Working with AI: Measuring the Occupational Implications of Generative AI*

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

[1]Microsoft Research
[2]Microsoft

Abstract

Given the rapid adoption of generative AI and its potential to impact a wide range of tasks, understanding the effects of AI on the economy is one of society's most important questions. In this work, we take a step toward that goal by analyzing the work activities people do with AI, how successfully and broadly those activities are done, and combine that with data on what occupations do those activities. We analyze a dataset of 200k anonymized and privacy-scrubbed conversations between users and Microsoft Bing Copilot, a publicly available generative AI system. We find the most common work activities people seek AI assistance for involve gathering information and writing, while the most common activities that AI itself is performing are providing information and assistance, writing, teaching, and advising. Combining these activity classifications with measurements of task success and scope of impact, we compute an AI applicability score for each occupation. We find the highest AI applicability scores for knowledge work occupation groups such as computer and mathematical, and office and administrative support, as well as occupations such as sales whose work activities involve providing and communicating information. Additionally, we characterize the types of work activities performed most successfully, how wage and education correlate with AI applicability, and how real-world usage compares to predictions of occupational AI impact.

1 Introduction

General purpose technologies [7], such as the steam engine and the computer, have historically been strong drivers of economic growth, impacting a broad range of sectors and accelerating this impact with each new technical advancement. In the last several years, generative AI has come to the fore as the next candidate general purpose technology [17], capable of improving or speeding up tasks as varied as medical diagnosis [27] and software development [14]. These capabilities are reflected in the astounding rate of AI adoption: nearly 40% of Americans report using generative AI at home or work, 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 are being most affected by AI and, by extension, which occupations.

We provide evidence towards answering this question by identifying the work activities performed in real-world usage of a mainstream large language model (LLM)-powered generative AI system, Microsoft Bing Copilot (now Microsoft Copilot). We analyze 200k anonymized user–AI conversations, which were automatically scrubbed for any personally identifiable information, sampled representatively from 9 months of Copilot usage in the U.S. during 2024. A key insight of our analysis is that there are two distinct ways in

*This study was approved by Microsoft IRB #11028. We thank Jennifer Neville, Ashish Sharma, Hancheng Cao, 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. Correspondence to: kitomlinson@microsoft.com, sojaffe@microsoft.com, suri@microsoft.com.

which a single conversation with an AI assistant can affect the workforce, corresponding to the two parties engaged in conversation. 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. In addition, 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 figure out how to print a document, the user goal is to operate office equipment, while the AI action is to train others to use equipment.

To measure how AI usage indicates potential occupational impact, we classify conversations into work activities as defined by the O*NET database [29], which decomposes occupations hierarchically into the work activities performed in those occupations. We measure how successfully different work activities are assisted or performed by AI, using both explicit thumbs up and down feedback from users and a task completion classifier. To distinguish between broad and narrow AI contributions towards work activities, we also classify the scope of AI impact demonstrated in each conversation toward each matching work activity. From these classifications, we compute an *AI applicability score* for each occupation. This score captures if there is non-trivial AI usage that successfully completes activities corresponding to significant portions of an occupation's tasks.

Our user goal vs. AI action distinction, combined with their classification into work activities, relates to a key question in the literature and public discourse around AI: to what extent is AI automating vs. augmenting work activities? The implication is that augmentation will raise wages and automation will lower wages or lead to job loss. However, this question often conflates the capability of a new technology with the downstream business choices made as a result of that technology. For example, if AI makes software developers 50% more productive, companies could raise their ambitions and hire more developers as they are now getting more output per developer, or hire fewer developers because they can get the same amount done with fewer of them. Our data is only about AI usage and we have no data on the downstream impacts of that usage, so we only weigh in on the automation vs. augmentation question by separately measuring the tasks that AI performs and assists.

We find that information gathering, writing, and communicating with others are the most common user goals in Copilot conversations. In addition to being the most common user goals, information gathering and writing activities receive the most positive thumbs feedback and are the most successfully completed tasks. On the AI action side, we see that AI often acts in a service role to the human as a coach, advisor, or teacher that gathers information and explains it to the user. Furthermore, the activities that AI performs are very different from the user goals the AI assists: in 40% of conversations, these sets are disjoint. To measure occupation-level impacts, we use the standard practice of decomposing an occupation into its constituent work activities [4]. The occupations with highest AI applicability scores are knowledge work and communication focused occupations, but we find that all occupational groups have at least some potential for AI impact (unsurprisingly, with much narrower effects on occupations with large physical components). More specifically, we find the major occupation categories with the highest AI applicability scores are Sales; Computer and Mathematical; Office and Administrative Support; Community and Social Service; Arts, Design, Entertainment, Sports, and Media; Business and Financial Operations; and Educational Instruction and Library. Overall, our measurements largely align with predictions of AI labor impact made by Eloundou et al. [17], with correlation $r = 0.73$ between their occupation-level impact predictions and our AI applicability score ($r = 0.91$ at the broadest occupation group level). We find a weak correlation between AI applicability scores and educational requirements, with occupations requiring a Bachelor's degree slightly more affected than jobs with lower requirements. In addition, we find only a slightly higher average AI applicability for high- (though not highest-) wage occupations.

2 Related work

A growing set of studies examine to what extent AI improves outcomes such as productivity in specific occupational tasks like programming [32, 14], customer support [10], medical diagnosis [22, 27, 23], writing [30], consulting [15], advertising [12], entrepreneurship [31], and legal analysis [13], among other settings. Rather

than measuring the effects of AI on productivity, the focus of our work is to understand what work activities are people using AI for. To that end, we measure how people use LLMs in the wild.

Our work draws from a common economic framework tracing its roots to Autor et al. [4], who decomposed an occupation into the tasks commonly done by that occupation and estimated how susceptible those tasks are to automation. This, in turn, lets one estimate job-level impacts. This technique has become a standard practice in the economics literature [1, 21, 20, 17, 9, 8, 34] and the business world [26]. Some of these papers decompose an occupation into tasks to explain how previous forms of automation affected the labor market [4, 1], while others use them to predict how future forms of automation [21, 26], such as AI [20, 17, 9, 8, 34, 18], will affect occupations. One notable recent work in this space is by Eloundou et al. [17], who predict (using both human and LLM judgments) which tasks and which jobs are most likely to be impacted by the recent advances in LLM technology. We contribute to this literature by analyzing actual conversations between humans and an LLM and showing which work activities those humans are using the LLM for. In addition, we compare our findings to the predictions of Eloundou et al. [17].

The study most similar to ours is a recent analysis by Handa et al. [24] of Claude conversations focused on the economic activities that users perform on that AI platform. Like us, Handa et al. [24] classify conversations according the O*NET taxonomy, although there are several distinguishing features of our approaches. First, we separately classify that activity that the user is seeking assistance with and the activity the AI is performing, which allows us to separate AI assistance from direct AI actions. Second, we incorporate task success and scope of impact into our AI applicability score, providing more nuanced estimates of potential for AI impact. Third, we use different parts of the O*NET taxonomy, focusing on *work activities* (which apply across occupations) rather than *tasks* (which are occupation-specific). This allows us to identify how a particular instance of AI usage impacts all occupations for which that activity is relevant rather than needing to assign a particular occupation to that conversation, which introduces noise to the data since people in different occupations often do indistinguishable tasks. The smaller number of work activities (332 compared to $>$ 18k tasks) also allows us to do exhaustive binary classification, finding all relevant work activities for every conversation, rather than the hierarchical classification approach of Handa et al. [24] that assigns a single task to each conversation (and, by association, a single occupation). Finally, we believe it is valuable for such analyses to be conducted across various AI platforms, as we find that the distribution of Copilot usage differs substantially from Claude, with considerably less focus on computer and mathematical tasks. By combining the results of various such studies, we can build a fuller picture of overall AI impact.

3 Data and methods

3.1 Bing Copilot data

We analyze two collections of anonymized U.S. conversation data from Microsoft Bing Copilot (henceforth, Copilot) gathered over a nine-month period 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. We denote our main data set Copilot-Uniform, which consists of approximately 100k conversations sampled uniformly from conversations in the United States over this time period. Copilot-Uniform provides a representative view of what tasks users perform with a mainstream, publicly available, free-to-use generative AI chatbot. This dataset underlies the majority of analyses in this work.

In Copilot, a user can provide feedback on an LLM response by clicking a thumbs-up or a thumbs-down icon. To take advantage of this valuable signal of user satisfaction, we use a supporting data set denoted Copilot-Thumbs consisting of 100k uniformly sampled conversations containing at least one thumbs up or thumbs down reaction. Copilot-Thumbs allows us to investigate what activities are performed more or less successfully, as measured by explicit user feedback. Note that Copilot-Thumbs may not be representative of overall task success, as some types of users may be more likely to provide feedback, or some types of tasks may be more likely to elicit feedback from users. This motivates our use of an LLM classifier to evaluate whether a conversation completed the user's task, as described in Section 3.3.

Table 1: Example occupation and work activities from O*NET.

Occupation	**Task** $\xrightarrow[\text{many–many}]{}$	**DWA** $\xrightarrow[\text{many–1}]{}$	**IWA** $\xrightarrow[\text{many–1}]{}$	**GWA**
Economists	Compile, analyze, and report data to explain economic phenomena and forecast market trends, applying mathematical models and statistical techniques.	Forecast economic, political, or social trends.	Analyze market or industry conditions.	Analyzing Data or Information.

3.1.1 User goals and AI actions

A key insight of our analysis is that there are two distinct ways in which a single conversation with an AI assistant can affect the workforce, corresponding to the two parties engaged in conversation. First, the user has some task in mind with which they are seeking assistance from the AI, which we call the *user goal*. If the user goal is described by some work activity, then the conversation provides evidence that people are seeking AI assistance with that work activity. On the other hand, the AI itself can perform a work activity in the conversation, which we call the *AI action*. The AI action represents work which may otherwise have been performed by a third party.

Even in successful conversations, the AI action and user goal may not be the same: for instance, in research-based tasks, the user's goal is to gather information (a work activity performed by journalists, scientists, etc), while the AI's action is to provide information (a work activity performed by receptionists, librarians, customer service agents, etc). Another common example of asymmetric user goal and AI action is resolving computer issues (user goal) and providing technical support (AI action). The user goal and AI action may also be the same, e.g., in the case of content generation.

3.2 O*NET and BLS data

To understand the structure and scope of labor in the United States, we draw on the O*NET 29.0 Database[1]. In particular, we use O*NET's hierarchical decomposition of occupations into their tasks and work activities. At the lowest level of the O*NET hierarchy, an *occupation* contains a set of *tasks* performed in that occupation. Each task is mapped to a set of *detailed work activities* (DWAs), which are more general descriptions of work that apply to tasks that span different occupations. Every DWA belongs to an *intermediate work activity* (IWA), which in turn belongs to a *generalized work activity* (GWA); these provide more and more general groupings of similar work activities. See Table 1 for an example. Our analysis focuses on IWAs, which map to multiple occupations through tasks. For instance, the IWA *Analyze market or industry conditions* from the example is also performed by Marketing Managers, Credit Analysts, and Political Scientists, among 29 total O*NET occupations. We combine O*NET with data on wages and employment from the *Occupational Employment and Wage Statistics* data published by the U.S. Bureau of Labor Statistics (BLS)[2].

3.3 Work activity classification

For each conversation in our datasets, we use a GPT-4o-based LLM classification pipeline to identify *all* intermediate work activities (IWAs) that match the user goal and the AI action. If the user goal or AI action is not related to any work activity, then it should be matched with zero IWAs. We validate our classifiers using labels from three human annotators who were blind to the output of the classifier; see Appendix B for details about the pipeline, prompts, and validation. We chose to classify at the IWA rather than task level for several reasons. First, classifying into IWAs is likely to be more accurate and reliable: there are 332 IWAs, most of which are fairly distinct and non-overlapping, but there are 18,796 tasks, with a lot of redundancy. For instance, exactly one IWA describes all programming work activities (*Program computer*

[1]Developed under U.S. Department of Labor sponsorship [29].
[2]From May 2024 [36]. See Appendix A.1 for details about merging the datasets.

systems or production equipment), whereas many O*NET occupations have (distinct) tasks that involve programming (e.g., Data Scientists, Web Developers, and Database Architects, among 30 others). Since we do not know the occupations of users, we cannot hope to reliably distinguish between different programming tasks. Second, since our research question is to understand the potential impact of AI on occupations, we need to understand, to the extent possible, all of the occupations that do a work activity. IWA-level classification allows us identify how capabilities demonstrated in one context translate to all occupations that perform that work activity.

Since each conversation can be assigned multiple IWAs, we focus on the *activity share* each IWA comprises, where we allocate an equal fraction of each conversation to each IWA it is labeled with, separately on the user and AI sides.

3.4 Occupational coverage and AI applicability score

To measure the potential for impact on occupations we define a holistic *AI applicability score* for each occupation, where a higher score for an occupation means it is more likely to be impacted than an occupation with a lower score. The score captures whether AI is being used (with sufficient activity share) for the work activities of an occupation and whether that usage tends to be successful (completion rate) and cover a moderate share of the work activity (scope), which we describe in turn.

We start by considering the work activities that are done a non-trivial amount with Copilot. We use a threshold of 0.05% activity share[3] above which we consider an IWA to appear in our data non-trivially often, which we refer to as "covered." We then use this as a signal that AI can potentially assist or perform that IWA. To account for the fact that some tasks are more central to a job than others, we use the task relevance and importance metrics in O*NET to get a weight w_{ij} for each occupation-IWA pair, with weights summing to one within an occupation (see Appendix A.3 for details). We define the *coverage* of an occupation to be the weighted fraction of its IWAs that are covered. Figure A12 shows how the average and standard deviation of the occupation coverage varies with the threshold, and Figure A11 shows the distribution of coverage scores. We chose the threshold 0.05% to minimize the number of occupations assigned coverage 0 or 1, thereby maximizing the usefulness of the measure for relative comparisons between occupations; see Figure A13. The ordering of occupations induced by our AI applicability score is robust to the chosen coverage threshold; see Figure A14.

Next, work activities that are completed more successfully with Copilot are more likely to experience AI impact. Thus, we also perform a task completion classification with an LLM. For each conversation, we ask GPT-4o-mini[4] if the AI completed the user's task in the conversation. We validated our completion prompt (see Appendix B.1.1) with our Copilot-Thumbs dataset telling us which work activities receive the most positive user feedback, which we find to be highly correlated with task completion (weighted $r > 0.75$; see Figure A15).

For each matching 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 *impact scope* (or simply scope), measured on a six-point Likert scale: none, minimal, limited, moderate, significant, complete. The goal of impact scope is to distinguish between cases where Copilot assists with a large fraction of the work in an IWA (e.g., *Edit written documents or materials* when Copilot edits a report) and a small portion (e.g., *Research biological or ecological phenomena* when the user ask what a mitochondrion is). As with the IWA classification, we validate the scope classifiers with human judges blind to the classifier outputs; see Appendix B for details.

