NBER WORKING PAPER SERIES

HOW PEOPLE USE CHATGPT

Aaron Chatterji
Thomas Cunningham
David J. Deming
Zoe Hitzig
Christopher Ong
Carl Yan Shan
Kevin Wadman

Working Paper 34255
http://www.nber.org/papers/w34255

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2025

We acknowledge help and comments from Joshua Achiam, Hemanth Asirvatham, Ryan Beiermeister, Rachel Brown, Cassandra Duchan Solis, Jason Kwon, Elliott Mokski, Kevin Rao, Harrison Satcher, Gawesha Weeratunga, Hannah Wong, and Analytics & Insights team. We especially thank Tyna Eloundou and Pamela Mishkin who in several ways laid the foundation for this work. This study was approved by Harvard IRB (IRB25-0983). A repository containing all code run to produce the analyses in this paper is available on request. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w34255

How People Use ChatGPT
Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman
NBER Working Paper No. 34255
September 2025
JEL No. J01, O3, O4

ABSTRACT

Despite the rapid adoption of LLM chatbots, little is known about how they are used. We document the growth of ChatGPT's consumer product from its launch in November 2022 through July 2025, when it had been adopted by around 10% of the world's adult population. Early adopters were disproportionately male but the gender gap has narrowed dramatically, and we find higher growth rates in lower-income countries. Using a privacy-preserving automated pipeline, we classify usage patterns within a representative sample of ChatGPT conversations. We find steady growth in work-related messages but even faster growth in non-work-related messages, which have grown from 53% to more than 70% of all usage. Work usage is more common for educated users in highly-paid professional occupations. We classify messages by conversation topic and find that "Practical Guidance," "Seeking Information," and "Writing" are the three most common topics and collectively account for nearly 80% of all conversations. Writing dominates work-related tasks, highlighting chatbots' unique ability to generate digital outputs compared to traditional search engines. Computer programming and self-expression both represent relatively small shares of use. Overall, we find that ChatGPT provides economic value through decision support, which is especially important in knowledge-intensive jobs.

Aaron Chatterji
Duke University
Fuqua School of Business
and OpenAI
ronnie@duke.edu

Thomas Cunningham
OpenAI
tom.cunningham@gmail.com

David J. Deming
Harvard University
Harvard Kennedy School
and NBER
david_deming@harvard.edu

Zoe Hitzig
OpenAI
and Harvard Society of Fellows
zhitzig@g.harvard.edu

Christopher Ong
Harvard University
and OpenAI
christopherong@hks.harvard.edu

Carl Yan Shan
OpenAI
cshan@openai.com

Kevin Wadman
OpenAI
kevin.wadman@c-openai.com

1 Introduction

ChatGPT launched in November 2022. By July 2025, 18 billion messages were being sent each week by 700 million users, representing around 10% of the global adult population.[1] For a new technology, this speed of global diffusion has no precedent (Bick et al., 2024).

This paper studies consumer usage of ChatGPT, the first mass-market chatbot and likely the largest.[2] ChatGPT is based on a Large Language Model (LLM), a type of Artificial Intelligence (AI) developed over the last decade and generally considered to represent an acceleration in AI capabilities.[3]

The sudden growth in LLM abilities and adoption has intensified interest in the effects of artificial intelligence on economic growth (Acemoglu, 2024; Korinek and Suh, 2024); employment (Eloundou et al., 2025); and society (Kulveit et al., 2025). However, despite the rapid adoption of LLMs, there is limited public information on how they are used. A number of surveys have measured self-reported adoption of LLMs (Bick et al., 2024; Pew Research Center, 2025); however there are reasons to expect bias in self-reports (Ling and Imas, 2025), and none of these papers have been able to directly track the quantity or nature of chatbot conversations.

Two recent papers do report statistics on chatbot conversations, classified in a variety of ways (Handa et al., 2025; Tomlinson et al., 2025). We build on this work in several respects. First, the pool of users on ChatGPT is far larger, meaning we expect our data to be a closer approximation to the average chatbot user.[4] Second, we use automated classifiers to report on the types of messages that users send using new classification taxonomies relative to the existing literature. Third, we report the diffusion of chatbot use across populations and the growth of different types of usage within cohorts. Fourth, we use a secure data clean room protocol to analyze aggregated employment and education categories for a sample of our users, lending new insights about differences in the types of messages sent by different groups while protecting user privacy.

Our primary sample is a random selection of messages sent to ChatGPT on consumer plans (Free, Plus, Pro) between May 2024 and June 2025.[5] Messages from the user to chatbot are classified automatically using a number of different taxonomies: whether the message is used for paid work, the topic of conversation, and the type of interaction (asking, doing, or expressing), and the O*NET task the user is performing. Each taxonomy is defined in a prompt passed to an LLM, allowing us to classify messages without any human seeing them. We give the text of most prompts in Appendix A along with details about how the prompts were validated in Appendix B.[6] The classification pipeline is protected by a series of privacy measures, detailed below, to ensure no leakage of sensitive information during the automated analysis. In a secure data clean room, we relate taxonomies of messages to aggregated employment and education categories.

Table 1 shows the growth in total message volume for work and non-work usage. Both types of

[1] Reuters (2025), Roth (2025)

[2] Bick et al. (2024) report that 28% of US adults used ChatGPT in late 2024, higher than any other chatbot.

[3] We use the term LLM loosely here and give more details in the following section.

[4] Wiggers (2025) reports estimates that in April 2025 ChatGPT was receiving more than 10 times as many visitors as either Claude or Copilot.

[5] Our sample includes the three consumer plans (Free, Plus, or Pro). OpenAI also offers a variety of other ChatGPT plans (Business fka. Teams, Enterprise, Education), which we do not include in our sample.

[6] Our classifiers take into account not just the randomly-selected user message, but also a portion of the preceding messages in that conversation.

Month	Non-Work (M)	(%)	Work (M)	(%)	Total Messages (M)
Jun 2024	238	53%	213	47%	451
Jun 2025	1,911	73%	716	27%	2,627

Table 1: ChatGPT daily message counts (millions), broken down by likely work-related or non-work-related. Total daily counts are exact measurements of message volume from all consumer plans. Daily counts of work and non-work related messages are estimated by classifying a random sample of conversations from that day. Sampling is done to exclude users who opt-out of sharing their messages for model training, users who self-report their age as under 18, logged-out users, deleted conversations, and accounts which have been deactivated or banned (details available in Section 3). Reported values are 7-day averages (to smooth weekly fluctuation) ending on the 26th of June 2024 and 26th of June 2025.

messages have grown continuously, but non-work messages have grown faster and now represent more than 70% of all consumer ChatGPT messages. While most economic analysis of AI has focused on its impact on productivity in paid work, the impact on activity outside of work (home production) is on a similar scale and possibly larger. The decrease in the share of work-related messages is primarily due to changing usage within each cohort of users rather than a change in the composition of new ChatGPT users. This finding is consistent with Collis and Brynjolfsson (2025), who use choice experiments to uncover willingness-to-pay for generative AI and estimate a consumer surplus of at least $97 billion in 2024 alone in the US.

