers none when it is unsure or there is not enough context to determine one. The prompt tells the classifier to use none "liberally." The appendix to the March 2026 report reproduces the question wording but not the full prompts (Anthropic, 2026c).

Withheld cells and exact recovery. Anthropic withholds every task-by-category cell with fewer than 15 conversations and reports the withheld conversations as the task's unclassified residual, which is distinct from the classifier's label none. For interaction

Table A.1: Variable Definitions and Sources

Variable	Definition	Construction and scale	Source
Panel A: AEI, Claude.ai task-level aggregates, November 2025 and February 2026			
Conversations, n_j	Conversations that the task classifier assigns to task j	Count	`onet_task`
Performance, S_j	Share of the task's conversations judged successful	Yes count over n_j, exact if $n_j \geq 29$	`task_success`
Delegation, D_j	Share of conversations with a pattern that are in directive use or feedback loops	Equation (1), published cells	`collaboration`
All-conversation delegation, D_j^{all}	Delegation with the conversations labeled none in the denominator	Equation (A.1), published cells	`collaboration`
Human-only ability	Share of conversations in which the user could do the task alone	Yes count over n_j, exact	`human_only_ability`
Human-only time	Estimated hours a competent professional needs without AI for the work the AI did	Task mean, in hours. Its logarithm enters regressions	`human_only_time`
Multitasking share	Share of conversations in which the user multitasks	Yes count over n_j, exact	`multitasking`
Work, coursework, and personal shares	Shares of conversations for work, coursework, and personal use	Count over n_j. A withheld cell gets an equal part of the residual, at most 14	`use_case`
AI discretion	Latitude the AI had to decide how the task was done	Task mean of a rating from 1 to 5. Enters only the main-sample rule	`ai_autonomy`
*Panel B: O*NET 20.1, independent of Anthropic's classifiers*			
Task statement and occupations	Assigned O*NET statement and the occupations listing it	Exact text match, ignoring case	Task Statements
Computer-occupation task	Task listed in a computer occupation	1 if any O*NET-SOC code is in SOC 2010 minor group 15-11	Task Statements
Work activities, DWA, IWA, GWA	Detailed, intermediate, and generalized work activities to which the task links	Shares of the task's DWA links by IWA and by GWA, the GWA shares $\omega_{j,g}$	Tasks to DWAs, DWA Reference
O*NET task characteristics	Interacting-with-others and mental-processes DWA shares, task frequency, four Work Context ratings, Job Zone	Averaged over the statement's task identifiers or occupations	Task Ratings, Work Context, Job Zones, Tasks to DWAs
Panel C: other external measures			
RPS adoption rate	Share of U.S. workers performing a DWA who regularly use GenAI for it	Survey-weighted, 682 DWAs with at least 20 observations	Bick et al. (2026b), adoption index, pooled
RPS task share	Share of U.S. workers' GenAI-assisted tasks in each of 332 IWAs	Recovered from the published figure	Bick et al. (2026b), Figure 9, pooled

Notes: Panel A: Claude.ai files of the AEI, rows aggregated over all countries (Anthropic, 2026a). In the Source column, a facet name such as `task_success` stands for its version by O*NET task, `onet_task::task_success`. Request groups use `request::task_success` and `request::collaboration`. Main samples are the tasks with $n_j \geq 100$ and with S_j, D_j, and AI discretion defined. All other AEI measures are defined for every main-sample task. Panel B: O*NET 20.1 files (National Center for O*NET Development, 2015). Panel C: "pooled" means survey waves from August 2025 to May 2026.

patterns and success labels, the sample-wide totals contain no unclassified conversations, so a task's residual consists exactly of its withheld cells, each between 0 and 14 conversations. The share of conversations judged successful is therefore exactly recoverable whenever at least one of its two cells is published, which is guaranteed for tasks with at least 29 conversations and so for every task in the main samples. For delegation D_j, I use the published-cell share of equation (1), which leaves both the none cell and the withheld conversations out of its numerator and denominator. I obtain the identified range of the true share by allocating each task's withheld conversations to delegation, augmentation, or no-pattern cells within their bounds. Human-only ability is published exactly for every main-sample task, the multitasking share is recovered exactly, and the work share is withheld for 2.2 and 3.0 percent of main-sample conversations, where it is bounded in $[0, 0.14]$. No task's bounds straddle the work-share thresholds of Section 4.2.

Signs despite withheld cells. A least-squares slope is linear in the dependent variable. For the slope of delegation D_j on an exactly measured regressor, the extreme values over all admissible allocations of the withheld cells are attained by setting each task's delegation at the lower or the upper end of its range, according to the sign of the task's centered regressor. A Pearson correlation between delegation and performance has the sign of the slope of delegation on performance for every admissible vector of shares, so its sign is identified whenever the range of that slope excludes zero. In the main samples, the slope of delegation on performance is -0.549 in November and -0.756 in February with equal weights, and -0.822 and -0.895 with conversation weights. Its range over admissible allocations is $[-0.636, -0.369]$ and $[-0.839, -0.547]$ with equal weights and $[-0.841, -0.777]$ and $[-0.913, -0.846]$ with conversation weights, so its sign is identified in all four cases. The same check identifies the sign of every correlation in Table C.4. Replacing the published-cell shares with the midpoint of each task's range changes every correlation in Panel B of Table C.4 by at most 0.007.[17]

Delegation over all conversations. Delegation D_j conditions on an identified interaction pattern. Write n_j^D and n_j^A for a task's published conversations in the delegation and the augmentation patterns, n_j^{none} for its published conversations labeled none, and n_j^{w} for its withheld conversations. Together, these four groups make up the task's n_j conversations. Equation (1) is $D_j = n_j^D / (n_j^D + n_j^A)$. Conversations labeled none are outside this denominator, and so, under the published-cell convention, are the withheld ones. Anthropic's reports for the two weeks instead state automation as a share of

[17] The bounds do not impose the sample-wide category totals, so they are conservative.

all sampled conversations (Appel et al., 2026; Massenkoff et al., 2026), and its published task-level percentages use all of a task's conversations as the base. As a robustness check, I therefore use

$$D_j^{\text{all}} = \frac{n_j^D}{n_j - n_j^{\text{w}}} = \frac{n_j^D}{n_j^D + n_j^A + n_j^{\text{none}}}, \tag{A.1}$$

a published-cell value of the all-conversation share $\tilde{D}_j$, the share of all n_j conversations in the two delegation patterns, which is known only within a range when pattern cells are withheld. D_j^{all} is the preferred robustness measure. This measure keeps conversations labeled none in the denominator with nothing in the numerator, without counting them as augmentation, and it treats withheld cells as equation (1) does. Two other conventions treat withheld cells differently: the literal share n_j^D / n_j, which counts every withheld conversation as not delegated, and the midpoint of the identified range of $\tilde{D}_j$. In the main samples, conversations labeled none are 2.0 and 2.1 percent of all conversations and withheld ones 1.4 and 1.6 percent, but these conversations are concentrated. In 23 November and 25 February tasks, with 0.5 and 0.6 percent of the conversations, more than 30 percent of conversations lie outside the five published pattern cells.

Table A.2 reports the results. D_j^{all} correlates 0.954 to 0.989 with D_j in both weeks under both weightings, and the negative relationship with performance holds under every convention. The correlations are weaker than with D_j by 0.028 to 0.063, a small difference that a paired bootstrap nonetheless distinguishes from zero for D_j^{all}, and the slopes differ from those with D_j by at most 0.034. With equal weights, the difference comes mostly from the few tasks in which more than 30 percent of conversations lie outside the published pattern cells, many of them dominated by the label none. Without these tasks, the correlation with equal weights is -0.435 with D_j and -0.426 with D_j^{all} in November and -0.549 and -0.534 in February. Under every convention, the sign of the relationship holds for every admissible allocation of the withheld cells.

