Working with AI: Measuring the Applicability of Generative AI to Occupations

Kiran Tomlinson[1,*], Sonia Jaffe[1,*], Will Wang[1], Scott Counts[2,*], and Siddharth Suri[1,*]

[1]Microsoft Research
[2]Microsoft
[*]Corresponding author: kitomlinson@microsoft.com, sojaffe@microsoft.com, counts@microsoft.com, suri@microsoft.com

Abstract

With generative AI emerging as a general-purpose technology, understanding its economic effects is among society's most pressing questions. Existing studies of AI impact have largely relied on predictions of AI capabilities or focused narrowly on individual firms. Drawing instead on real-world AI usage, we analyze a dataset of 200k anonymized conversations with Microsoft Bing Copilot to measure AI applicability to occupations. We use an LLM-based pipeline to classify the O*NET work activities assisted or performed by AI in each conversation. We find that the most common and successful AI-assisted work activities involve information work—the creation, processing, and communication of information. At the occupation level, we find widespread AI applicability cutting across sectors, as most occupations have information work components. Our methodology also allows us to predict which occupations are more likely to delegate tasks to AI and which are more likely to use AI to assist existing workflows.

Short title Measuring the Applicability of AI to Occupations

Teaser A dataset of 200k conversations with generative AI reveals which work activities—and by extension, occupations—the technology is most useful for.

1 Introduction

General purpose technologies (1), such as the steam engine and the computer, have historically been strong drivers of economic growth, impacting a broad range of sectors. In the last several years, generative artificial intelligence (AI) has come to the fore as the next general purpose technology (2, 3), capable of improving or speeding up tasks as varied as medical diagnosis (4) and software development (5). These capabilities are reflected in the astounding rate of AI adoption: nearly 40% of Americans report using generative AI, outpacing the early diffusion of the personal computer and the internet (6). Given this widespread adoption and potential for economic impact, a crucial question is *which* work activities AI is most useful for and which occupations perform them, as this will shape where any productivity benefits of AI fall.

We answer these questions by identifying the work activities performed with a mainstream large language model (LLM)-powered generative AI system, Microsoft Bing Copilot (now Microsoft Copilot). A key insight of our analysis is that there are two distinct ways in which a single user–AI conversation is relevant to work activities, corresponding to the two conversing parties. First, the user is seeking assistance with a task they are trying to accomplish; we call this the *user goal*. Analyzing user goals allows us to measure how generative AI is *assisting* different work activities. Second, the AI itself performs a task in the conversation, which we call the *AI action*. Classifying AI actions separately lets us measure which work activities generative AI is

performing. To illustrate the distinction, if the user is trying to learn how to print a document, the user goal is to operate office equipment, while the AI action is to train others to use equipment. This distinction allows to separately identify the types of work activities that workers may delegate to AI, shifting to focus to other tasks, from activities workers may perform in collaboration with AI.

To identify the occupations where AI might be most useful, we draw from a common economic methodology tracing its roots to Autor, Levy, and Murnane (7), which decomposes an occupation into its common tasks and aggregates task-level measurements to occupation-level metrics (3, 8–15). The O*NET database (16) decomposes occupations hierarchically into their work activities, and we use an LLM-based pipeline to map the user goals and AI actions in each conversation to the set of work activities being performed. We measure how successfully different work activities are assisted or performed by AI, using both explicit user feedback and a task completion classifier. To distinguish between broad and narrow AI contributions towards work activities, we also classify the scope of AI capability towards each work activity matched to a conversation. From these classifications, we compute an *AI applicability score* for each occupation. This score captures if there is a non-trivial amount of AI usage that successfully completes activities relevant to an occupation.

Our work connects to research predicting occupational AI impact using various proxies for AI capabilities, including expert forecasts (3, 8), progress on AI benchmarks (9), and AI patents (17). Before the advent of generative AI, similar techniques were used to predict the occupational impacts of robotics, machine learning, and computerization more broadly (7, 10–15, 18). We contribute to this literature by analyzing actual conversations between humans and an LLM, identifying which work activities those humans are using the LLM for and what the associated occupations are. The study most similar to ours is a recent analysis by Handa et al. (19) that classifies Claude conversations using the O*NET taxonomy, but focuses more narrowly on which tasks people use AI for. As a result, they classify each conversation into one O*NET *task*, which is occupation-specific and thereby prevents identification of cross-occupation AI capabilities. Our research goal is to go beyond usage to an estimate of occupational applicability of AI. Thus, we exhaustively classify a conversation into all relevant O*NET *work activities*, which apply across occupations. Furthermore, our measures of conversational success allow us to identify the efficacy of AI across work activities, and our user goal/AI action split allows us to identify how occupations may change in response to AI. More recent work using ChatGPT data adopted our approach of classifying by work activities, but does not use them to measure relevance for occupations (20).

2 Results

2.1 Work activities

We focus our analyses on *intermediate work activities* (IWAs) from the O*NET 29.0 Database (16) (see materials and methods for classification details). We do not know the occupation of a user, but if we observe a work activity being done with AI (regardless of context) then we know it is possible for AI to help with that kind of work. The most common Copilot user goals fall into four broad categories: learning, communicating, teaching/explaining, and writing—with learning and communicating being the most prominent (Figure 1A). The IWAs reflected in AI actions tell a complementary story. Figure 1B shows that the AI plays a service role: some common IWA verbs include *Provide*, *Explain*, *Teach*, *Assist*, and *Respond*. A majority of AI-side activity is devoted to communicating and teaching/explaining information to the user (Figure S2 and table S1 highlight additional differences between user goals and AI actions).

The conversation-level asymmetry between user goals and AI actions is pronounced: these sets of IWAs are disjoint in 40% of conversations, and in 96%, there are more IWAs unique to a side than in common. Furthermore, for every work activity matched to a user goal the AI performs an average of two work activities in service of the user's request (Figure S1).

In O*NET, IWAs are grouped into *generalized work activities* (GWAs); we use GWAs to identify how Copilot usage for work activities compares to the overall frequencies of work activities across the U.S. workforce. Figure 1C shows the activity shares in Copilot aggregated to 37 GWAs and a roughly estimated fraction of total U.S. labor falling under each GWA (described in the supplementary text). GWAs focused

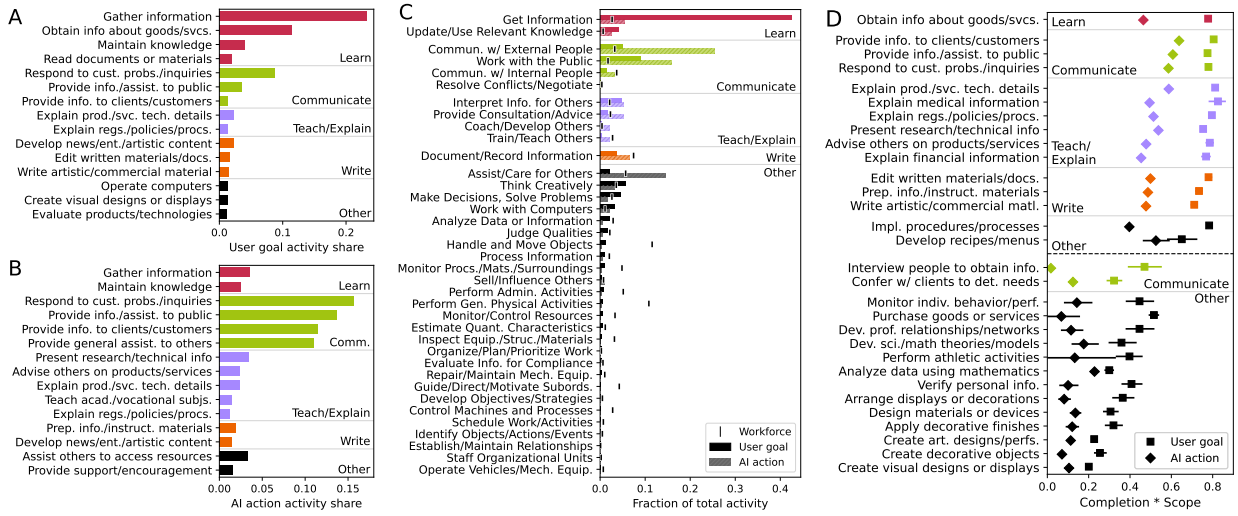


Figure 1: Frequency and success of work activities in Bing Copilot usage. (**A**) The 15 IWAs that occur most frequently as user goals. (**B**) The 15 IWAs that occur most frequently as AI actions. (**C**) The frequencies of GWAs as user goals and AI actions, compared to a rough estimate of GWA workforce frequency (described in the supplementary text). (**D**) The top and bottom 15 IWAs (with user goal or AI action activity share at least 0.05%) by the product of their completion rate and the share of conversations where Bing Copilot demonstrated at least a moderate scope of capability. All panels show researcher-generated IWA groupings to highlight patterns.

on learning and communication are most prevalent, echoing the results in Figure 1A,B. The GWAs more prevalent in Copilot usage than in the workforce include *Getting Information*, *Updating and Using Knowledge*, *Communicate with External People*, and *Interpreting Information for Others*. These align with *information work* (21, 22), concerning the creation, processing, and communication of information, and the related concept of *knowledge work*, which is typically refers to work requiring non-routine and complex problem-solving (23–27). Our data show that the user goals that people use Copilot for, and the activities AI performs in service of those goals, center around information work, and these activities occur disproportionately often relative to the fraction of information work in the workforce.

2.1.1 Metrics of success

Beyond identifying which activities generative AI is being used for, we have three measures of its efficacy: task completion, user feedback, and scope. *Completion* is an LLM-based evaluation of whether the AI completed the user's goal in the conversation validated by real-world *user feedback*, in the form of a thumbs up or down (see materials and methods and Figures S3 to S7). *Scope* is an LLM-based evaluation of the fraction of work in an IWA that Copilot demonstrates the ability to assist or perform, measured via a 6-point ordinal scale (none, minimal, limited, moderate, significant, complete; see materials and methods and Figures S8 and S9). For each IWA, we measure the share of conversations for which the scope is moderate or higher.

Figure 1D shows the top and bottom IWAs ranked by the product of completion rate and scope share. Copilot performs best at activities related to communicating, teaching/explaining, and writing, largely corresponding to the most frequent AI actions in Figure 1B. One the other hand, it performs worse at image generation and data analysis tasks (these patterns are confirmed by thumbs feedback; Figures S3 and S4). Scope in particular is highly correlated with (log) share of user activity ($r = 0.64$; Figure S10). This suggests that users have discovered and are exploiting the informational tasks where AI can have broadest impact. Finally, AI action completion $\times$ scope is consistently lower than on the user goal side, in Figure 1D, indicating that AI can help users with a broader fraction of their work than it can perform directly.

Table 1: **SOC minor groups by AI applicability score.**

Minor Group Title (Abbr.)	Score	Minor Group Title (Abbr.)	Score	Minor Group Title (Abbr.)	Score
Media and Communication Workers*	0.38	Food and Beverage Serving Workers**	0.20	Woodworkers	0.12
Sales Representatives, Services**	0.35	Other Educational and Library Workers*	0.19	Top Executives**	0.11
Information and Record Clerks**	0.33	Material Recording and Distrib. Workers**	0.19	Construction and Extraction Sups.*	0.11
Mathematical Science Occupations	0.32	Prim., Second., and Special Ed. Teachers**	0.18	Assemblers and Fabricators*	0.11
Tour and Travel Guides	0.32	Media and Comms. Equipment Workers	0.18	Printing Workers	0.11
Postsecondary Teachers*	0.31	Advertising, Mktg., PR, and Sales Managers*	0.18	Health Technologists and Technicians**	0.10
Sales Reps., Wholesale and Manufacturing*	0.31	Air Transportation Workers	0.18	Metal Workers and Plastic Workers*	0.10
Communications Equipment Operators	0.30	Lawyers, Judges, and Related Workers*	0.17	Vehicle and Mobile Equipment Mechs.*	0.10
Baggage Porters, Bellhops, and Concierges	0.30	Transportation and Material Moving Sups.*	0.17	Other Maintenance Workers**	0.10
Retail Sales Workers**	0.30	Architects, Surveyors, and Cartographers	0.17	Water Transportation Workers	0.09
Other Sales and Related Workers	0.30	Electrical and Electronic Equipment Mechs.	0.17	Other Production Occupations**	0.08
Computer Occupations**	0.29	Other Healthcare Pracs. and Tech. Occs.	0.16	Bldg. Cleaning and Pest Ctrl. Workers**	0.08
Personal Care and Service Sups.	0.27	Sports Entertainers and Related*	0.16	Firefighting and Prevention Workers	0.07
Entertainment Attendants and Related*	0.27	Other Personal Care and Service Workers*	0.16	Protective Service Sups.	0.07
Religious Workers	0.26	Cooks and Food Preparation Workers**	0.15	Material Moving Workers**	0.07
Social Scientists and Related	0.26	Funeral Service Workers	0.14	Construction Trades Workers**	0.07
Librarians, Curators, and Archivists	0.25	Other Management Occupations**	0.14	Personal Appearance Workers*	0.06
Counselors, Social Workers, and Related**	0.25	Building, Grounds, and Maintenance Sups.	0.14	Agricultural Workers	0.06
Supervisors of Production Workers*	0.25	Other Protective Service Workers*	0.14	Other Construction and Related Workers	0.06
Other Office and Admin. Support Workers**	0.25	Operations Specialties Managers**	0.14	Legal Support Workers	0.06
Office and Administrative Support Sups.*	0.25	Occupational Health and Safety Specialists	0.14	Other Healthcare Support Occupations*	0.06
Financial Clerks**	0.24	Supervisors of Sales Workers*	0.14	Helpers, Construction Trades	0.06
Secretaries and Administrative Assistants**	0.24	Law Enforcement Workers*	0.13	Farming, Fishing, and Forestry Sups.	0.05
Business Operations Specialists**	0.24	Life, Phys., and Social Science Tech'ns	0.13	Occupational and Physical Therapy Assts.	0.05
Animal Care and Service Workers	0.24	Motor Vehicle Operators**	0.13	Textile, Apparel, and Furnishings Workers	0.05
Financial Specialists**	0.23	Healthcare Diagnosing or Treating Pracs.**	0.13	Extraction Workers	0.05
Other Teachers and Instructors*	0.23	Installation, Maintenance, and Repair Sups.*	0.13	Home Health Aides and Nursing Assts.**	0.04
Engineers*	0.22	Rail Transportation Workers	0.12	Grounds Maintenance Workers*	0.04
Physical Scientists	0.21	Other Food Prep. and Serving Workers*	0.12	Plant and System Operators	0.04
Drafters and Eng. and Mapping Tech'ns*	0.21	Food Preparation and Serving Sups.*	0.12	Forest and Conservation Workers	0.03
Life Scientists	0.20	Food Processing Workers*	0.12		
Art and Design Workers*	0.20	Other Transportation Workers	0.12		

Score is the employment-weighted average AI applicability score for each specific occupation in the SOC minor group, averaging the mean of the user goal and AI action scores. Asterisks indicate U.S. employment (none: 0 to 500k, one: 500k to 2M, two: greater than 2M). Minor groups in black have a majority of workers performing information work (classified as described in the supplementary text); minor groups in gray have only a minority. Titles have been abbreviated for space (Sups: Supervisors, Pracs: Practitioners).

2.2 AI applicability to occupations

Using O*NET task importance and relevance as weights, we aggregate the IWA-level measures into an *AI applicability score* for each occupation in the Standard Occupational Classification (SOC) (28). The score combines whether Copilot assists or performs an occupation's associated work activities (frequency $\geq 0.05\%$) successfully (completion rate) and in ways that cover a broad share of the work activity (scope $\geq$ moderate). We compute AI applicability score separately for user goals and AI actions, averaging the two when we need a single summary metric. (See materials and methods for a full definition.) Treating all work activities with activity share of at least 0.05% in the same way mitigates potential distributional bias from considering only consumer Copilot data. The use of a frequency threshold means relative comparisons are more meaningful than absolute score values (see materials and methods and Figures S19 and S20).

The use of AI for information work activities observed in Figure 1 translates directly to the occupations where AI is most applicable. Figure 2 shows the 25 occupations with the highest AI applicability score and the work activities that contribute the most to those occupations' scores. The top IWAs involve the creation (e.g., *Edit written materials, Write commercial material*), gathering (e.g., *Maintain knowledge*), and dissemination (e.g., *Respond to customer inquiries, Provide information to customers*) of information. In fact, almost all of the top-contributing IWAs fall into one or more of these categories. These IWAs contribute to the AI applicability scores of a number of occupations involving information creation and processing, such as Interpreters and Translators, Writers, Editors, and Data Scientists. Occupations involving the communication and dissemination of information also stand out with high AI applicability, such as Sales Representatives, Customer Service Representatives, and Concierges.