We next report on a classification of messages using a taxonomy developed at OpenAI for understanding product usage ("conversation classifier"). Nearly 80% of all ChatGPT usage falls into three broad categories, which we call *Practical Guidance*, *Seeking Information*, and *Writing*. *Practical Guidance* is the most common use case and includes activities like tutoring and teaching, how-to advice about a variety of topics, and creative ideation.[7] *Seeking Information* includes searching for information about people, current events, products, and recipes, and appears to be a very close substitute for web search. *Writing* includes the automated production of emails, documents and other communications, but also editing, critiquing, summarizing, and translating text provided by the user. *Writing* is the most common use case at work, accounting for 40% of work-related messages on average in June 2025. About two-thirds of all *Writing* messages ask ChatGPT to modify user text (editing, critiquing, translating, etc.) rather than creating new text from scratch. About 10% of all messages are requests for tutoring or teaching, suggesting that education is a key use case for ChatGPT.

Two of our findings stand in contrast to other work. First, we find the share of messages related to computer coding is relatively small: only 4.2% of ChatGPT messages are related to computer programming, compared to 33% of work-related Claude conversations Handa et al. (2025).[8] Second, we find the share of messages related to companionship or social-emotional issues is fairly small: only 1.9% of ChatGPT messages are on the topic of *Relationships and Personal Reflection* and 0.4% are related

[7] The difference between *Practical Guidance* and *Seeking Information* is that the former is highly customized to the user and can be adapted based on conversation and follow-up, whereas the latter is factual information that should be the same for all users. For example, users interested in running might ask ChatGPT for the Boston Marathon qualifying times by age and gender (*Seeking Information*), or they might ask for a customized workout plan that matches their goals and current level of fitness (*Practical Guidance*).

[8] Handa et al. (2025) report that 37% of conversations are mapped to a "computer and mathematical" occupation category, and their Figure 12 shows 30% or more of all imputed tasks are programming or IT-related. We believe the discrepancy is partly due to the difference in types of users between Claude and ChatGPT, additionally Handa et al. (2025) only includes queries that "possibly involve an occupational task".

to *Games and Role Play*. In contrast, Zao-Sanders (2025) estimates that *Therapy/Companionship* is the most prevalent use case for generative AI.[9]

We also document several important facts about demographic variation in ChatGPT usage. First, we show evidence that the gender gap in ChatGPT usage has likely narrowed considerably over time, and may have closed completely. In the few months after ChatGPT was released about 80% of active users had typically masculine first names.[10] However, that number declined to 48% as of June 2025, with active users slightly more likely to have typically feminine first names. Second, we find that nearly half of all messages sent by adults were sent by users under the age of 26, although age gaps have narrowed somewhat in recent months. Third, we find that ChatGPT usage has grown relatively faster in low- and middle-income countries over the last year. Fourth, we find that educated users and users in highly-paid professional occupations are substantially more likely to use ChatGPT for work.

We introduce a new taxonomy to classify messages according to the kind of output the user is seeking, using a simple rubric that we call *Asking, Doing,* or *Expressing.*[11] *Asking* is when the user is seeking information or clarification to inform a decision, corresponding to problem-solving models of knowledge work (e.g., Garicano (2000); Garicano and Rossi-Hansberg (2006); Carnehl and Schneider (2025); Ide and Talamas (2025)). *Doing* is when the user wants to produce some output or perform a particular task, corresponding to classic task-based models of work (e.g., Autor et al. (2003)). *Expressing* is when the user is expressing views or feelings but not seeking any information or action. We estimate that about 49% of messages are *Asking*, 40% are *Doing*, and 11% are *Expressing.* However, as of July 2025 about 56% of work-related messages are classified as *Doing* (e.g., performing job tasks), and nearly three-quarters of those are *Writing* tasks. The relative frequency of writing-related conversations is notable for two reasons. First, writing is a task that is common to nearly all white-collar jobs, and good written communication skills are among the top "soft" skills demanded by employers (National Association of Colleges and Employers, 2024). Second, one distinctive feature of generative AI, relative to other information technologies, is its ability to produce long-form outputs such as writing and software code.

We also map message content to work activities using the Occupational Information Network (O*NET), a survey of job characteristics supported by the U.S. Department of Labor. We find that about 81% of work-related messages are associated with two broad work activities: 1) obtaining, documenting, and interpreting information; and 2) making decisions, giving advice, solving problems, and thinking creatively. Additionally, we find that the work activities associated with ChatGPT usage are highly similar across very different kinds of occupations. For example, the work activities *Getting Information* and *Making Decisions and Solving Problems* are in the top five of message frequency in nearly all occupations, ranging from management and business to STEM to administrative and sales occupations.

Overall, we find that information-seeking and decision support are the most common ChatGPT use cases in most jobs. This is consistent with the fact that almost half of all ChatGPT usage is either *Practical Guidance* or *Seeking Information*. We also show that *Asking* is growing faster than

[9] Zao-Sanders (2025) is based on a manual collection and labeling of online resources (Reddit, Quora, online articles), and so we believe it likely resulted in an unrepresentative distribution of use cases.

[10] Among those with names commonly associated with a particular gender.

[11] Appendix A gives the full prompt text and Appendix B gives detail about how the prompts were validated against public conversation data.

Doing, and that *Asking* messages are consistently rated as having higher quality both by a classifier that measures user satisfaction and from direct user feedback.

How does ChatGPT provide economic value, and for whom is its value the greatest? We argue that ChatGPT likely improves worker output by providing *decision support*, which is especially important in knowledge-intensive jobs where better decision-making increases productivity (Deming, 2021; Caplin et al., 2023). This explains why *Asking* is relatively more common for educated users who are employed in highly-paid, professional occupations. Our findings are most consistent with Ide and Talamas (2025), who develop a model where AI agents can serve either as *co-workers* that produce output or as *co-pilots* that give advice and improve the productivity of human problem-solving.

2 What is ChatGPT?

Here we give a simplified overview of LLMs and chatbots. For more precise details, refer to the papers and system cards that OpenAI has released with each model e.g., (OpenAI, 2023, 2024a, 2025b). A chatbot is a statistical model trained to generate a text response given some text input, so as to maximize the "quality" of that response, where the quality is measured with a variety of metrics.

In a prototypical interaction, a user submits a plain-text message ("prompt") and ChatGPT returns the message ("response") generated from an underlying LLM. A large set of additional features have been added to ChatGPT—including the possibility for the LLM to search the web or external databases, and generate images based on text—but the exchange of text-based messages remains the most typical interaction.