We aggregate these measures into an occupational AI applicability score a_i^{user}, which for occupation i calculated from user goals is

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

[3]Approximately equal to appearing in 100-300 conversations in our 100k samples, before converting to activity share.

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

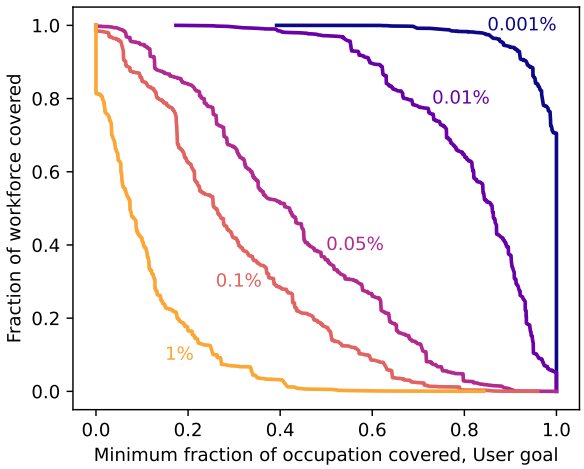


Figure 1: Effect of coverage threshold on absolute impact estimates

Note: 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.

where IWAs(i) is the set of IWAs performed by occupation i, $w_{ij} \in [0,1]$ is the importance- and relevance-weighted fraction of work in i composed of IWA j, $f_j^{\text{user}} \in [0,1]$ 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 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, and report $a_i = (a_i^{\text{user}} + a_i^{\text{AI}})/2$ unless otherwise specified.

We briefly contrast our approach of using a score for relative comparisons with a common metric in the literature, a measurement [24] or prediction [17] of the fraction of occupations or of the workforce that have at least $x\%$ of their tasks impacted by AI. For instance, Eloundou et al. [17] predict that 80% of the U.S. workforce could have at least 10% of their tasks affected by LLMs and 19% could have 50% of their tasks affected.[5] 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 1 shows that by picking different usage thresholds, we can conclude that either $\sim 0\%$ of the workforce has 50% of its importance-weighted tasks represented in our data (if we require 1% of chat activity for a task to be covered) or $\sim 100\%$ of the workforce (if we only require .01% of activity). As such, we believe it is much more meaningful to make relative statements about different kinds of occupations (who is more or less impacted, which is robust to arbitrary thresholds; see Figure A14) from this kind of usage data, which is what our AI applicability score is designed to do.

4 Results

4.1 Generalized Work Activities

Since GWAs are at the highest level of the O*NET work activity hierarchy, we use them for a macroscopic understanding of our data before focusing the rest of our analyses on the more specific IWAs. Figure 2 shows the activity shares we see in Bing Copilot aggregated to GWAs, alongside the estimated fractions of

[5]Similarly, Handa et al. [24] report that 36% of occupations have at least 25% of their tasks with Claude usage, with a threshold of 15 or more conversations across 5 or more user accounts in their sample (approximately 0.0015% of conversation). This type of number is sensitive to the chosen threshold.

Figure 2: Frequency of O*NET Generalized Work Activities (GWAs) in Copilot usage

Note: This Figure shows the share of user goals and AI actions mapping to each GWA, alongside our estimate of how much of the total work in the U.S. falls under each GWA. See Appendix A.2 for how the workforce share is calculated.

the GWAs that appear in the workforce, computed from O*NET and BLS statistics (see Appendix A.2 for how we estimate the total fraction of work in the U.S. falling under each IWA/GWA).

The GWAs where the amount of work in the workforce substantially exceeds the fractions we see in our data generally align with types of work activities for which an LLM chatbot is ill-suited. These fall into three broad clusters: physical activities (e.g., *Handling and Moving Objects*, *Performing General Physical Activities*), monitoring (e.g., *Monitoring Processes*, *Monitoring Resources*, *Inspecting Equipment*), and guiding people or machines (e.g., *Controlling Machines*, *Guiding Subordinates*).

The GWAs more prevalent in Copilot data than in the workforce include GWAs such as *Getting Information*, *Interpreting Information*, *Thinking Creatively*, *Updating and Using Knowledge*, and *Working with Computers*. These align with *knowledge work* [16], which concerns ideas and information rather than physical goods or services, typically involving non-routine and creative problem-solving [28, 33, 35]. These GWAs show a focus of generative AI users on knowledge work activities, in line with findings from prior research [35, 24].

The GWAs that are more prevalent as an AI action (blue) than as a user goal (red) largely fall into two

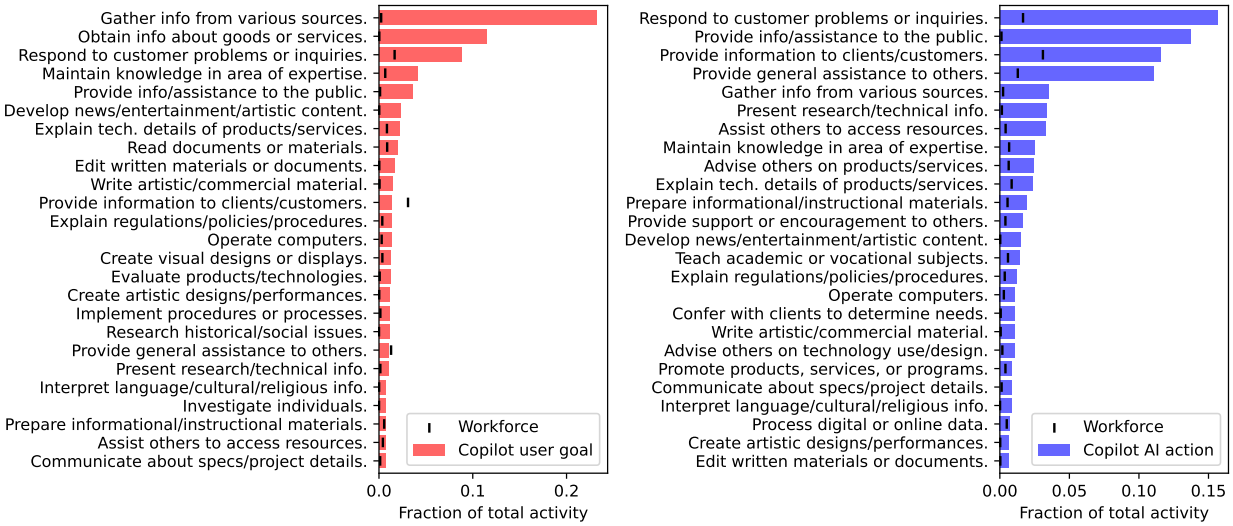


Figure 3: Frequency of top IWAs

Note: This Figure shows the share of user goals (left) and AI actions (right) mapping to each IWA in the top 25 on each side, alongside our estimate of how much of the total work in the U.S. falls under each IWA. See Appendix A.2 for how the workforce share is calculated. IWA titles have been shortened for space.

clusters: service to the user (e.g., *Assisting/Caring for Others*, *Providing Advice*, *Coaching*, *Training*) and communication (e.g., *Communicating with People*, *Communicating with Supervisors*). Conversely, the GWAs more prevalent as a user goal than AI action are mostly related to knowledge work (e.g., *Getting Information*, *Thinking Creatively*, *Updating and Using Knowledge*, *Making Decisions*, *Analyzing Data*). Thus, we find that people are using Copilot to provide services for the execution of knowledge work activities, and do so disproportionately often relative to the fraction of knowledge work in the workforce.

4.2 Intermediate Work Activities

We next turn to the data at the disaggregated IWA level. Figure 3 (left) shows which IWAs are most common as Copilot user goals; these fall into three broad categories: gathering information (e.g., *Gather information*, *Obtain information*, *Maintain knowledge*, *Read documents*), writing, editing, or developing content (e.g., *Develop content*, *Write material*, *Create visual designs*), and communicating to others (e.g., *Provide information*, *Provide assistance*, *Explain technology*, *Explain regulations*).

The IWAs reflected in the AI actions tell a complementary story. Figure 3 (right) shows that the AI plays a service role: some common IWA verbs include *Respond*, *Provide*, *Present*, and *Assist*. More specifically, Figure 3 shows that the most frequent IWAs fall into three broad categories: gathering and reporting information (e.g., *Gather information*, *Prepare informational materials*, *Develop content*), explaining information (e.g., *Present research*, *Explain technical details*, *Explain regulations*), and communicating with the user (e.g., *Respond to customer problems*, *Provide assistance*, *Provide information*, *Advise others*). Combining the user goal and AI action IWAs again shows that humans are using AI to gather, process, and disseminate information while the AI is helping by gathering, explaining, and communicating information to the user.

Figure 3 shows that there is overlap between the activities on user and AI sides, but also some interesting differences. At the conversation level, the asymmetry is surprisingly pronounced: 40% of conversations have disjoint sets of user goal and AI action IWAs, and 96% have more IWAs unique to each side than in common (i.e., Jaccard index < 0.5). Overall, the AI tends to do more advising and teaching whereas the user side involves more obtaining information, reading, and researching. Table 2 further investigates these differences

Table 2: 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 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 clients/customers. (8.6x)
Obtain formal documentation or authorization. (12.5x)	Advise others on workplace health/safety. (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. (6.5x)

*N*ote: Only includes IWAs with user or AI activity share $\geq 0.05\%$. Numbers show IWA overrepresentation factors.

by listing the IWAs where we see the biggest (relative) differences in user and AI activity shares. Naturally, the AI is much more likely to assist (rather than perform) activities that involve a physical component, such as athletic activities and operating equipment, as well as activities that require interacting with other entities, such as purchasing goods and executing financial transactions (here, IWA verbs are very active: *Purchase*, *Execute*, *Perform*, *Obtain*, etc.). On the other hand, the AI is much more likely to perform activities related to training, coaching, teaching, and advising.

4.2.1 Satisfaction, task completion, and scope

To go beyond mere usage and map out potential impact on occupations, we need to understand if the LLM is actually helpful for these work activities. We use three different metrics to measure different aspects of that question, one based on user feedback and two based on LLM analysis of the conversations.

Satisfaction and completion. To measure how successfully different work activities are assisted and performed by Copilot, we use user thumbs feedback as a signal of satisfaction and an LLM task completion classifier, as described in Section 3.4. For satisfaction, we report the share of feedback on conversations in Copilot-Thumbs matched to an IWA that is positive, i.e., the number of conversations with thumbs up over the total number of conversations with thumbs feedback. Figure 4 highlights the top and bottom 15 IWAs by the fraction of feedback which is positive, after removing rare IWAs. All common IWAs have a positive feedback share of 50% or higher showing that, overall, people find Copilot helpful. More specifically, we find that three types of work activities tend to have particularly positive feedback: those involving writing and editing text (*Edit documents*, *Write material*), researching information (e.g., *Research healthcare issues*, *Research laws*, *Maintain knowledge*), and evaluating or purchasing goods (e.g., *Purchase goods*, *Evaluate characteristics of products*, *Select materials*). In contrast, we find that work activities involving data analysis (e.g., *Process data*, *Calculate financial data*, *Analyze scientific data*) or visual design (e.g., *Create visual/artistic designs*, *Arrange displays*) have the worst feedback. These results suggest that Copilot is better at the writing and researching parts of knowledge work than its analysis and visual components. If we do the same analysis aggregating to the GWA level (see Figure A3), we see that lower-satisfaction GWAs reveal a similar pattern, including *Thinking Creatively*, which the visual design IWAs map up to, and *Processing/Analyzing Information*.

There are a few IWAs that have a noticeably large gap between the fraction of positive feedback when they are a user goal vs. an AI action. Interestingly, the two largest are *Provide support or encouragement to others* and *Advise others on products or services*. When the AI tries to directly provide support or advice, people are less satisfied than when it helps them provide support or advice to others. The GWA-level analysis (Figure A3) also shows that activities involving doing things for others (coaching, providing advice,

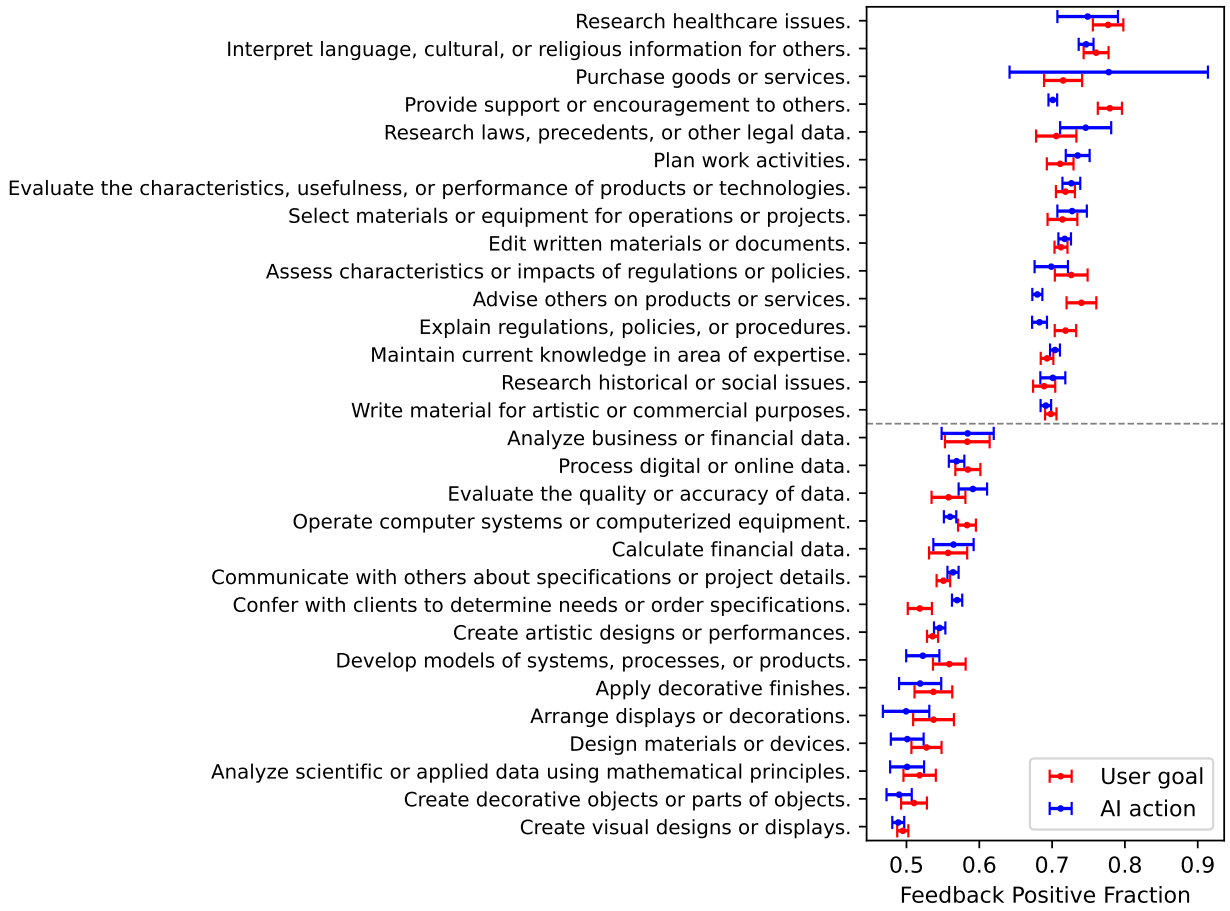


Figure 4: IWAs with the highest and lowest shares of positive feedback

Note: 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.

and interpreting things) stand out with high shares of positive feedback, all with even higher satisfaction when the AI helps the user do them than when it tries to do them itself.