To get a broader view of all occupations in the economy, Table 1 reports the AI applicability of all SOC minor groups, highlighting occupational groups containing a majority of workers in occupations that consist primarily of information work (the supplementary text describes how we classify occupations as information

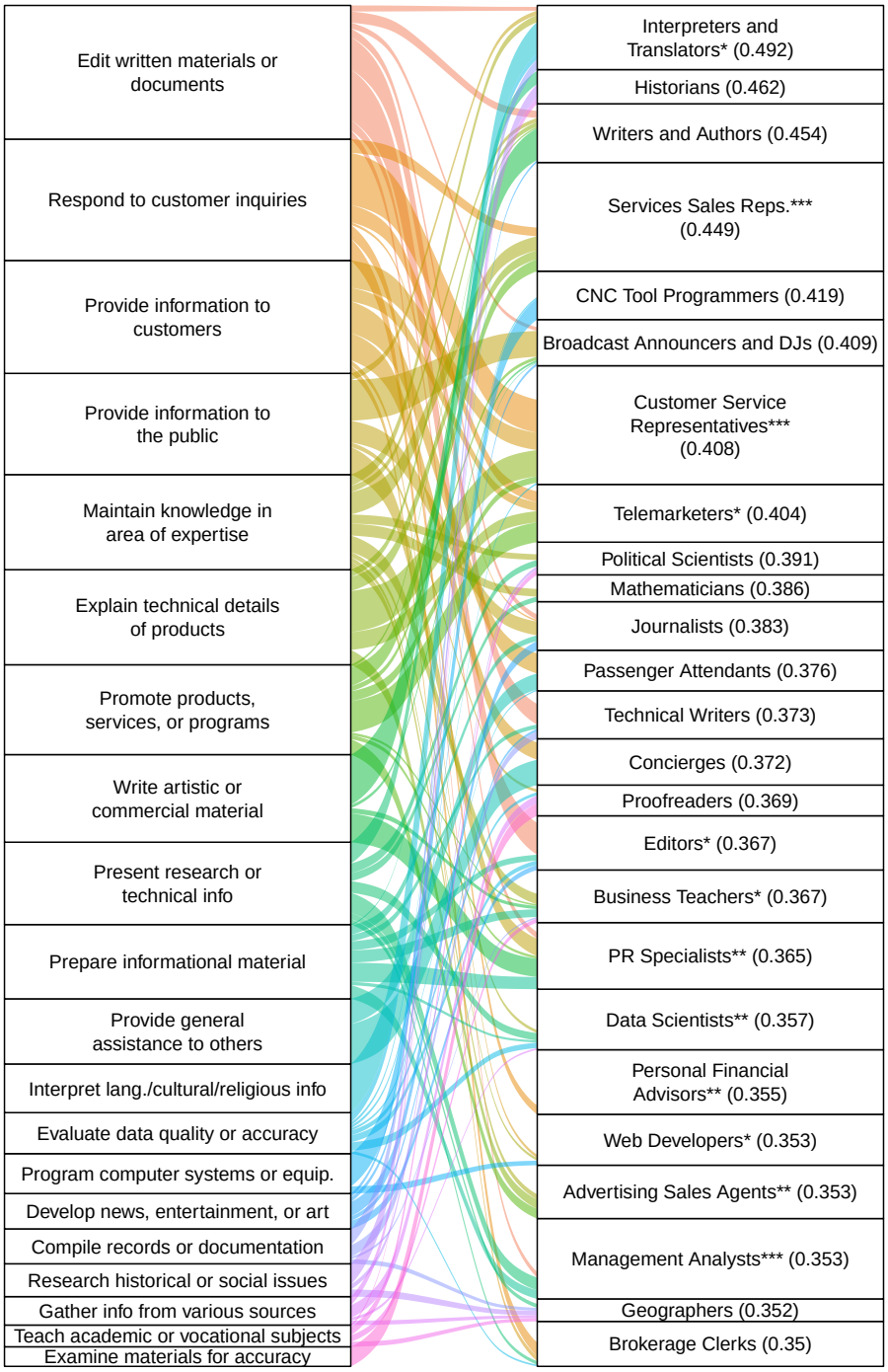


Figure 2: Top occupations by AI applicability score and their contributing IWAs. The 25 SOC occupations with the highest AI applicability scores (right) along with the 20 IWAs that provide the greatest contributions to those scores (left). Asterisks indicate 2023 U.S. occupational employment (none: up to 50k, one: 50k to 100k, two: 100k to 500k, three: more than 500k); the heights of each stratum indicate the same. Occupations' applicability scores are in parentheses, and occupations are sorted by their score in decreasing order. Portions of the occupational strata not connected to an IWA by a colored flow represent IWAs not present in the Figure that still contribute to the occupation's applicability score. Both occupation and IWA titles have been shortened for space.

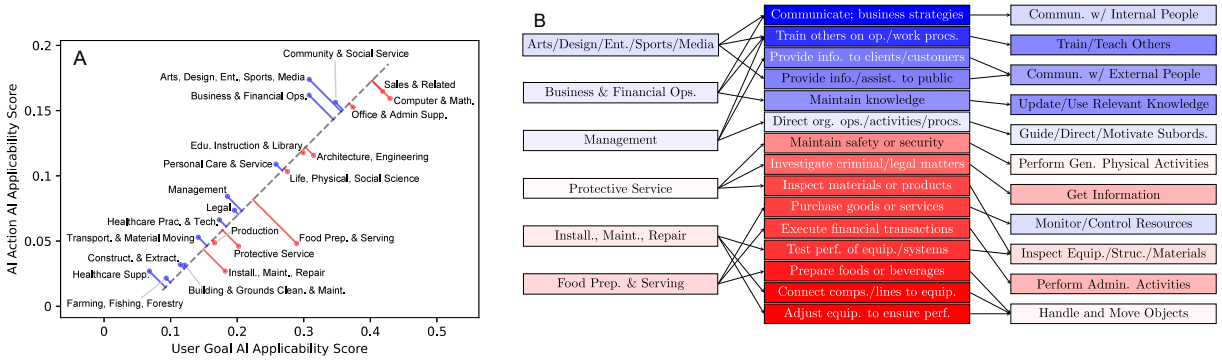


Figure 3: Differences between occupations' user-goal and AI-action applicability and the contributing work activities. (**A**) Scatterplot of each major occupation group's (employment-weighted) average AI action applicability score and average user goal applicability score. The dashed gray line shows the first principal component, capturing overall applicability. The blue and red lines show the extent to which a major group has higher applicability from one metric versus the other (based on the second principal component, which captures the extent to which, compared to the overall relationship between user score and AI score, the score skews more toward the red user-goal or blue AI-action side). (**B**) A tripartite graph showing six prominent occupation groups that are skewed toward user goal or AI action applicability (left), the three IWAs contributing the most to each group's skew (middle) ordered by the extent to which those IWAs skew more to the AI action side, and the corresponding GWAs. As in (A), blue indicates a larger skew toward AI action applicability, while red indicates a larger skew toward user goal applicability. White is neutral. See materials and methods for further details.

work). The top groups are mostly information workers, with Media and Communication, Information and Record Clerks, and Sales Representatives of Services at the very top. The groups with the lowest scores include occupations that require physically working with people, operating machinery, and other manual labor. However, we find that most occupations have at least some AI applicability, reflecting that most work has some information processing component (21).

At a high level, these observations of AI applicability based on real-world usage align with predictions of AI capabilities. Eloundou et al. (3) used six annotators to predict the share of tasks within each occupation for which access to an LLM alone would lead to time savings. The employment-weighted occupation-level correlation between their predictions and our AI applicability score is $r = 0.73$ (Figure S12). But our data do not fully confirm expectations: for instance, we find only a weak relationship between AI applicability score and wage (employment-weighted $r = 0.13$; Figure S13A), while previous work found strong positive correlations (3, 9) (some of this may be due to a lack of employment-weighting in prior work; see the supplementary text). Similarly, we find a broad range of AI applicability across educational requirements (Figure S13B). Beyond wage and education, AI may have differential effects across demographic groups to the extent that some demographics are more or less likely to be employed in AI-applicable occupations (Figure S18).

2.2.1 User goals and AI actions

Our approach of separating user goals from AI actions allows us to not only identify where AI is most useful, but also separate out the different ways AI may change the way occupations do their work. There are at least two different ways occupations could change as a result of AI: (1) delegating some tasks to AI, freeing up workers to focus on different parts of the job, corresponding to high AI action applicability, and (2) performing the same tasks but in collaboration with AI, corresponding to high user goal applicability score. Figure 3A plots the average AI-action and user-goal applicability scores for each major SOC group,

highlighting occupational groups that lean towards either side of this dichotomy. The lengths of the red (resp. blue) lines indicate the extent to which the occupations in that group are more associated with user goals (resp. AI actions); these are derived from the second principal component of the user goal and AI action applicability scores (see materials and methods). To see which activities are driving these differences, Figure 3B takes three of the groups with the largest difference in each direction (left column), maps them to the IWAs that contribute most to that difference (middle column), and then shows which GWAs those IWAs are part of (right column). Figure 3 indicates that media and financial operations occupations are more likely to delegate tasks to AI, such as those involving communication and training. On the other hand, to the extent that AI is used for more physical occupations like those in the food preparation and serving category, it is likely to be in a more assistive role.

3 Discussion

This work provides an early signal indicating where and how AI is likely to change the way people work. Our data shows that people use AI to do a variety of work activities related to the creation, analysis, and processing of information, i.e., information work. Most occupations have at least some information work component, reflected in our finding that AI is somewhat relevant to most occupations. This helps to explain the rapid rate of adoption of AI in the workplace (6). Our analysis also highlights the different ways AI may be used in different roles. Occupations with high AI applicability to user goals, such as computer and mathematical occupations, may see more active AI usage by workers as they use AI to improve existing workflows. Meanwhile, occupations with high AI action applicability, like those in business and financial operations, may delegate tasks to AI and focus more on the parts of their jobs that require a human touch.

Understanding which occupations have tasks within the current frontier of AI capability is important because these jobs are most likely to see productivity boosts or a shift in their core tasks. In addition, understanding the current frontier of AI applicability allows system designers to investigate where AI is currently falling short and see if it is due to an addressable technical shortcoming (such as low satisfaction for data analysis), a mismatch in modality (such as moving objects), or a societal factor (such as laws or regulations). Looking ahead, as AI systems gain functionality and adoption continues to evolve, we can use our AI applicability scores to see how the AI frontier shifts over time.

Our results show that AI is most applicable to information work occupations and often plays the role of an advisor or teacher, providing expertise to the user. In the same way that word processors turned typing into a widespread skill rather than a core task for typists and secretaries, AI too may democratize skills around information work. To the extent that people can successfully apply AI expertise, shrinking the performance gap between low- and high-skilled workers (5, 29–31), this provides some empirical backing to Autor's theoretical argument (32) that AI could reduce income inequality: if AI can broaden access to expertise, and people have sufficient foundational knowledge to apply and evaluate AI assistance, then AI could provide more workers access to previously niche expert work, potentially shrinking economic inequality.

Our user goal versus AI action distinction relates to a key question in the literature and public discourse around AI: to what extent is AI automating versus augmenting work activities? It is tempting to conclude that occupations that have high AI action applicability score will be automated and thus experience job or wage loss, and that occupations with high user goal applicability score will be augmented and raise wages. This would be a mistake, as downstream consequences of new technologies are very hard to predict and often counterintuitive (33). Take the example of ATMs, which automated a core task of bank tellers, but led to an *increase* in the number of bank teller jobs as banks opened more branches at lower costs and tellers focused on more valuable relationship-building rather than processing deposits and withdrawals (34). Our measures of AI applicability indicate where AI is likely to change work, which is a part of the automation vs. augmentation question, but AI's effects on employment and wages will depend on hard-to-predict business decisions.

There are some natural limitations to the conclusions that can be drawn from our data. Our task completion and scope metrics are measures of AI's utility towards an IWA, but they do no capture the productivity impact on people doing that IWA. Another gap is the difference between the way work activities

are performed in occupations compared to in our data (for instance, *Provide general assistance* means something different for a passenger attendant and for Copilot). Our data also represents only one slice of the AI market: there are many other AI platforms, including more task- or occupation-specific LLMs, which are not represented in our data. While decomposing an occupation into its work activities is standard practice in the literature, it does not provide a complete representation of every occupation: the connecting glue between tasks also contributes to the value of work. Finally, our use of O*NET means our results are shaped by its U.S.-centric view, may lag behind current workplace activities, and do not capture valuable tasks performed outside of occupations (e.g., work in the home or volunteering). Modernizing our understanding of workplace activities will be crucial as generative AI continues to change how work is done.

This work gives rise to a number of future research questions of high importance to society. We measured how AI capabilities overlap with work activities, but it remains to be seen how different occupations refactor their work responsibilities in response to AI's rapid progress. In addition, entirely new occupations may emerge, performing new types of work activities (35, 36). This is not a new phenomenon: the *majority* of employment today is in occupations that arose in the last 100 years as a result of new technologies (37). Which new jobs emerge, and how old ones are reconstituted, is a crucial future research direction.

Information work has become central to the global economy, embedded in the majority of occupations. Technological advances of the last several decades, such as the computer, the internet, and the search engine, have provided valuable tools for storing, processing, and finding information more efficiently. Our results highlight how LLMs can contribute to broader parts of the information life cycle—including creation, interpretation, and communication, in more flexible ways than earlier technologies. As generative AI continues to progress, it will be important to map the frontier of AI functionality to understand what additional capabilities AI is providing people and how that new functionality can be applied in the workplace.

4 Materials and methods

4.1 Bing Copilot data

We analyze anonymized and privacy-scrubbed U.S. conversation data from Microsoft Bing Copilot (henceforth, Copilot) gathered from January 1, 2024 to September 30, 2024. We focus only on conversations in the United States to align with occupation and work activity information from O*NET. Our main dataset is a uniform sample of approximately 100k conversations, providing a representative view of what tasks users perform with a mainstream, publicly available, free-to-use generative AI chatbot. We use a supporting dataset of 100k conversations uniformly sampled from those that received at least one thumbs up or thumbs down reaction from the user. This dataset allows us to investigate what activities are performed more or less successfully, as measured by explicit user feedback. Our use of Copilot data was reviewed and approved by the Microsoft IRB (under ID # 11028). While we cannot release conversational data due to legal and privacy obligations, all of our aggregated IWA- and occupation-level metrics have been made public on GitHub (`https://github.com/microsoft/working-with-ai`).

4.2 O*NET and Bureau of Labor Statistics data

We use O*NET's (16) hierarchical decomposition of occupations into their tasks and work activities. Our analysis focuses on *intermediate work activities* (IWAs), which map to multiple occupations through tasks. We combine O*NET with data on wages and employment from the 2023 Occupational Employment and Wage Statistics data published by the U.S. Bureau of Labor Statistics (BLS) (38). We use task frequency and relevance data from O*NET to get a rough frequency of how often each IWA is done by the U.S. labor force (details in the supplementary text). We also use wage, education, and demographic data from the Current Population Survey (CPS) data from 2024 (39), which is gathered by the U.S. Census Bureau for the BLS. All of our occupational analysis is performed on occupations from the 2018 Standard Occupational Classification (SOC) (28). The supplementary text describes how we map the O*NET and CPS occupational

taxonomies to SOC codes and the supplementary text describes which SOC codes we drop due to missing data.

4.3 Work activity classification

For each conversation in our datasets, we use an LLM classification pipeline to identify *all* IWAs that match the user goal and the AI action. In most conversations, there are multiple IWA matches on both sides (on average, about 3 user goal IWAs and 6 AI actions IWAs; see Figure S1). If the user goal or AI action is not related to any work activity, then it should be matched with zero IWAs. All prompts, pipeline architecture, model details, and human validation metrics are provided in the supplementary text.

4.3.1 Activity share and coverage

Since each conversation can be assigned multiple IWAs, instead of counting raw IWA occurrences in our conversation data, we focus on each IWA's *activity share*, where we allocate an equal fraction of each conversation to its labeled IWAs, separately on the user and AI sides. We consider the work activities that appear at non-trivial frequencies in Copilot chat data to be "covered." We define non-trivial usage using a threshold of 0.05% activity share and use this as a signal that AI can potentially assist or perform that IWA. We chose the threshold 0.05% to minimize the number of occupations with no or all IWAs covered, thereby maximizing the usefulness of the measure for relative comparisons between occupations (Figure S19A). The ordering of occupations induced by our AI applicability score is robust to the chosen coverage threshold (Figure S19B). We define an occupation's *coverage* to be the weighted fraction of its work activities that are covered (using the occupation–IWA weights defined in Section 4.4.1; see Figures S21 and S22).

4.3.2 Completion and Scope

We use an conversation-level LLM completion classification (details and prompts in the supplementary text), to measure the share of conversations labeled with each IWA where the AI completed the user's task in the conversation. Completion is highly correlated with direct user feedback (weighted $r > 0.75$; Figure S5), and has the advantage of being available for all conversations, so it is not subject to the selection bias of thumbs feedback.

For each relevant IWA in a conversation, we also perform an LLM classification of the fraction of work in the IWA that Copilot demonstrates the ability to assist or perform, which we call the *scope*, measured on a six-point ordinal scale: none, minimal, limited, moderate, significant, complete. The goal of scope is to distinguish between cases where Copilot assists with a large or small fraction of the work in an IWA. Scope classification and validation are detailed in the supplementary text.

4.4 Occupational AI applicability score

4.4.1 Aggregating from IWAs to occupations

We compute a weight for every task using the importance and relevance scores in O*NET, which allows us to capture which work activities are more or less important for each occupation. For each task k in SOC occupation i, we say $\text{weight}_{ik} = 2^{\text{importance}_{ik}} \cdot \text{relevance}_{ik}$. We sum weights for tasks mapping to the same IWA. When a task maps to multiple IWAs, its weight contributes to them equally. If an occupation has no ratings for any of its tasks, we assign them all weight $w_{ik} = 1$. If an occupation has ratings for only some of its tasks, we ignore the tasks with missing ratings. Dividing IWA weights by the total weight for an occupation then gives us a proxy measure for how much of a job consists of each of its IWAs. We use w_{ij} to denote this normalized weight of IWA j in job i.

In addition, we also compute a second IWA weight w_{ij}^{nonphys} to account for the fact that some IWAs like *Provide general assistance* map to both physical and nonphysical tasks, but applicability to the IWA does not imply AI action applicability to physical tasks. To compute w_{ij}^{nonphys}, we use GPT-5 to label every O*NET task according to whether it requires touching or moving people or objects, validated by two human

annotators (prompt and validation details in the supplementary text). When summing task weights w_{ik} to get w_{ij}^{nonphys}, we only include those tasks labeled as nonphysical, but normalize by the weight of all tasks; this makes w_{ij}^{nonphys} as estimate of the fraction of an job i's work that is both nonphysical and described by IWA j.