Since its launch ChatGPT has used a variety of different underlying LLMs e.g., GPT-3.5, GPT-4, GPT-4o, o1, o3, and GPT-5.[12] In addition there are occasional updates to the model's weights and to the model's system prompt (text instructions sent to the model along with all the queries).

An LLM can be thought of as a function from a string of words to a probability distribution over the set of all possible words (more precisely, "tokens," which very roughly correspond to words[13]). The functions are implemented with deep neural nets, typically with a transformer architecture (Vaswani et al., 2017), parameterized with billions of model "weights". We will refer to all of ChatGPT's models as language models, though most can additionally process tokens representing images, audio, or other media.

The weights in an LLM-based chatbot are often trained in two stages, commonly called "pre-training" and "post-training". In the first stage ("pre-training"), the LLMs are trained to predict the next word in a string, given the preceding words, over an enormous corpus of text. At that point the models are purely predictors of the likelihood of the next word given a prior context, and as such they have a relatively narrow application. In the second stage ("post-training"), the models are trained to produce words that comprise "good" responses to some prompt. This stage often consists of a variety of different strategies: fine-tuning on a dataset of queries and ideal responses, reinforcement learning against another model that is trained to grade the quality of a response (Ouyang et al., 2022), or reinforcement learning against a function that knows the true response to queries (OpenAI (2024b),

[12] For a timeline of model launches, see Appendix C.

[13] Tokenization is a way of cutting a string of text into discrete chunks, chosen to be statistically efficient. In many tokenization schemes, one token corresponds to roughly three-quarters of an English word.

Lambert et al. (2024)). This second stage also typically includes a number of "safety" constraints to avoid certain classes of response, especially those which are deemed harmful or dangerous (OpenAI, 2025a).

This two-stage process has a common statistical interpretation: the first stage teaches the model a latent representation of the world; the second stage fits a function using that representation (Bengio et al., 2014). Pre-training the model to predict the next word effectively teaches the model a low-dimensional representation of text, representing only the key semantic features, and therefore rendering the prompt-response problem tractable with a reasonable set of training examples.

Two common ways of evaluating chatbots are with benchmarks (batteries of questions with known answers, e.g. Measuring Massive Multitask Language Understanding (Hendrycks et al., 2021)) and comparisons of human preferences over two alternative responses to the same message (e.g. Chatbot Arena (Chiang et al., 2024)).

3 Data and Privacy

In this section, we describe the data used in the paper and the privacy safeguards we implemented. No member of the research team ever saw the content of user messages, and all analyses were conducted in accordance with OpenAI's Privacy Policy (OpenAI, 2025c).

The analysis in this paper is based on the following datasets:

1. **Growth:** total daily message volumes from consumer ChatGPT users between November 2022 and September 2025, along with basic self-reported demographic information. This dataset is primarily used in Section 4.
2. **Classified messages:** messages classified into coarse categories.
    - **Sampled from all ChatGPT users:** a random sample of approximately one million de-identified messages from logged-in consumer ChatGPT users between May 2024 and June 2025.[14] This dataset is primarily used in Section 5.
    - **Sampled from a subset of ChatGPT users:** two random samples of messages sent between May 2024 and July 2025 by a subset of consumer ChatGPT users (one sample at the conversation level, one sample at the user level).[15] These datasets are primarily used in Section 6.
3. **Employment:** aggregated employment and education categories based on publicly available data for a subset of consumer ChatGPT users. This data is only used in Section 6.

We describe the contents of each dataset, the sampling procedures that produced them, and the privacy protections we implemented in constructing and employing them in analysis.

3.1 Growth Dataset

We compiled a dataset covering all usage on consumer ChatGPT Plans (Free, Plus, Pro) since ChatGPT's launch in November 2022. We exclude users on non-consumer plans (Business f.k.a. Teams,

[14]The exact beginning and end dates of this sample are May 15, 2024 and June 26, 2025.
[15]The exact beginning and end dates of this sample are May 15, 2024 and July 31, 2025.

Enterprise, Education).

For each user and day, this dataset reports the total number of messages sent by the user on that day. It also reports, for each message, de-identified user metadata, including the timestamp of their first interaction with ChatGPT, the country from which their account is registered, their subscription plan on each day, and their self-reported age (reported in coarse 5–7-year buckets to protect user privacy).

3.2 Classified Messages

To understand usage while preserving user privacy, we construct message-level datasets without any human ever reading the contents of a message. See Figure 1 for an overview of the privacy-preserving classification pipeline. Messages are categorized according to 5 different LLM-based classifiers. The classifiers are introduced in more detail in Section 5, their exact text is reproduced in Appendix A, and our validation procedure is described in Appendix B.

Sampled From All ChatGPT Users. We uniformly sampled approximately 1.1 million conversations, and then sampled one message within each conversation, with the following restrictions:

1. We only include messages from May 2024 to July 2025.
2. We exclude conversations from users who opted out of sharing their messages for model training.
3. We exclude users who self-report their age as under 18.
4. We exclude conversations that users have deleted and from users whose accounts have been deactivated or banned.
5. We exclude logged-out users,[16] which represented a minority share of ChatGPT users over the sample period.

Our sample is drawn from a table that is itself sampled, where the sampling rate varied over time. We thus adjust our sampling weights to maintain a fixed ratio with aggregate messages sent.

Sampled From a Subset of ChatGPT Users. We construct two samples of classified messages from a subset of ChatGPT users (approximately 130,000 users). This sample of users does not include any users who opted out of sharing their messages for training, nor does it include users whose self-reported age is below 18, nor does it include users who have been banned or deleted their accounts.

The first sample contains classifications of 1.58 million messages from this subset of users, sampled at the conversation level (a conversation is a series of messages between the user and chatbot). This sample is constructed such that the user's representation in the data is proportional to overall message volume. The second sample contains messages sent from this subset of users, sampled at the user level with up to six messages from each user in the group.

[16] ChatGPT became available to logged-out users in April 2024, i.e., users could use ChatGPT without signing up for an account with an email address. However, messages from logged-out users are only available in our dataset from March 2025, thus for consistency we drop all messages from logged-out users.

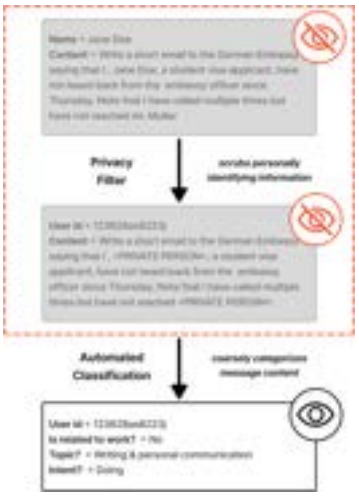


Figure 1: Illustration of Privacy-Preserving Automated Classification Pipeline (Synthetic Example). Messages are first stripped of PII via an internal LLM-based tool called *Privacy Filter*. Then they are classified by LLM-based automated classifiers, described in detail in Appendices A and B. Humans do not see raw messages or PII-scrubbed messages, only the final classifications of messages.