To supplement thumbs feedback, we also look at which work activities have the highest and lowest completion rates, as described in Appendix B.1.1. Relative to the thumbs data, this has the disadvantage of not reflecting the user's opinion, but the advantage of avoiding selection in which users give feedback (which is why we use completion in our AI applicability score). We find that there is a strong correlation between the positive feedback fraction for an IWA and its completion rate (weighted $r = 0.83$ for user goal IWAs and $r = 0.76$ for AI action IWAs, filtering out IWAs below activity share 0.05%; see Figure A15). Moreover, we find very high consistency between IWA completion rates measured in Copilot-Uniform and Copilot-Thumbs (weighted $r > 0.9$, see Figure A17). At the conversation level in Copilot-Thumbs, the correlation between whether a conversation received a thumbs up and whether it was classified as completing the user's task is $r = 0.28$, indicating they are related but that the relationship is noisier before aggregating to IWAs. Figure A4 shows the top and bottom IWAs by completion rate, which mostly shows similar patterns as the top and bottom IWAs by thumbs feedback, with the addition of advice and explanation IWAs having high completion rates.

Scope of impact. In addition to success within a conversation, another crucial aspect of work impact is the extent to which the AI capability demonstrated in the conversation translates to the work represented by an IWA. As described in Section 3.4, we use our measure of impact scope to identify which IWAs are most deeply affected by demonstrated AI capabilities. Figure A5 shows the IWAs with highest and lowest impact scope; as with satisfaction and completion, the most deeply impacted IWAs include gathering information and writing, as well as providing information, advising, and explaining on the AI side. Low impact scope IWAs again include data analysis and visual design, but also others about interacting with external people (e.g., *Confer with clients*, *Coordinate with others*, *Investigate individuals*, *Verify personal information*). Notably, we find consistently lower impact scope on the AI action side than the user goal side: our data indicates that AI can help users with a broader fraction of their work than it can perform directly. Supporting the notion that scope measures something different from completion, we find that scope is much less correlated with completion than satisfaction is (weighted IWA-level $r = 0.45$ and $r = 0.22$; see Figure A16). On the other hand, of these three measures, IWA scope is the best predictor of which activities people seek AI assistance with most often ($r = 0.64$ with log user goal activity share; see Figure A18). That is, people are using LLMs for the tasks for which the LLM can have broadest impact (but not necessarily the ones the LLM completes most successfully).

4.3 Occupations

Table 3 shows the 40 occupations with the highest AI applicability score as defined by Equation (1). Recall that our AI applicability score combines, for each occupation, whether Copilot users are performing its associated work activities (frequency $> .05\%$) successfully (completion rate) and covering a broad share of the work activity (scope $\geq$ moderate). (See Section 3.4 and Equation (1) for more details.) Interpreters and Translators are at the top of the list, with 98% of their work activities overlapping with frequent Copilot tasks with fairly high completion rates and scope scores. Other occupations with high applicability scores include those related to writing/editing, sales, customer service, programming, and clerking. Along with Interpreters and Translators, there are myriad other knowledge work occupations such as Historians, Writers and Authors, CNC Tool Programmers, Brokerage Clerks, Political Scientists, Reporters and Journalists, Mathematicians, Proofreaders, Editors, PR Specialists, etc. By contrast, Table 4 shows the 40 occupations with the lowest AI applicability scores. The least-impacted occupations include occupations that require physically working with people (e.g., Nursing Assistants, Massage Therapists), operating or monitoring machinery (e.g., Water Treatment Plant and Systems Operators, Pile Driver Operators, Truck and Tractor Operators), and other manual labor (e.g., Dishwashers, Roofers, Maids and Housekeeping Cleaners). Note that our measurement is purely about LLMs: other applications of AI could certainly affect occupations involving operating and monitoring machinery, such as truck driving.

Table 3: Top 40 occupations with highest AI applicability score.

Job Title (Abbrv.)	**Coverage**	**Cmpltn.**	**Scope**	**Score**	**Employment**
Interpreters and Translators	0.98	0.88	0.57	0.49	51,560
Historians	0.91	0.85	0.56	0.48	3,040
Passenger Attendants	0.80	0.88	0.62	0.47	20,190
Sales Representatives of Services	0.84	0.90	0.57	0.46	1,142,020
Writers and Authors	0.85	0.84	0.60	0.45	49,450
Customer Service Representatives	0.72	0.90	0.59	0.44	2,858,710
CNC Tool Programmers	0.90	0.87	0.53	0.44	28,030
Telephone Operators	0.80	0.86	0.57	0.42	4,600
Ticket Agents and Travel Clerks	0.71	0.90	0.56	0.41	119,270
Broadcast Announcers and Radio DJs	0.74	0.84	0.60	0.41	25,070
Brokerage Clerks	0.74	0.89	0.57	0.41	48,060
Farm and Home Management Educators	0.77	0.91	0.55	0.41	8,110
Telemarketers	0.66	0.89	0.60	0.40	81,580
Concierges	0.70	0.88	0.56	0.40	41,020
Political Scientists	0.77	0.87	0.53	0.39	5,580
News Analysts, Reporters, Journalists	0.81	0.81	0.56	0.39	45,020
Mathematicians	0.91	0.74	0.54	0.39	2,220
Technical Writers	0.83	0.82	0.54	0.38	47,970
Proofreaders and Copy Markers	0.91	0.86	0.49	0.38	5,490
Hosts and Hostesses	0.60	0.90	0.57	0.37	425,020
Editors	0.78	0.82	0.54	0.37	95,700
Business Teachers, Postsecondary	0.70	0.90	0.52	0.37	82,980
Public Relations Specialists	0.63	0.90	0.60	0.36	275,550
Demonstrators and Product Promoters	0.64	0.88	0.53	0.36	50,790
Advertising Sales Agents	0.66	0.90	0.53	0.36	108,100
New Accounts Clerks	0.72	0.87	0.51	0.36	41,180
Statistical Assistants	0.85	0.84	0.49	0.36	7,200
Counter and Rental Clerks	0.62	0.90	0.52	0.36	390,300
Data Scientists	0.77	0.86	0.51	0.36	192,710
Personal Financial Advisors	0.69	0.88	0.52	0.35	272,190
Archivists	0.66	0.88	0.49	0.35	7,150
Economics Teachers, Postsecondary	0.68	0.90	0.51	0.35	12,210
Web Developers	0.73	0.86	0.51	0.35	85,350
Management Analysts	0.68	0.90	0.54	0.35	838,140
Geographers	0.77	0.83	0.48	0.35	1,460
Models	0.64	0.89	0.53	0.35	3,090
Market Research Analysts	0.71	0.90	0.52	0.35	846,370
Public Safety Telecommunicators	0.66	0.88	0.53	0.35	97,820
Switchboard Operators	0.68	0.86	0.52	0.35	43,830
Library Science Teachers, Postsecondary	0.65	0.90	0.51	0.34	4,220

Note: Metrics reported as mean of user goal and AI action score.

Table 4: Bottom 40 occupations with lowest AI applicability score.

Job Title (Abbrv.)	**Coverage**	**Cmpltn.**	**Scope**	**Score**	**Empl.**
Phlebotomists	0.06	0.95	0.29	0.03	137,080
Nursing Assistants	0.07	0.85	0.34	0.03	1,351,760
Hazardous Materials Removal Workers	0.04	0.95	0.35	0.03	49,960
Helpers–Painters, Plasterers, ...	0.04	0.96	0.38	0.03	7,700
Embalmers	0.07	0.55	0.22	0.03	3,380
Plant and System Operators, All Other	0.05	0.93	0.38	0.03	15,370
Oral and Maxillofacial Surgeons	0.05	0.89	0.34	0.03	4,160
Automotive Glass Installers and Repairers	0.04	0.93	0.34	0.03	16,890
Ship Engineers	0.05	0.92	0.39	0.03	8,860
Tire Repairers and Changers	0.04	0.95	0.35	0.02	101,520
Prosthodontists	0.10	0.90	0.29	0.02	570
Helpers–Production Workers	0.04	0.93	0.36	0.02	181,810
Highway Maintenance Workers	0.03	0.96	0.32	0.02	150,860
Medical Equipment Preparers	0.04	0.96	0.31	0.02	66,790
Packaging and Filling Machine Op.	0.04	0.91	0.39	0.02	371,600
Machine Feeders and Offbearers	0.05	0.89	0.36	0.02	44,500
Dishwashers	0.03	0.95	0.30	0.02	463,940
Cement Masons and Concrete Finishers	0.03	0.92	0.39	0.01	203,560
Supervisors of Firefighters	0.04	0.88	0.39	0.01	84,120
Industrial Truck and Tractor Operators	0.03	0.94	0.28	0.01	778,920
Ophthalmic Medical Technicians	0.04	0.89	0.33	0.01	73,390
Massage Therapists	0.10	0.91	0.32	0.01	92,650
Surgical Assistants	0.03	0.78	0.29	0.01	18,780
Tire Builders	0.03	0.93	0.40	0.01	20,660
Helpers–Roofers	0.02	0.94	0.37	0.01	4,540
Gas Compressor and Gas Pumping Station Op.	0.01	0.96	0.47	0.01	4,400
Roofers	0.02	0.94	0.38	0.01	135,140
Roustabouts, Oil and Gas	0.01	0.95	0.39	0.01	43,830
Maids and Housekeeping Cleaners	0.02	0.94	0.34	0.01	836,230
Paving, Surfacing, and Tamping Equipment Op.	0.01	0.96	0.29	0.01	43,080
Logging Equipment Operators	0.01	0.95	0.36	0.01	23,720
Motorboat Operators	0.01	0.93	0.39	0.00	2,710
Orderlies	0.00	0.76	0.18	0.00	48,710
Floor Sanders and Finishers	0.00	0.94	0.34	0.00	5,070
Pile Driver Operators	0.00	0.98	0.24	0.00	3,010
Rail-Track Laying and Maintenance Equip. Op.	0.00	0.96	0.27	0.00	18,770
Foundry Mold and Coremakers	0.00	0.95	0.36	0.00	11,780
Water Treatment Plant and System Op.	0.00	0.92	0.44	0.00	120,710
Bridge and Lock Tenders	0.00	0.93	0.39	0.00	3,460
Dredge Operators	0.00	0.99	0.22	0.00	940

Note: Metrics reported as mean of user goal and AI action score.

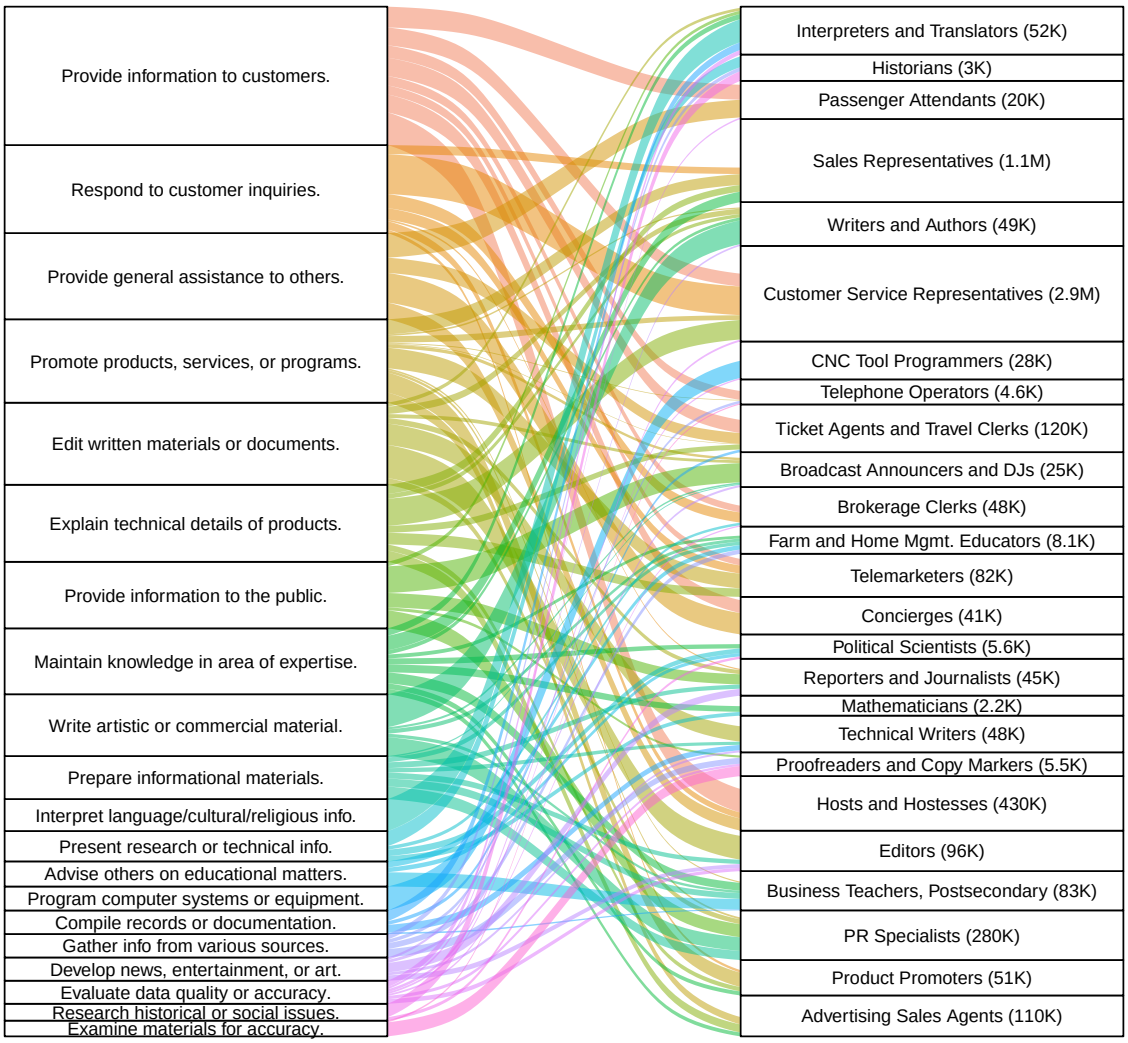


Figure 5: Top occupations by AI applicability score and their contributing IWAs

Note: This Figure shows the 25 occupations with the greatest AI applicability scores along with the 20 IWAs that provide the greatest contributions to those scores. Occupations with higher employment have taller strata on the right. 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. Occupations are sorted by their score, decreasing. Both occupation and IWA titles have been shortened for space.