Panel C shows that the all-conversation measures rise across the waves by 0.47 to 0.69 percentage points with equal weights and by 0.27 to 0.34 points with conversation weights, against 0.39 and 0.20 points for D_j. With equal weights, the three increases are significant at 5 percent with heteroskedasticity-robust standard errors and with standard errors clustered by occupation, the first O*NET-SOC occupation listed for each task (378 clusters). With conversation weights, no increase is significant with either. Part of the increase with equal weights is not a change in how execution is allocated. Of the 0.69-point rise in the literal share, an exact decomposition attributes 0.34 points to a shift among the five patterns, 0.17 points to fewer conversations labeled none, and 0.18 points to fewer withheld cells. The last component is consistent with larger task

Table A.2: Alternative Delegation Denominators

	November 2025		February 2026	
	Equal	Conversation	Equal	Conversation
Panel A: correlation between the delegation measure and S_j, main samples				
D_j, equation (1)	−0.455	−0.586	−0.558	−0.640
D_j^{all}, equation (A.1)	−0.403	−0.558	−0.514	−0.598
Literal share n_j^D / n_j	−0.392	−0.552	−0.504	−0.594
Midpoint of the range of $\tilde{D}_j$	−0.402	−0.558	−0.519	−0.599
Panel B: slope of S_j on the delegation measure, main samples				
D_j	−0.378	−0.418	−0.412	−0.458
	(0.030)	(0.031)	(0.024)	(0.038)
D_j^{all}	−0.344	−0.403	−0.394	−0.439
	(0.031)	(0.034)	(0.025)	(0.037)
Literal share n_j^D / n_j	−0.362	−0.399	−0.418	−0.439
	(0.032)	(0.033)	(0.026)	(0.038)
Midpoint of the range of $\tilde{D}_j$	−0.380	−0.409	−0.444	−0.449
	(0.034)	(0.034)	(0.028)	(0.039)
Panel C: mean change in percentage points, February minus November, 883 balanced-panel tasks				
	Equal weights		Conversation weights	
D_j	0.39		0.20	
	(0.25)		(0.39)	
D_j^{all}	0.54		0.29	
	(0.26)		(0.39)	
Literal share n_j^D / n_j	0.69		0.34	
	(0.24)		(0.39)	
Midpoint of the range of $\tilde{D}_j$	0.47		0.27	
	(0.23)		(0.39)	
Performance S_j, for reference	4.12		3.49	
	(0.24)		(0.36)	

Notes: Notation as in equation (A.1). The range of $\tilde{D}_j$ comes from the withheld pattern cells. Panels A and B use the main samples of 947 and 1,023 tasks. Conversation weights are n_j in Panels A and B and the two-week averages $\bar{n}_j$ in Panel C. Slopes are least squares, and mean changes are intercepts of the change on a constant. Parentheses hold HC1 standard errors.

counts, which lift previously withheld cells above the publication threshold. Changing the denominator does not make the success and pattern labels jointly observed, so the argument of Appendix B.3 applies unchanged.

The two waves. Table A.3 collects what Anthropic's public documentation says about the two samples and the model environment in each week, together with what Anthropic reported to me about the classifier and the sampling code. Two facts matter for the paper. First, the public documentation names the classifier model only for the November sample, Claude Sonnet 4.5, and Anthropic's dataset description associates the March 2026 release "with Opus 4.5/4.6." In a personal communication in October

Table A.3: Documentation of the Two Waves

	November 2025	February 2026
Sample	999,875 Claude.ai conversations, November 13 to 20, 2025; file labeled Free and Pro plans, report names Free, Pro, and Max	1,000,000 Claude.ai conversations, February 5 to 12, 2026; Free, Pro, and Max plans
Official statement on model use	The November sample, Claude.ai and API together, was "predominantly using Claude Sonnet 4.5"; no share is given	Among paid Claude.ai accounts, 51 percent of usage was on the Opus model class; the model is user-selected
Success classifier	Claude Sonnet 4.5	Claude Sonnet 4.5 with the same prompt, according to Anthropic; not named in the public documentation, and the dataset description associates the release "with Opus 4.5/4.6"
Other changes		The week overlapped Anthropic's Super Bowl advertising, which brought many first-time users; coursework use fell and personal use rose within tasks (Table 3, Panel B); the surrounding sampling code changed slightly, according to Anthropic

Notes: Sources are the January and March 2026 reports and their appendices (Appel et al., 2026; Anthropic, 2026e; Massenkoff et al., 2026; Anthropic, 2026c), the overview of the January report, which gives the November statement on model use (Anthropic, 2026b), and the dataset description (Anthropic, 2026a). The entries marked "according to Anthropic" come from a personal communication in October 2026.

2026, Anthropic confirmed that the same classifier model, Claude Sonnet 4.5, labeled both samples and that the success classifier used the same prompt in both, although the surrounding sampling code changed slightly. The association with Opus 4.5 and 4.6 therefore does not refer to the classifier. Second, no released file records which model served a conversation. The table also lists the plans that each sample covers, but no release reports how many sampled conversations come from each plan, so it is not established whether the two samples cover the same plans.

B Proofs and Identification Limits

This appendix proves the two propositions of Section 3 and states what the published data identify about the success probabilities of the two modes and their changes between the two weeks.

B.1 Proof of Proposition 1

Proof. By assumption, $\Delta B_j = \Delta C_j = 0$ and $\delta_j^D = \delta_j^A = \delta_j$, so equation (10) gives $\Lambda_{j,1} - \Lambda_{j,0} = V_j(\delta_j - \delta_j) = 0$. With $F_{j,1} = F_{j,0}$, equation (8) then gives $D_{j,1}^* = F_{j,1}(\Lambda_{j,1}) = F_{j,0}(\Lambda_{j,0}) = D_{j,0}^*$, which is part (a). The condition $s_{j,0}^m + \delta_j \leq 1$ makes $s_{j,1}^m = s_{j,0}^m + \delta_j$ a probability for both modes, and with delegation unchanged at $D_{j,0}^*$, equation (9) gives part (b):

$$S_{j,1}^* - S_{j,0}^* = D_{j,0}^* \, \delta_j + (1 - D_{j,0}^*) \, \delta_j = \delta_j. \tag{B.1}$$

For part (c), let $\nu_j > 0$ be the fixed weights of the tasks in J, normalized to sum to one, which leaves the fitted line unchanged, and let $\bar{D} = \sum_{j \in J} \nu_j D_{j,0}^*$ and $\bar{S}_t = \sum_{j \in J} \nu_j S_{j,t}^*$. By parts (a) and (b), $D_{j,1}^* = D_{j,0}^*$ and $S_{j,1}^* = S_{j,0}^* + \delta$ for every task in J, so $\bar{S}_1 = \bar{S}_0 + \delta$. The weighted least-squares slope in environment t is:

$$\theta_t = \frac{\sum_{j \in J} \nu_j \left(D_{j,0}^* - \bar{D} \right) \left(S_{j,t}^* - \bar{S}_t \right)}{\sum_{j \in J} \nu_j \left(D_{j,0}^* - \bar{D} \right)^2}, \tag{B.2}$$

whose denominator is positive because delegation varies across the tasks in J. Because $S_{j,1}^* - \bar{S}_1 = S_{j,0}^* - \bar{S}_0$ for every task, $\theta_1 = \theta_0$, and the intercept $\bar{S}_t - \theta_t \bar{D}$ rises by δ. □