Privacy via Automated Classifiers. No one looked at the content of messages while conducting analysis for this paper. All analysis of message content was performed via automated LLM-based classifiers run on de-identified and PII-scrubbed message data (see Figure 1). The messages are first scrubbed of PII using an internal LLM-based tool,[17] and then classified according to classifiers defined over a controlled label space—the most precise classifier we use on the message-level data set is the O*NET Intermediate Work Activities taxonomy, which we augment to end up with 333 categories. We introduce technical and procedural frictions that prevent accidental access to the underlying text (for example, interfaces that do not render message text to researchers).

Our classifications aim to discern the intent of a given message, and thus we include the prior 10 messages in a conversation as context.[18] For an example, see Table 2.

Stand-Alone Message	**Message with Prior Context**
[user]: "10 more"	[user]: "give me 3 cultural activities to do with teens" [assistant]: "1. Visit a museum . . . " (*truncated*) [user]: "10 more"

Table 2: Illustration of Context-Augmented Message Classifications (Synthetic Example). The left column shows a standalone message to be classified, and the right column shows the prior context included in the classification of the message on the left.

We truncate each message to a maximum of 5,000 characters, because long context windows could induce variability in the quality of the classification (Liu et al., 2023). We classify each message with the "gpt-5-mini" model, with the exception of *Interaction Quality,* which uses "gpt-5," using the prompts listed in Appendix A.

[17] Internal analyses show that the tool, *Privacy Filter*, has substantial alignment with human judgment.
[18] In the case of *Interaction Quality,* we additionally include the next two messages in the conversation as context.

We validated each of the classification prompts by comparing model classification decisions against human-judged classifications of a sample of conversations from the publicly available WildChat dataset (Zhao et al., 2024), a set of conversations with a third-party chatbot which users affirmatively gave their assent to share publicly for research purposes.[19] Appendix B provides detail on our validation approach and performance relative to human judgment. For additional transparency, we classify a sample of 100,000 public WildChat messages and provide those data in this paper's replication package.

3.3 Employment Dataset

We conduct limited analyses of aggregated employment categories based on publicly available data for a sample of consumer ChatGPT users. This sample included approximately 130,000 Free, Plus, and Pro users, and the employment categories were aggregated by a vendor working through a secure Data Clean Room (DCR). For this analysis, we use the same exclusion criteria as for the message-level datasets: we exclude deactivated users, banned users, users who have opted out of training, and users whose self-reported age is under 18. Because the data was only available for a subset of users the results may not be representative of the full pool of users.

Description. The employment data, which is aggregated from publicly available sources, includes industry, occupations coarsened to O*NET categories, seniority level, company size, and education information that is limited to the degree attained. A vendor working within a DCR procured this dataset, restricted us to running only aggregated queries against it through the DCR, and deleted it upon the study's completion.

Privacy via a Data Clean Room. We never directly accessed user-level demographic records. All analysis of employment data was executed exclusively within a secure DCR that permits only pre-approved aggregate computations across independently held datasets; neither party can view or export the other party's underlying records. We governed the DCR with strict protocols: To execute any query that touched the external demographic data, we first obtained explicit sign-off from a committee of 6 coauthors and then submitted the notebook to our data partner for approval; only approved notebooks could run in the DCR (see Figure 2).

Our partner enforced strict aggregation limits: they only approved code that returned cells meeting a threshold of 100 users. Consequently, no individual rows or narrowly defined categories were ever visible to researchers. For example, if 99 users had the occupation "anesthesiologist," any occupation-level output would place those users into a "suppressed" category, or place these observations in a coarsened category (e.g. "medical professionals") rather than reporting a separate cell of anesthesiologists.

[19] The dataset was collected from a third party chatbot using OpenAI's LLMs via their API.

Figure 2: Illustration of Aggregated Employment Category Analysis via a Data Clean Room. All queries run in the Data Clean Room must be approved by our data partner, enforcing a strict aggregation threshold (100 observations). As a result, researchers cannot access user-level employment data, only aggregated employment categories.

3.4 Summarizing Our Approach to Privacy

We took measures to safeguard user privacy at every stage of analysis. To summarize, the key elements of our approach are:

> *Automated classification of messages.* In the course of analysis, no one ever looked directly at the content of user messages: all of our analysis of the content of user messages is done through output of automated classifiers run on de-identified and PII-scrubbed usage data.
>
> *Aggregated employment data via a data clean room.* We analyze and report aggregated employment data through a secure data clean room environment: no one on the research team had direct access to user-level demographic data and none of our analyses report aggregates for groups with less than 100 users.

In following these measures, we aim to match or exceed the privacy protection precedents set by other social scientists studying chatbots and those linking digital platform data to external sources.

We follow the precedent established in recent analyses of chatbot conversations (Phang et al. (2025), Eloundou et al. (2025), Handa et al. (2025), Tomlinson et al. (2025)) that rely on automated classification rather than human inspection of raw transcripts. In particular, Phang et al. (2025)'s study of affective use of ChatGPT and Eloundou et al. (2025) investigation of first-person fairness in chatbots both analyze ChatGPT message content via automated classifiers and emphasize classifier-based labeling as a scalable, privacy-preserving approach. Anthropic's Handa et al. (2025) used a similar approach: their *Clio* methodology applies automated classifiers to large collections of conversations, classifying conversations into thousands of topics, and in their appendix they describe manual validation on sampled conversations (100 user conversations flagged for review and 100 randomly sampled calibrations). Like Eloundou et al., we validate our classifiers using WildChat, a public dataset of user conversations.

Other papers have analyzed digital behavior and demographic data; we mention a few relevant precedents here. Humlum and Vestergaard (2025b) and Humlum and Vestergaard (2025a), for example, analyze large-scale surveys on chatbot use along with Danish administrative labor market data. Chetty et al. (2022) analyze de-identified Facebook friendship graphs and anonymized IRS tax records, aggregated at the zip code level.