Figure 5 shows which work activities are contributing to the high applicability scores of the occupations in Table 3. The right side of Figure 5 shows the 25 occupations with the highest AI applicability score. Occupations are in descending order of applicability score and the height of the boxes is indicative of employment (also shown in the labels). The left side shows the work activities that contribute most to the scores for those occupations. The top IWAs involve delivering information to people such as *Provide information to customers*, *Respond to customer inquiries*, *Provide general assistance to others*, and *Provide information to the public*. These IWAs flow into occupations such as Passenger Attendants, Sales Representatives, Customer Service Representatives, Broadcast Announcers, Concierges, Hosts and Hostesses, etc. While it may have been surprising at first glance to see these occupations with high AI applicability scores in Table 3, this is explained by AI's ability to communicate information, which is a substantial component

Table 5: SOC Major groups sorted by AI Applicability Score

Major Group	Coverage	Completion	Scope	Score	Employment
Sales and Related	0.56	0.89	0.51	0.32	13,266,370
Computer and Mathematical	0.64	0.86	0.48	0.30	5,177,390
Office and Administrative Support	0.56	0.89	0.49	0.29	18,163,760
Community and Social Service	0.51	0.88	0.44	0.25	2,216,930
Arts, Design, Entertainment, Sports, Media	0.59	0.80	0.49	0.25	2,039,830
Business and Financial Operations	0.49	0.89	0.47	0.24	10,087,850
Educational Instruction and Library	0.46	0.89	0.46	0.23	8,328,920
Architecture and Engineering	0.49	0.84	0.46	0.22	2,523,090
Personal Care and Service	0.39	0.90	0.45	0.20	2,959,620
Life, Physical, and Social Science	0.39	0.88	0.46	0.20	1,381,930
Food Preparation and Serving Related	0.32	0.91	0.43	0.18	13,142,870
Management	0.27	0.90	0.45	0.14	10,445,050
Protective Service	0.33	0.84	0.40	0.14	3,484,710
Legal	0.33	0.89	0.42	0.13	1,196,870
Healthcare Practitioners and Technical	0.25	0.91	0.39	0.12	9,251,930
Installation, Maintenance, and Repair	0.22	0.92	0.41	0.11	5,979,150
Production	0.23	0.91	0.41	0.11	8,419,460
Transportation and Material Moving	0.21	0.92	0.38	0.11	13,664,940
Building, Grounds Cleaning, Maintenance	0.15	0.94	0.38	0.08	4,403,350
Construction and Extraction	0.16	0.92	0.40	0.08	6,188,720
Farming, Fishing, and Forestry	0.11	0.92	0.39	0.06	422,740
Healthcare Support	0.13	0.90	0.38	0.05	7,063,540

*N*ote: Metrics reported as mean of user goal and AI action

of these occupations.

There are also a number of IWAs related to knowledge work such as *Edit written materials*, *Maintain knowledge*, *Write artistic or commercial material*, *Interpret language/cultural information*, and *Program computers* that flow into knowledge work occupations such as Technical Writers, Editors, Brokerage Clerks, Political Scientists, Mathematicians, Writers, PR Specialists, Interpreters and Translators, and CNC Tool Programmers.

To get a broader view of the applicability of AI to occupations, we aggregate occupations to their Standard Occupational Classification (SOC) major groups, which are 22 broad categories under which every occupation code falls[6] [37]. Aggregating occupations highlights the trend of current AI applicability to knowledge work and communication-oriented occupations. Table 5 shows that Sales and Related, Computer and Mathematical, and Office and Administrative Support occupations have the highest AI applicability scores, with Sales and Office/Administrative Support also being two of the largest groups by employment. Similarly, groups with a large communication component such as Community and Social Service and Educational Instruction also have high AI applicability scores. Conversely, Healthcare Support has the lowest score, along with occupations that involve physical labor or operating machinery such as Farming and Construction. Table A2 provides a more granular view at the SOC minor group level (one level down in the SOC classification hierarchy), where the highest score groups are Media and Communication, Mathematical Science, Sales Representatives of Services, Communications Equipment Operators, and Information and Record Clerks.

Finally, we identify which occupations differ most in the AI applicability scores computed only from user goal IWAs and only from AI action IWAs (all results discussed above combine the two). Table A3 shows occupations that are ranked highly by AI applicability score on one side but not the other. Occupations with potential for AI assistance but not AI performance (high a_i^{user}, low a_i^{AI}) include occupations with

[6]Excluding military occupations, which are not fully represented in O*NET.

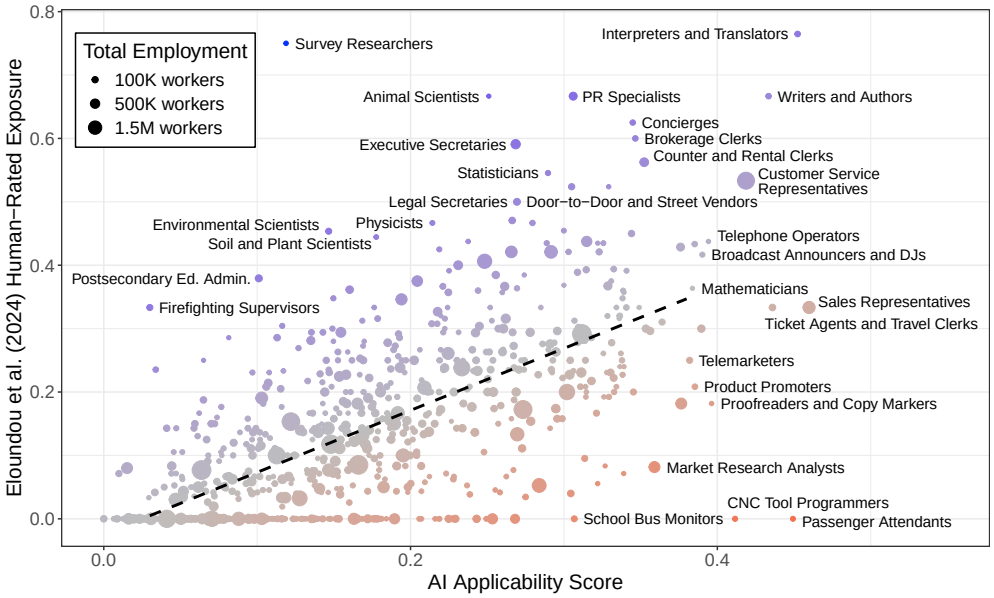


Figure 6: Comparing the AI applicability score to the human-rated E1 exposure from Eloundou et al. [17].

*N*ote: The AI applicability score is the fraction of work activity in an occupation that appears in Copilot data, adjusted by completion rate and impact scope. The Eloundou et al. [17] E1 metric comes from asking human raters whether LLM technology could allow a task to be completed at least 50% faster and then computing the fraction of an occupation's tasks labeled as E1. Occupations are colored by their distance from the regression line: blue points have higher E1 than AI applicability score, and vice versa for red points.

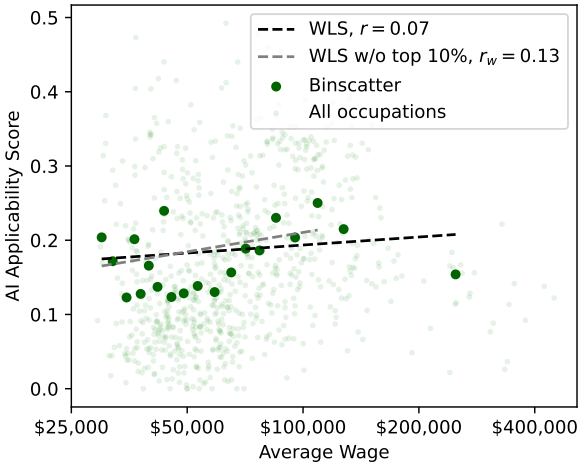
physical components, especially cooking and working with animals, tasks which are commonly assisted but not performed by Copilot (e.g., Cooks and Animal Breeders). Conversely, occupations with potential for AI performance but not assistance (low a_i^{user}, high a_i^{AI}) focus on teaching, training, managing, and communicating (e.g., Training and Development Managers, Coaches and Scouts, and HR Specialists).

4.3.1 Comparing to predictions

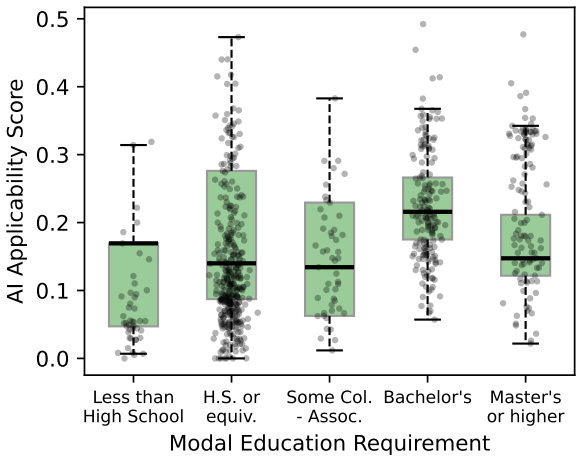
We now examine how our measurements from real-world AI usage data compare to predictions of occupational AI impact. Eloundou et al. [17] asked both people and GPT-4 to predict which tasks would be impacted by LLM technology. For each occupation they then calculated a metric they call E1, "the share of an occupation's tasks where access to an LLM alone or with a simple interface would lead to 50% time savings" [17]. Figure 6 plots E1 against our AI applicability score.[7] We would not necessarily expect alignment between the two metrics, since we cannot assess how much time people are saving on their tasks. However, the occupation-level correlation (weighted by employment) between their predictions and our measurements of occupational AI applicability is $r = 0.73$; this increases to a remarkably high $r = 0.91$ when aggregating occupations to their SOC major groups.

Figure 6 labels some of the occupations where the two metrics diverge. Some red-colored occupations in the lower-right where our estimate is high relative to theirs, such as Market Research Analysts and CNC Tool Programmers, seem like they may have missed some of the potential uses of the technology. Others, such as Passenger Attendants and School Bus Monitors, seem like places where our method is potentially over-extrapolating the tool's ability to *Provide information* to occupations where LLMs may be less relevant. For the blue-colored occupations in the upper-left, where our metric is surprisingly low, we find their low

[7]See Section 3.4. For this comparison only, we use a uniform weighting over tasks to compute AI applicability score to align with the approach of Eloundou et al. [17].



(a) Occupation wage



(b) Education requirements

Figure 7: AI applicability scores across socioeconomic correlates.

*N*ote: These Figures plot occupation AI applicability scores against the (a) average wage and (b) the most common education requirement reported in O*NET surveys. In (a), binscatters and weighted least squares (WLS) fit lines are weighted by employment. Excluding the top 10% of highest-paid workers slightly increases the correlation with wage. In (b), boxplots show medians and quartiles weighted by employment (i.e., the median worker rather than the median occupation)

employment and specialization means that their work activities are rare. Thus, even if an LLM may be well-suited to them, these activities are not done sufficiently often to meet our .05% coverage threshold.

4.3.2 Socioeconomic correlates

It is natural to ask how AI applicability score correlates with wage and education. Some prior work predicts that higher-wage occupations will be substantially more affected by generative AI [17, 19], while other prior work predicts no correlation for pre-LLM machine learning [9]. Figure 7a shows a scatter and binscatter plot of AI applicability score and average occupation wage (computed using BLS data), with a dot for each occupation and darker points for the average of each ventile weighted by employment. Despite looking at this relationship several different ways, we do not find a strong and consistent relationship between AI applicability score and wage. The employment-weighted correlation between AI applicability score and wage is only 0.07 (Figure A6 separates this into user goals, with a correlation of 0.05, and AI actions, with a correlation of 0.10). Since others have found a decrease in AI exposure at the highest-wage occupations [17], we also calculate the employment-weighted correlation omitting these occupations, which is still only 0.13. Figure A8 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 0.21 for AI actions.[8] 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 will be much more affected than others, but the overall relationship between wage and AI applicability is weak.

[8]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.

O*NET also provides the education required for each occupation, from surveys of incumbents. Figure 7b shows the distribution of AI applicability score by the modal education requirement, weighted by employment. Occupations requiring a Bachelor's degree tend to have higher AI applicability score than occupations with lower educational requirements: the employment-weighted mean score for Bachelor's is 0.27, compared to 0.19 for all groups below Bachelor's, a significant difference (weighted t-test $p < 10^{-14}$). Splitting out the user and AI applicability scores, we find the difference to be more pronounced on the AI action side (Figure A7). However, there is still substantial overlap between applicability scores across education requirements. Without employment-weighting, the trend appears monotonic (Figure A9), again due to Sales and Office and Administrative Support that have high AI applicability score and employment but low modal education requirements.

5 Discussion

We analyzed Bing Copilot conversations to see what work activities users are seeking AI assistance with, what activities the AI performs, and what this means about occupations. A work activity seen in current AI interaction data demonstrates an AI capability being leveraged by some users that could extend to other uses and to occupations which perform that activity. We combine this evidence of demonstrated capability with measures of task success and scope of impact into an AI applicability score for occupations, which allows us to track the frontier of AI's relevance to work. The current capabilities of generative AI align most strongly with knowledge work and communication occupations, though most occupations have at least some potential for AI collaboration. Occupations for which the potential is small or non-existent include those involving manual labor, operating machinery, or other physical activities. Turning to socioeconomic correlates, we find a very small positive correlation between our AI applicability measure and occupational wage. In terms of education requirements, we find higher AI applicability for occupations requiring a Bachelor's degree than occupations with lower requirements. However, our data indicate a wide range of potential impact across the wage and education distributions. When comparing to predictions of occupational AI impact [17], we find that these are largely borne out in usage data, especially at the most general, coarsest aggregation levels. However, the magnitude of this impact (if not its direction) remains to be seen.

Our data do not indicate that AI is performing all of the work activities of any one occupation. That being said, the overlap between AI capabilities and various occupations is very uneven. There are definitely some occupations for which many—perhaps even most—work activities have some overlap with demonstrated AI capabilities. But even when there is overlap, the task completion rate is not 100% and the scope of impact is usually moderate. Thus, even when there is overlap between an AI capability and a work activity, it does not mean the work activity is done to its full extent all of the time. Furthermore, there are a few limitations to these analyses that prevent us from assessing the total fraction of work being done with AI. First, we are only able to analyze the data from one widely used, publicly available LLM. Different people use different LLMs for different purposes. Second, decomposing an occupation into its work activities, while standard practice in the literature, does not provide a complete representation of every occupation: the connecting glue between tasks also contributes to the value of work. Finally, this decomposition can only be as accurate and up-to-date as the O*NET database.

One of the key aspects of our analysis is our classification of work activities into actions the AI performs versus user goals the AI assists with. In terms of AI performing actions, we show that it often does so in a supporting role to the human acting as a coach, trainer, or advisor [25]. The most common user goals that Copilot assists with involve gathering information, writing, and communicating. The relatively high prevalence of information gathering may be due to Copilot's connection to the Bing search engine at the time our data originates. Information gathering and writing are also the most successful work activities, as measured by thumbs up, task completion, and impact scope, indicating that Copilot is providing significant useful input to these activities. We also saw that it can be helpful beyond the boundaries of what AI can physically do. For example, it can help people cook by providing recipe and nutritional suggestions without actually performing the cooking activities. Compared to a similar analysis of Claude conversations, Copilot usage is much less focused on programming and mathematical tasks, which comprises more than a third of

"occupationally relevant" Claude usage [24]. As discussed above, this may be due to the different population of users who choose to use one AI assistant versus another.