4.4.2 AI applicability score

We aggregate IWA coverage, completion, and scope into an AI applicability score for each occupation. For the user-goal side,

$$a_i^{\text{user}} = \sum_{j \in \text{IWAs}(i)} \mathbf{1}[f_j^{\text{user}} \geq 0.0005] c_j^{\text{user}} s_j^{\text{user}} w_{ij}, \tag{1}$$

where IWAs(i) is the set of IWAs performed by occupation i, $f_j^{\text{user}}$is the user goal activity share of j, c_j^{user} is the task completion rate of conversations with IWA j as a user goal, and s_j^{user} is the fraction of conversations with user goal j in which the scope classification is moderate or higher. We define a_i^{AI} similarly for AI actions but using AI action IWA measures and, for a_i^{AI} only, using w_{ij}^{nonphys} instead of w_{ij}. We report $a_i = (a_i^{\text{user}} + a_i^{\text{AI}})/2$ unless otherwise specified.

Our approach contrasts with other research that uses a measurement (19) or prediction (3) of the fraction of occupations or of the workforce that have at least x% of their tasks impacted by AI. Such measurements cannot be made reliably from usage data alone, as the selected threshold for usage has a significant impact on the resulting numbers, whose apparent straightforwardness belies this issue. Figure S20 shows that by picking different usage thresholds, we can conclude that either ~ 0% or ~ 100% of the workforce has 50% of its importance-weighted tasks represented in our data, depending on whether we require 1% of chat activity for a task to be covered or only .01% of activity. As such, we believe it is much more meaningful to make relative statements about different kinds of occupations, which is what our AI applicability score is designed to do.

4.4.3 User goal vs. AI action applicability

We use principal component analysis (PCA) to measure the extent to which work activities or occupations are more successfully represented in user goals or AI actions. We either perform PCA on occupational AI applicability scores (as in Figure 3A) or on IWA applicability scores (as in Figure 3B), where an IWA j's (user goal) applicability score is $\mathbf{1}[f_j^{\text{user}} \geq 0.0005] c_j^{\text{user}} s_j^{\text{user}}$ (i.e., its contribution to Equation (1), before occupation-specific weighting). As the user goal and AI scores are strongly correlated, the first principal component captures overall applicability, while the second perpendicular component measures skew towards AI actions or user goals. To identify which IWAs contribute most to an occupation group's skew in Figure 3B, we take the second component magnitudes in IWA-level PCA, multiply them by IWA weights w_{ij} for each occupation, and then take an employment-weighted average to get a score for each occupational group $\times$ IWA. IWA-level scores are also averaged to get GWA-level scores.

References

[1] Timothy F Bresnahan and Manuel Trajtenberg. General purpose technologies: "engines of growth"? *Journal of Econometrics*, 65(1):83–108, 1995.

[2] Avi Goldfarb, Bledi Taska, and Florenta Teodoridis. Could machine learning be a general purpose technology? a comparison of emerging technologies using data from online job postings. *Research Policy*, 52(1):104653, 2023.

[3] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308, 2024.

[4] Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R. Webster, Ewa Dominowska, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Jake Sunshine, Alan Karthikesalingam, and Vivek Natarajan. Towards accurate differential diagnosis with large language models. *Nature*, 642:451–457, 2025.

[5] Zheyuan Kevin Cui, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz. The effects of generative AI on high skilled work: Evidence from three field experiments with software developers. SSRN 4945566, 2024.

[6] Alexander Bick, Adam Blandin, and David J Deming. The rapid adoption of generative AI. NBER 32966, 2024.

[7] David H Autor, Frank Levy, and Richard J Murnane. The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, 118(4):1279–1333, 2003.

[8] Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. Future of work with AI agents: Auditing automation and augmentation potential across the U.S. workforce. arXiv:2506.06576 [cs.CY], 2025.

[9] Ed Felten, Manav Raj, and Robert Seamans. How will language modelers like chatgpt affect occupations and industries? arXiv:2303.01157 [econ.GN], 2023.

[10] Erik Brynjolfsson, Tom Mitchell, and Daniel Rock. What can machines learn, and what does it mean for occupations and the economy? *AEA Papers and Proceedings*, 108:43–47, May 2018.

[11] Carl Benedikt Frey and Michael A. Osborne. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114:254–280, 2017.

[12] James Manyika, Michael Chui, Mehdi Miremadi, Jacques Bughin, Katy George, Paul Willmott, and Martin Dewhurst. A future that works: Automation, employment, and productivity. Technical report, McKinsey Global Institute, 2017.

[13] Songül Tolan, Annarosa Pesole, Fernando Martínez-Plumed, Enrique Fernández-Macías, José Hernández-Orallo, and Emilia Gómez. Measuring the occupational impact of AI: tasks, cognitive abilities and AI benchmarks. *Journal of Artificial Intelligence Research*, 71:191–236, 2021.

[14] Edward W. Felten, Manav Raj, and Robert Seamans. A method to link advances in artificial intelligence to occupational abilities. *AEA Papers and Proceedings*, 108:54–57, May 2018.

[15] Michael Webb. The impact of artificial intelligence on the labor market. SSRN 3482150, January 2019.

[16] National Center for O*NET Development. O*NET Database Version 29.0, 2024. Accessed: 2025-05-29.

[17] Ali Akbar Septiandri, Marios Constantinides, and Daniele Quercia. The potential impact of AI innovations on US occupations. *PNAS Nexus*, 3(9):pgae320, 2024.

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

[19] Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. Which economic tasks are performed with AI? evidence from millions of Claude conversations. arXiv:2503.04761 [cs.CY], 2025.

[20] Aaron Chatterji, Tom Cunningham, Christopher Ong, Carl Shan, David Deming, Z"oe Hitzig, and Kevin Wadman. How people use chatgpt, September 2025.

[21] Isto Huvila. Ecological framework of information interactions and information infrastructures. *Journal of Information Science*, 35(6):695–708, 2009.

[22] Isto Huvila, Annemaree Lloyd, John M Budd, Carole Palmer, and Elaine Toms. Information work in information science research and practice. *Proceedings of the Association for Information Science and Technology*, 53(1):1–5, 2016.

[23] Peter Ferdinand Drucker. *Landmarks of tomorrow: a report on the new "post-modern" world*. Harper, New York, 1st edition, 1959.

[24] Pasi Pyöriä. The concept of knowledge work revisited. *Journal of Knowledge Management*, 9(3):116–127, 2005.

[25] C. McKercher and V. Mosco. *Knowledge Workers in the Information Society*. Lexington Books, 2008.

[26] Wolfgang Reinhardt, Benedikt Schmidt, Peter Sloep, and Hendrik Drachsler. Knowledge worker roles and actions—results of two empirical studies. *Knowledge and Process Management*, 18(3):150–174, 2011.

[27] Siddharth Suri, Scott Counts, Leijie Wang, Chacha Chen, Mengting Wan, Tara Safavi, Jennifer Neville, Chirag Shah, Ryen W. White, Reid Andersen, Georg Buscher, Sathish Manivannan, Nagu Rangan, and Longqi Yang. The use of generative search engines for knowledge work and complex tasks. arXiv:2404.04268 [cs.IR], 2024.

[28] U.S. Office of Management and Budget. *Standard Occupational Classification Manual, 2018*, 2017. Accessed: 2025-07-07.

[29] Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023.

[30] Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. The impact of AI on developer productivity: Evidence from GitHub copilot. *arXiv preprint arXiv:2302.06590*, 2023.

[31] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work. *The Quarterly Journal of Economics*, pages 889–942, 2025.

[32] David Autor. AI could actually help rebuild the middle class. *Noema Magazine* [Internet], 2 2024.

[33] David H Autor. Why are there still so many jobs? the history and future of workplace automation. *Journal of Economic Perspectives*, 29(3):3–30, 2015.

[34] James Bessen. Toil and technology: Innovative technology is displacing workers to new jobs rather than replacing them entirely. *Finance & Development*, 52(001):16, 2015.

[35] Daron Acemoglu and Pascual Restrepo. Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30, May 2019.

[36] Robert Capps. A.I. might take your job. Here are 22 new ones it could give you. *The New York Times Magazine* [Internet], 6 2025.

[37] David Autor, Caroline Chin, Anna Salomons, and Bryan Seegmiller. New frontiers: The origins and content of new work, 1940–2018. *The Quarterly Journal of Economics*, 139(3):1399–1465, 2024.

[38] U.S. Bureau of Labor Statistics. Occupational Employment and Wage Statistics (OEWS), May 2023, 2024. Accessed: 2025-10-17.

[39] U.S. Census Bureau and U.S. Bureau of Labor Statistics. Current population survey (cps). https://www.bls.gov/cps/tables.htm, 2024. Accessed 2025-08-18.

[40] Lawrence R James, Robert G Demaree, and Gerrit Wolf. Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, 69(1):85, 1984.

Acknowledgments

We thank Jennifer Neville, Ashish Sharma, Hancheng Cao, David Holtz, Carolyn Tsao, the Microsoft Research AI Interaction and Learning Group, and the Microsoft Research Computational Social Science Working Group for helpful discussions and feedback, and David Tittsworth, Jonathan McLean, Patrick Bourke, Nick Caurvina, and Bryan Tower for software and data engineering support.

Author contributions
Conceptualization: KT, SJ, SC, SS
Methodology: KT, SJ, SC, SS
Validation: KT, SC, SS
Formal Analysis: KT, SJ, WW
Data Curation: KT
Writing - Original Draft: KT, SJ, SC, SS
Writing - Review & Editing: KT, SJ, WW, SC, SS
Visualization: KT, SJ, SC, SS, WW

Competing interests KT, SJ, SC, SS are employed by Microsoft. WW is a contracted vendor for Microsoft.

Data and materials availability All occupation- and IWA-level metrics are available at https://github.com/microsoft/working-with-ai. Conversation texts cannot be made public due to legal and privacy obligations.

Supplementary Materials for Working with AI: Measuring the Applicability of Generative AI to Occupations

Kiran Tomlinson et al.
kitomlinson@microsoft.com

This PDF file includes:
Supplementary Text (Sections A–G)
Figures S1–S23 (Section H)
Tables S1–S5 (Section I)

A Merging O*NET with BLS and CPS data

The BLS Occupational Employment and Wage Statistics data identifies occupations by Standard Occupational Classification (SOC) codes, which differ slightly from the O*NET-SOC codes used in O*NET data. We use the BLS-provided mapping between the codes (`https://www.bls.gov/emp/documentation/crosswalks.htm`) and present all of our results in terms of SOC occupations. When multiple O*NET-SOC occupations share the same SOC code (e.g., Tour Guides and Travel Guides share the SOC code for "Tour and travel guides"), we take the union over O*NET data mapping to the SOC Code (i.e., tasks and DWA/IWA/GWAs) [1].

To measure how AI applicability varies across demographic groups, we use the Current Population Survey (CPS) data from 2024 (39), which is gathered by the U.S. Census Bureau for the BLS. In particular, we use tables 11 ("Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity") and 11b ("Employed persons by detailed occupation and age") from the 2024 CPS annual averages. These provide demographic statistics for detailed occupation codes, although CPS uses a different occupational taxonomy than either SOC or O*NET. We use the BLS-provided crosswalk from SOC codes to CPS codes (also at the above link) to propagate SOC-level AI applicability scores to CPS-level AI applicability scores. When multiple CPS codes map to the same SOC code, we apply the SOC-level score to each CPS code. When multiple SOC codes map to the same CPS code (which is much more common), we use the employment-weighted average of the SOC-level AI applicability scores.

A.1 Handling missing data

We omit all military occupations (SOC Codes 55-xxxx), as they have no task data in O*NET, no employment data in the BLS OEWS data, and are not included in the O*NET-SOC to SOC crosswalk. We also omit fishing and hunting workers (SOC Code 45-3031), as they are missing from the 2023 OEWS data. Finally, we omit 74 SOC codes mapping to O*NET occupations for which there is no task data. This leaves us with 785 SOC codes covering 149.8 million workers in the 2023 OEWS data (total US employment in 2023 OEWS data is 151.9 million).

B Estimating real-world IWA frequency

For each task, O*NET provides the overall share of survey respondents in an occupation who say the task is relevant to their job, as well as the share of respondents who report performing that task at each of various frequency levels (e.g., hourly, weekly, yearly). To convert these into annual total workforce counts for each IWA, we perform the following procedure:

[1] When comparing our AI applicability scores to E1 exposure scores from Eloundou et al. (3), we take the mean of O*NET-SOC occupational exposure scores mapped to the same SOC code.

1. Convert O*NET task frequency categories into annual counts based on 260 workdays/year and 8 hours/workday: "Yearly or less": 1, "More than yearly": 4, "More than monthly": 24, "more than weekly": 104, "Daily": 260, "Several times daily": 780, "Hourly or more": 2080.

2. For each task, compute its average annual frequency by averaging the above counts (weighted by surveyed percentages) and multiplying by relevance.

3. To get GWA-level frequencies, sum over tasks mapping to the same GWA.

4. To compute the total annual counts of an GWA in the workforce, sum over all occupations performing the IWA, multiplying by employment of each occupation.

C Work activities and scope classifier

Our pipeline consists of two stages of LLM prompts. In the first-stage prompts, we give an LLM (specifically, `gpt-4o-2024-08-06`, API version `2024-08-01-preview`) the entire conversation and ask it to summarize (a) the user goal and (b) the AI action in the style of an O*NET IWA, as well as four rewordings of each statement (for this "generate" prompt, we use temperature 1).[2] We then use these summaries to sort all IWA statements in order of relevance to the user task and AI goal (creating two rankings) through cosine similarity of their OpenAI `text-embedding-3-large` embeddings. More specifically, we sort by average similarity between true IWAs and the five alternate phrasings of the LLM-generated summaries to average out differences caused by word choice rather than meaning. In the second-stage prompts, we use GPT-4o to do a binary classification for every IWA as to whether it matches the user goal or AI action in the conversation (for these "classify" prompts, we use temperature 0).[3] The user and AI classifications are done in separate prompts, with each prompt containing 20 IWAs for classification (taking the sorted order from stage one and splitting into contiguous blocks of 20 IWAs). In validation against human labels (discussed later), we found that GPT-4o could perform 20 IWA classifications in a single prompt without degrading accuracy, but that more led to worse classification; we also found that grouping IWAs by level of similarity as described led to higher classification reliability. As another measure to improve agreement between human and LLM labels, we provide the first GPT-4o-generated summary from stage one as an additional "IWA" in each prompt, which serves as a point of reference against which other IWAs are measured. Compared to alternative approaches (e.g., hierarchical clustering-based classification (19)), our pipeline sacrifices efficiency for thoroughness.

C.1 Validation

We tuned and validated our prompts using independent annotations by three of the authors on a sample of 195 anonymized English conversations which were already automatically scrubbed of personally identifiable information (the sensitive nature of the data precluded external annotators). For each conversation, the three annotators were shown the conversation text, 20 candidate user goal IWAs, and 20 candidate AI action IWAs. These sets of 20 consisted of the 10 most similar according to cosine similarity to stage one summaries (where matches are dramatically more likely) and 10 uniformly sampled from the next 90 most similar IWAs, all shuffled together; the same IWAs were sampled across annotators. The annotators independently listed all matching IWAs for the user goal and all matching IWAs for the AI action. We randomly split the conversations into a validation set of 95 used for prompt and pipeline tuning and a test set of 100, which was not touched until all full-scale pipeline runs had completed. The binary classification task over IWA matches was challenging but still had moderate agreement, with Cohen's kappa inter-rater

[2] We found strong evidence that GPT-4o includes O*NET data in its pretraining corpus, as it exhibits strong knowledge of O*NET structure, occupational information, and work activities.

[3] We had initially intended to only classify the top-k most relevant IWAs in the sorted order generated by the stage one prompt, but decided to classify every IWA for completeness. We kept the stage one sorting since we found that grouping IWAs by similarity to the generated summaries led to better agreement with human labels.

reliabilities of 0.51, 0.58, 0.41 between the three pairs of annotators for user goal classification and 0.50, 0.49, 0.57 for AI action (on the test set, $n = 100$). The inter-rater reliabilities of our human annotators was higher than the $\kappa = 0.27$ for the pair of human annotators in (20) on a similar task. Our final classification pipeline achieves Cohen's kappas with our three annotators on the test set that are only slightly lower: 0.44, 0.35, 0.38 for user goal and 0.53, 0.34, 0.39 for AI action (with very similar scores on validation, $0.48, 0.38, 0.42$ for user goals and $0.52, 0.38, 0.32$ for AI actions, indicating that our prompt tuning did not result in overfitting). These kappa scores are generally low due to the high degree of uncertainty around whether a particular IWA accurately describes the intent of the user or the action of the AI; in many cases, it is easy to make compelling arguments both that an IWA does and does not apply to a conversation, so we found even moderate agreement encouraging. Additionally, the overall match rate is very low (single-digit percentages), so the overall accuracies of all raters (including our LLM pipeline) with respect to each other are well over 90%. Our final prompts are included below.