4 The Growth of ChatGPT

ChatGPT was released to the public on November 30, 2022 as a "research preview," and by December 5 it had more than one million registered users. Figure 3 reports the growth of overall weekly active users (WAU) on consumer plans over time. ChatGPT had more than 100 million logged-in WAU after one year, and almost 350 million after two years. By the end of July 2025, ChatGPT had more than 700 million total WAU, nearly 10% of the world's adult population.[20]

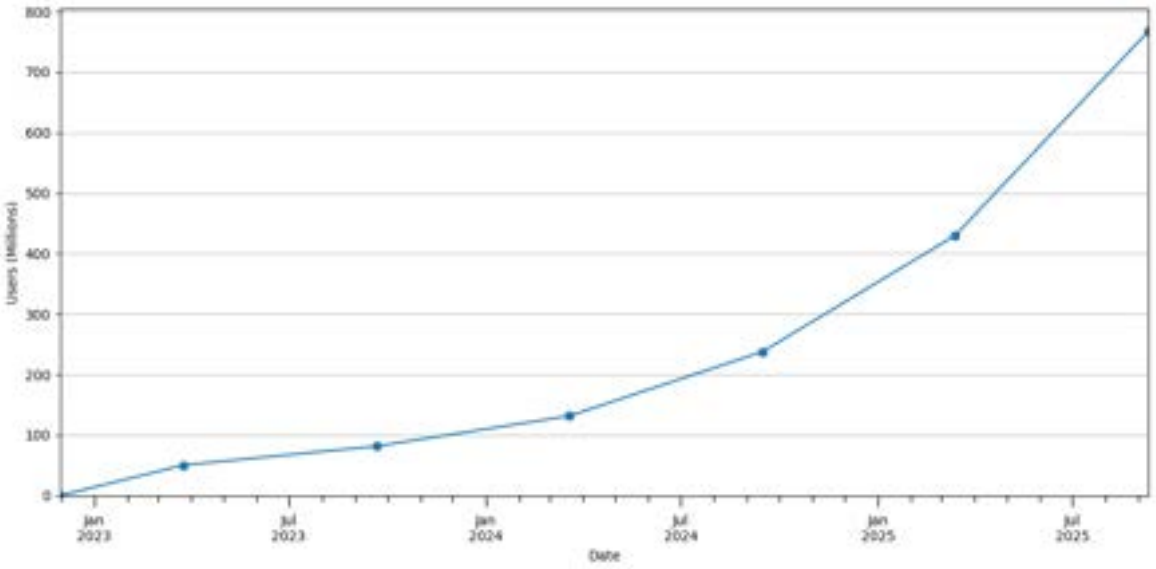


Figure 3: Weekly active ChatGPT users on consumer plans (Free, Plus, Pro), shown as point-in-time snapshots every six months, November 2022–September 2025.

Figure 4 presents growth in the total messages sent by users over time. The solid line shows that between July 2024 and July 2025, the number of messages sent grew by a factor of more than 5.

Figure 4 also shows the contribution of individual cohorts of users to aggregate message volume. The yellow line represents the first cohort of ChatGPT users: their usage declined somewhat over 2023, but started growing again in late 2024 and is now higher than it has ever been. The pink line represents messages from users who signed up in Q3 of 2023 or earlier, and so the *difference* between

[20]Note that we expect our counts of distinct accounts to somewhat exceed distinct people when one person has two accounts (or, for logged-out users, one person using two devices). For logged-in users, the count is based on distinct login credentials (email addresses), and one person may have multiple accounts. For logged-out users, the count is based on distinct browser cookies; this would double-count people if someone returns to ChatGPT after clearing their cookies, or if they access ChatGPT with two different devices in the same week.

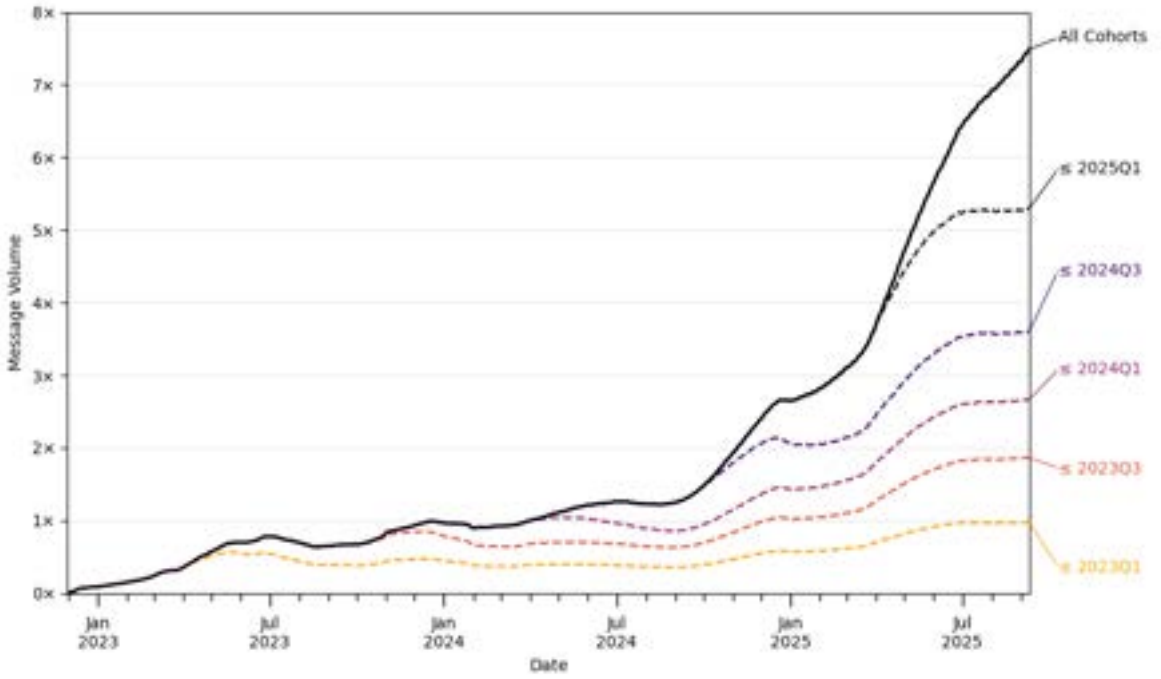


Figure 4: Daily message volumes from ChatGPT consumer plans (Free, Plus, Pro), split by sign-up date of the requesting user. Reported values are moving averages of the past 90 days. Y-axis is an index normalized to the reported value for "All Cohorts" at the end of Q1 2024 (April 1, 2024).

the yellow and pink lines represents the messages sent by users who signed up in Q2 and Q3 of 2023. There has been dramatic growth in message volume both by new cohorts of users, and from growth in existing cohorts.

Figure 5 normalizes each cohort, plotting daily messages per weekly active user. Each line represents an individual cohort (instead of a cumulative cohort, as in Figure 4). The figure shows that earlier sign-ups have consistently had higher usage, but that usage has also consistently grown within every cohort, which we interpret as due to both (1) improvements in the capabilities of the models, and (2) users slowly discovering new uses for existing capabilities.

5 How ChatGPT is Used

We next report on the *content* of ChatGPT conversations using a variety of different taxonomies. For each taxonomy we describe a "prompt" which defines a set of categories, and then apply an LLM to map each message to a category. Our categories often apply to the user's *intention*, rather than the text of the conversation, and as such we never directly observe the ground truth. Nevertheless the classifier results can be interpreted as the best-guess inferences that a human would make: the guesses from the LLM correlate highly with human guesses from the same prompt, and we get similar qualitative results when the prompt includes a third category for "uncertain."