It is tempting to conclude that occupations that have high overlap with activities AI performs will be automated and thus experience job or wage loss, and that occupations with activities AI assists with will be augmented and raise wages. This would be a mistake, as our data do not include the downstream business impacts of new technology, which are very hard to predict and often counterintuitive [3]. 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 [5].

This work gives rise to a number of future research questions of extremely 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. It could be that jobs change which activities they encompass, as in the case of bank tellers and ATMs. In addition, entirely new occupations may emerge due to the rise of AI, performing new types of work activities [11]. 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 [2]. Exactly which new jobs emerge, and how old ones are reconstituted, is an important future research direction in the AI age. At the same time, the technology itself will continue to evolve; our measurement of AI applicability is only a snapshot in time. An important research question going forward is to understand how the frontier of AI capabilities is shifting, and which occupations have more or less overlap with that moving frontier. Measuring changes in AI usage over time will help reveal how these new capabilities are exploited.

There are some natural limitations, in addition to the ones already stated, to the conclusions that can be drawn from our data. It is very difficult (or impossible) to determine what conversations are performed in a work context or for leisure.[9] As such, we looked for work activities performed in any conversation to find evidence that AI can impact tasks of that type. It is also difficult to determine the magnitude of impact that AI might have on different work activities based only on this conversation data; we attempted to address this issue with measures of task completion and scope of impact, but these are imperfect and approximate. 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). We reiterate that 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. Finally, our use of O*NET means our results are shaped by its U.S.-centric view, may lag behind current actual 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.

References

[1] 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. doi: 10.1257/jep.33.2.3. URL `https://www.aeaweb.org/articles?id=10.1257/jep.33.2.3`.

[2] 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.

[3] 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.

[9] Is someone asking for a recipe a chef brainstorming their new menu or just someone cooking dinner at home? Is someone asking for information about a video game a QA tester, a game developer, or just a gamer? We can make (potentially high-probability) guesses in these cases, but consistently making the highest-probability guess may lead us to conclude that video game QA testers and chefs have no AI impact on their occupations.

[4] 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.

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

[6] Alexander Bick, Adam Blandin, and David J Deming. The rapid adoption of generative AI. Technical report, National Bureau of Economic Research, 2024.

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

[8] Erik Brynjolfsson and Tom Mitchell. What can machine learning do? workforce implications. *Science*, 358(6370):1530–1534, December 2017.

[9] 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. doi: 10.1257/pandp.20181019. URL `https://www.aeaweb.org/articles?id=10.1257/pandp.20181019`.

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

[11] Robert Capps. A.I. Might Take Your Job. Here Are 22 New Ones It Could Give You. *The New York Times Magazine*, June 2025. URL `https://www.nytimes.com/2025/06/17/magazine/ai-new-jobs.html`.

[12] Zenan Chen and Jason Chan. Large language model in creative work: The role of collaboration modality and user expertise. *Management Science*, 70(12):9101–9117, 2024.

[13] Jonathan H Choi, Amy B Monahan, and Daniel Schwarcz. Lawyering in the age of artificial intelligence. *Minn. L. Rev.*, 109:147, 2024.

[14] 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. Technical report, Available at SSRN 4945566, 2024.

[15] Fabrizio Dell'Acqua, Saran Rajendran, Edward McFowland III, Lisa Krayer, Ethan Mollick, François Candelon, Hila Lifshitz-Assaf, Karim R. Lakhani, and Katherine C. Kellogg. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality, 2023.

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

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

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

[19] Ed Felten, Manav Raj, and Rob Seamans. Generative ai requires broad labor policy considerations. *Communications of the ACM*, 67(8):29–32, 2024.

[20] 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. doi: 10.1257/pandp.20181021. URL `https://www.aeaweb.org/articles?id=10.1257/pandp.20181021`.

[21] 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. ISSN 0040-1625. doi: https://doi.org/10.1016/j.techfore.2016.08.019. URL https://www.sciencedirect.com/science/article/pii/S0040162516302244.

[22] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969, 2024.

[23] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

[24] 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 preprint arXiv:2503.04761*, 2025.

[25] Jake M Hofman, Daniel G Goldstein, and David M Rothschild. A sports analogy for understanding different ways to use AI. *Harvard Business Review*, 4, 2023.

[26] 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.

[27] Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7, 2025.

[28] C. McKercher and V. Mosco. *Knowledge Workers in the Information Society*. Critical media studies. Lexington Books, 2008. ISBN 9780739117811. URL https://books.google.com/books?id=_MeCr31C9S8C.

[29] National Center for O*NET Development. O*NET Database Version 29.0, 2024. URL https://www.onetcenter.org/db_releases.html. Accessed: 2025-05-29.

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

[31] Nicholas G Otis, Solène Delecourt, Katelyn Cranney, and Rembrand Koning. *Global Evidence on Gender Gaps and Generative AI*. Harvard Business School, 2024.

[32] 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.

[33] 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. doi: https://doi.org/10.1002/kpm.378. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/kpm.378.

[34] 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, 2025. URL https://arxiv.org/abs/2506.06576.

[35] Siddharth Suri, Scott Counts, Leijie Wang, Chacha Chen, Mengting Wan, Tara Safavi, Jennifer Neville, Chirag Shah, Ryen W White, Reid Andersen, et al. The use of generative search engines for knowledge work and complex tasks. *arXiv preprint arXiv:2404.04268*, 2024.

[36] U.S. Bureau of Labor Statistics. Occupational Employment and Wage Statistics (OEWS), May 2024, 2024. URL `https://www.bls.gov/oes/tables.htm`. Accessed: 2025-05-29.

[37] U.S. Office of Management and Budget. *Standard Occupational Classification Manual, 2018*, 2017. URL `https://www.bls.gov/soc/2018/soc_2018_manual.pdf`. Accessed: 2025-07-07.

A Data Details

A.1 Merging O*NET and BLS data

The 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 (e.g., tasks and DWA/IWA/GWAs).

A.2 Calculating real world IWA frequency

For each task, O*NET provides the share of respondents in an occupation that perform that task at 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. 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 multiplied by relevance.
3. To get IWA-level frequencies, sum over tasks mapping to the same IWA.
4. To compute the total annual counts of an IWA in the workforce, sum over all occupations performing the IWA, multiplying by employment of each occupation.

A.3 Aggregating from IWAs to Occupations

First, we merge all O*NET-SOC occupations into SOC occupations, taking the union of their tasks. Then, we compute a weight for every task using the importance and relevance score in O*NET.[10] More precisely, for each task i in SOC occupation j, we say $\text{weight}_{ij} = 2^{\text{importance}_{ij}} \cdot \text{relevance}_{ij}$. If an occupation has no ratings for any of its tasks, assign them all weight 1. If an occupation has ratings for only some of its tasks, we ignore the tasks with missing ratings. We propagate these task weights to IWAs through the DWAs that each task maps to, summing weights for tasks mapping to the same IWA. Dividing by the total weight for an occupation then gives us a proxy measure for how much of a job consists of each of its work activities.

B Classification pipeline and validation

We developed a two-stage LLM-based pipeline classifying user goals and AI actions in a conversation. In the first-stage prompts, we give an LLM (specifically, GPT-4o) 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.[11] 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

[10] The *relevance* of a task to an occupation is the fraction of surveyed incumbents who said the task was relevant to their job. The *importance* of a task is a score from 1 to 5 representing the average response to a five-point Likert question about how important the task is to the incumbent's job (if they said the task was relevant).

[11] 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.

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.[12] 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 [24]), our pipeline sacrifices efficiency for thoroughness.

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 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). 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, 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 can be found in Appendix B.1.

B.1 Prompts

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.
```

[12] 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.

```
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 summary
↪  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|>
```

B.1.1 Completion

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[13] if the AI completed the user's task in the conversation. For comparison with the E1 measure of Eloundou et al. [17], we also ask if the AI reduced the time it takes to complete the task by at least 50%.

Task 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|>
```

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

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

<|end of prompt|>
```

```python
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
```

(a) Generate

```python
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]
```

(b) Classify user

```python
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]
```

(c) Classify bot