B.2 Proof of Proposition 2

Proof. For part (a), a user delegates if and only if $\eta_{i,j} \leq \Lambda_j$, that is, if and only if $\xi_{i,j} \leq \Lambda_j - a(h_{i,j})$. Because a is nondecreasing, $h'_{i,j} \geq h_{i,j}$ implies $\Lambda_j - a(h'_{i,j}) \leq \Lambda_j - a(h_{i,j})$, so with Λ_j and every $\xi_{i,j}$ unchanged, every user who delegates with skill $h'_{i,j}$ also delegates with skill $h_{i,j}$. The event that a conversation is delegated is therefore no larger with the more skilled users, and neither is its probability D_j^*. For part (b), equations (4) and (5) give $D_j^* = F_j(\Lambda_j)$ with $\Lambda_j = V_j \, g_j^* + B_j - C_j$. Because F_j does not vary with ℓ_j and its density f_j is continuous at Λ_j, the chain rule gives $dD_j^*/d\ell_j = f_j(\Lambda_j) \, d\Lambda_j/d\ell_j$, and $d\Lambda_j/d\ell_j$ is the bracket in equation (11). Because $f_j(\Lambda_j) > 0$, the derivative has the sign of the bracket. □

B.3 Why mode-specific success is not identified

Fix a task and a week and drop both indices. In the framework of Section 3, each conversation is in the delegation mode, $z = 1$, or in the augmentation mode, $z = 0$, and

$y = 1$ if the success classifier judges it successful, which this appendix treats as success. The published data give the two marginal probabilities but not the joint probability:

$$D^* = \Pr(z = 1), \qquad S^* = \Pr(y = 1), \qquad q^* = \Pr(y = 1, z = 1). \tag{B.3}$$

If $0 < D^* < 1$, then $s^D = q^*/D^*$ and $s^A = (S^* - q^*)/(1 - D^*)$, so the mode-specific success gap is:

$$g^* = s^D - s^A = \frac{q^* - D^*S^*}{D^*(1 - D^*)}. \tag{B.4}$$

The marginals restrict q^* only to its Fréchet bounds, which are sharp (Manski, 2003, p. 161):

$$\max(0, D^* + S^* - 1) \leq q^* \leq \min(D^*, S^*). \tag{B.5}$$

Because $(1 - D^*)(1 - S^*) \geq 0$ and $0 \leq D^*, S^* \leq 1$, these bounds always contain $q^* = D^*S^*$, at which success is independent of the mode and $g^* = 0$. If also $0 < S^* < 1$, D^*S^* lies strictly inside the bounds, so g^* can take either sign. The published marginals therefore cannot exclude a zero gap, and because performance lies strictly between zero and one in every main-sample task, the sign of the within-task gap is identified for no main-sample task in either week. In particular, the negative delegation–performance relationship across tasks does not imply that delegated conversations succeed less often than augmented ones within a task. Recovering a joint distribution from its marginals is an ecological inference problem (Cross and Manski, 2002). Even the joint distribution would identify only the success rates of the modes that users selected, and interpreting their difference as the causal effect of delegating would require further assumptions about how users select into modes.

The published shares measure these probabilities only approximately. S_j is a share of all of a task's conversations, whereas D_j leaves out the conversations labeled none and the withheld conversations, whose treatment Appendix A describes. Equation (6) is therefore not an accounting identity in the measured shares, and treating D_j as D_j^* assumes that the conversations it leaves out split between the modes like the others. The argument above holds at every value of D_j^* strictly between zero and one, and the data admit such values for every main-sample task, even where D_j itself is zero or one, so its conclusion does not depend on that assumption.

B.4 Changes in mode-specific success

Let $t = 0$ denote November and $t = 1$ February. For the observed weeks, $s_{j,t}^m$ is the success rate of mode m in week t, which equals $s_j^m(M_t)$ in the framework but can also change with the users and the requests between the observed weeks. Let $\delta_j^m =$

$s^m_{j,1} - s^m_{j,0}$, as in Section 3, and let $g^*_{j,t} = s^D_{j,t} - s^A_{j,t}$ be the gap in week t. Equation (9) holds for these rates by the law of total probability. Writing it as $S^*_{j,t} = s^A_{j,t} + D^*_{j,t}\, g^*_{j,t}$, differencing, and using $g^*_{j,1} = g^*_{j,0} + \delta^D_j - \delta^A_j$ gives, for every task:

$$\Delta S^*_j = \delta^A_j + \Delta D^*_j\, g^*_{j,0} + D^*_{j,1}\left(\delta^D_j - \delta^A_j\right). \tag{B.6}$$

The first term is the change in success under augmentation, the second is the reallocation of conversations between the modes at the November gap, and the third is the differential improvement under delegation, weighted by the February delegation probability. If delegation is unchanged, $D^*_{j,1} = D^*_{j,0}$, the change in performance is:

$$\Delta S^*_j = D^*_{j,0}\,\delta^D_j + \left(1 - D^*_{j,0}\right)\delta^A_j. \tag{B.7}$$

A rise in performance at unchanged delegation therefore fixes only this weighted average of the two improvements. Equal improvements with the other payoff terms unchanged, the conditions of Proposition 1, produce it, but so do unequal improvements, which, as Section 3 notes, can also leave delegation unchanged. Because the published data identify neither the two improvements nor the gap $g^*_{j,0}$, they do not establish equal improvements under the two modes.

For the same reason, the within-task slopes of Section 5.2 are not estimates of these terms. Reading the November slope as the gap and the change in slope as $\delta^D_j - \delta^A_j$ would require restrictions that this paper does not impose: that the November gap and both improvements are the same in every task, and that, in both weeks, $D^*_{j,t} = D_{j,t}$ and $S_{j,t}$ measures $S^*_{j,t}$ up to sampling error. Under these restrictions, equation (B.6) would take the first-difference form of equation (12), with the gap as the coefficient on ΔD_j and the differential improvement as the coefficient on February delegation.

C Cross-Task Robustness

This appendix collects the checks of the cross-task relationship of Section 4, in the order in which that section uses them. Appendix C.1 reports bootstrap intervals for the composition checks of Tables 4 and 5, describes the task groups, varies the sample threshold, groups conversations by Anthropic's request groups instead of O*NET tasks, and measures delegation and performance in different weeks. Appendix C.2 decomposes the relationship between and within GWAs, adds controls for task characteristics that O*NET measures independently of Anthropic's classifiers, and shows which GWAs have high or low delegation given performance. Appendix C.3 reweights Claude's

IWAs toward U.S. workers' use of GenAI and defines the activity groups of the adoption comparison.

C.1 Bootstrap intervals, task groups, and measurement

Sampling uncertainty of the composition checks. Table C.1 reports 95 percent intervals for the correlations of the three composition checks, in Panels B and C of Table 4 and in Table 5. They come from the percentile bootstrap that gives the intervals of Panel A of Table 4, with 2,000 draws of tasks, or of IWAs at the IWA level. The intervals are wider for the work-share subsamples and at the IWA level, but every one lies below zero, and every 99 percent interval excludes zero as well. The 99 percent upper bound closest to zero is -0.152, for the equal-weight IWA-level correlation in November.

Task groups. Table C.2 splits the main samples at each week's medians of delegation and performance into four groups and names the largest tasks in each, and Panel A of Table 4 reports the shares of tasks and conversations in the two opposite groups. The groups describe the samples and do not classify tasks into types. The high-delegation, low-performance group is led by software work but is not confined to it: only 19 and 18 percent of its tasks belong to computer occupations.

Sample thresholds. Table C.3 reports the correlation between delegation and performance at thresholds of 15, 30, 50, and 100 conversations. It is negative at every threshold, in both weeks, and under both weightings, and its sign holds for all allocations of the withheld conversations.