In addition to validating the IWA classifications, we also validated the scope of impact ordinal classification. Note that the sample size for scope of impact validation is limited by only being able to compare scope classifications on IWAs that both raters labeled as a match. The test set user goal and AI action r_{wg} scores (40) (where 1 indicates perfect agreement and 0 only agreement due to chance) between the human rater pairs were 0.48 ($n = 55$), 0.32 ($n = 66$), 0.55 ($n = 51$) and 0.66 ($n = 74$), 0.12 ($n = 89$), 0.30 ($n = 88$), respectively. This indicates some agreement, but with substantial variance. Between human raters and LLM classifications, the corresponding r_{wg} scores were 0 ($n = 43$), 0.35 ($n = 44$), 0.49 ($n = 34$) and 0.19 ($n = 90$), 0.04 ($n = 64$), 0.28 ($n = 61$). As another measure, the average mean absolute error (MAE) between pairs of human raters on the test set was 0.94 for user goals and 1.05 for AI actions, while the average human-LLM MAEs were 1.06 and 1.23. In comparison, the expected MAE against uniform random ratings on a six-point scale is $35/18 \approx 1.94$. This indicates that the scope of impact LLM classifications have some agreement with human ratings, although the gap between human-human and human-LLM agreement is larger for scope classification than for IWA classification. An independent indication that, while noisy, the scope of impact classifier captures real signal is its correlation with IWA share ($r = 0.64$ and $r = 0.54$ for user goals and AI action; Figure S10).

C.2 Prompts

Figure S23 shows the structure of each of the prompt's outputs.

Generate prompt
<|Instruction|>
Task overview You will be given a conversation between a User and an AI chatbot.
You have two primary goals:
(1) summarize the main goal that the user is trying to accomplish in the style of an O*NET Intermediate Work Activity (IWA).
(2) summarize the action that the bot is performing in the conversation in the style of an O*NET IWA.
For example, if the user asks for help with a computer issue and the bot provides suggestions to resolve the issue, the user's IWA is "Resolve computer problems" and the bot's IWA is "Advise others on the design or use of technologies."
Sometimes, the user intent and bot action may be the same.
For instance, if the user asks the bot to spellcheck a research paper and the bot corrects a few misspelled words, the user's IWA is "Edit written materials or documents" and the bot's IWA is also "Edit written materials or documents"
For both the user and bot IWA summaries, you will generate several variations of the summary to capture the same intent using different wordings.
To aid your analysis, you will also summarize the conversation.
Finally, you will also determine whether the User is a student trying to do homework.

Task details
Your task is to fill out the following fields:
summary: Summarize User's queries in 3 sentences or fewer in **English**.
user_iwa: Summarize the task the user is trying to accomplish in the style of an O*NET IWA. Ensure that the sum-

mary accurately describes the goal of the User as directly evidenced in the conversation. Ensure that the summary matches the level of generality of an O*NET IWA: it should general enough to be an activity performed in a large number of occupations across multiple job families, but specific enough to capture the essence of the User's goal. Provide exactly one succinct IWA-style summary.
user_iwa_variations: Generate 4 variations of the user IWA summary that capture the same intent using different wordings.
bot_iwa: Summarize the task that the bot is performing in the style of an O*NET IWA. Ensure that the summary matches the level of generality of an O*NET IWA: it should general enough to be an activity performed in a large number of occupations across multiple job families, but specific enough to capture the essence of the bot's actions. Provide exactly one succinct IWA-style summary.
bot_iwa_variations: Generate 4 variations of the bot IWA summary that capture the same action using different wordings.
is_homework_explanation: Determine whether the User is a student trying to do homework. This may be obvious if they have pasted in assignment instructions, or it may be clear from the type of question they are asking. Explain in one sentence.
is_homework: Based on your explanation, provide the label 0 (not homework) or 1 (homework).

Hints
Provide your answers in **English** using the given structured output format.
<|end Instruction|>

<|Conversation between User and AI|>
{convo}
<|end Conversation|>

<|end of prompt|>

Classify user prompt
<|Instruction|>
Task overview
You will be given a conversation between a User and an AI chatbot as well as a summary of the conversation and a list of Candidate Intermediate Work Activity (IWA) statements from O*NET.
The IWAs will be numbered with numerical IDs to help you reference them in your responses.
Your primary task is to determine for each of the Candidate IWAs whether the user is trying to perform that IWA, according to the meaning of the IWA in the context of O*NET. The conversation must provide direct evidence that the user is themself trying to accomplish the IWA.
For example, a user asking for tech support does not match a IWA about providing tech support, but does match a IWA about resolving technical issues.
As another example, a user seeking information about a product does not match a IWA about providing product information, but does match a IWA about researching product information.
Additionally, you will determine the level of assistance that the bot provides to the user in the conversation for each matching IWA.

Task details
Your reply to iwa_analyses should be a list of UserIWAAnalysis objects, one for each Candidate IWA in the order below. For each Candidate IWA, you will analyze the user's intent relative to that IWA and fill out the fields of UserIWAAnalysis as follows:
iwa (str): Copy the current Candidate IWA verbatim into this field. All of the following fields will be based on this IWA.
iwa_explanation (str): Explain in one sentence what the IWA means in the context of O*NET and what kinds of occupations perform this IWA.
is_match_explanation (str): Explain in one sentence whether the user is seeking to perform an activity described by the IWA, according to the meaning of the IWA in O*NET. To be considered a match, the user's intent must be to perform the action themselves, so if the IWA mentions or implies assisting clients or customers, for instance, there must be evidence in their query that the user is seeking to assist a client or customer.
is_match (bool): Based on your explanation, provide the label True if the user is seeking to perform an activity

described by the IWA, according to the meaning of the IWA in O*NET, and False otherwise. To be considered a match, the user's intent must be to perform the action themselves.
assistance_level_explanation (str): Consider the full scope of the work performed under this IWA across all occupations. What fraction of this work can the bot assist users with by applying only the capability it demonstrates in this conversation? Pay careful attention to the fact that the IWA might encompass many more subtasks than represented in this conversation. Explain in one sentence, or reply N/A if the IWA does not match the user's intent (i.e., when is_match is False).
assistance_level (IWAAssistanceLevel): Based on your explanation, label the bot's capability to assist with the IWA using the IWAAssistanceLevel enum, which has the following options:
- none: The user is not seeking to perform the IWA, or the conversation does not indicate that the bot is capable of assisting with the IWA.
- minimal: With this demonstrated capability, the bot can assist with a minimal portion of the work in the IWA.
- limited: With this demonstrated capability, the bot can assist with a limited portion of the work in the IWA.
- moderate: With this demonstrated capability, the bot can assist with a moderate portion of the work in the IWA.
- significant: With this demonstrated capability, the bot can assist with a significant portion of the work in the IWA.
- complete: With this demonstrated capability, the bot can assist with all of the work in the IWA.

Hints
- Provide your answers in **English** using the given structured output format.
<|end Instruction|>

<|Conversation between User and AI|>
{convo}
<|end Conversation|>

<|Conversation Summary|>
{summary}
<|end Conversation Summary|>

<|Candidate IWAs|>
{iwas}
<|end Candidate IWAs|>

<|end of prompt|>

Classify bot prompt
<|Instruction|>
Task overview
You will be given a conversation between a User and an AI chatbot as well as a summary of the conversation and a list of Candidate Intermediate Work Activity (IWA) statements from O*NET.
The IWAs will be numbered with numerical IDs to help you reference them in your responses.
Your task is to determine for each of the Candidate IWAs whether the bot is performing that IWA in the conversation, based on the meaning of the IWA in the context of O*NET.
For example, if the user asks for help with a computer issue and the bot provides suggestions to resolve the issue, this matches an IWA about providing tech support, as that is the task that the bot is performing.
However, if the user asks the bot to spellcheck a research paper and the bot corrects a few misspelled words, this does not match an IWA about writing research papers: while the **user's** overarching goal may be writing research papers, that does not match the **bot's** task in the conversation.
Additionally, you will assess whether this conversation demonstrates the bot's ability to automate each matching IWA in the conversation.

Task details
Your reply to iwa_analyses should be a list of BotIWAAnalysis objects, one for each candidate IWA in the order below. For each candidate IWA, you will analyze the bot's actions relative to that IWA and fill out the fields of BotIWAAnalysis as follows:
iwa (str): Copy the current Candidate IWA verbatim into this field. All of the following fields will be based on this

IWA.
iwa_explanation (str): Explain in one sentence what the IWA means in the context of O*NET and what kinds of occupations perform this IWA.
is_match_explanation (str): Explain in one sentence whether the action that the bot is performing in the conversation is an example of a work activity described by the IWA, given the meaning of the IWA in the context of O*NET.
is_match (bool): Based on your explanation, provide the label True if the action that the bot is performing in the conversation is an example of a work activity described by the IWA, given the meaning of the IWA in the context of O*NET, and False otherwise.
automation_level_explanation (str): Consider the full scope of the work performed under this IWA across all occupations. What fraction of this work can the bot perform by applying only the capability it demonstrates in this conversation? Pay careful attention to the fact that the IWA might encompass many more subtasks than represented in this conversation. Explain in one sentence, or reply N/A if the IWA does not match the bot's action (i.e., when is_match is False).
automation_level (IWAAutomationLevel): Based on your explanation, label the bot's capability to perform the IWA using the IWAAutomationLevel enum, which has the following options:
- none: The bot does not perform the IWA, or the conversation does not indicate that the bot is capable of performing the IWA.
- minimal: With this demonstrated capability, the bot can perform a minimal portion of the work in the IWA.
- limited: With this demonstrated capability, the bot can perform a limited portion of the work in the IWA.
- moderate: With this demonstrated capability, the bot can perform a moderate portion of the work in the IWA.
- significant: With this demonstrated capability, the bot can perform a significant portion of the work in the IWA.
- complete: With this demonstrated capability, the bot can perform all of the work in the IWA.

Hints
- Provide your answers in **English** using the given structured output format. <|end Instruction|>

<|Conversation between User and AI|>
{convo}
<|end Conversation|>

<|Conversation Summary|>
{summary}
<|end Conversation Summary|>

<|Candidate IWAs|>
{iwas}
<|end Candidate IWAs|>

<|end of prompt|>

D Completion classifier

While the COPILOT-THUMBS dataset tells us which work activities receive the most positive user feedback, thumbs feedback may not reflect the success of AI across tasks, as not all types of users give feedback at the same rate (e.g., suppose users who perform some tasks are inherently more critical than those who perform others). To supplement the thumbs feedback data, we therefore also perform task completion classification with an LLM. For each conversation, we ask GPT-4o-mini (version `gpt-4o-mini-2024-07-18`, API version `2024-08-01-preview`, temperature 0) [4] if the AI completed the user's task in the conversation. For comparison with the E1 measure of Eloundou et al. (3), we also ask if the AI reduced the time it takes to complete the task by at least 50%, though we did not end up using that.

[4]This task is much simpler than the difficult and ambiguous IWA classification task, hence our use of the smaller model.

Completion prompt <|Instruction|>
Task overview
You will be given a conversation between a User and an AI chatbot.
You will summarize the main task that the user is trying to accomplish in the conversation.
You will also determine whether the AI chatbot is able to complete the task, and if so, whether it reduced the time it takes to complete the task with equivalent quality by at least half.

Task details
Your task is to fill out the following fields:
task_summary: Summarize the task the User is trying to accomplish in **English**.
completed_explanation: Explain in one sentence whether the AI chatbot is able to complete the User's task, based on the conversation.
completed: Based on your explanation, provide one of the following labels:
- not_complete: The AI chatbot did not make substantive progress towards completing the User's task.
- partially_complete: The AI chatbot made progress towards completing the User's task, but did not complete it.
- complete: The AI chatbot completed the User's task.
speedup_50pct_explanation: Explain in one sentence whether the AI chatbot reduced the time it takes to complete the task with equivalent quality by at least half. This includes tasks that can be reduced to:
- Writing and transforming text and code according to complex instructions,
- Providing edits to existing text or code following specifications,
- Writing code that can help perform a task that used to be done by hand,
- Translating text between languages,
- Summarizing medium-length documents,
- Providing feedback on documents,
- Answering questions about a document, or
- Generating questions a user might want to ask about a document.
Assume the user is a worker with an average level of expertise in their role trying to complete the given task.
speedup_50pct: Based on your explanation, provide the label True if the AI chatbot reduced the time it takes to complete the task with equivalent quality by at least half, and False otherwise.

Hints
Provide your answers in **English** using the given structured output format.
<|end Instruction|>

<|Conversation between User and AI|>
{convo}
<|end Conversation|>

<|end of prompt|>

E Physical tasks classifier

We used `GPT-5-2025-08-07`[5] to label whether each of 18796 O*NET tasks requires touching or moving physical objects or people. We also had two human annotators label a random sample of 200 tasks. The two annotators agreed with the model on 183 and 184 of the 200 tasks (Cohen's κ 0.82 and 0.84), while they agreed with each other on 181 of the tasks (Cohen's κ 0.80).

Physical task prompt
<|Instruction|>
Instructions
You will be given an occupation and task from the O*NET database. Determine whether performing that task

[5]Conversation-level classifications were performed before GPT-5 was released; here, we can afford to use a larger, slower, more expensive model as the number of task-level classifications is small compared to the conversation-level classifiers.

requires physical action, defined as touching or moving objects or people.
Interacting with a computer by typing on a keyboard or using a mouse does not count as physical action.
To help illustrate the task, we have provided some examples below.

Examples
Occupation: Electrician
Task: Work from ladders, scaffolds, or roofs to install, maintain, or repair electrical wiring, equipment, or fixtures.
Is physical action required? Yes, this requires touching and moving objects.

Occupation: Software Developer
Task: Modify existing software to correct errors, adapt it to new hardware, or upgrade interfaces and improve performance.
Is physical action required? No, this does not require touching or moving objects or people.

Occupation: Chief Executives
Task: Direct, plan, or implement policies, objectives, or activities of organizations or businesses to ensure continuing operations, to maximize returns on investments, or to increase productivity.
Is physical action required? No, this does not require touching or moving objects or people.

Occupation: Nurse Practitioners
Task: Perform primary care procedures such as suturing, splinting, administering immunizations, taking cultures, and debriding wounds.
Is physical action required? Yes, this requires touching people.

Output format
Provide your answer using the given structured output, filling out the fields as follows:
- is_physical_explanation (str): A brief explanation of your reasoning about whether the task requires touching or moving objects or people.
- is_physical (bool): true if physical action is required, false otherwise.

Your task
Occupation: {occupation}
Task: {task}
Is physical action, defined as touching or moving objects or people, required?
<|end Instruction|>

F Information work classifier

To highlight information work occupations in Table 1 and Figure S11, we used `GPT-5-2025-08-07` to label whether each SOC occupation consisted primarily of information work and/or knowledge work. See the prompt below for the definitions we used. As a robustness check, we also repeated the classification with the explicit definitions replaced by an instruction to use "standard economic definitions" of information and knowledge work. Agreement between the two classification runs was 86% and 91% for information work and knowledge work, respectively.

Information work prompt
Instructions
You will be given a detailed SOC occupation. You tasks are to determine if that occupation primarily consists of information work and, separately, whether it primarily consists of knowledge work.
Information work is labor in which workers create, process, or communicate information rather than produce physical goods or provide physical services.
Knowledge work is labor in which workers apply knowledge, expertise, and judgment in non-routine and creative

ways to solve complex problems, make decisions, or generate new ideas.

Output format
Provide your answer using the given structured output, filling out the fields as follows:
- is_information_work_explanation (str): A brief explanation of your reasoning about whether the occupation primarily consists of information work.
- is_information_work (bool): true if the occupation primarily consists of information work, false otherwise.
- is_knowledge_work_explanation (str): A brief explanation of your reasoning about whether the occupation primarily consists of knowledge work.
- is_knowledge_work (bool): true if the occupation primarily consists of knowledge work, false otherwise.

Task
The occupation in question: {soc_code} {occupation}

G Employment weighting and socioeconomic correlates

Most prior work did not weight occupations by employment when examining the relationship between AI exposure and wage. Since occupations vary a lot in size and the boundaries are somewhat subjective (e.g., Cooks get separate occupations for Short Order, Restaurant, Institution and Cafeteria, and Fast Food, but Maids and Housekeeping Cleaners are one category), results weighted by occupation better answer the research question about overall workforce relationship between wages and occupational applicability of AI. Nonetheless, Figure S16 shows the correlation between AI applicability score and average occupation wage without employment weighting, which increases the correlation to 0.17 for user goals and 0.32 for AI actions. This higher correlation without employment weighting indicates that prior findings may be due in part to a lack of employment weighting. The difference between the weighted and unweighted results is primarily due to high-employment Sales and Office and Administrative Support occupations that have relatively low wages, but high AI applicability. There is a lot of variation across occupations and some occupations have much higher AI applicability than others, but the overall relationship between wage and AI applicability is weak.

The same pattern also occurs with educational requirements: without employment weighting, the relationship between AI applicability score and higher education requirements appears stronger (Figure S17). Figures S14 and S15 separate out the (employment-weighted) wage and education plots from Figure S13 into user goals and AI actions; the patters are similar on each side, although the relationships are slightly stronger on the AI side.

H Supplementary figures

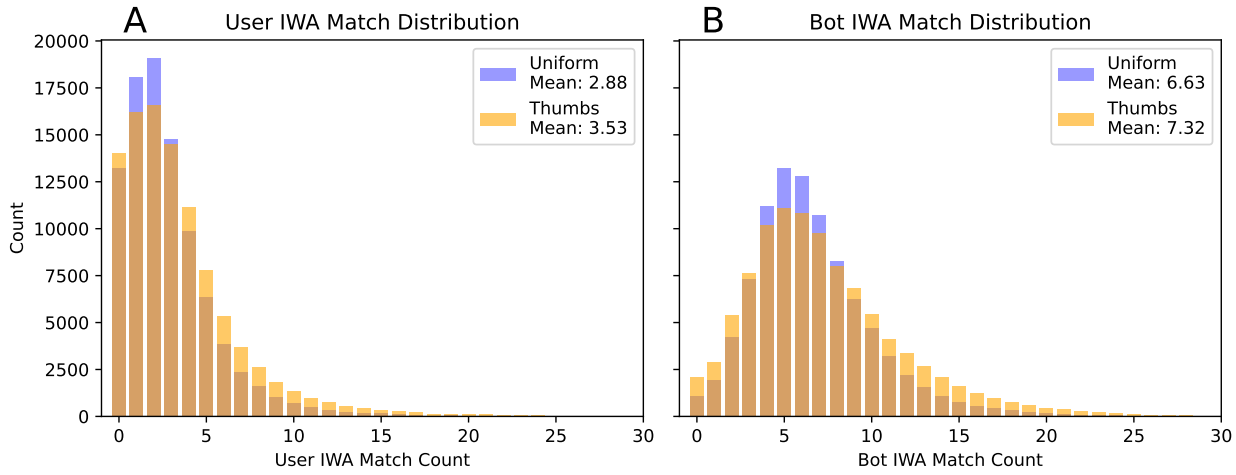


Figure S1: **Distribution of IWAs per conversation.** (**A**) The distribution of the number of IWAs a conversation is matched to for user goal in both the uniform and thumbs Copilot datasets. (**B**) The same distributions for AI actions IWAs.