```python
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
```

(d) Task completion

Figure A1: Structured output formats for our four LLM prompts.

Table A1: LLM details

Prompts	**Model**	**API version**	**Temperature**
Generate, Classify	`gpt-4o-2024-08-06`	`2024-08-01-preview`	1 (generate), 0 (classify)
Completion	`gpt-4o-mini-2024-07-18`	`2024-08-01-preview`	0

C Additional figures and tables

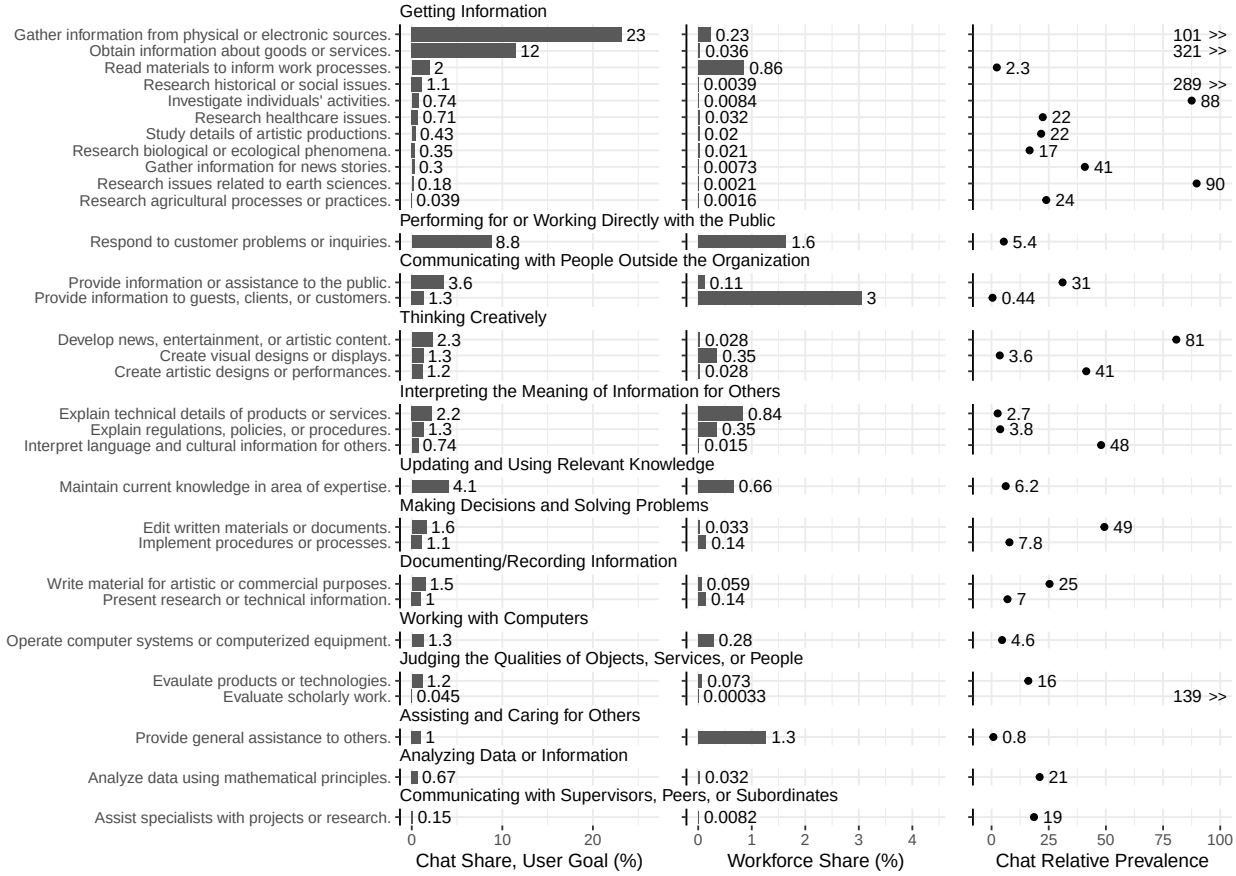


Figure A2: Frequency of top IWAs

Note: This Figure shows the share of total user chat activity (left), estimated share of work activity (center), and their ratio (right) for IWAs that are either in the top 20 for their share of user activity (high prevalence) or for the ratio of their activity share to their estimated share of tasks done in the workforce (high relative prevalence). For chat activity, when a conversation is labeled with multiple IWAs, that share of Copilot activity is evenly distributed among the IWAs; for both chat activity and work activity, the sum across all IWAs is 1. See Appendix A.2 for how the workforce share is calculated. IWAs are grouped by GWA and titles have been shortened for space.

Figure A2 shows all the IWAs, grouped by GWA, that are in the top 20 of either of those lists and plots the share of conversations categorized as that IWA (left), the share of workforce activity categorized as that IWA (center), and the ratio between them (right). Two IWAs, including *Provide information to guests, clients, or customers*, appear less frequently in the data than in the workforce, suggesting they rank highly in the chat data because of how often they are often performed in the world.

The remaining top IWAs are all overrepresented in Copilot conversations. Notably, seven, such as *Write material for artistic or commercial purposes*, are common in the data but rank in the bottom half of workforce activities, implying people are relatively likely to use the LLM for those activities. Also of interest are the eleven, such as *Evaluate scholarly work*, that are somewhat less common in the data but still highly overrepresented, again suggesting tendency for LLM use. The most common IWA in the data, *Gather information from physical or electronic sources*, is in the middle third of workplace activities but appears so frequently in conversations that its ratio ranks fourth.

Many of the IWAs in Figure A2 fall under the GWA *Getting Information*. This may be partially because

people see Copilot as a substitute for a search engine, but even once normalized for how often the activities are done in the workforce, many information-based tasks, including research, appear relatively frequently. The other GWAs where the IWAs stand out on their relative frequency are *Thinking Creatively* and *Judging the Qualities of Object, Services, or People*. *Thinking Creatively* is consistent with the writing abilities of LLMs, but it is perhaps more surprising the extent to which people are using the tools for evaluation (*Judging... or people*).

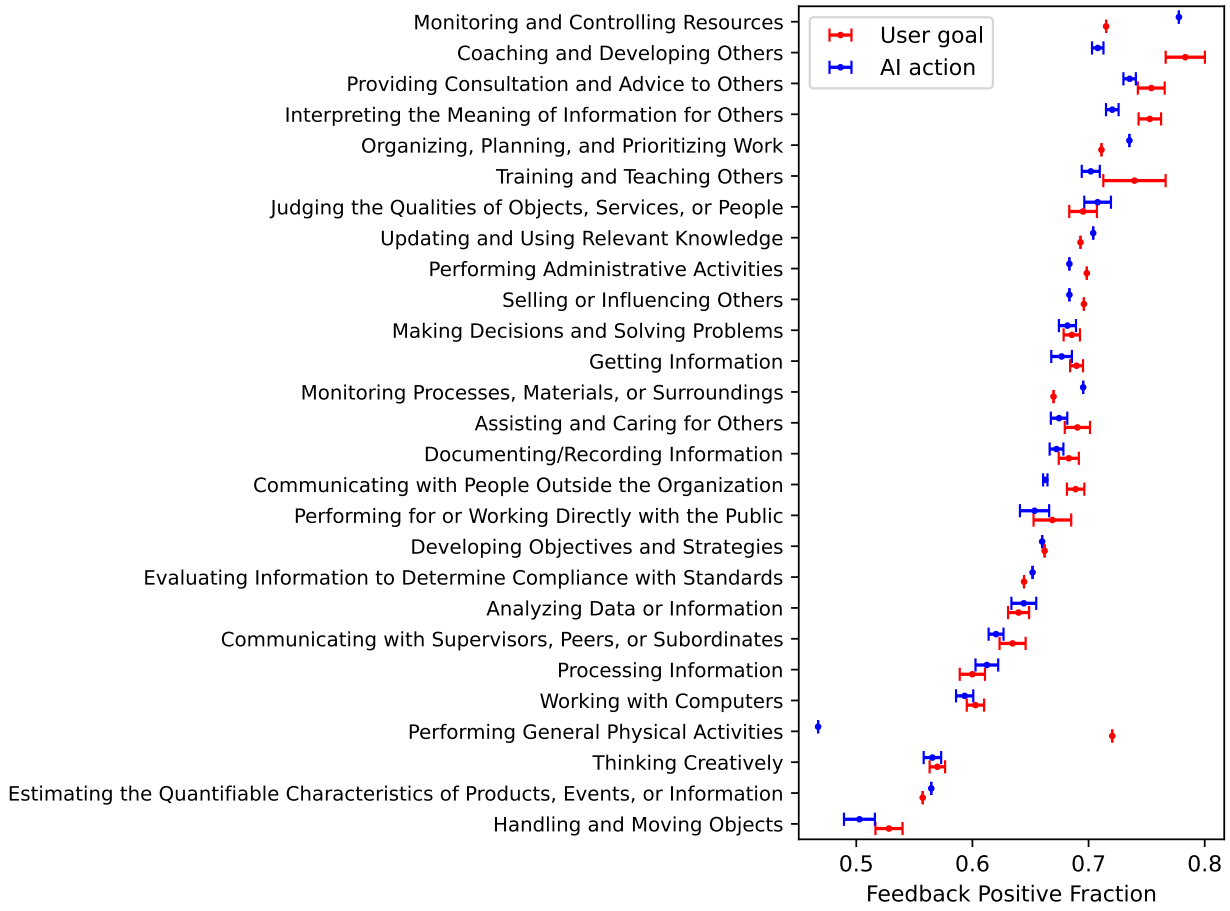


Figure A3: Share of positive feedback by GWA

*N*ote: 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.

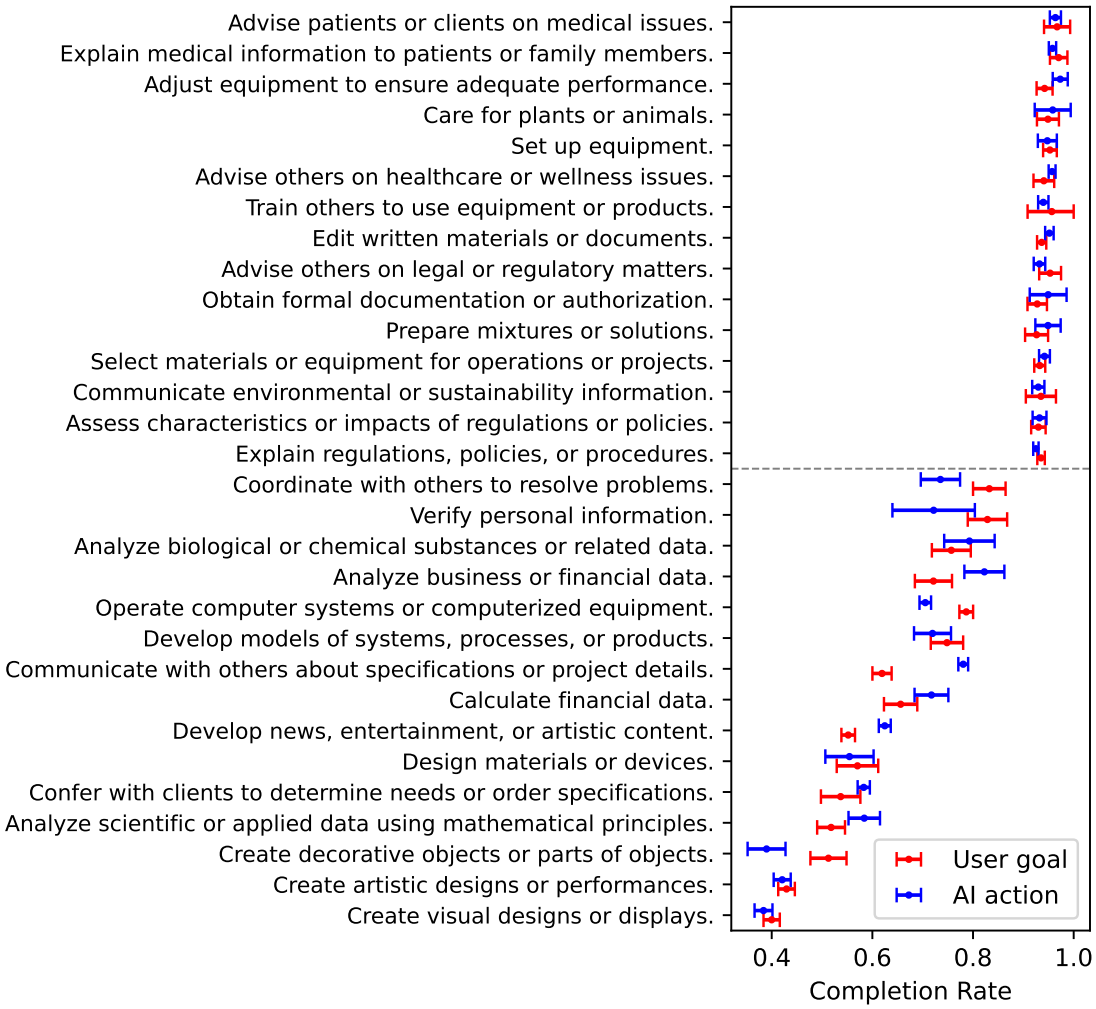


Figure A4: IWAs with the highest and lowest completion rates

Note: 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 4), 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.

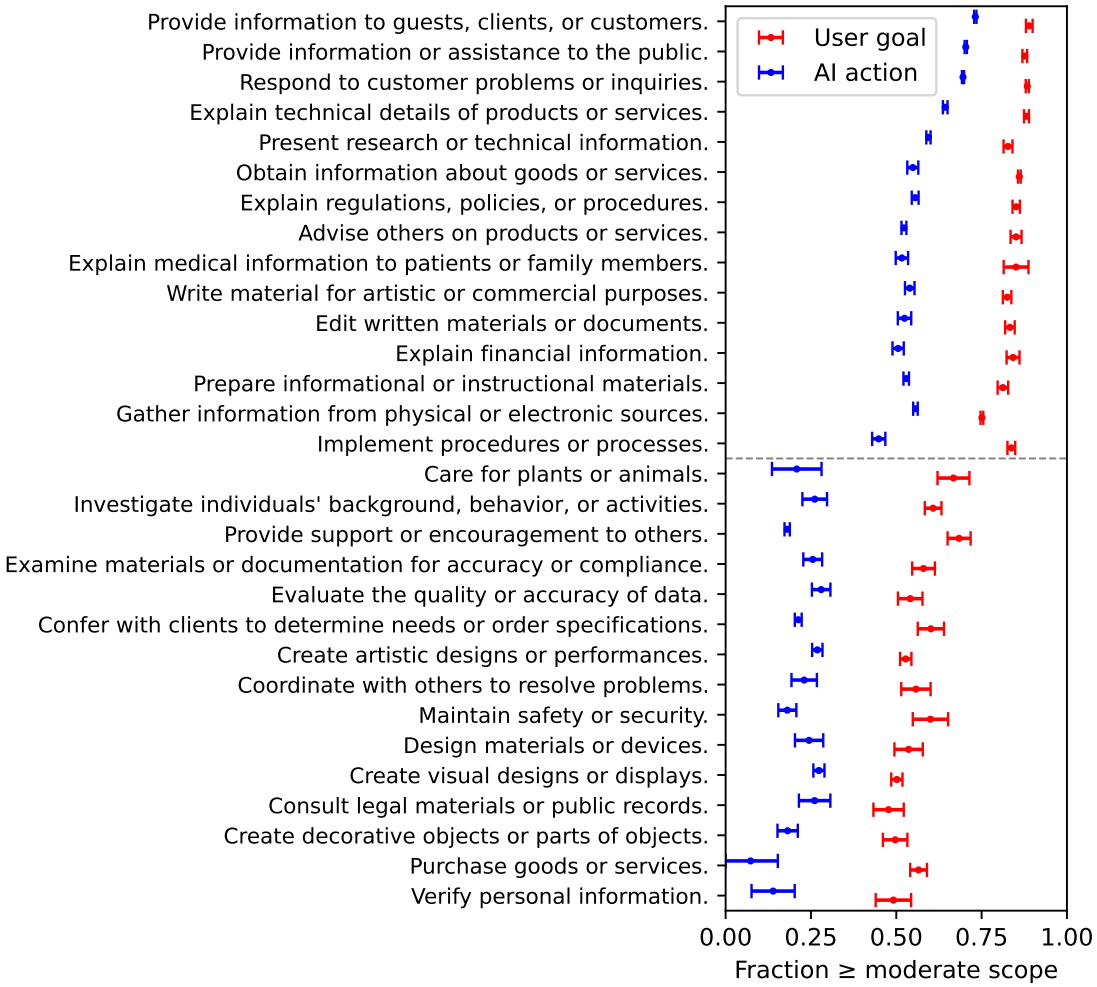


Figure A5: IWAs with the highest and lowest fraction of conversations at moderate or higher impact scope

Note: 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.

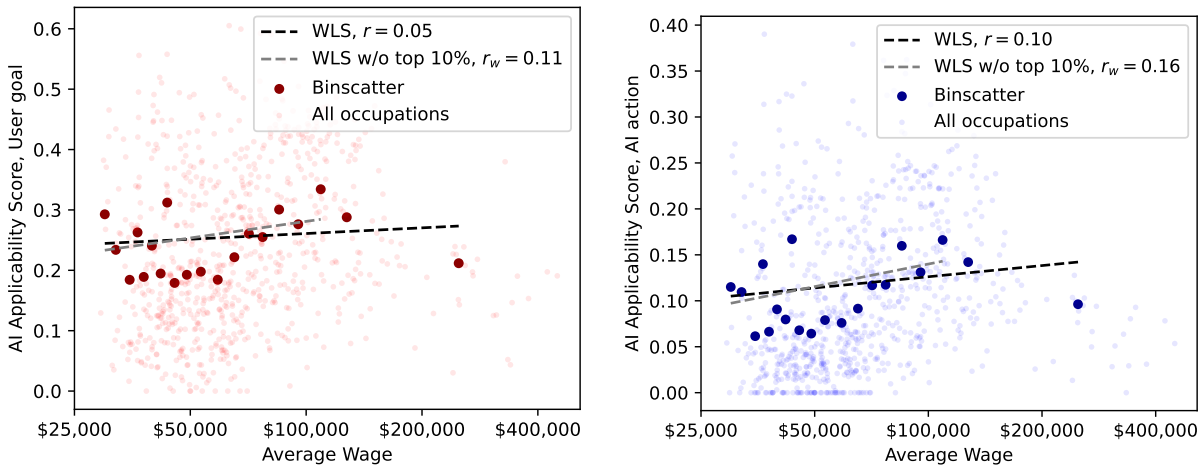


Figure A6: AI applicability score by wage, separating out user goals and AI actions

Note: Wage plot from Figure 7, but showing user goal and AI action separately instead of averaging the two. The correlation is marginally higher on the AI side.

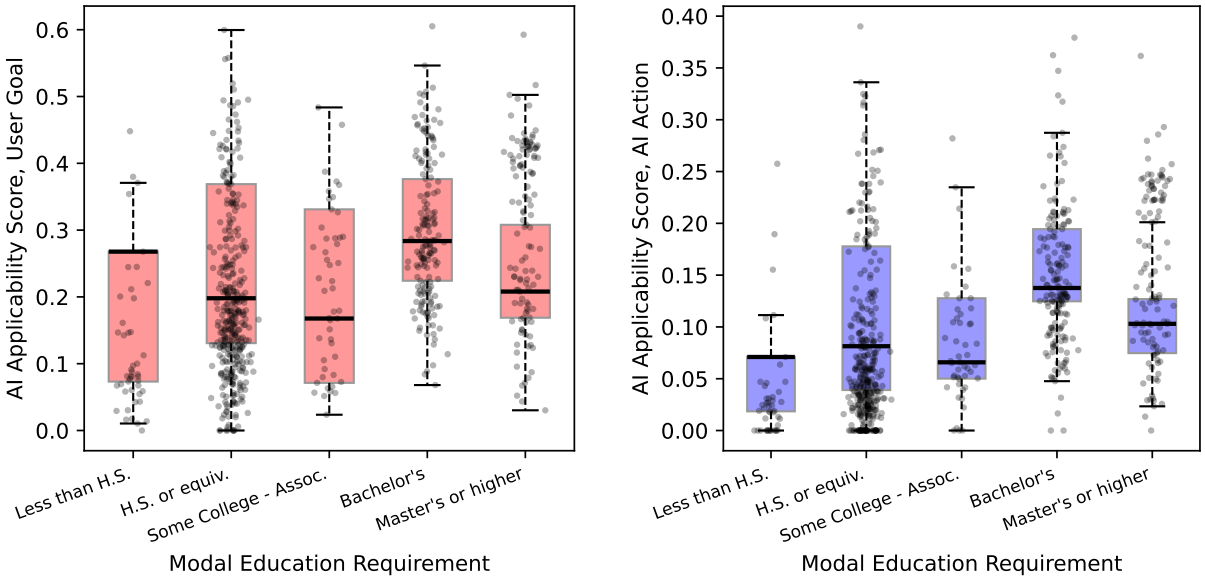


Figure A7: AI applicability score by educational requirement, separating out user goals and AI actions

Note: Education plot from Figure 7, but showing user goal and AI action separately instead of averaging the two. As with wage, the relationship with education is slightly stronger on the AI side.

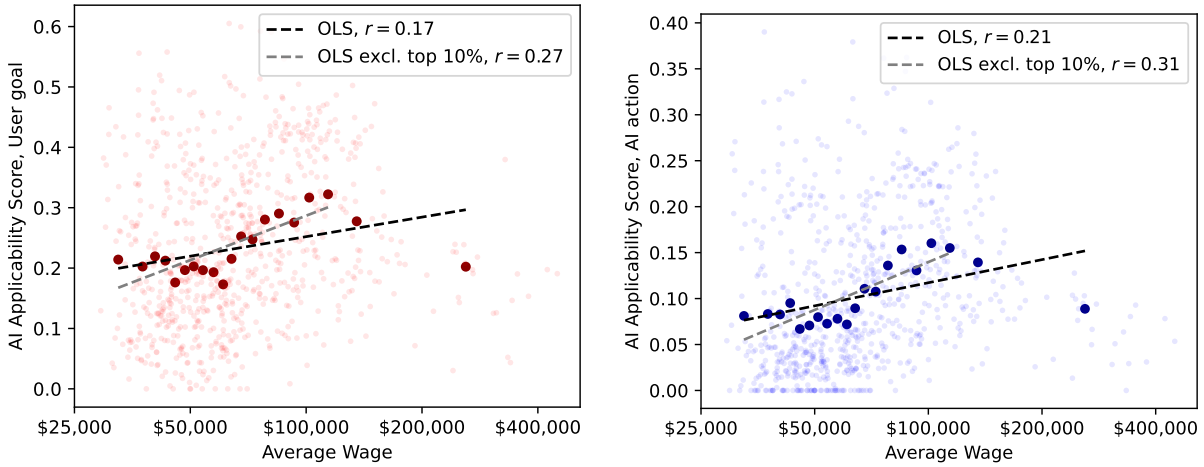


Figure A8: AI applicability score by wage, unweighted

*N*ote: These Figures are the same as Figure A6 without weighting the binscatters. They graph each occupation's AI applicability score calculated over user goals (left) or AI actions (right) against the occupation's average wage. 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.

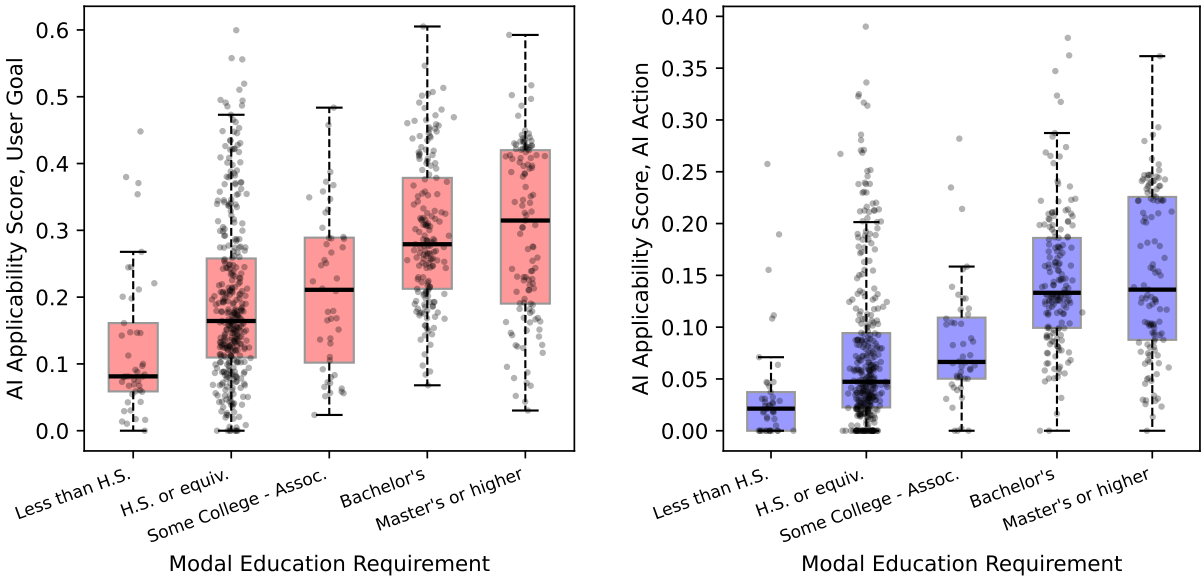


Figure A9: AI applicability score by educational requirement, unweighted

*N*ote: These Figures are the same as Figure A7 without weighting by employment. As in Figure A8, this reduces the noise in the relationship.