An alternative grouping of user requests. Anthropic also provides a bottom-up grouping of the requests in the same conversations, built without O*NET (Appel et al., 2026; Anthropic, 2026a). I construct performance and delegation for each request group by the same rules as for tasks. Panel A of Table C.4 shows that the correlation is negative at both of the two finest levels in both weeks, from -0.410 to -0.710, and that every confidence interval excludes zero. Use is far less concentrated across request groups than across tasks. Kish's effective number of units under conversation weights is $(\sum_j n_j)^2 / \sum_j n_j^2$, which equals the number of units when all weights are equal and falls as the weights become more concentrated (Kish, 1992, pp. 190–191). It is 534 and 528 for finest-level groups, against 73 and 109 for tasks. The relationship under conversation weights therefore does not rest on a few programming tasks. The

Table C.1: Bootstrap Intervals for the Composition Checks

	November 2025		February 2026	
	Equal	Conversation	Equal	Conversation
Use composition, Panel B of Table 4				
Tasks with work share ≥ 0.50	−0.475 [−0.562, −0.374]	−0.642 [−0.764, −0.413]	−0.476 [−0.560, −0.384]	−0.559 [−0.686, −0.377]
Tasks with work share ≥ 0.75	−0.548 [−0.688, −0.373]	−0.583 [−0.720, −0.427]	−0.436 [−0.555, −0.306]	−0.525 [−0.666, −0.357]
Adjusted for work and coursework shares	−0.445 [−0.505, −0.384]	−0.559 [−0.641, −0.439]	−0.542 [−0.592, −0.489]	−0.634 [−0.700, −0.543]
Work-type composition, Panel C of Table 4				
Within-GWA correlation	−0.487 [−0.544, −0.421]	−0.547 [−0.634, −0.419]	−0.566 [−0.619, −0.507]	−0.629 [−0.700, −0.529]
Platform composition, Table 5				
	Equal	Conversation or RPS	Equal	Conversation or RPS
Excluding computer-occupation tasks	−0.460 [−0.522, −0.395]	−0.475 [−0.554, −0.385]	−0.565 [−0.618, −0.507]	−0.657 [−0.730, −0.560]
Tasks on RPS-covered IWAs, conversation weights		−0.590 [−0.687, −0.455]		−0.639 [−0.706, −0.554]
Same tasks, RPS weights		−0.531 [−0.639, −0.414]		−0.478 [−0.569, −0.379]
Same, excluding computer-occupation tasks, RPS weights		−0.517 [−0.638, −0.385]		−0.450 [−0.555, −0.329]
RPS-covered IWAs, equal weights	−0.451 [−0.616, −0.233]		−0.456 [−0.578, −0.311]	
Same IWAs, conversation weights		−0.615 [−0.776, −0.353]		−0.640 [−0.759, −0.484]
Same IWAs, RPS weights		−0.533 [−0.670, −0.352]		−0.441 [−0.565, −0.297]

Notes: Correlations between D_j and S_j from Panels B and C of Table 4 and from the composition checks of Table 5, on the same samples, with 95 percent percentile intervals in brackets from 2,000 bootstrap draws. In the platform rows, the second column of each week uses conversation weights unless the row names RPS weights. Draws are of tasks, except at the IWA level, where they are of IWAs and so treat IWAs that share tasks as independent.

Table C.2: Task Groups by Delegation and Performance

	November 2025		February 2026	
	Share of tasks	Share of conversations	Share of tasks	Share of conversations
High delegation, low performance	33.6	46.4	35.0	43.9
Largest tasks. Modifying and debugging software, writing programs, network and system troubleshooting, and building websites				
Low delegation, high performance	33.6	31.0	35.0	33.0
Largest tasks. Tutoring and coursework help, instruction, advising, and copy editing				
High delegation, high performance	16.5	13.2	15.1	10.9
Largest tasks. Creative writing, technical editing and documentation, supporting web code, and translation				
Low delegation, low performance	16.4	9.4	15.0	12.2
Largest tasks. Consulting on computing needs, requirements analysis, technical problem analysis, and reviewing software documentation				
Tasks and conversations	947 and 846,472		1,023 and 838,965	

Notes: Main samples. Tasks are split at each week's medians of delegation, 0.405 in November and 0.416 in February, and of performance, 0.693 and 0.721. "High" means at or above the median. Shares are percentages of tasks and of the conversations in the sample. "Largest tasks" summarizes the tasks with the most conversations in each group.

Table C.3: Delegation–Performance Correlation by Sample Threshold

	November 2025		February 2026	
Minimum number of conversations	Equal	Conversation	Equal	Conversation
$n \geq 15$	−0.364	−0.546	−0.505	−0.614
$n \geq 30$	−0.356	−0.547	−0.504	−0.616
$n \geq 50$	−0.409	−0.567	−0.539	−0.629
$n \geq 100$	−0.455	−0.586	−0.558	−0.640

Notes: Pearson correlations between D_j and S_j across tasks with at least the indicated number of conversations and both measures defined, for example 1,551 tasks in November and 1,678 in February at 50 conversations. At 100 conversations, the samples are the main samples.

check shows that the result does not depend on the O*NET assignment, but it does not address how success is judged, because the groups contain the same conversations with the same labels.

Delegation and performance measured in different weeks. In any one week, a task's performance and its delegation come from the same conversations. Chance in which conversations are sampled, and errors the classifiers make in reading a given conversation, can move both measures at once and could produce a negative correlation across tasks even if tasks differed in nothing else. The two weekly samples contain different conversations, labeled separately, so performance measured in one week and delegation measured in the other do not share such conversation-specific errors. Panel B of Table C.4 reports the resulting cross-week correlations on the balanced panel. They are −0.468 and −0.483 with equal weights and −0.614 and −0.549 with conversation weights. In absolute value, the average of the two cross-week

Table C.4: An Alternative Grouping of Requests and a Cross-Week Check

	Equal weights		Conversation weights	
	Correlation	95% interval	Correlation	95% interval
Panel A: request groups, each week separately				
November, finest level, 617 groups	−0.416	[−0.508, −0.317]	−0.410	[−0.501, −0.313]
February, finest level, 620 groups	−0.527	[−0.624, −0.421]	−0.584	[−0.662, −0.495]
November, second level, 111 groups	−0.523	[−0.639, −0.378]	−0.524	[−0.661, −0.364]
February, second level, 103 groups	−0.710	[−0.815, −0.539]	−0.679	[−0.795, −0.525]
*Panel B: O*NET tasks in both main samples, 883 tasks*				
Same week				
Delegation Nov., performance Nov.	−0.471	[−0.521, −0.421]	−0.565	[−0.658, −0.443]
Delegation Feb., performance Feb.	−0.546	[−0.594, −0.495]	−0.649	[−0.715, −0.557]
Cross week				
Delegation Feb., performance Nov.	−0.468	[−0.517, −0.420]	−0.614	[−0.695, −0.505]
Delegation Nov., performance Feb.	−0.483	[−0.530, −0.431]	−0.549	[−0.644, −0.433]

Notes: Pearson correlations between delegation and performance. Confidence intervals are percentiles of 2,000 nonparametric bootstrap draws of request groups or tasks. Panel A uses Anthropic's bottom-up grouping of user requests at its two finest levels, excluding the group of conversations not classified at a level, with performance and delegation built as for tasks. Every group has more than 500 conversations, and a group's conversation weight is its number of conversations. In Panel B, the conversation weight is the two-week average $\bar{n}_j$.

covariances is only 2 percent smaller than the same-week covariance in November and 10 percent smaller in February with equal weights. With conversation weights, it is less than 1 percent smaller in November and 7 percent smaller in February. The delegation–performance relationship therefore does not mainly reflect conversation-specific errors. The check does not remove persistent measurement components, such as a tendency in both weeks for success to be judged more harshly in some forms of interaction.

C.2 GWA decomposition and O*NET controls

This subsection uses the main-sample tasks linked to a GWA, 893 in November and 968 in February. The 54 and 55 tasks without a DWA link hold about 7 percent of conversations. As in Panel C of Table 4, each task's conversations are split across its GWAs in proportion to its GWA shares $\omega_{j,g}$.