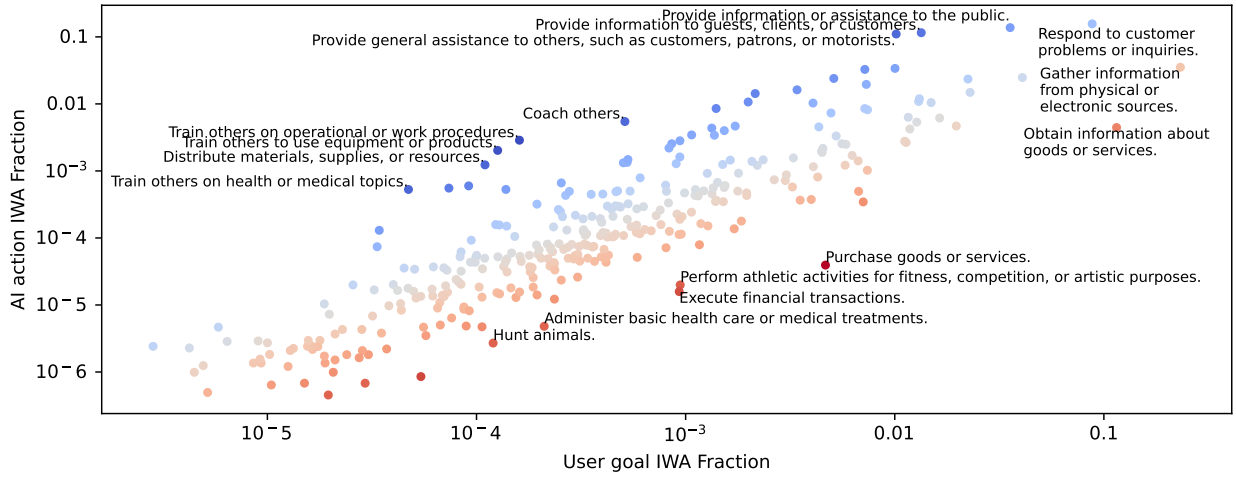


Figure S2: **IWA frequency in AI actions and user goals.** For each IWA, this Figure plots the fraction of user intents (x axis) and AI actions (y axis) described by that IWA. Outliers show which IWAs are more likely to be performed (blue) or assisted (red) by Bing Copilot. When a conversation is labeled with multiple IWAs, that share of Copilot activity is evenly distributed among the IWAs so that the sum of all IWA fractions is 1. Table S1 lists the relative frequencies for the most off-diagonal IWAs.

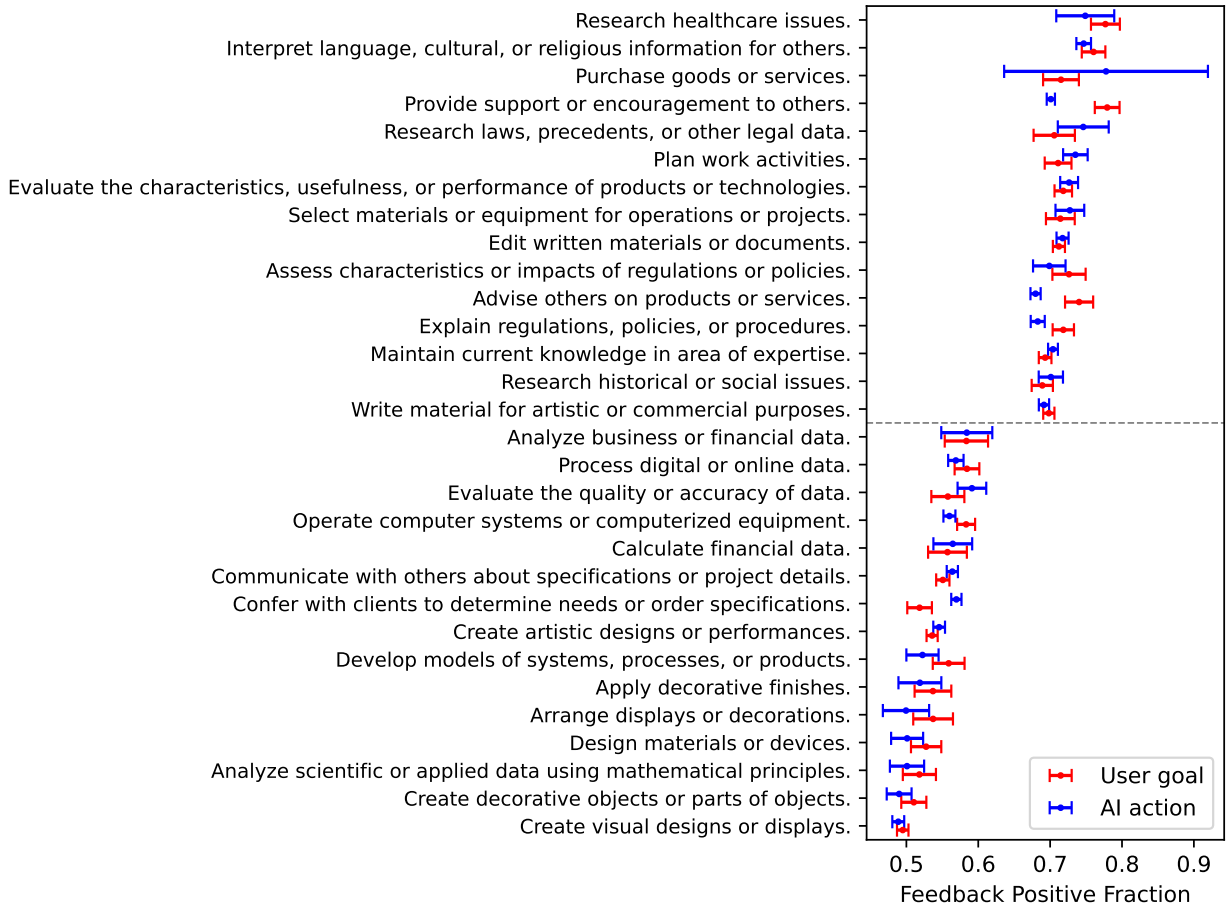


Figure S3: **IWAs with the highest and lowest shares of positive feedback.** This Figure shows the top and bottom 15 IWAs by the share of positive feedback, filtered to common IWAs matched in at least 1% of conversations in our feedback dataset, with bootstrapped 95% confidence intervals. The common IWAs with highest positive feedback share include two about writing and editing; four about evaluating or purchasing goods and services; and six about researching information about health, culture, law, policy, and society. Meanwhile, the common IWAs with the lowest positive feedback share include five visual design and five data analysis IWAs.

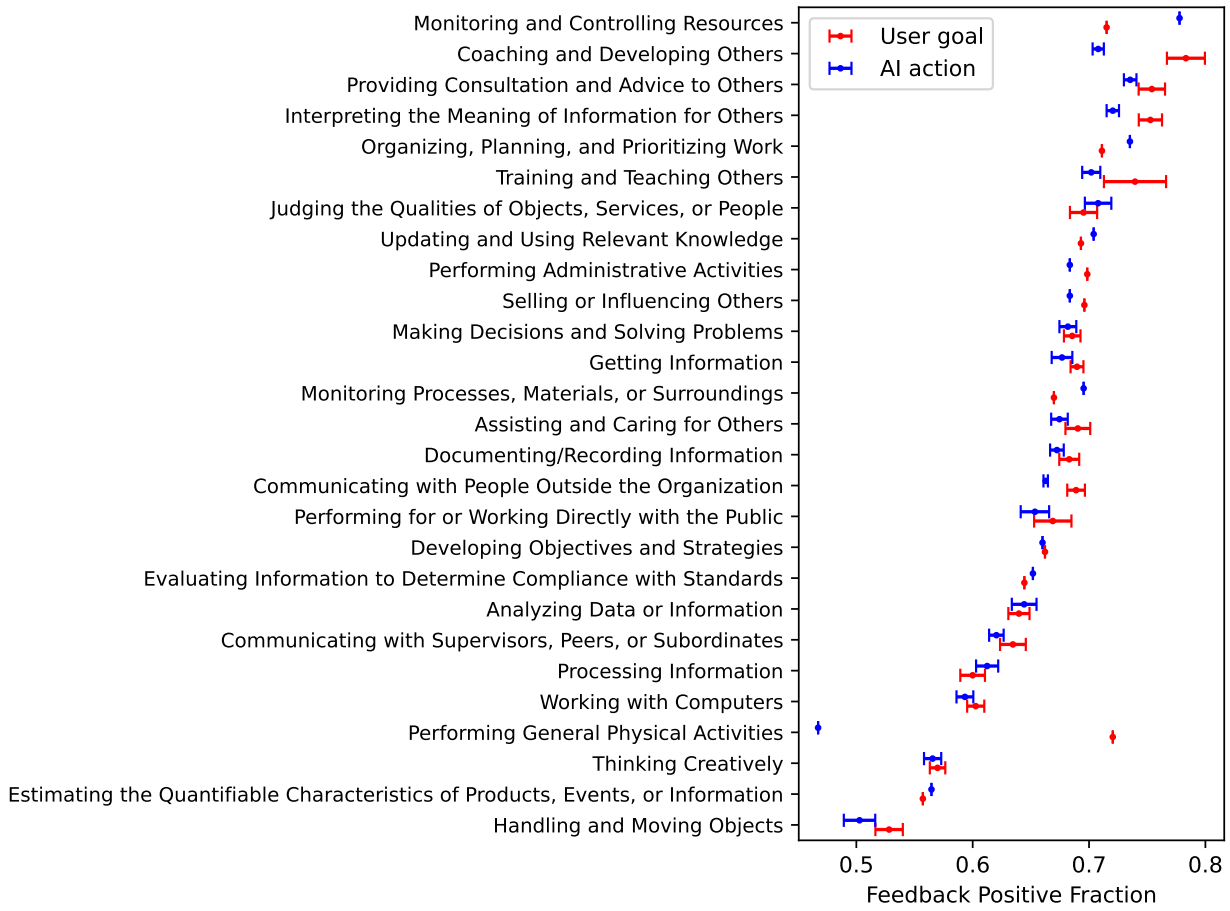


Figure S4: **Share of positive feedback by GWA.** This Figure plots the positive feedback share for each GWA, aggregating common IWAs into their GWAs and with bootstrapped 95% confidence intervals; any IWA appearing in less than 1% of our feedback data is ignored. 14 GWAs have no common IWAs and are thus excluded from this plot, including those relating to operating vehicles, repairing equipment, and management tasks like hiring, negotiation, and guiding subordinates.

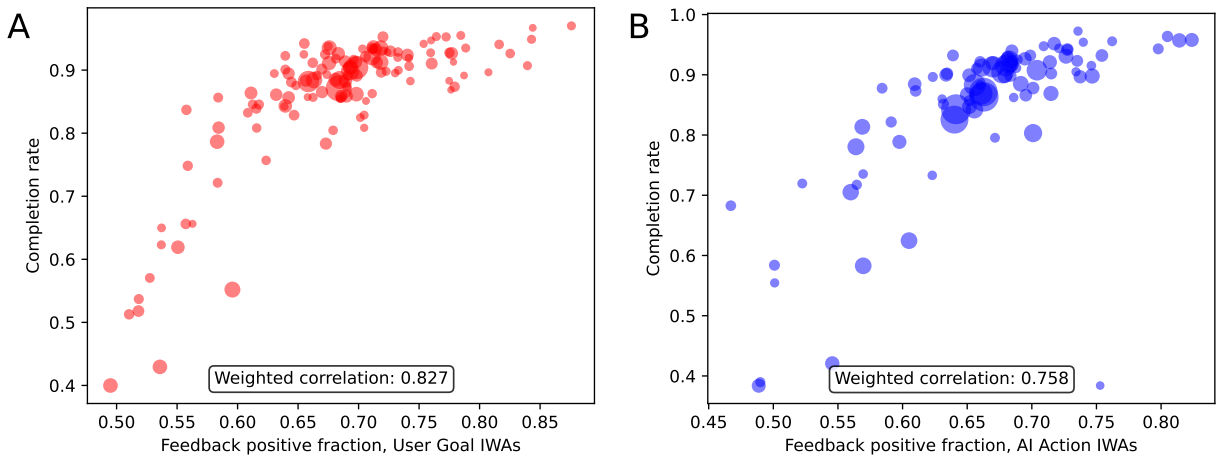


Figure S5: **Relationship between thumbs feedback and task completion rate for each IWA.** (**A**) Scatterplot of user goal IWA feedback positive rate (measured in Copilot-Thumbs) and completion rate (measured in Copilot-Uniform). (**B**) The same scatterplot for AI action IWAs. There is a strong correlation between thumbs feedback rate and GPT-4o-mini task completion rate for each IWA, indicating that both are capturing real signal about AI success in assisting or performing an IWA. Point size proportional to square root of IWA match count in Copilot-Uniform. Weighted correlations also weighted by match count in Copilot-Uniform.

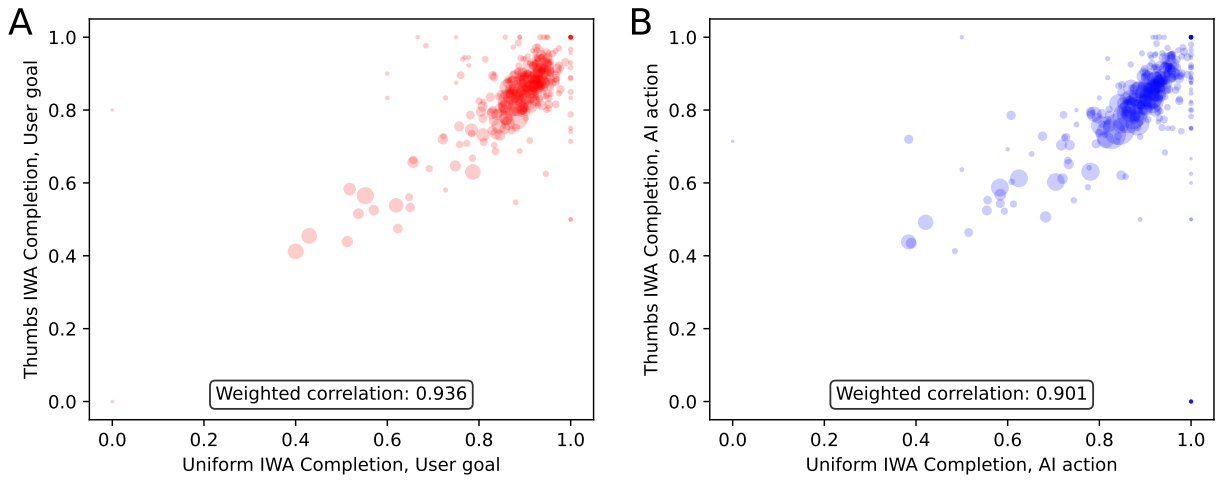


Figure S6: **Correlation between IWA completion rates in Copilot-Uniform and Copilot-Thumbs.** (**A**) Scatterplot of user goal IWA completion rate in our two datasets. (**B**) The same scatterplot for AI action IWAs. IWA-level completion rates are consistent between Copilot-Uniform and Copilot-Thumbs. Point size proportional to square root of IWA match count in Copilot-Uniform. Weighted correlations also weighted by match count in Copilot-Uniform.