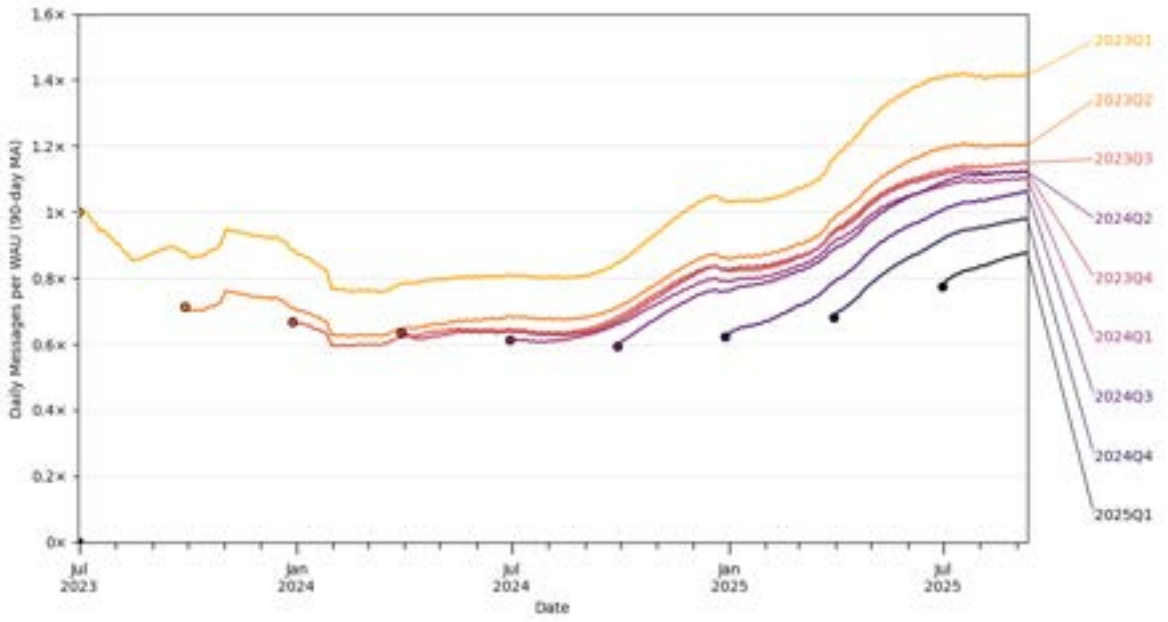


Figure 5: Daily messages sent per weekly active user, split by sign-up cohort. Sample only considers users of ChatGPT consumer plans (Free, Plus, Pro). Reported values are moving averages of the past 90 days and are reported starting 90 days after the cohort is fully formed. Y-axis is an index normalized to the first reported value for the Q1 2023 cohort.

5.1 What share of ChatGPT queries are related to paid work?

We label each user message in our dataset based on whether it appears to be related to work, using an LLM classifier. The critical part of the prompt is as follows:[21]

> *Does the last user message of this conversation transcript seem likely to be related to doing some work/employment? Answer with one of the following:*
>
> *(1) likely part of work (e.g., "rewrite this HR complaint")*
>
> *(0) likely not part of work (e.g., "does ice reduce pimples?")*

Table 1 shows that both types of queries grew rapidly between June 2024 and June 2025, however non-work-related messages grew faster: 53% of messages were not related to work in June 2024, which climbed to 73% by June 2025.

Figure 6 plots the share of non-work messages decomposed by cumulative sign-up cohorts. Successive cohorts have had a higher share of non-work messages, but also within each cohort their non-work use has increased. Comparing the share among all users (black line) to the share among the earliest cohort of users (yellow line), we can see that they track very closely.

[21] See Appendix A for the full prompt, see Appendix B for validation.

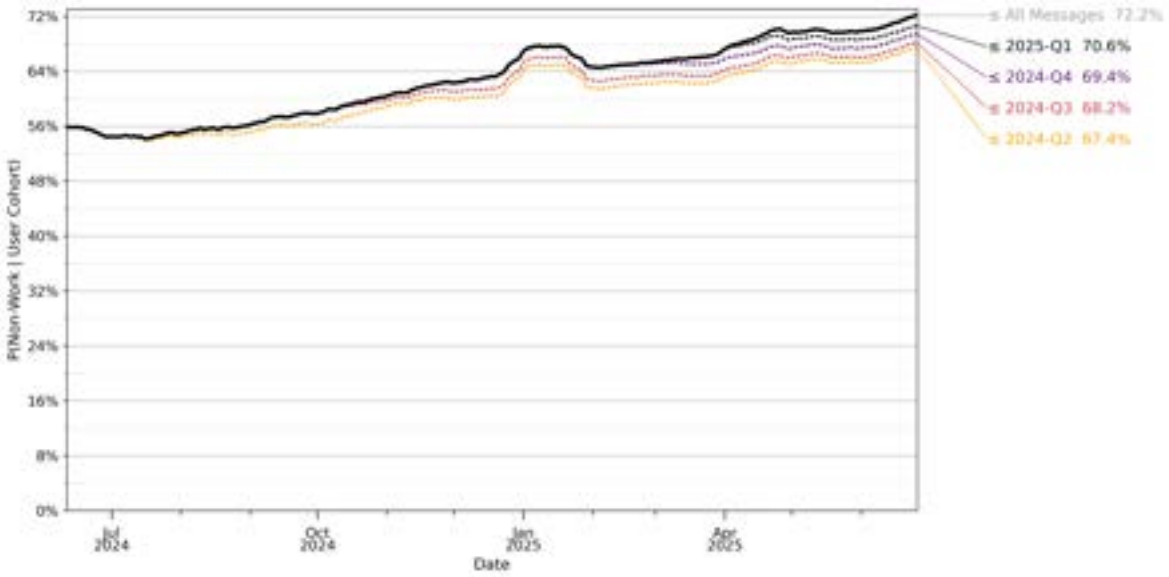


Figure 6: The solid black line represents the probability that a messages on a given day is not related to work, as determined by an automated classifier. Values are averaged over a 28-day lagging window. The dotted orange line shows the same calculation, but conditioned on messages being from users who first used ChatGPT during or before Q2 of 2024. The remaining lines are defined similarly for successive quarters, with coloring cooling for more recent cohorts. Counts are calculated from a sample of approximately 1.1 million sampled conversations from May 15, 2024 through June 26, 2025. Observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3.