Decomposition. By the law of total covariance, the covariance between delegation and performance across tasks splits exactly into a between-GWA component and a within-GWA component. Table C.5 reports the split. Both components are negative. Without the ten largest tasks, which are mostly programming and hold 30 and 24 percent of the linked tasks' conversations, the within-GWA share under conversation weights is 64 and 68 percent. Within GWAs, the correlation between delegation and

Table C.5: Delegation and Performance within and between GWAs

	November 2025		February 2026	
	Equal	Conversation	Equal	Conversation
Decomposition of the covariance across 35 GWAs				
Total correlation between D_j and S_j, linked tasks	−0.473	−0.585	−0.562	−0.634
Between-GWA share of the covariance, percent	23	53	23	43
Within-GWA share of the covariance, percent	77	47	77	57
Within-GWA correlation	−0.487	−0.547	−0.566	−0.629
Within-GWA share without the ten largest tasks, percent		64		68
Nested shares: categories, subcategories, GWAs within subcategories, percent		29, 1, 23		13, 9, 21
Tasks	893	893	968	968

Notes: Main-sample tasks linked to a GWA, with each task's conversations split across its GWAs in proportion to its GWA shares, one row per task and GWA. The decomposition applies the law of total covariance. The equal-weight columns give each task the same weight, split across its GWAs in the same proportions.

performance is −0.456 to −0.629 in the four cells and without the ten largest tasks. With conversation weights, in the nested hierarchy, the four categories account for 29 and 13 percent of the covariance, subcategories within categories for 1 and 9 percent, and GWAs within subcategories for 23 and 21 percent.

O*NET task characteristics as controls. Table C.6 asks whether task characteristics that O*NET measures independently of Anthropic's classifiers account for the negative relationship. It uses the eight O*NET 20.1 characteristics of Table A.1: the shares of the task's DWA links in the categories Interacting With Others and Mental Processes, the frequency of the task, and, averaged over the task's occupations, four Work Context ratings and the Job Zone. Within each week, it regresses performance on delegation alone, with the eight characteristics, with the task's GWA shares, and with both, always on the same tasks. The slope keeps its sign and size in every specification. Adding the characteristics moves it by between 4 percent toward zero and 6 percent away from zero, and with the GWA shares, with or without the characteristics, it is 6 to 13 percent steeper than without controls, −0.437 to −0.510 against −0.402 to −0.462. The characteristics raise the R^2 of performance by only 0.02 to 0.04 without the GWA shares, and the six that are not sums of GWA shares raise it by 0.01 to 0.02 with them. They are imperfect proxies for task requirements. They describe O*NET tasks and occupations, not conversations, and they do not measure the difficulty of a particular request, the skill of the user, or how fully the request is specified. The table therefore does not rule out differences in difficulty as a source of the relationship.

Table C.6: Delegation and Performance with O*NET Controls

	November 2025		February 2026	
Slope of S_j on D_j	Equal	Conversation	Equal	Conversation
M0: delegation only	−0.402	−0.416	−0.423	−0.462
	(0.031)	(0.035)	(0.024)	(0.042)
M1: plus eight O*NET characteristics	−0.398	−0.398	−0.450	−0.479
	(0.036)	(0.047)	(0.028)	(0.047)
M2: plus GWA shares	−0.440	−0.460	−0.450	−0.508
	(0.038)	(0.060)	(0.029)	(0.052)
M3: plus both	−0.437	−0.471	−0.474	−0.510
	(0.041)	(0.062)	(0.030)	(0.051)
R^2: M0	0.23	0.34	0.33	0.41
R^2: M1	0.25	0.38	0.37	0.44
R^2: M2	0.42	0.61	0.49	0.64
R^2: M3	0.43	0.62	0.51	0.65
Tasks	855	855	924	924
GWAs with a positive share	33	33	34	34

Notes: Least-squares slopes of performance on delegation, with heteroskedasticity-robust (HC3) standard errors in parentheses (MacKinnon and White, 1985). Within each week, all four specifications use the same tasks: the main-sample tasks with all eight O*NET characteristics defined, less two tasks in November and one in February, each the only task with a positive share in its GWA. M1 and M3 add the eight characteristics, standardized within each week and weighting, and M2 and M3 add the task's GWA shares. The interacting-with-others and mental-processes shares are sums of GWA shares, so in M3 they enter through the GWA shares. With standard errors clustered by occupation, every slope is more than nine standard errors from zero.

Delegation by type of work given performance. Conditional on performance, GWAs differ systematically and stably in delegation. The residual of delegation given performance is highest for computing and information-technology work, which O*NET files under the GWAs Interacting With Computers, Thinking Creatively, and two smaller ones, and lowest for Getting Information, Communicating with Supervisors, Peers, or Subordinates, and Provide Consultation and Advice to Others. These residuals come from a conversation-weighted regression of delegation on performance across the linked tasks, aggregated to GWAs. With each GWA weighted equally, they correlate 0.988 across the two weeks among the 16 GWAs with at least one percent of conversations in both, and 0.978 when the regression also includes the work share, human-only ability, the multitasking share, and log human-only time. Adding human-only ability alone halves the residual of Interacting With Computers and nearly removes that of Thinking Creatively. The residuals adjust for performance, and in their second version for these four measures, but not for the full specification of equation (13), whose GWA shares capture the contrast between GWAs given performance and all five measures.

C.3 Reweighting toward U.S. workers' GenAI use and adoption

The RPS measure. Bick et al. (2026b) ask employed respondents of the RPS which of the ten most important O*NET DWAs of their occupation they perform and, if they use GenAI at work, which of those they regularly use it for, in waves from August 2025 to May 2026. I use two objects from their work. The first is the distribution of U.S. workers' GenAI-assisted tasks over the 332 O*NET IWAs, their Figure 9, which they compare with chat-based usage shares from Anthropic, OpenAI, and Microsoft. No released file contains this vector. I recovered it from the vector drawing of their figure and accepted it only because it reproduces all 100 statistics printed in their paper that can be recomputed from it, within rounding. The figure carries no identifiers, so I attach the O*NET 20.1 IWA identifiers to the recovered values by the alphabetical order of the O*NET IWA titles, the order in which the figure draws them. The second object is their released task adoption index, defined in Table A.1, whose identifiers and titles all match O*NET 20.1. Every merge with either object is on these identifiers.

Claude's task composition. Claude's IWA distribution is far more concentrated than the RPS distribution, as Panel A of Table C.7 shows. Claude's largest IWA, designing computer or information systems or applications, holds 15 to 20 percent of Claude's conversations and 2 percent of RPS tasks. Reweighting Claude's IWAs to the RPS distribution, while holding Claude's task composition within each IWA fixed, lowers the computer-occupation share of conversations from 40.0 and 35.6 percent to 19.2 and 18.1 percent. Applying the same within-IWA task composition to the OpenAI and Microsoft distributions that Bick et al. (2026b) map to IWAs from the data of Chatterji et al. (2025) and Tomlinson et al. (2025) gives 17 to 18 percent, so the concentration in programming is characteristic of Claude. Bick et al. (2026b) document the same concentration, with the same largest IWA, in Anthropic's data of March 2025.

The reweighted relationship. The RPS weights are the RPS shares restricted to and renormalized over the IWAs on which Claude has main-sample tasks and the RPS share is positive, 181 IWAs in November and 184 in February, holding 99 percent of Claude's linked conversations and 82 to 83 percent of RPS tasks. Each task's weight is its conversations reallocated in proportion to the ratio of the RPS share to Claude's share of each IWA it links to. Table 5 reports the correlations, and Table C.7 the slopes and further samples. Under RPS weights, the task-level correlation keeps 90 and 75 percent of its size under conversation weights. Excluding computer-occupation tasks, the ten largest tasks, or the ten largest DWAs changes no conclusion. The weakest correlation under conversation or RPS weights is -0.422. Capping the IWA ratios at their