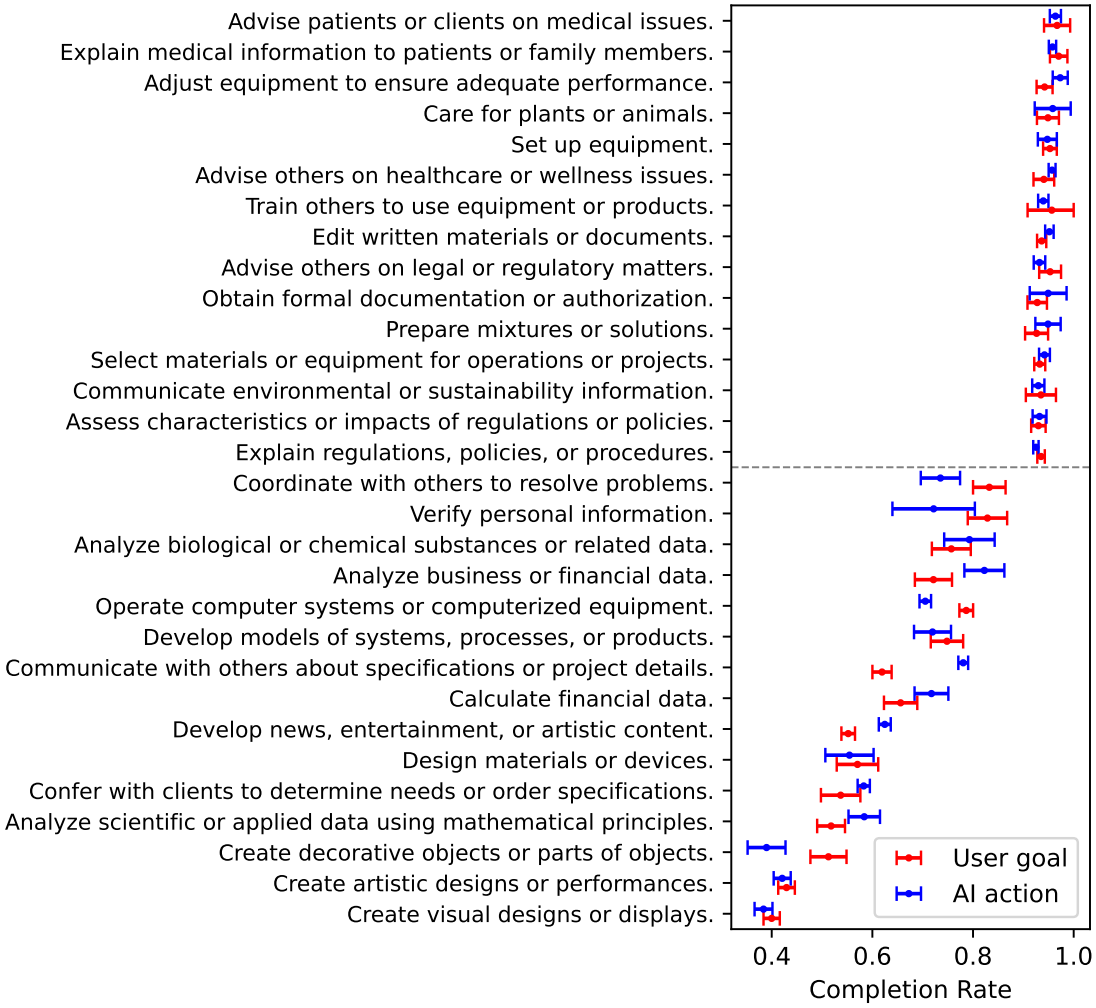


Figure S7: **IWAs with the highest and lowest completion rates.** This Figure shows the top and bottom 15 IWAs by completion rate, filtered to 'common' IWAs with activity share at least 0.1% in Copilot-Uniform, with 95% confidence intervals (using the normal approximation to binomial confidence intervals). Task completion shows some of the same patterns as positive feedback fraction (Figure S3), with visual design and data analysis on the low end and writing on the high end. One notable difference is that the highest completion rate IWAs include 7 about explaining, advising, or training.

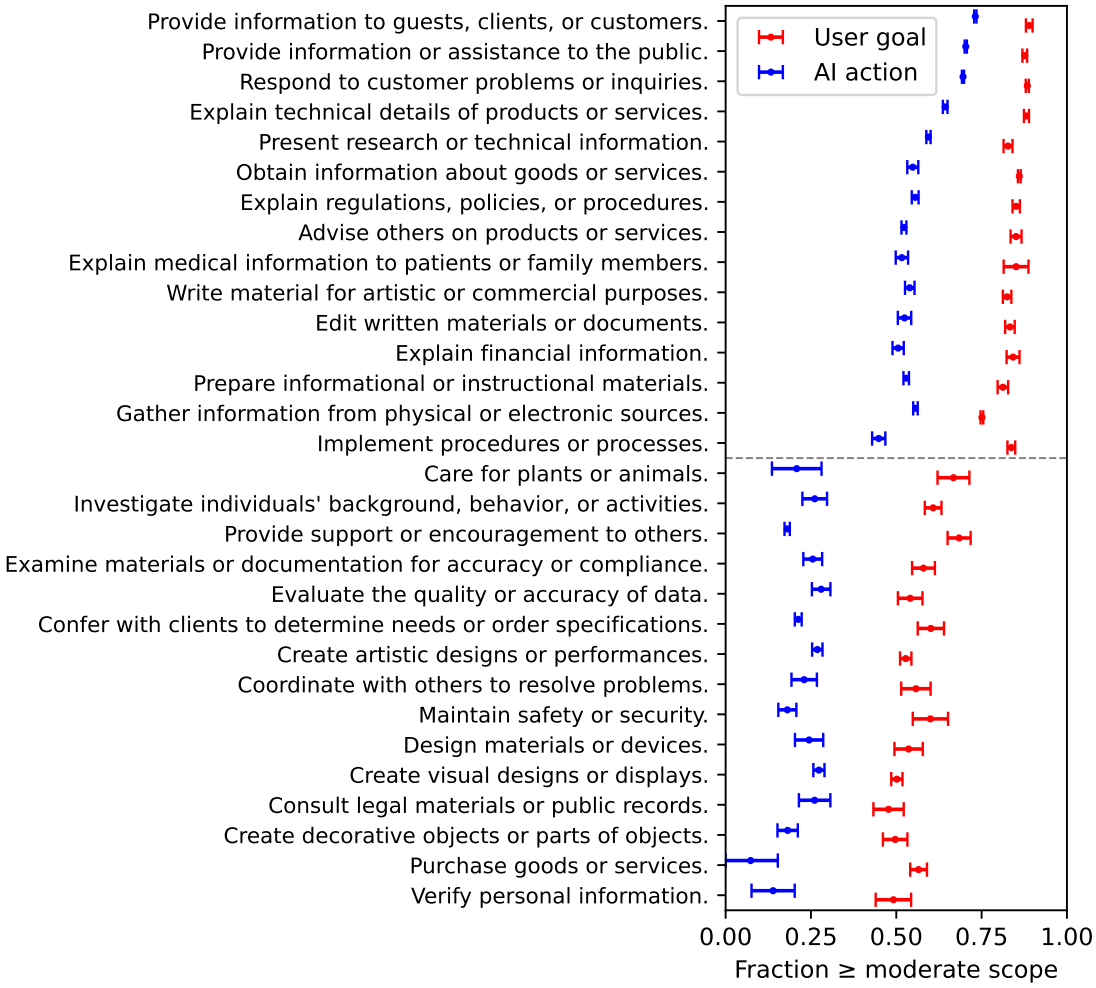


Figure S8: **IWAs with the highest and lowest fraction of conversations at moderate or higher scope.** This Figure shows the top and bottom 15 IWAs by how often they are assigned scope of impact at least moderate, filtered to "common" IWAs with activity share at least 0.1% in Copilot-Uniform, with 95% confidence intervals (using the normal approximation to binomial confidence intervals). Some patterns mirror those of thumbs feedback and completion, with research and writing IWAs having high scope, and data analysis and visual design IWAs having low scope. AI performance consistently has lower impact scope than user assistance.

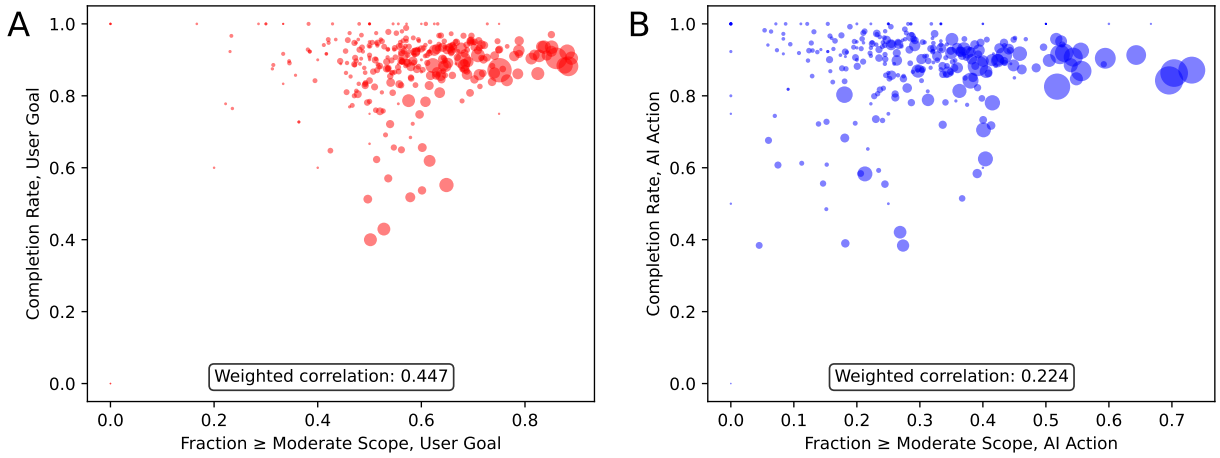


Figure S9: **Relationship between scope and task completion rate for each IWA.** (**A**) Scatterplot of user goal IWA scope above moderate share and completion rate. (**B**) The same scatter plot for AI action IWAs. The relationship between scope and completion is much weaker than between completion and thumbs feedback, as they are capturing different questions. Point size proportional to square root of IWA match count in Copilot-Uniform. Weighted correlations also weighted by match count in Copilot-Uniform.

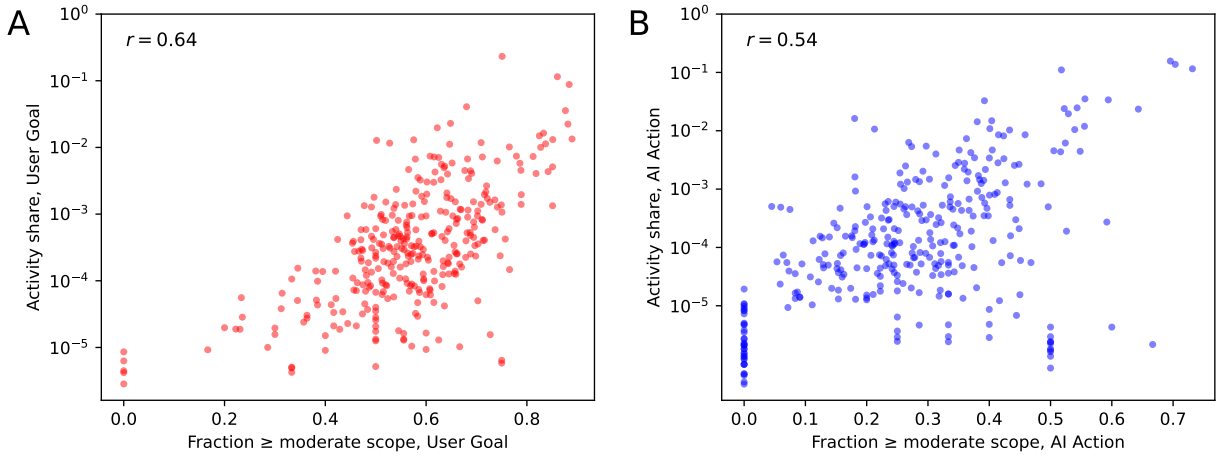


Figure S10: **Correlation between scope and activity share in Copilot-Uniform for each IWA.** (**A**) Scatterplot of user goal IWA activity share and the fraction of conversations in which it classified at moderate impact scope or higher. (**B**) The same scatterplot for AI action IWAs. On both sides, impact scope is a good predictor of what activities people seek AI assistance with (better than completion or satisfaction).

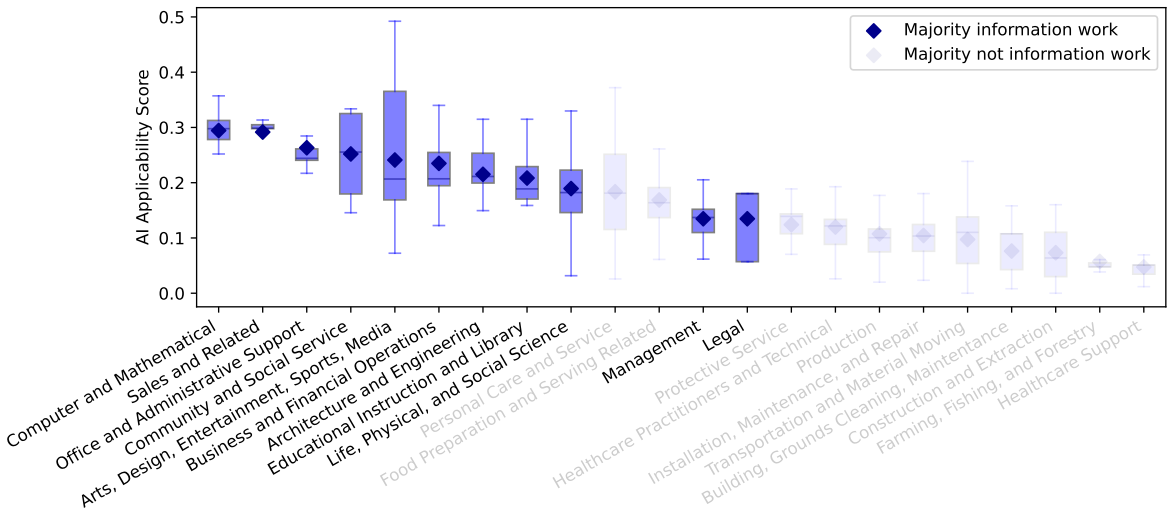


Figure S11: **AI applicability score for major occupation groups.** The employment-weighted mean, median, and inter-quartile range of the AI applicability score for each major occupation group. Occupational groups with a majority of workers in occupations classified as information work.

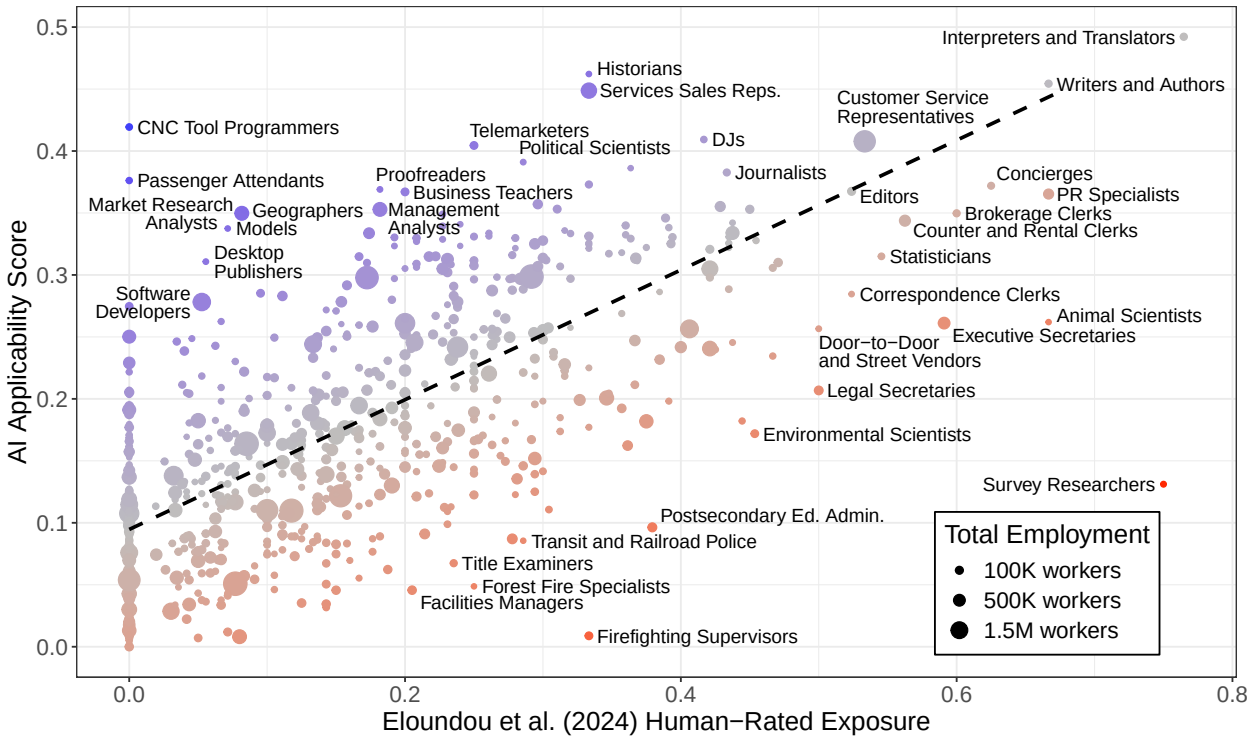


Figure S12: **AI applicability scores by predicted exposure** AI applicability score versus E1 exposure from Eloundou et al. (3), the fraction of an occupation's tasks that human raters predicted LLMs could speed up by 50%. Occupations are colored by their distance from the employment-weighted regression line: blue points have higher AI applicability scores than would be expected from their E1 exposure.

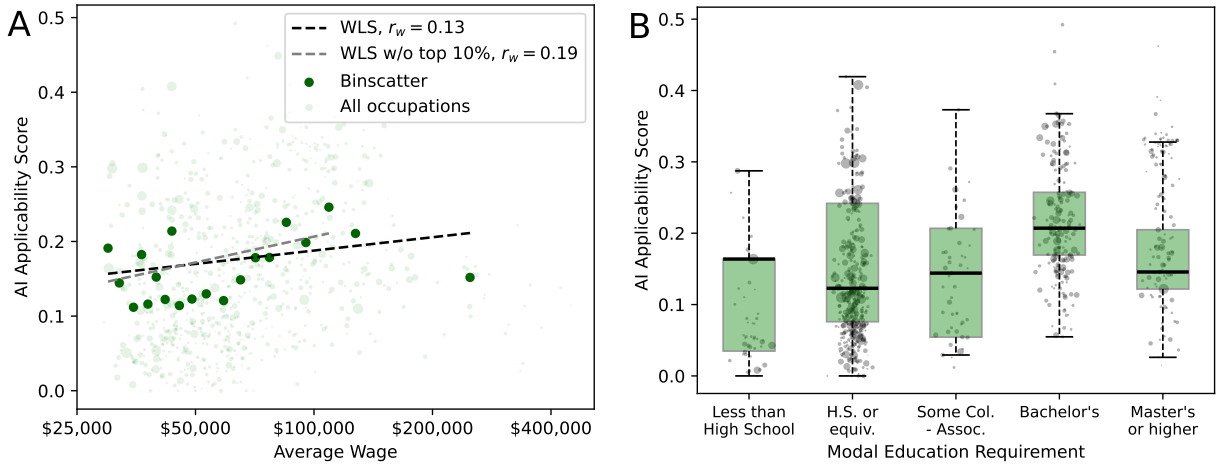


Figure S13: **AI applicability scores by average wage and educational requirement.** (A) AI applicability score by average occupational wage. There is a point for each occupation and darker points for the employment-weighted average of each ventile. Binscatter conditional means and least squares fits are weighted by employment. (B) AI applicability score by modal education requirement reported to O*NET. Boxplot medians and quartiles are weighted by employment (e.g., they show the score of the median worker rather than that of the median occupation).