5.2 What are the topics of ChatGPT conversations?

We modify a classifier used by internal research teams at OpenAI that identifies which capabilities the user is requesting from ChatGPT. The classifier itself directly assigns the user's query into one of 24 categories. We aggregate these 24 categories into seven topical groupings (the full conversation-categorization prompt is given in Appendix A):

Topic	**Conversation Category**
Writing	Edit or Critique Provided Text
	Personal Writing or Communication
	Translation
	Argument or Summary Generation
	Write Fiction
Practical Guidance	How-To Advice
	Tutoring or Teaching
	Creative Ideation
	Health, Fitness, Beauty, or Self-Care
Technical Help	Mathematical Calculation
	Data Analysis

Topic	Conversation Category
	Computer Programming
Multimedia	Create an Image
	Analyze an Image
	Generate or Retrieve Other Media
Seeking Information	Specific Info
	Purchasable Products
	Cooking and Recipes
Self-Expression	Greetings and Chitchat
	Relationships and Personal Reflection
	Games and Role Play
Other/Unknown	Asking About the Model
	Other
	Unclear

Table 3: Coarse Conversation Topics and Underlying Classifier Categories

Figure 7 shows the composition of user messages over time. The three most common Conversation Topics are *Practical Guidance*, *Seeking Information*, and *Writing*, collectively accounting for about 77% of all ChatGPT conversations. *Practical Guidance* has remained constant at roughly 29% of overall usage. *Writing* has declined from 36% of all usage in July 2024 to 24% a year later. *Seeking Information* has grown from 14% to 24% of all usage over the same period. The share of *Technical Help* declined from 12% from all usage in July 2024 to around 5% a year later – this may be because the use of LLMs for programming has grown very rapidly through the API (outside of ChatGPT), for AI assistance in code editing and for autonomous programming agents (e.g. Codex). *Multimedia* grew from 2% to just over 7%, with a large spike in April 2025 after ChatGPT released new image-generation capabilities: the spike attenuated but the elevated level has persisted.

Figure 8 shows Conversation Topics, restricting the sample to only work-related messages. About 40% of all work-related messages in July 2025 are *Writing*, by far the most common Conversation Topic. *Practical Guidance* is the second most common use case at 24%. *Technical Help* has declined from 18% of all work-related messages in July 2024 to just over 10% in July 2025.

Figure 9 disaggregates four of the seven Conversation Topics into smaller groups and sums up messages of each type over a one-year period. For example, the five sub-categories within *Writing* are (in order of frequency) *Editing or Critiquing Provided Text*, *Personal Writing or Communication*, *Translation*, *Argument or Summary Generation*, and *Writing Fiction*. Three of those five categories (*Editing or Critiquing Provided Text*, *Translation*, and *Argument or Summary Generation*) are requests to modify text that has been provided to ChatGPT by the user, whereas the other two are requests to produce novel text. The former constitute two thirds of all *Writing* conversations, which

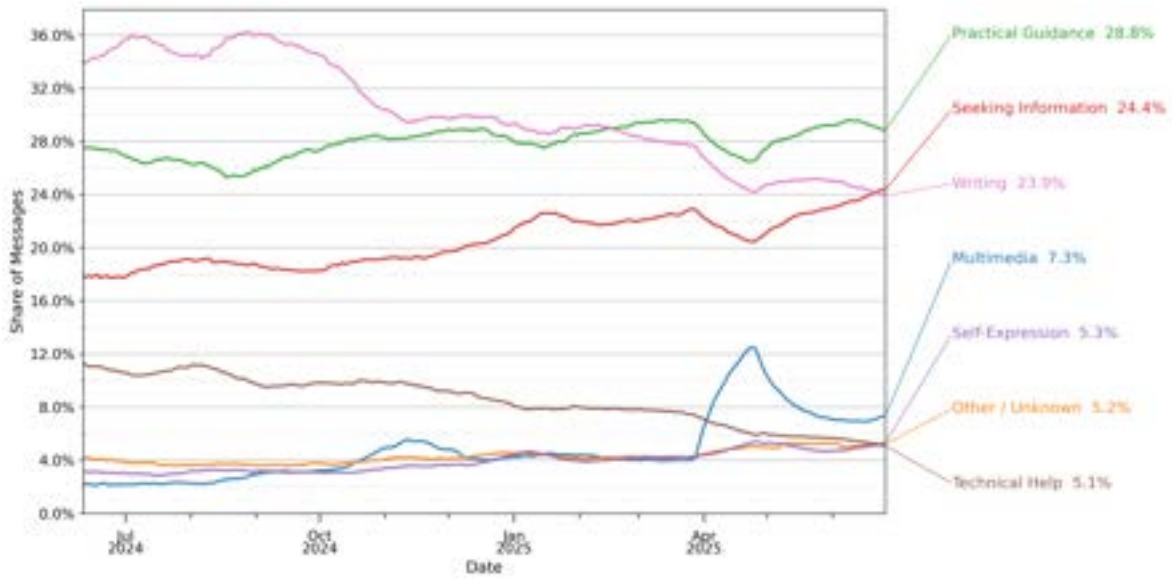


Figure 7: Share of consumer ChatGPT messages broken down by high level conversation topic, according to the mapping in Table 3. Values are averaged over a 28 day lagging window. Shares are calculated from a sample of approximately 1.1 million sampled conversations from May 15, 2024 through June 26, 2025. Observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3.

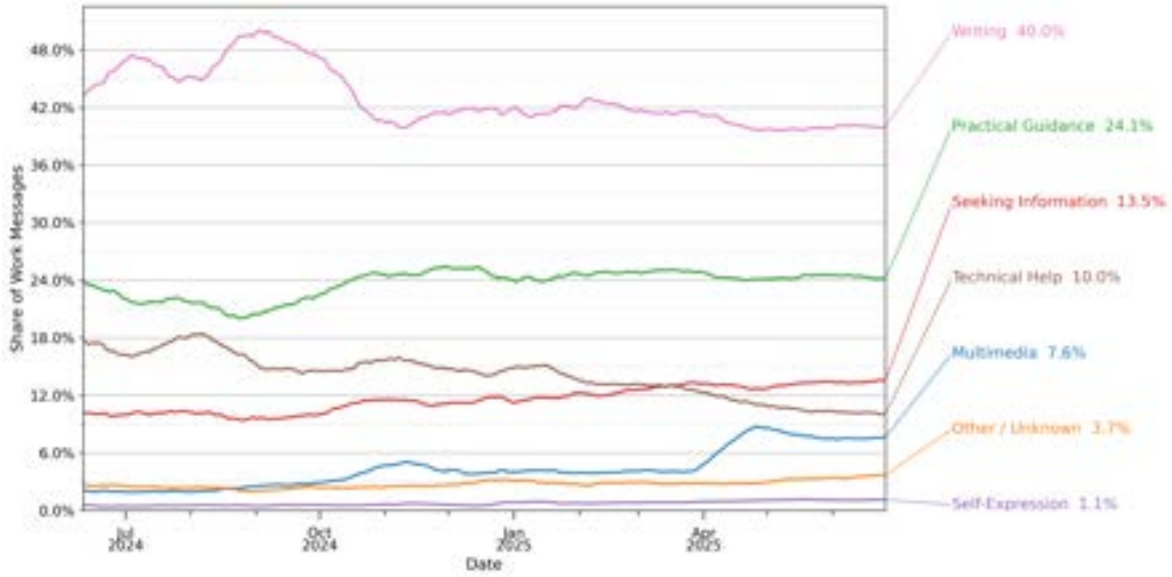


Figure 8: Share of **work related** consumer ChatGPT messages broken down by high level conversation topic, according to the mapping in Table 3. Values are averaged over a 28 day lagging window. Shares are calculated from a sample of approximately 1.1 million sampled conversations from May 15, 2024 through June 26, 2025. Observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3.