Table C.7: Claude's IWA Composition and Correlations under RPS Weights

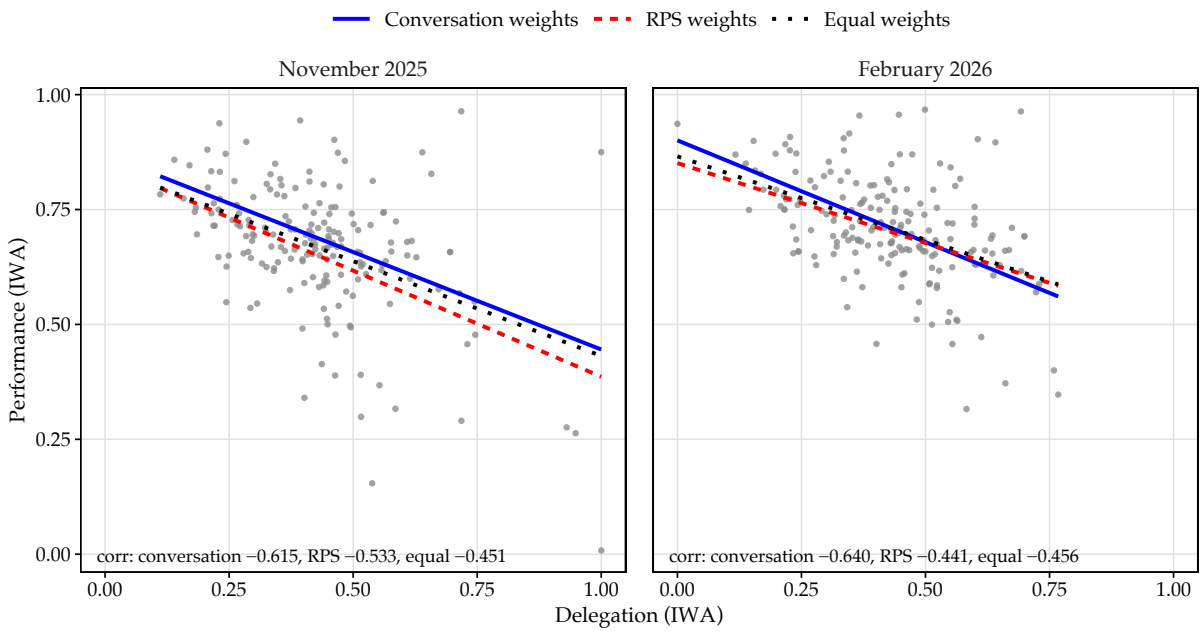
	November 2025		February 2026	
Panel A: composition over 332 IWAs				
	Claude	RPS	Claude	RPS
Effective number of IWAs	16.8	99.9	24.5	99.9
Share of the ten largest IWAs, percent	55	22	47	22
Computer-occupation share of conversations, percent	40.0	19.2	35.6	18.1
Total variation distance between the two distributions	0.65		0.62	
Panel B: slope of S_j on D_j and correlations in further samples				
	Conversation weights	RPS weights	Conversation weights	RPS weights
IWA level: slope	−0.424	−0.462	−0.443	−0.348
Task level: slope	−0.420	−0.477	−0.457	−0.370
	(0.031)	(0.065)	(0.040)	(0.046)
Task-by-IWA rows, GWA fixed effects: slope	−0.456	−0.581	−0.504	−0.401
	(0.047)	(0.067)	(0.042)	(0.040)
Task level: correlation excluding computer-occupation tasks	−0.479	−0.517	−0.661	−0.450
Task level: correlation excluding the ten largest tasks	−0.438	−0.523	−0.592	−0.471
Tasks in the baseline sample	882	882	958	958
Tasks excluding computer-occupation tasks	739	739	808	808
Tasks excluding the ten largest tasks	873	873	950	950

Notes: Panel A compares Claude's main-sample conversations, allocated to IWAs through the O*NET crosswalk, with the RPS distribution of U.S. workers' GenAI-assisted tasks from Bick et al. (2026b). The effective number of IWAs is the inverse of the sum of squared IWA shares. The total variation distance is the share of Claude's mass that would have to move to match the RPS distribution. The RPS computer-occupation share holds Claude's task composition within each IWA fixed. Panel B's baseline sample consists of the tasks on the 181 and 184 IWAs that both sources cover. Table 5 reports their correlations. The ten largest tasks are the week's ten main-sample tasks with the most conversations. One of them in November and two in February link to no DWA and are not in the baseline sample. Conversation weights are Claude's conversation counts. RPS weights reallocate each task's conversations in proportion to the ratio of the RPS share to Claude's share of its IWAs. Standard errors are HC1 at the task level and clustered by task in the fixed-effects row, which has one row per task and linked IWA and fixed effects for the IWA's GWA.

99th or 95th percentile changes the task-level correlation by at most 0.017. Reweighting toward the flatter RPS distribution spreads the weight across tasks. The effective number of tasks rises from 67 and 101 to 232 and 240. Figure C.1 shows the IWA-level relationship under the three weightings.

Limits. The reweighting changes only the IWA composition. It keeps Claude's users, task composition, and outcomes within each IWA, so it does not show what the relationship would be among the workers the RPS describes. It cannot represent the 17.9 and 16.5 percent of RPS tasks on IWAs without a main-sample Claude task, about a third of which is physical and manual work. The RPS measure covers only each occupation's ten most important DWAs, is pooled over a period that brackets both Claude

Figure C.1: IWA-Level Delegation and Performance under Three Weightings



Notes: Each point is an IWA on which both Claude and the RPS have positive mass, 181 in November and 184 in February, plotted by its conversation-weighted mean delegation and performance over Claude's main-sample tasks. Lines are least-squares fits weighted by Claude conversations, by RPS task shares, and equally. The corr values are the correlations under each weighting.

weeks, and was recovered from a published figure. The task-level standard errors ignore that the weights are estimated.

Adoption and delegation. The programming DWAs of Section 6.2 are the matched DWAs with more than half of their O*NET task links in computer occupations, and the teaching and advice DWAs are the other matched DWAs in the GWAs Training and Teaching Others and Provide Consultation and Advice to Others. Without both groups, the conversation-weighted correlation between adoption and delegation is -0.018 in November and -0.161 in February. At the IWA level, I use the RPS index's rate for the IWA, which pools the observations of its DWAs. The correlation of adoption with delegation is then small under equal weights, RPS observation weights, and the RPS weights of Section 4.3, 0.045 to 0.113, with intervals that include zero. It is positive under conversation weights, 0.608 and 0.474.

D Cross-Wave Robustness

This appendix checks the cross-wave comparison of Section 5. Appendix D.1 reports the alternative samples, composition controls, sampling noise, and the weekly cross-task lines, and Appendix D.2 reports the nonlinear specifications.

D.1 Samples, composition, noise, and cross-task lines

Estimation. I estimate the regressions in first differences with HC1 standard errors. With the same task weights in both weeks, the coefficient estimates equal those from the corresponding task-fixed-effects regressions. The HC1 standard errors agree with task-clustered standard errors to within 0.03 percent.

Robustness samples. Table D.1 reports the intercept γ of the regression of ΔS_j on ΔD_j, its slope b, the change in the within-task slope τ, and the mean change in delegation for the alternative samples. The intercept and the slope keep their sign and a similar size in every sample, and they do not rest on computer-occupation tasks. The slope b, -0.259 with equal weights and -0.356 with conversation weights in the full panel, describes how the changes co-move across tasks. Sampling alone can produce it, as the paragraph on sampling noise explains, so I report it without interpretation.

Work-heavy tasks. Two of the six samples, tasks whose two-week mean work share is at least 0.50 or 0.75, show a larger and significant flattening, τ of 0.051 to 0.106. In a fully interacted first-difference specification on all 883 tasks, the extra flattening of these tasks has p-values of 0.03 to 0.22 without controls. Its p-values are 0.12 to 0.23 once the November human-only ability, human-only time, and multitasking labels, the computer-occupation flag, the GWA shares, and slopes by O*NET category share are controlled for. At both thresholds, it is significant at 5 percent at some intermediate steps of adding these controls. Work-heavy tasks are more often computer-occupation tasks than the other tasks. The evidence that work-heavy tasks flatten more than other tasks is therefore weak.