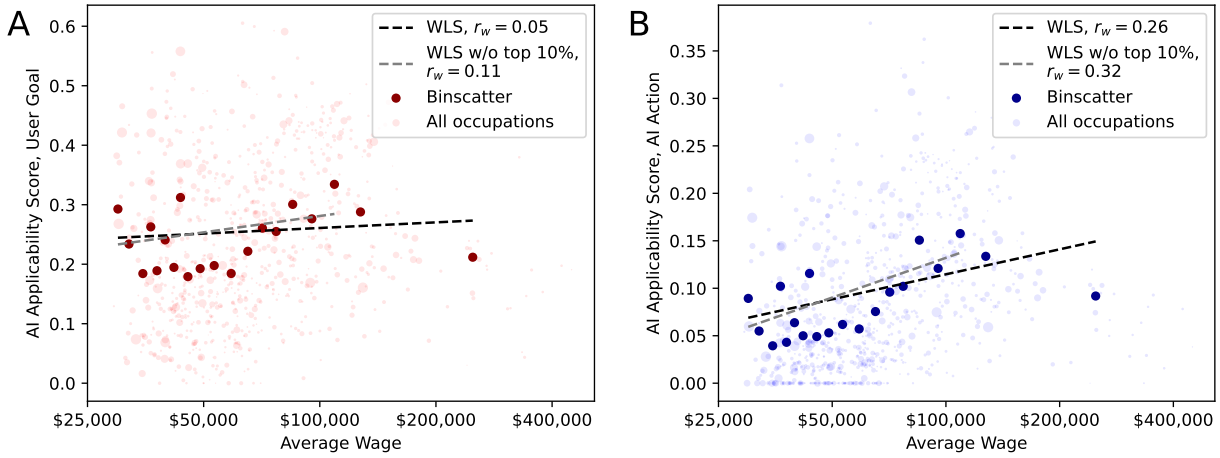


Figure S14: **AI applicability score by wage, separating out user goals and AI actions.** (**A**) Wage plot as in Figure S13A, but showing user goal AI applicability score only. (**B**) AI action applicability score only. The correlation is higher on the AI side.

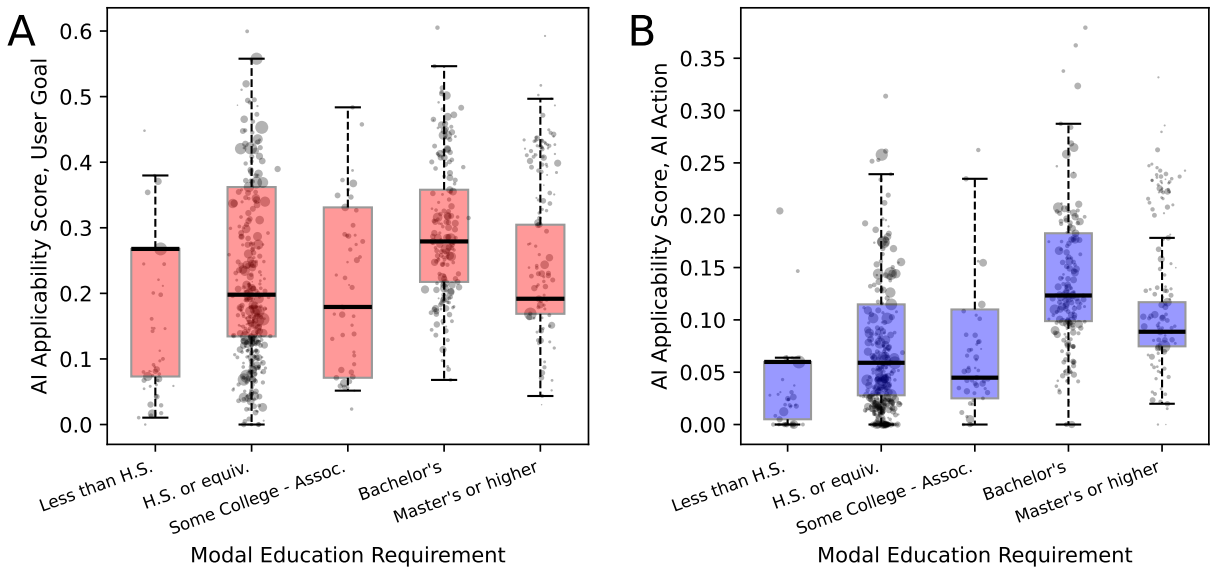


Figure S15: **AI applicability score by educational requirement, separating out user goals and AI actions.** (**A**) Education plot from Figure S13B, but showing only user goal AI applicability score. (**B**) Using AI action scores. As with wage, the relationship with education is stronger on the AI side.

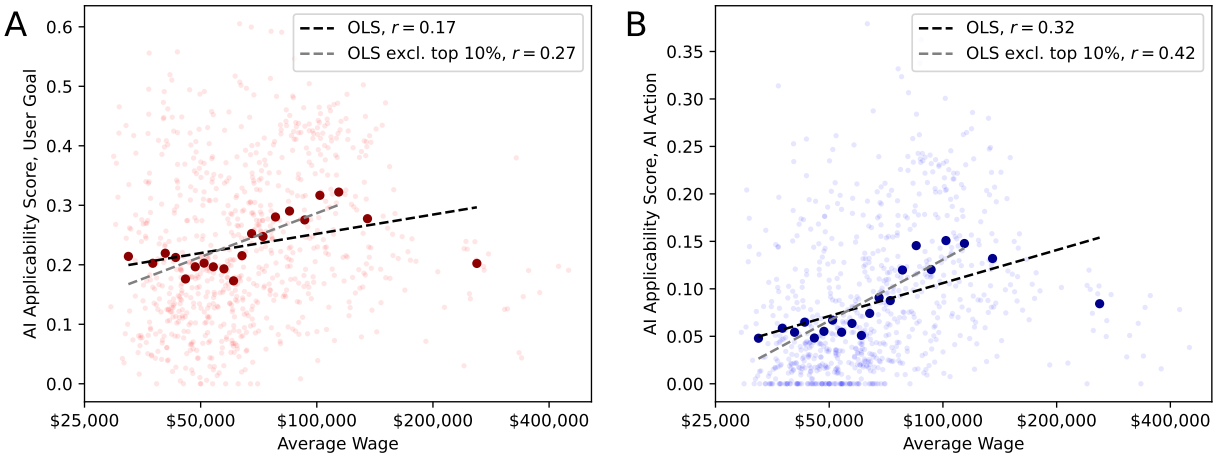


Figure S16: **AI applicability score by wage, unweighted**. This Figure is the same as Figure S14 without weighting the binscatters by employment. (**A**) Scatterplot with unweighted binscatter of each occupation's AI applicability score calculated over user goals only against the occupation's average wage. (**B**) Using AI action AI applicability score. The relationship is much less noisy without employment weighting, likely because weighting by employment causes the binscatters to be influenced dramatically by a small number of high-employment occupations, increasing the variance due to noise in the coverage metric and in the mapping between occupations and IWAs. Excluding the top 10% of highest-paid workers strengthens the correlation.

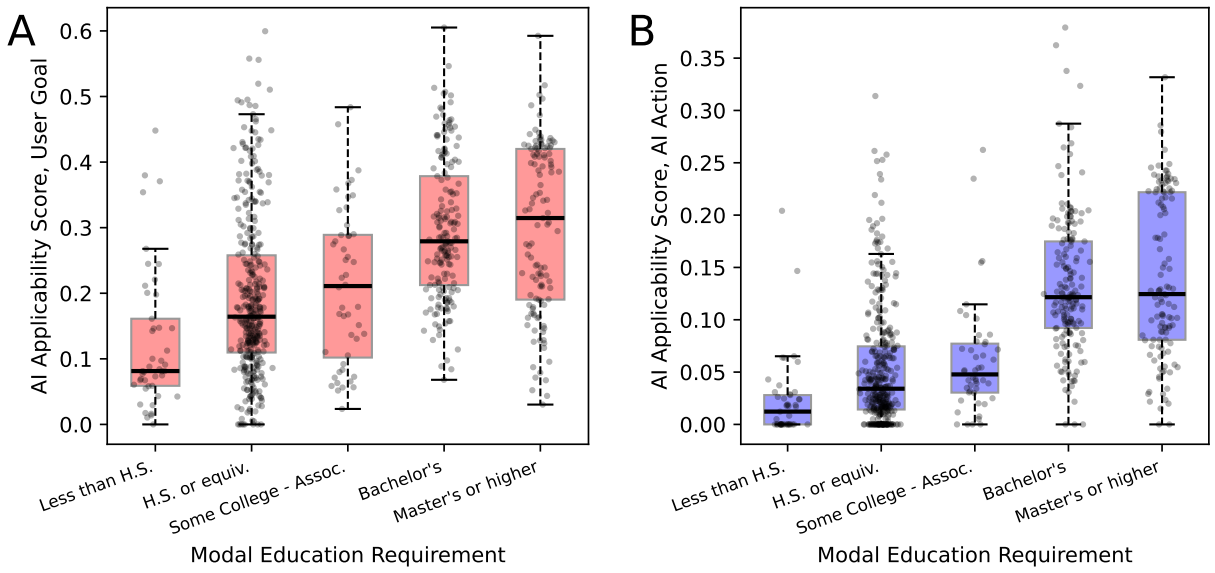


Figure S17: **AI applicability score by educational requirement, unweighted.** This Figure is the same as Figure S15 without weighting by employment. (**A**) User goal AI applicability scores over occupations with different educational requirements, with boxplots over occupations (not weighting by employment). (**B**) Using AI action scores. As in Figure S16, this reduces the noise in the relationship.

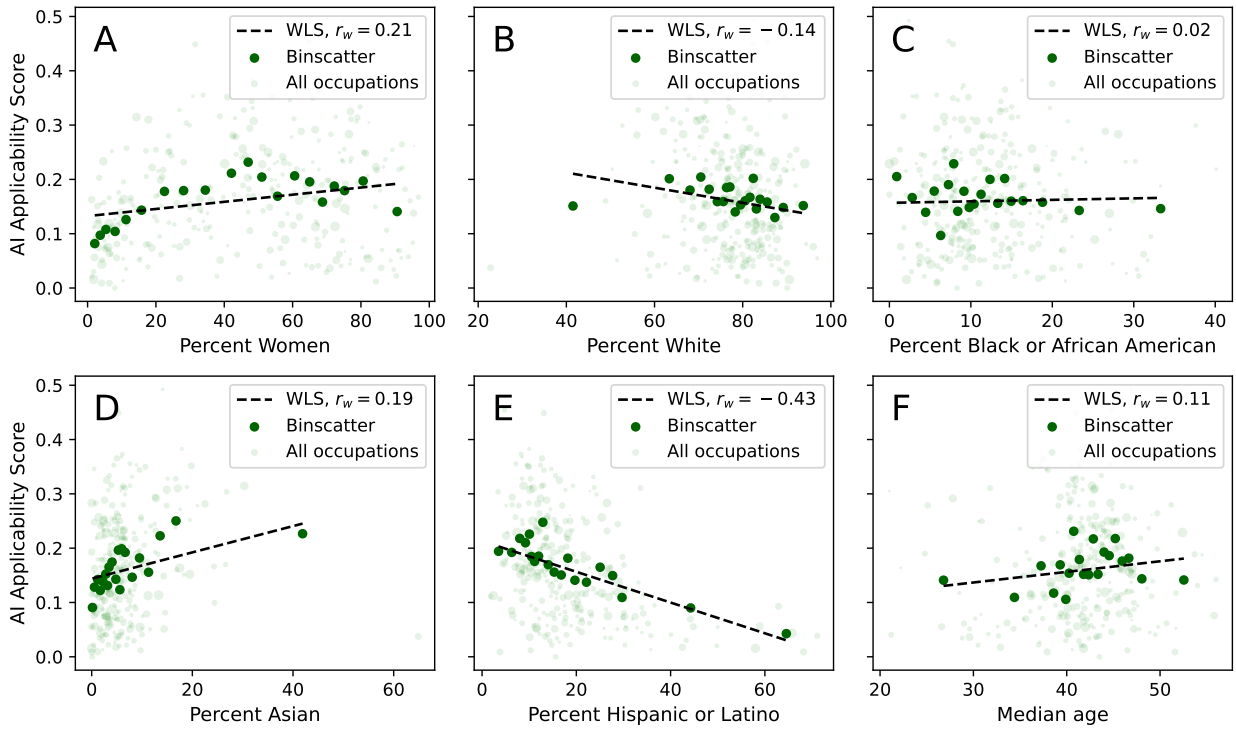


Figure S18: **Correlations between AI applicability score and 2024 Current Population Survey occupational demographics, weighted by employment.** (**A**) Correlation with percent women. (**B**) Correlation with percent white. (**C**) Correlation with percent black or African American. (**D**) Correlation with percent Asian. (**E**) Correlation with percent Hispanic or Latino. (**F**) Correlation with median age. For each panel, we drop occupation codes with missing demographic data in the 2024 CPS Annual Averages.

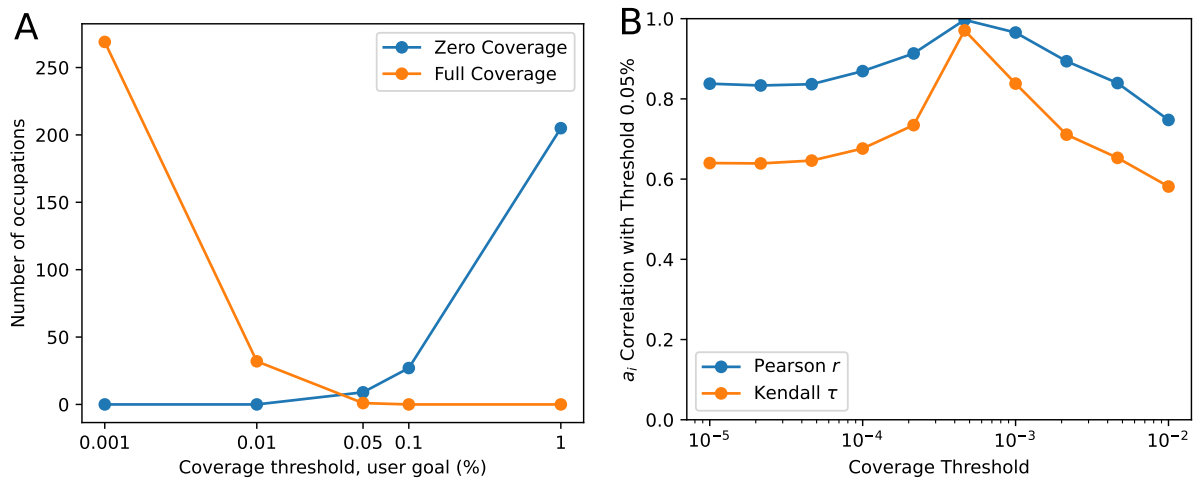


Figure S19: **Rationale for our chosen coverage threshold.** (A) Effects of coverage threshold on number of occupations with user goal coverage 0 and 1. Our threshold of 0.05% approximately minimizes the number of occupations assigned user goal coverage 0 or 1. (B) The correlation between AI applicability scores defined using different coverage thresholds and the one we report with threshold 0.05%. Across thresholds spanning three orders of magnitude, we get strongly correlated rankings of occupations by AI applicability. Contrast this robustness of relative applicability score with the absolute measure in Figure S20, which is highly sensitive to arbitrary choice of threshold.

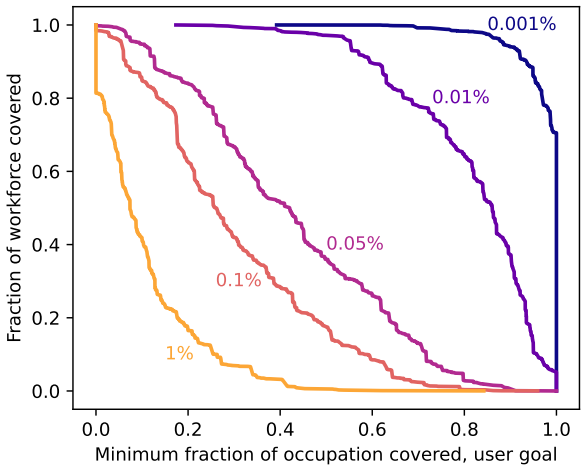


Figure S20: **Effect of activity share threshold on absolute impact estimates.** The share of workers who have at least $x\%$ of their work in covered IWAs for different definitions of an IWA being covered ($.001\%, \ldots, 1\%$ of user chat activity). The resulting numbers depend significantly on the selected threshold, making relative statements more meaningful than absolute coverage numbers.