Table A2: All SOC minor groups by AI applicability score estimated from Copilot data

Minor Group Title (Abbr)	Coverage	Cmpltn.	Scope	Score	Empl.
Media and Communication Workers	0.73	0.86	0.58	0.39	602,710
Information and Record Clerks	0.64	0.89	0.55	0.37	5,385,660
Sales Representatives, Services	0.66	0.90	0.52	0.36	2,245,510
Communications Equipment Operators	0.69	0.86	0.52	0.35	48,430
Tour and Travel Guides	0.57	0.88	0.53	0.34	46,760
Retail Sales Workers	0.57	0.88	0.52	0.33	7,655,030
Sales Representatives, Wholesale and Manufacturing	0.60	0.88	0.52	0.33	1,600,700
Mathematical Science Occupations	0.71	0.85	0.50	0.32	372,550
Baggage Porters, Bellhops, and Concierges	0.55	0.89	0.48	0.32	69,800
Other Sales and Related Workers	0.58	0.89	0.52	0.32	450,090
Postsecondary Teachers	0.61	0.90	0.49	0.31	1,210,240
Entertainment Attendants and Related Workers	0.52	0.88	0.51	0.30	592,140
Computer Occupations	0.63	0.86	0.48	0.30	4,804,840
Other Office and Administrative Support Workers	0.59	0.89	0.49	0.29	3,041,920
Librarians, Curators, and Archivists	0.59	0.89	0.47	0.29	242,760
Religious Workers	0.57	0.88	0.49	0.27	79,910
Supervisors of Personal Care and Service Workers	0.48	0.91	0.50	0.27	219,680
Secretaries and Administrative Assistants	0.53	0.89	0.49	0.27	3,171,290
Financial Clerks	0.60	0.86	0.47	0.27	2,695,230
Other Teachers and Instructors	0.54	0.88	0.47	0.26	915,830
Social Scientists and Related Workers	0.50	0.88	0.47	0.26	273,230
Counselors, Social Workers, ...	0.50	0.88	0.44	0.25	2,137,020
Supervisors of Production Workers	0.56	0.91	0.45	0.25	671,160
Supervisors of Office and Administrative Support Workers	0.48	0.89	0.46	0.25	1,504,570
Business Operations Specialists	0.48	0.90	0.48	0.24	7,048,360
Animal Care and Service Workers	0.41	0.93	0.46	0.24	288,070
Financial Specialists	0.51	0.86	0.47	0.24	3,039,490
Engineers	0.48	0.86	0.47	0.23	1,703,700
Other Educational Instruction and Library Occupations	0.47	0.89	0.44	0.22	1,698,660
Physical Scientists	0.44	0.88	0.46	0.21	254,400
Drafters, Engineering/Mapping Technicians	0.55	0.80	0.44	0.21	624,780
Life Scientists	0.41	0.88	0.48	0.21	344,490
Food and Beverage Serving Workers	0.37	0.91	0.44	0.21	6,893,410
Air Transportation Workers	0.39	0.90	0.45	0.21	313,070
Art and Design Workers	0.69	0.68	0.44	0.21	658,340
Material Recording, Dispatching, and Distributing Workers	0.38	0.91	0.43	0.20	2,316,660
Media and Communication Equipment Workers	0.43	0.85	0.47	0.20	223,820
Teachers, Preschool-Seconday and Special Education	0.39	0.90	0.45	0.19	4,261,430
Supervisors of Transportation and Material Moving Workers	0.36	0.91	0.46	0.18	603,350
Sales, Marketing, PR Managers	0.34	0.90	0.44	0.18	1,070,020
Electrical/Electronic Equip. Mechanics/Installers/Repairers	0.36	0.91	0.44	0.18	494,540
Architects, Surveyors, and Cartographers	0.48	0.77	0.44	0.17	194,610
Lawyers, Judges, and Related Workers	0.42	0.89	0.42	0.17	792,220
Other Personal Care and Service Workers	0.39	0.90	0.44	0.17	1,147,350
Entertainers and Performers, Sports and Related Workers	0.38	0.86	0.46	0.17	554,960
Other Healthcare Practitioners and Technical Occupations	0.34	0.88	0.41	0.16	121,640
Other Transportation Workers	0.28	0.91	0.44	0.16	294,450
Cooks and Food Preparation Workers	0.27	0.91	0.40	0.16	3,528,200
Funeral Service Workers	0.32	0.83	0.36	0.15	63,420
Other Protective Service Workers	0.36	0.84	0.42	0.15	1,676,910
Other Food Preparation and Serving Related Workers	0.25	0.92	0.39	0.15	1,372,350
Supervisors; Building, Grounds Cleaning, Maintenance	0.30	0.91	0.42	0.15	297,140
Law Enforcement Workers	0.36	0.81	0.37	0.15	1,136,430
Other Management Occupations	0.28	0.90	0.45	0.14	3,109,640
Occupational Health/Safety Specialists	0.32	0.90	0.42	0.14	149,570
Supervisors of Food Preparation and Serving Workers	0.26	0.92	0.45	0.14	1,348,910
Supervisors of Installation, Maintenance, and Repair Workers	0.29	0.89	0.39	0.14	589,880
Life, Physical, and Social Science Technicians	0.29	0.88	0.43	0.14	360,240
Operations Specialties Managers	0.28	0.89	0.44	0.14	2,513,890
Motor Vehicle Operators	0.29	0.92	0.40	0.14	4,302,220
Supervisors of Sales Workers	0.27	0.88	0.42	0.14	1,315,040
Healthcare Diagnosing or Treating Practitioners	0.26	0.91	0.39	0.13	6,119,630
Rail Transportation Workers	0.26	0.91	0.40	0.13	109,780
Food Processing Workers	0.22	0.87	0.42	0.12	784,660
Top Executives	0.23	0.90	0.48	0.12	3,751,500
Woodworkers	0.26	0.91	0.42	0.12	208,510
Supervisors of Construction and Extraction Workers	0.25	0.89	0.47	0.11	777,420
Health Technologists and Technicians	0.24	0.89	0.37	0.11	3,010,660
Assemblers and Fabricators	0.25	0.92	0.42	0.11	1,924,980
Metal Workers and Plastic Workers	0.23	0.92	0.41	0.11	1,584,800
Printing Workers	0.21	0.91	0.42	0.11	213,920
Other Installation, Maintenance, and Repair Occupations	0.20	0.93	0.42	0.10	3,186,610
Vehicle and Mobile Equip. Mechanics/Installers/Repairers	0.19	0.93	0.40	0.10	1,708,120