Composition controls. Table D.2 adds the within-task changes in the work and coursework shares, human-only ability, the multitasking share, and log human-only time to the regression of ΔS_j on ΔD_j. The controls raise the intercept γ rather than absorbing it. Relative to personal use, a shift toward work or coursework goes with larger performance gains, and personal use rose between the weeks. The slope b moves by at most 0.019. By contrast, among the 466 tasks with fully published work and coursework cells in both weeks, adding the controls reduces the intercept from 0.035 to 0.031 with equal weights and from 0.032 to 0.026 with conversation weights. The corresponding slopes change from -0.309 to -0.314 and from -0.397 to -0.358.

Sampling noise and same-conversation coupling. Performance and delegation are shares of a weekly random sample, and both come from labels of the same conversa-

Table D.1: Cross-Wave Estimates across Alternative Samples

Sample	Tasks	Intercept γ	Slope b of ΔS_j on ΔD_j	Change in slope τ	Mean ΔD_j
Panel A: equal weights					
All tasks	883	0.042	−0.259	0.020	0.004
		(0.002)	(0.047)	(0.015)	(0.003)
Excluding computer-occupation tasks	741	0.042	−0.250	0.020	0.002
		(0.003)	(0.052)	(0.017)	(0.003)
Excluding the ten largest tasks	873	0.042	−0.260	0.021	0.004
		(0.002)	(0.047)	(0.015)	(0.003)
Computer-occupation tasks only	142	0.046	−0.343	0.011	0.013
		(0.004)	(0.074)	(0.026)	(0.005)
Two-week mean work share ≥ 0.50	396	0.036	−0.282	0.051	0.014
		(0.003)	(0.094)	(0.024)	(0.004)
Two-week mean work share ≥ 0.75	169	0.036	−0.354	0.106	0.014
		(0.004)	(0.112)	(0.040)	(0.005)
Panel B: conversation weights					
All tasks	883	0.036	−0.356	0.030	0.002
		(0.003)	(0.078)	(0.014)	(0.004)
Excluding computer-occupation tasks	741	0.032	−0.353	0.023	0.005
		(0.003)	(0.082)	(0.019)	(0.005)
Excluding the ten largest tasks	873	0.036	−0.370	0.030	0.006
		(0.003)	(0.075)	(0.015)	(0.005)
Computer-occupation tasks only	142	0.042	−0.325	0.028	−0.003
		(0.005)	(0.184)	(0.024)	(0.005)
Two-week mean work share ≥ 0.50	396	0.035	−0.351	0.058	0.003
		(0.004)	(0.112)	(0.016)	(0.004)
Two-week mean work share ≥ 0.75	169	0.032	−0.358	0.085	0.004
		(0.004)	(0.112)	(0.027)	(0.005)

Notes: First-difference regressions on the balanced panel with HC1 standard errors in parentheses. γ and b are from $\Delta S_j = \gamma + b\Delta D_j + u_j$. The change in slope τ is the coefficient on February delegation in $\Delta S_j = \phi + \beta\,\Delta D_j + \tau D_{j,\text{Feb}} + u_j$, the first-difference form of equation (12). Conversation weights are the two-week average conversation counts $\bar{n}_j$, and the ten largest tasks are those with the largest $\bar{n}_j$.

Table D.2: Composition Controls in the Cross-Wave Regression

	Equal weights		Conversation weights	
Controls added to $\Delta S_j = \gamma + b\Delta D_j + u_j$	γ	b	γ	b
None	0.042	−0.259	0.036	−0.356
	(0.002)	(0.047)	(0.003)	(0.078)
Δ work share, Δ coursework share	0.049	−0.274	0.041	−0.375
	(0.004)	(0.045)	(0.004)	(0.071)
Also Δ human-only ability, Δ multitasking share,	0.052	−0.275	0.041	−0.358
Δ log human-only time	(0.004)	(0.041)	(0.004)	(0.059)

Notes: Balanced panel of 883 tasks, first differences, HC1 standard errors in parentheses. Conversation weights are the two-week average conversation counts $\bar{n}_j$. The personal share is omitted because the work, coursework, and personal shares sum to about one.

tions. Binomial sampling noise accounts for 30 and 15 percent of the cross-task variance of ΔS_j and for 33 and 17 percent of that of ΔD_j, with equal and conversation weights. If, within a task and week, delegated conversations are judged successful less often, sampling alone makes ΔS_j and ΔD_j co-move negatively. The within-task joint distribution of the success label and the interaction pattern is not published, so this sampling-induced co-movement cannot be removed, and b mixes co-movement of task means across weeks with the conversation-level association inside task-weeks.

Weekly cross-task lines. Part (c) of Proposition 1 concerns each week's least-squares line of performance on delegation across a fixed set of tasks with fixed weights. On the balanced panel, with the two-week averages $\bar{n}_j$ as conversation weights, the slope of this line is -0.403 in November and -0.399 in February with equal weights and -0.419 and -0.441 with conversation weights, and the February line lies 0.041 higher with equal weights and 0.046 higher with conversation weights at $D = 0$. The line therefore shifted up while its slope changed little, the pattern that part (c) gives for a common improvement, although the pattern does not establish one. These lines describe the cross-task relationship of Section 4 in each week. They are a different estimand from the within-task profiles of Section 5.2, which compare each task with itself, and their vertical distance is not the February shift of Table 7.

D.2 Nonlinear delegation–performance specifications

I estimate quadratic and linear-spline versions of the within-task regression in Section 5.2. Both specifications allow the delegation–performance profile to differ between weeks. The spline has one knot at $D = 0.5$. Both are estimated in first differences on the balanced panel, which give the task-fixed-effects estimates. Table D.3 reports the slopes and the February shifts at points of the common support, with standard errors for the spline slopes and the shifts. The curvature terms are jointly insignificant in every version. The November slope above the knot differs from the slope below it by 9 and 31 percent of the latter, with p-values of 0.84 and 0.50. The February shift is flat in delegation with equal weights and 0.018 to 0.020 larger at $D = 0.8$ than at $D = 0.2$ with conversation weights, the weak flattening of Section 5.2. The checks give no evidence that the profile is nonlinear or that February's gains are concentrated in a region of delegation.

Table D.3: Quadratic and Spline Profiles with Task Fixed Effects

	Weights	
	Equal	Conversation
Panel A: linear spline, knot at $D = 0.5$		
November slope below the knot	-0.283	-0.323
	(0.043)	(0.061)
November slope above the knot	-0.258	-0.424
	(0.115)	(0.138)
February slope below the knot	-0.248	-0.306
	(0.042)	(0.060)
February slope above the knot	-0.269	-0.378
	(0.103)	(0.142)
p-value: November slope change at the knot	0.84	0.50
p-value: linear profile in both weeks	0.71	0.74
February shift at $D = 0.2$	0.036	0.029
	(0.003)	(0.004)
February shift at $D = 0.5$	0.047	0.034
	(0.004)	(0.006)
February shift at $D = 0.8$	0.043	0.048
	(0.012)	(0.009)
Panel B: quadratic		
November slope at $D = 0.2$	-0.277	-0.331
November slope at $D = 0.5$	-0.273	-0.374
November slope at $D = 0.8$	-0.268	-0.418
February slope at $D = 0.2$	-0.233	-0.329
February slope at $D = 0.5$	-0.260	-0.342
February slope at $D = 0.8$	-0.288	-0.354
p-value: November curvature	0.96	0.68
p-value: linear profile in both weeks	0.64	0.76
February shift at $D = 0.2$	0.037	0.030
	(0.003)	(0.004)
February shift at $D = 0.5$	0.045	0.035
	(0.003)	(0.004)
February shift at $D = 0.8$	0.044	0.050
	(0.012)	(0.009)
Panel C: February shift at $D = 0.8$ minus at $D = 0.2$		
Linear	0.012	0.018
	(0.009)	(0.009)
Quadratic	0.007	0.020
	(0.012)	(0.010)
Spline	0.007	0.019
	(0.012)	(0.009)

Notes: Balanced panel of 883 tasks, first differences with HC1 standard errors in parentheses. Conversation weights are the two-week average conversation counts $\bar{n}_j$. Slopes are the derivative of the fitted profile with respect to delegation. The February shift is the fitted February profile minus the November profile at the given value of D. The joint tests are F tests that both weeks' profiles are linear, that is, that the slope changes at the knot or the squared terms are zero in both weeks.