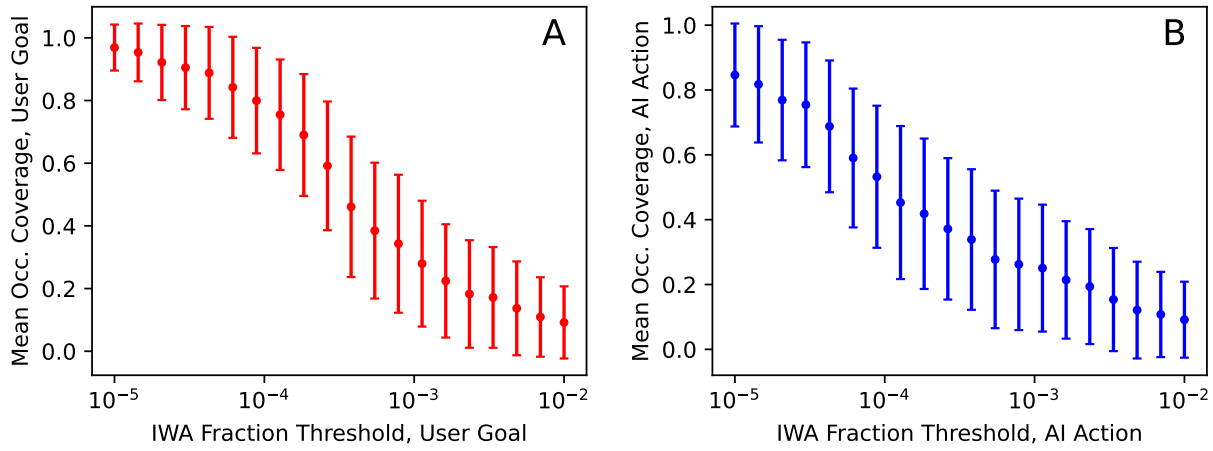


Figure S21: **Mean and standard deviation of occupation coverage by threshold.** (**A**) The averages and standard deviations of occupation user goal coverage for different thresholds for the share of chat activity an IWA must have to "be done" with the LLM. (**B**) The same means and standard deviations for coverage derived from AI action IWAs. We use a threshold of 0.0005, which results in 127 and 87 IWAs being considered covered (of 332) on the user goal and AI action sides, respectively.

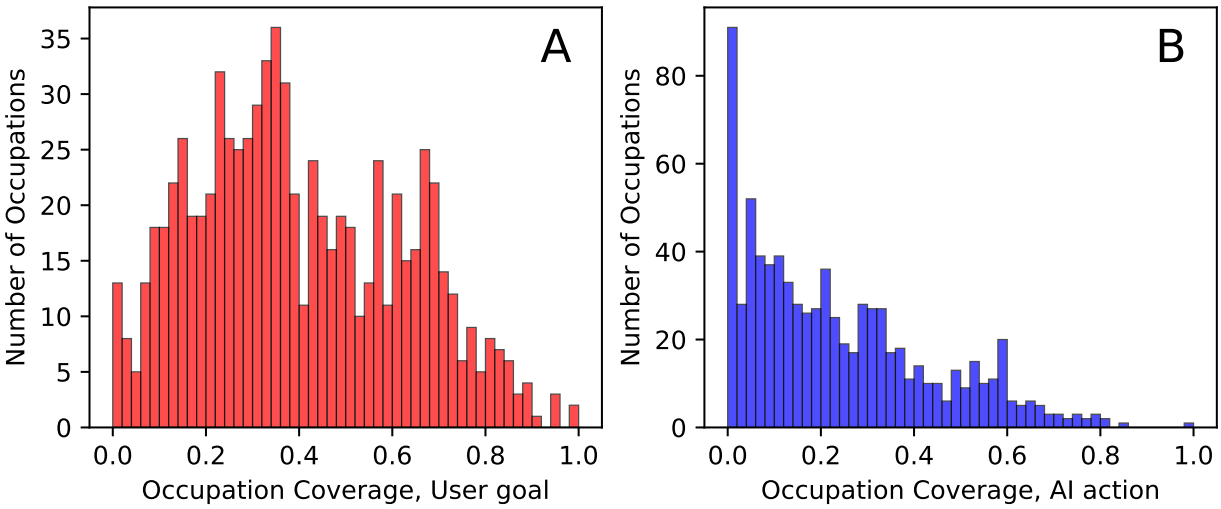


Figure S22: **Distribution of occupation coverage.** (**A**) The distribution of occupation user goal coverage scores (fraction of importance-weighted work in an occupation with user goal activity share at least 0.05% in Copilot-Uniform). (**B**) The same distribution for AI action coverage.

A

```
class GenerationAnswer:
    summary: str
    user_iwa: str
    user_iwa_variations: list[str]
    bot_iwa: str
    bot_iwa_variations: list[str]
    is_homework_explanation: str
    is_homework: int
```

B

```
class CompletionLevel(Enum):
    NOT_COMPLETE = "not_complete"
    PARTIAL = "partially_complete"
    COMPLETE = "complete"

class CompletionAnswer(BaseModel):
    task_summary: str
    completed_explanation: str
    completed: CompletionLevel
    speedup_50pct_explanation: str
    speedup_50pct: bool
```

C

```
class IWAAssistanceLevel(Enum):
    NONE = "none"
    MINIMAL = "minimal"
    LIMITED = "limited"
    MODERATE = "moderate"
    SIGNIFICANT = "significant"
    COMPLETE = "complete"

class UserIWAAnalysis:
    iwa: str
    iwa_explanation: str
    is_match_explanation: str
    is_match: bool
    assistance_level_explanation: str
    assistance_level: IWAAssistanceLevel

class UserClassificationAnswer:
    iwa_analyses: list[UserIWAAnalysis]
```

D

```
class IWAAutomationLevel(Enum):
    NONE = "none"
    MINIMAL = "minimal"
    LIMITED = "limited"
    MODERATE = "moderate"
    SIGNIFICANT = "significant"
    COMPLETE = "complete"

class BotIWAAnalysis:
    iwa: str
    iwa_explanation: str
    is_match_explanation: str
    is_match: bool
    automation_level_explanation: str
    automation_level: IWAAutomationLevel

class BotClassificationAnswer:
    iwa_analyses: list[BotIWAAnalysis]
```

Figure S23: **Structured output formats for our four conversation-level LLM prompts.** (**A**) Generation prompt. (**B**) Completion. (**C**) Classify, user goals. (**D**) Classify, AI actions.

I Supplementary tables

Table S1: **Work activities with the most extreme ratios between user goal and AI action activity share.**

More often assisted by AI	More often performed by AI
Purchase goods or services. (118.4x)	Train others on operational or work procedures. (17.9x)
Execute financial transactions. (58.8x)	Train others to use equipment or products. (16.0x)
Perform athletic activities. (47.3x)	Distribute materials, supplies, or resources. (11.2x)
Obtain information about goods or services. (25.9x)	Train others on health or medical topics. (11.2x)
Research healthcare issues. (20.5x)	Provide general assistance to others. (10.9x)
Prepare foods or beverages. (14.7x)	Coach others. (10.6x)
Research technology designs or applications. (13.5x)	Provide information to guests, clients, or customers. (8.6x)
Obtain formal documentation or authorization. (12.5x)	Advise others on workplace health or safety issues. (7.5x)
Operate office equipment. (11.4x)	Teach academic or vocational subjects. (6.6x)
Investigate incidents or accidents. (11.3x)	Teach safety procedures or standards to others. (6.5x)

Numbers show IWA over-representation factors. Only includes IWAs with user or AI activity share $\geq 0.05\%$.

Table S2: **SOC Major groups sorted by AI Applicability Score.**

Major Group	Coverage	Completion	Scope	Score	Employment
Computer & Mathematical	0.64	0.86	0.48	0.29	5,177,390
Sales & Related	0.56	0.89	0.51	0.29	13,266,370
Office & Admin. Supp.	0.56	0.89	0.49	0.26	18,163,760
Community & Social Service	0.51	0.88	0.44	0.25	2,216,930
Arts/Design/Ent./Sports/Media	0.59	0.80	0.49	0.24	2,039,830
Business & Financial Ops.	0.49	0.89	0.47	0.23	10,087,850
Architecture & Engineering	0.49	0.84	0.46	0.22	2,523,090
Edu. Instruction & Library	0.46	0.89	0.46	0.21	8,328,920
Life, Physical, Social Science	0.39	0.88	0.46	0.19	1,381,930
Personal Care & Service	0.39	0.90	0.45	0.18	2,959,620
Food Prep. & Serving	0.32	0.91	0.43	0.17	13,142,870
Management	0.27	0.90	0.45	0.13	10,445,050
Legal	0.33	0.89	0.42	0.13	1,196,870
Protective Service	0.33	0.84	0.40	0.12	3,484,710
Healthcare Prac. & Tech.	0.25	0.91	0.39	0.12	9,251,930
Production	0.23	0.91	0.41	0.11	8,419,460
Install., Maint., Repair	0.22	0.92	0.41	0.10	5,979,150
Transport. & Material Moving	0.21	0.92	0.38	0.10	13,664,940
Building/Grounds Clean./Maint.	0.15	0.94	0.38	0.08	4,403,350
Construction & Extraction	0.16	0.92	0.40	0.07	6,188,720
Farming, Fishing, & Forestry	0.11	0.92	0.39	0.06	422,740
Healthcare Support	0.13	0.90	0.38	0.05	7,063,540

Coverage, completion, scope, and score are first calculated at the occupation-level and then averaged across occupations in the major group, weighted by employment; each is the mean of user goal and AI action. Coverage, completion and scope are all importance- and relevance-weighted averages of the metrics for IWAs done by a occupation. Coverage is whether the IWA has activity share of at least .05%. Completion is the share of conversations in an IWA where the LLM completed the user's task. Scope is the share of conversations where the scope was at least moderate. Score is the AI applicability score (see Equation (1)). Employment is the total employment for occupations in that group according to BLS.

Table S3: **Top 40 occupations with highest AI applicability score.** See note in Table S2 for column definitions

Job Title (Abbrv.)	Coverage	Cmpltn.	Scope	Score	Employment
Interpreters and Translators	0.980	0.883	0.568	0.492	51,560
Historians	0.906	0.852	0.564	0.462	3,040
Writers and Authors	0.850	0.841	0.603	0.454	49,450
Sales Representatives of Services	0.842	0.902	0.573	0.449	1,142,020
Computer Numerically Controlled Tool Programmers	0.899	0.869	0.528	0.419	28,030
Broadcast Announcers and Radio Disc Jockeys	0.745	0.841	0.605	0.409	25,070
Customer Service Representatives	0.720	0.901	0.591	0.408	2,858,710
Telemarketers	0.656	0.893	0.603	0.404	81,580
Political Scientists	0.774	0.873	0.528	0.391	5,580
Mathematicians	0.905	0.744	0.538	0.386	2,220
News Analysts, Reporters, and Journalists	0.815	0.811	0.560	0.383	45,020
Passenger Attendants	0.798	0.875	0.616	0.376	20,190
Technical Writers	0.826	0.816	0.538	0.373	47,970
Concierges	0.699	0.880	0.556	0.372	41,020
Proofreaders and Copy Markers	0.911	0.860	0.486	0.369	5,490
Editors	0.780	0.818	0.537	0.367	95,700
Business Teachers, Postsecondary	0.700	0.904	0.517	0.367	82,980
Public Relations Specialists	0.627	0.900	0.600	0.365	275,550
Data Scientists	0.765	0.863	0.506	0.357	192,710
Personal Financial Advisors	0.689	0.875	0.524	0.355	272,190
Web Developers	0.733	0.858	0.506	0.353	85,350
Advertising Sales Agents	0.660	0.901	0.526	0.353	108,100
Management Analysts	0.676	0.896	0.540	0.353	838,140
Geographers	0.771	0.828	0.478	0.352	1,460
Brokerage Clerks	0.742	0.892	0.574	0.350	48,060
Market Research Analysts and Marketing Specialists	0.707	0.897	0.517	0.350	846,370
Economics Teachers, Postsecondary	0.676	0.899	0.512	0.349	12,210
Public Safety Telecommunicators	0.660	0.878	0.534	0.346	97,820
Counter and Rental Clerks	0.622	0.900	0.523	0.344	390,300
Telephone Operators	0.804	0.859	0.567	0.342	4,600
Library Science Teachers, Postsecondary	0.654	0.904	0.512	0.341	4,220
Tax Examiners and Collectors, and Revenue Agents	0.534	0.889	0.557	0.340	50,250
Political Science Teachers, Postsecondary	0.653	0.898	0.503	0.339	17,090
Philosophy and Religion Teachers, Postsecondary	0.654	0.897	0.509	0.338	20,320
Models	0.642	0.888	0.533	0.337	3,090
Mathematical Science Occupations, All Other	0.646	0.882	0.527	0.336	4,320
Computer User Support Specialists	0.720	0.875	0.475	0.334	689,700
Criminal Justice and Law Enforcement Teachers, Postsecondary	0.648	0.898	0.506	0.334	13,390
Child, Family, and School Social Workers	0.691	0.878	0.469	0.334	352,160
Foreign Language and Literature Teachers, Postsecondary	0.638	0.899	0.509	0.333	20,820

Table S4: **Bottom 40 occupations with lowest AI applicability score.** See note in Table S2 for column definitions

Job Title (Abbrv.)	Coverage	Cmpltn.	Scope	Score	Empl.
Helpers–Extraction Workers	0.089	0.956	0.336	0.028	7,360
Helpers–Painters, Paperhangers, Plasterers, and Stucco Masons	0.042	0.957	0.376	0.027	7,700
Plant and System Operators, All Other	0.053	0.927	0.375	0.026	15,370
Oral and Maxillofacial Surgeons	0.053	0.894	0.340	0.026	4,160
Embalmers	0.066	0.545	0.220	0.026	3,380
Automotive Glass Installers and Repairers	0.045	0.930	0.342	0.025	16,890
Ship Engineers	0.050	0.918	0.386	0.025	8,860
Tire Repairers and Changers	0.044	0.947	0.352	0.023	101,520
Phlebotomists	0.057	0.952	0.286	0.022	137,080
Hazardous Materials Removal Workers	0.045	0.953	0.347	0.022	49,960
Helpers–Production Workers	0.037	0.925	0.362	0.021	181,810
Highway Maintenance Workers	0.032	0.963	0.319	0.021	150,860
Medical Equipment Preparers	0.038	0.961	0.309	0.021	66,790
Packaging and Filling Machine Operators and Tenders	0.037	0.914	0.389	0.020	371,600
Machine Feeders and Offbearers	0.046	0.889	0.360	0.019	44,500
Dishwashers	0.029	0.947	0.301	0.018	463,940
Cement Masons and Concrete Finishers	0.033	0.921	0.391	0.015	203,560
Prosthodontists	0.099	0.896	0.288	0.015	570
Industrial Truck and Tractor Operators	0.031	0.944	0.281	0.013	778,920
Massage Therapists	0.102	0.908	0.321	0.012	92,650
Surgical Assistants	0.029	0.784	0.288	0.012	18,780
Tire Builders	0.028	0.927	0.399	0.012	20,660
Helpers–Roofers	0.016	0.943	0.368	0.010	4,540
Gas Compressor and Gas Pumping Station Operators	0.015	0.957	0.472	0.010	4,400
Roofers	0.018	0.942	0.376	0.009	135,140
First-Line Supervisors of Firefighting and Prevention Workers	0.040	0.883	0.385	0.009	84,120
Roustabouts, Oil and Gas	0.013	0.952	0.392	0.009	43,830
Maids and Housekeeping Cleaners	0.022	0.935	0.336	0.008	836,230
Ophthalmic Medical Technicians	0.037	0.890	0.329	0.007	73,390
Paving, Surfacing, and Tamping Equipment Operators	0.009	0.958	0.293	0.007	43,080
Logging Equipment Operators	0.012	0.955	0.364	0.005	23,720
Motorboat Operators	0.007	0.934	0.387	0.003	2,710
Orderlies	0.000	0.761	0.181	0.000	48,710
Floor Sanders and Finishers	0.000	0.937	0.337	0.000	5,070
Pile Driver Operators	0.000	0.976	0.236	0.000	3,010
Rail-Track Laying and Maintenance Equipment Operators	0.000	0.962	0.271	0.000	18,770
Foundry Mold and Coremakers	0.000	0.952	0.361	0.000	11,780
Water and Wastewater Treatment Plant and System Operators	0.000	0.919	0.439	0.000	120,710
Bridge and Lock Tenders	0.000	0.931	0.386	0.000	3,460
Dredge Operators	0.000	0.986	0.223	0.000	940

Table S5: **Occupations with the largest difference in user goal and AI action applicability score percentiles.**

AI assistance, not performance	AI performance, not assistance
Cooks, Fast Food (83, 6)	Exercise Trainers and Group Fitness Instructors (17, 75)
Meat Cutters and Trimmers (79, 6)	Choreographers (34, 81)
Butchers and Meat Cutters (83, 11)	Training and Development Managers (45, 88)
Animal Breeders (76, 6)	Coaches and Scouts (43, 80)
Cooks, Private Household (97, 32)	Actuaries (42, 76)
Cooks, Restaurant (76, 12)	Environmental Engineers (55, 88)
Photographers (95, 33)	Human Resources Specialists (53, 86)
Archivists (98, 45)	First-Line Supervisors of Non-Retail Sales Workers (48, 77)
Lighting Technicians (82, 35)	First-Line Sups. of Office and Admin. Supp. Workers (66, 92)
Door-to-Door Sales Workers and Related (94, 49)	Air Traffic Controllers (50, 75)

This Table shows the 10 occupations on each side with the largest difference in their AI applicability score percentile computed from user goals and AI actions, filtered for occupations in the top quartile on their higher-ranked side. Occupation title abbreviated. Numbers in parentheses are (user goal AI applicability score percentile, AI action applicability score percentile). The occupations on the left focus on physical occupations involving cooking and working with animals, while many of the occupations on the right involve teaching, training, or coaching.