Water Transportation Workers	0.17	0.92	0.41	0.09	76,050
Firefighting and Prevention Workers	0.19	0.89	0.40	0.09	331,930
Other Production Occupations	0.17	0.91	0.40	0.08	2,289,050
Building Cleaning and Pest Control Workers	0.16	0.94	0.37	0.08	3,102,490
Personal Appearance Workers	0.19	0.89	0.40	0.08	532,400
Supervisors of Protective Service Workers	0.19	0.86	0.39	0.08	339,440
Supervisors of Farming, Fishing, and Forestry Workers	0.18	0.91	0.39	0.08	27,150
Material Moving Workers	0.14	0.92	0.36	0.08	7,966,020
Occupational and Physical Therapy Assistants	0.20	0.87	0.35	0.07	196,910
Construction Trades Workers	0.15	0.92	0.40	0.07	4,588,630
Other Construction and Related Workers	0.11	0.93	0.38	0.06	455,520
Agricultural Workers	0.11	0.92	0.39	0.06	357,680
Legal Support Workers	0.15	0.89	0.42	0.06	404,650
Helpers, Construction Trades	0.11	0.94	0.38	0.06	164,440
Other Healthcare Support Occupations	0.13	0.90	0.35	0.06	1,744,500
Textile, Apparel, and Furnishings Workers	0.12	0.93	0.43	0.05	458,900
Extraction Workers	0.11	0.95	0.36	0.05	202,710
Home Health Aides, Nursing Assistants, Orderlies, ...	0.12	0.91	0.40	0.05	5,122,130
Grounds Maintenance Workers	0.09	0.95	0.39	0.05	1,003,720
Plant and System Operators	0.09	0.93	0.41	0.04	283,480
Forest, Conservation, and Logging Workers	0.06	0.94	0.37	0.03	37,910

Note: Metrics reported as mean of user goal and AI action.

Table A3: 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, 4)	Exercise Trainers (17, 79)
Butchers and Meat Cutters (83, 8)	Choreographers (34, 78)
Cooks, Private Household (97, 24)	Training and Development Managers (45, 83)
Cooks, Restaurant (76, 8)	Coaches and Scouts (43, 77)
Meat Cutters (79, 12)	Environmental Engineers (55, 82)
Animal Breeders (76, 18)	Human Resources Specialists (53, 80)
Lighting Technicians (82, 27)	Health Education Specialists (53, 76)
Animal Control Workers (79, 37)	Lodging Managers (57, 79)
Athletes (82, 41)	Coatroom, Locker Room Attendants (60, 82)
Animal Caretakers (89, 49)	Taxi Drivers (56, 78)

Note: 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.

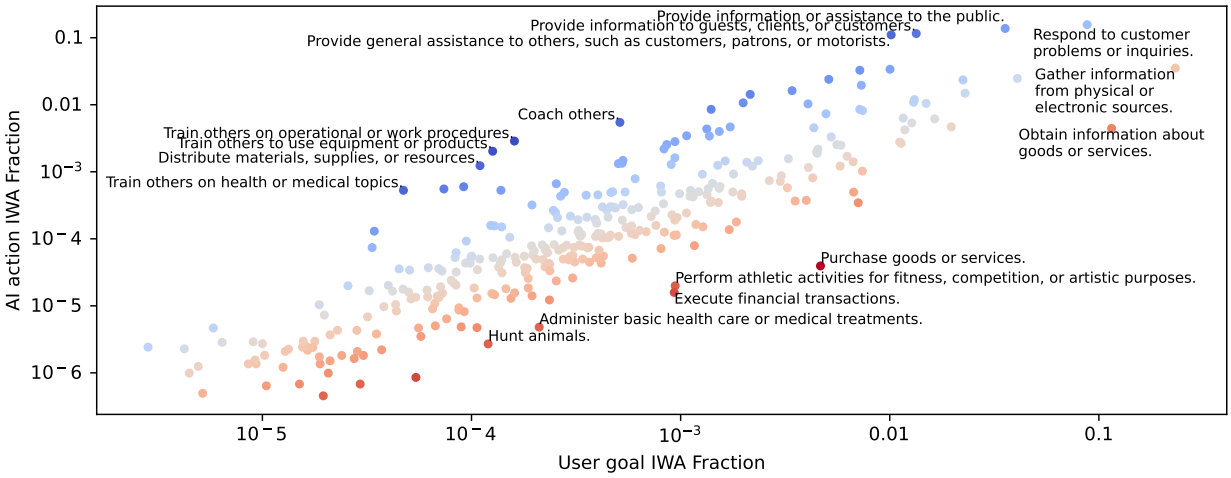


Figure A10: IWA frequency in AI actions and user goals

Note: 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.

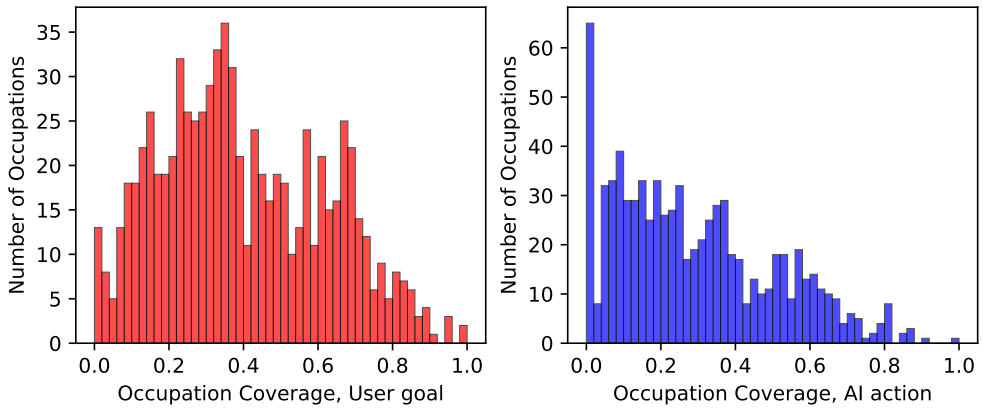


Figure A11: Distribution of occupation coverage

Note: These Figures show the distribution of occupation coverage scores (fraction of importance-weighted work in an occupation with user goal or AI action rate at least 0.05% in COPILOT-UNIFORM).

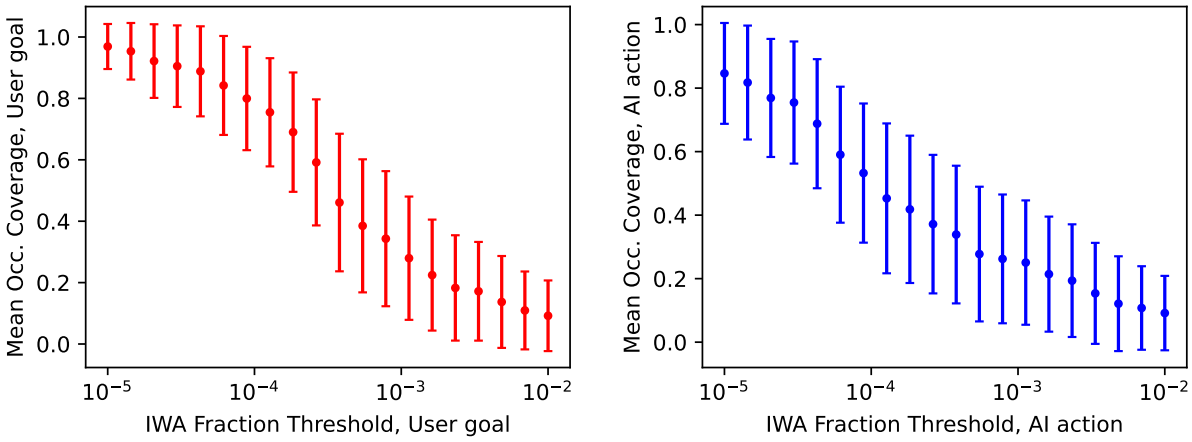


Figure A12: Mean and s.d of occupation coverage by threshold

Note: These Figures show the average and standard deviation of occupation coverage for different thresholds for the share of chat activity an IWA must have to "be done" with the LLM, for user goal IWAs (left) and AI action IWAs (right). 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.

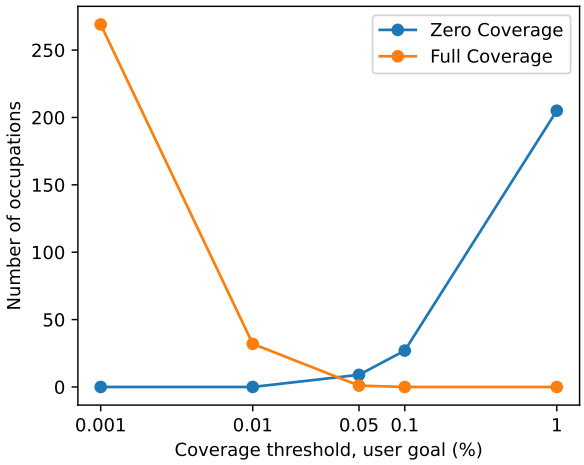


Figure A13: Effect of coverage threshold on occupations with coverage 0 and 1

Note: Our threshold of 0.05% approximately minimizes the number of occupations assigned user goal coverage 0 or 1.

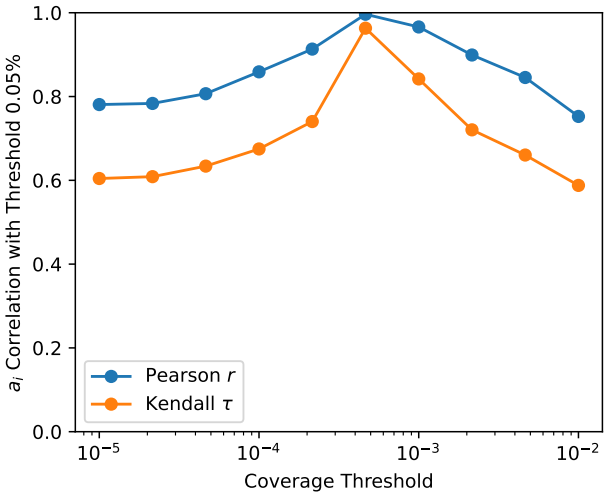


Figure A14: Robustness of AI applicability score to different coverage thresholds

Note: This Figure shows 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 1, which is highly sensitive to arbitrary choice of threshold.

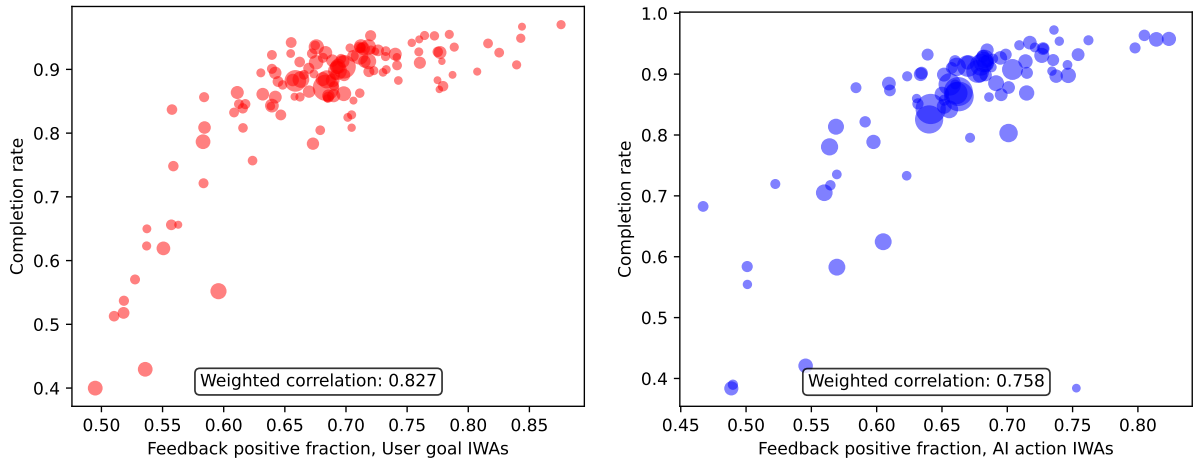


Figure A15: Relationship between thumbs feedback and task completion rate for each IWA

Note: There is a strong correlation between thumbs feedback rate (measured in Copilot-Thumbs) and GPT-4o-mini task completion rate (measured in Copilot-Uniform) 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.

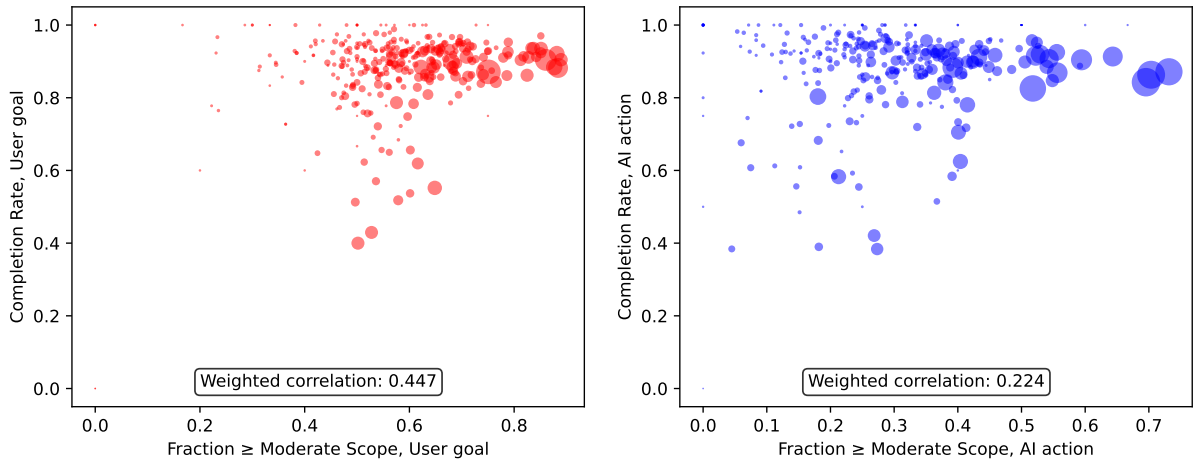


Figure A16: Relationship between impact scope and task completion rate for each IWA

Note: 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.

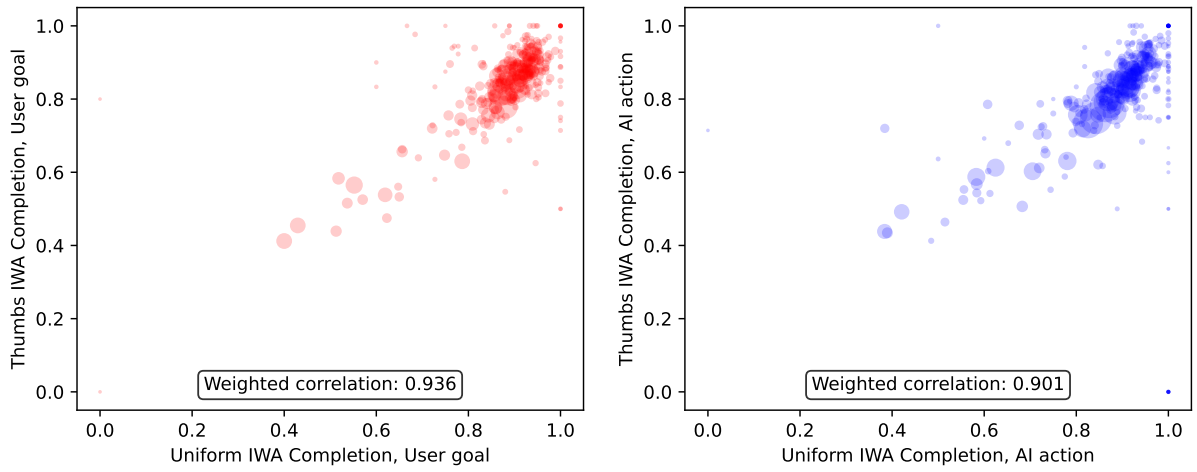


Figure A17: Correlation between IWA completion rates in Copilot-Uniform and Copilot-Thumbs

Note: 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.

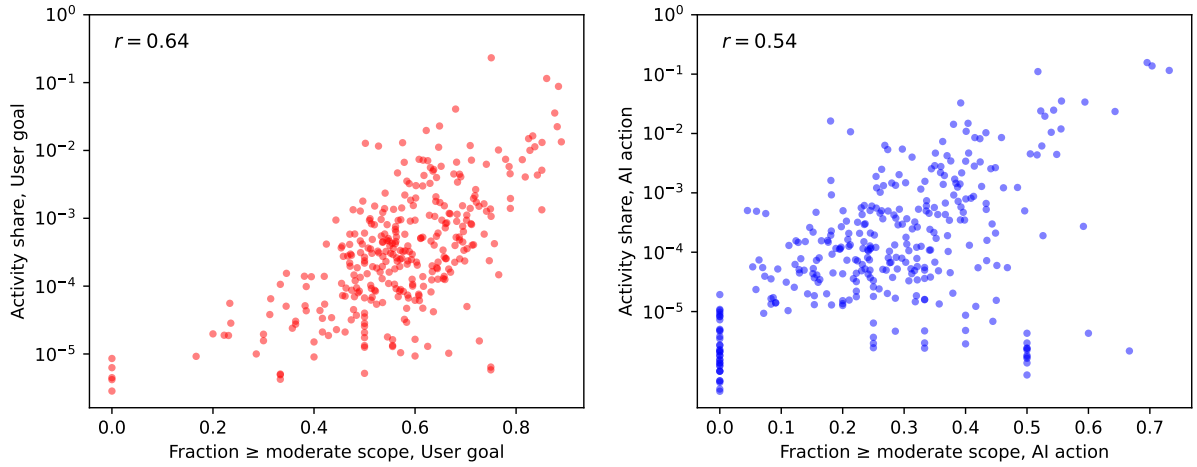


Figure A18: Correlation between impact scope and activity share in Copilot-Uniform for each IWA

Note: This Figure shows the relationship between the activity share of an IWA and the fraction of conversations in which it classsified at moderate impact scope or higher. Impact scope is a good predictor of what activities people seek AI assistance with (better than completion or satisfaction).

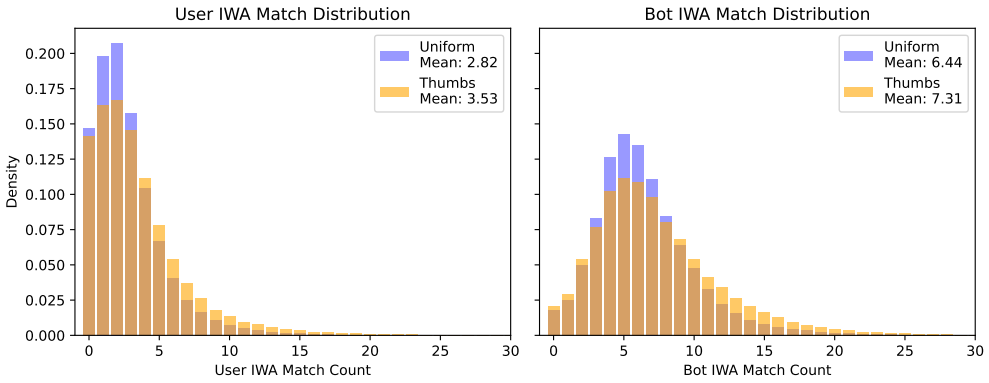


Figure A19: Distribution of IWAs per conversation

*N*ote: These Figures show the distribution of the number of IWAs a conversation is matched to for user goal (left) and bot activity (right) in both the in uniform and thumbs Copilot datasets.