E Correlates of Delegation

This appendix reports the estimates behind Section 6.1. Appendix E.1 gives the full estimates of equation (13) and their robustness checks. Appendix E.2 relates delegation and performance to task characteristics that O*NET measures independently of Anthropic's classifiers.

E.1 Full estimates and robustness checks

Full estimates. Table E.1 reports the standardized coefficients of equation (13) for performance and the five measures, the coefficient of performance alone, and the R^2 of the nested specifications.

Robustness checks. Four measures have coefficients with the same sign that are significant at 5 percent, with HC1 standard errors, in both weeks and under both weightings: performance, human-only ability, log human-only time, and the work share. I check whether these associations rest on one O*NET category by leaving out, in turn, each of the four categories into which O*NET sorts the GWAs, Information Input, Mental Processes, Work Output, and Interacting With Others. The check for a category drops the tasks whose largest share of DWA links falls in that category, including those in which another category has an equally large share. The four associations keep their sign in all 24 check fits: the four leave-one-category-out samples, the tasks linked to a GWA, and the 883 balanced-panel tasks within each week, with conversation weights given by that week's conversation counts, each in both weeks and under both weightings. Performance, human-only ability, and human-only time also keep 5 percent significance in every check. The work share does so in every check with equal weights but in only four of the 12 with conversation weights, where its p-values reach 0.32. Standard errors clustered by occupation use the first O*NET-SOC occupation listed for each task, which gives 391 clusters in November and 407 in February. With these standard errors, performance, human-only ability, and human-only time keep $p < 0.001$ in every cell, and the work share keeps $p < 0.001$ with equal weights, whereas with conversation weights its p-values are 0.031 in November and 0.058 in February.

E.2 Task characteristics measured by O*NET

O*NET characteristics. This subsection asks how delegation and performance relate to task characteristics that O*NET measures independently of Anthropic's classifiers (Peterson et al., 2001; Hansen et al., 2014; National Center for O*NET Development,

Table E.1: Correlates of Delegation Conditional on Performance, Full Estimates

	November 2025		February 2026	
Standardized measure	Equal	Conversation	Equal	Conversation
Performance	−0.079	−0.077	−0.102	−0.095
	(0.005)	(0.007)	(0.005)	(0.007)
Human-only ability	−0.024	−0.072	−0.022	−0.075
	(0.005)	(0.007)	(0.006)	(0.007)
Log human-only time	−0.078	−0.053	−0.080	−0.060
	(0.007)	(0.007)	(0.007)	(0.006)
Work share	0.037	0.023	0.032	0.016
	(0.006)	(0.010)	(0.006)	(0.007)
Multitasking share	−0.010	−0.013	−0.004	−0.004
	(0.006)	(0.006)	(0.005)	(0.005)
Coursework share	0.021	0.015	0.008	0.008
	(0.005)	(0.009)	(0.005)	(0.007)
Performance alone	−0.079	−0.108	−0.103	−0.118
	(0.006)	(0.016)	(0.005)	(0.011)
R^2: performance only	0.21	0.34	0.31	0.41
R^2: all six measures, no GWA shares	0.40	0.65	0.46	0.67
R^2: plus GWA shares	0.53	0.77	0.58	0.77
Adjusted R^2: plus GWA shares	0.51	0.76	0.56	0.76
Tasks	947	947	1,023	1,023

Notes: Least-squares estimates of equation (13) by week on the main samples. Except in the performance-alone row and the first two R^2 rows, every regression also includes the task's GWA shares, with Thinking Creatively omitted, and an indicator for tasks without a DWA link. One task per week is the only task with a positive share in its GWA and is fit exactly, which leaves the other coefficients unaffected. Performance and the five measures are standardized within each week and weighting, so coefficients are differences in delegation per standard deviation. Standard errors in parentheses are HC1.

2015). I use the eight O*NET 20.1 characteristics that Table C.6 uses as controls. Unlike equation (13), each regression has standardized delegation or standardized performance as its outcome and the eight standardized characteristics as its only regressors. I estimate it within each week, with equal weights and with conversation weights, with HC3 standard errors. I classify an association as robust if it has the same sign, is significant at 5 percent, and has an absolute value of at least 0.10 in both weeks under both weightings. Table E.2 reports the estimates. One characteristic is robustly associated with delegation, and none with performance. Tasks whose DWAs involve more interaction with other people are less delegated, by 0.330 to 0.463 standard deviations per standard deviation of the share. Explanation-seeking conversations are likely to be assigned to teaching and advising tasks, so part of this association may be definitional. With standard errors clustered by occupation, the association remains significant at 5 percent, and it survives the exclusion of computer occupations under both weightings. The eight characteristics account for 8 percent of the variance of performance with equal weights, against 21 to 24 percent for delegation.

Table E.2: O*NET Characteristics, Delegation, and Performance

	November 2025		February 2026	
	Equal	Conversation	Equal	Conversation
Panel A: delegation				
Task level				
Interacting-with-others share	−0.370	−0.463	−0.330	−0.418
	(0.038)	(0.091)	(0.038)	(0.088)
Mental-processes share	−0.082	0.013	−0.125	−0.070
	(0.037)	(0.184)	(0.035)	(0.149)
Task frequency	0.046	0.111	0.021	0.105
	(0.035)	(0.067)	(0.033)	(0.063)
Occupation level				
Freedom to make decisions	−0.105	−0.185	−0.095	−0.185
	(0.038)	(0.102)	(0.038)	(0.084)
Repeating same tasks	−0.011	0.081	−0.021	0.064
	(0.039)	(0.082)	(0.039)	(0.076)
Being exact or accurate	0.044	−0.046	0.072	−0.010
	(0.037)	(0.103)	(0.037)	(0.083)
Consequence of error	−0.057	−0.019	−0.031	−0.007
	(0.033)	(0.078)	(0.037)	(0.072)
Job Zone	−0.269	−0.078	−0.283	−0.175
	(0.040)	(0.076)	(0.041)	(0.075)
R^2	0.24	0.36	0.21	0.33
Panel B: performance				
Task level				
Interacting-with-others share	0.171	0.262	0.201	0.180
	(0.044)	(0.083)	(0.043)	(0.124)
Mental-processes share	−0.030	−0.031	−0.004	0.039
	(0.043)	(0.138)	(0.039)	(0.106)
Task frequency	0.046	0.076	0.101	0.044
	(0.036)	(0.064)	(0.036)	(0.072)
Occupation level				
Freedom to make decisions	−0.008	0.104	0.010	0.102
	(0.043)	(0.084)	(0.045)	(0.092)
Repeating same tasks	−0.079	−0.191	−0.013	−0.164
	(0.047)	(0.098)	(0.045)	(0.089)
Being exact or accurate	−0.036	0.104	−0.065	0.059
	(0.045)	(0.105)	(0.042)	(0.096)
Consequence of error	0.028	−0.115	0.011	−0.048
	(0.038)	(0.072)	(0.043)	(0.090)
Job Zone	0.132	0.056	0.100	0.161
	(0.046)	(0.080)	(0.052)	(0.119)
R^2	0.08	0.18	0.08	0.14
Tasks	857	857	925	925
Occupations	367	367	380	380
Effective number of tasks		69		106

Notes: Each column of a panel is one regression of the standardized outcome on all eight standardized O*NET 20.1 characteristics of Table A.1, on the main-sample tasks with all eight defined. Coefficients are standard-deviation differences, with HC3 standard errors in parentheses. Conversation weights are the week's conversation counts, and the effective number of tasks is Kish's.