suggests that most user *Writing* conversations with ChatGPT are requests to modify user inputs rather than to create something new. Education is a major use case for ChatGPT. 10.2% of all user messages and 36% of *Practical Guidance* messages are requests for *Tutoring or Teaching*. Another large share - 8.5% in total and 30% of *Practical Guidance* - is general how-to advice on a variety of topics. *Technical Help* includes *Computer Programming* (4.2% of messages), *Mathematical Calculations* (3%), and *Data Analysis* (0.4%). Looking at the topic of *Self-Expression*, only 2.4% of all ChatGPT messages are about *Relationships and Personal Reflection* (1.9%) or *Games and Role Play* (0.4%).

While users can seek information and advice from traditional web search engines as well as from ChatGPT, the ability to produce writing, software code, spreadsheets, and other digital products distinguishes generative AI from existing technologies. ChatGPT is also more flexible than web search even for traditional applications like *Seeking Information* and *Practical Guidance*, because users receive customized responses (e.g., tailored workout plans, new product ideas, ideas for fantasy football team names) that represent newly generated content or novel modification of user-provided content and follow-up requests.

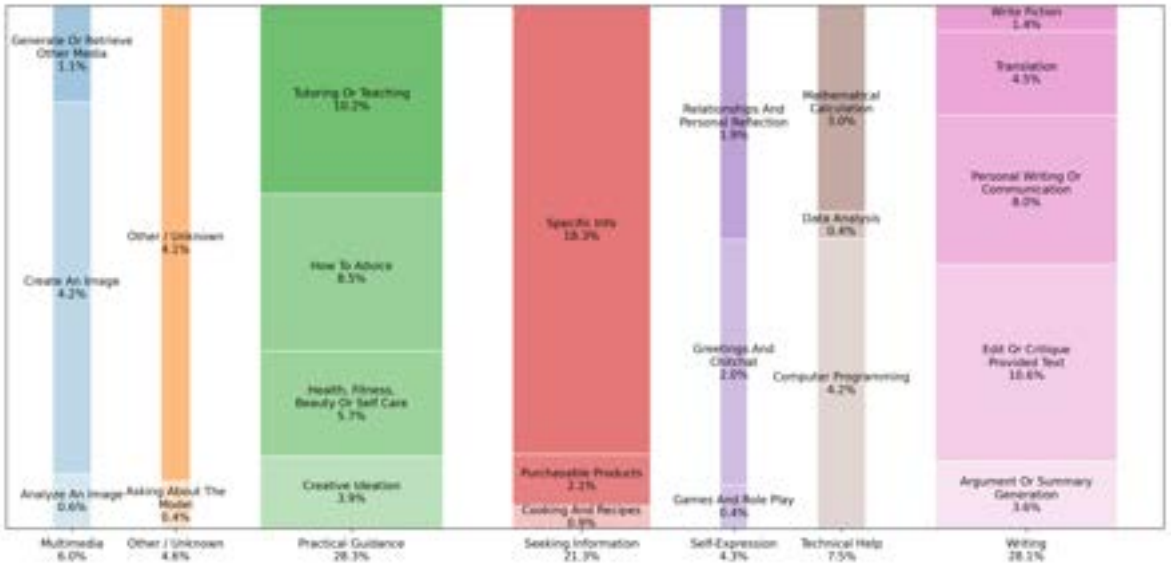


Figure 9: Breakdown of granular conversation topic shares within the coarse mapping defined in Table 3. The underlying classifier prompt is available in Appendix A. Each bin reports a percentage of the total population. Shares are calculated from a sample of approximately 1.1 million sampled conversations from May 15, 2024 through June 26, 2025. Observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3.

5.3 User Intent

Existing studies of the economic impacts of generative AI focus almost exclusively on the potential for AI to perform workplace tasks, either augmenting or automating human labor (e.g. Eloundou et al. (2025), Handa et al. (2025), Tomlinson et al. (2025)). However, generative AI is a highly flexible

technology that can be used in many different ways. In order to learn more about how people seek to use generative AI at work and outside of work, we introduce a classifier that is designed to measure the type of output the user hopes to receive. Specifically, we classify messages according to user intent, coding up conversations according to a simple *Asking, Doing, or Expressing* rubric. The critical part of our classification prompt is as follows:

Intent	**Prompt**
Asking	*Asking is seeking information or advice that will help the user be better informed or make better decisions, either at work, at school, or in their personal life. (e.g. "Who was president after Lincoln?", "How do I create a budget for this quarter?", "What was the inflation rate last year?", "What's the difference between correlation and causation?", "What should I look for when choosing a health plan during open enrollment?").*
Doing	*Doing messages request that ChatGPT perform tasks for the user. User is drafting an email, writing code, etc. Classify messages as "doing" if they include requests for output that is created primarily by the model. (e.g. "Rewrite this email to make it more formal", "Draft a report summarizing the use cases of ChatGPT", "Produce a project timeline with milestones and risks in a table", "Extract companies, people, and dates from this text into CSV.", "Write a Dockerfile and a minimal docker-compose.yml for this app.")*
Expressing	*Expressing statements are neither asking for information, nor for the chatbot to perform a task.*

Conceptually, *Doing* conversations are delivering output that can be plugged into a production process, while *Asking* conversations support decision-making but do not produce output directly, and *Expressing* conversations have little or no economic content.

Figure 10 shows the share of messages by each intent type in our sample. 49% of user messages are *Asking*, 40% are *Doing*, and 11% are *Expressing*. The figure also shows the relationship with our Topic classification: the two taxonomies are correlated but not redundant: *Asking* queries are more likely to be *Practical Guidance* and *Seeking Information*. *Doing* queries are disproportionately *Writing* and *Multimedia*. *Expressing* queries are disproportionately *Self-Expression*. However, the overlap is imperfect. For example, within the *Practical Guidance* topic, an *Asking* message might be advice about how to recover from a sports injury given a user's personal history, while a *Doing* message might request ChatGPT to produce a customized recovery and training plan that could be printed or saved. Within *Technical Help*, an *Asking* message might request help understanding how to debug some code, while a *Doing* message might ask ChatGPT to write code for the user directly.

Figure 11 presents shares of *Asking/Doing/Expressing* just for work-related messages. *Doing* constitutes nearly 56% of work-related queries, compared to 35% for *Asking* and 9% for *Expressing*. Nearly 35% of all work-related queries are *Doing* messages related to *Writing*. *Doing* and *Asking* comprise equal shares of *Technical Help* queries.

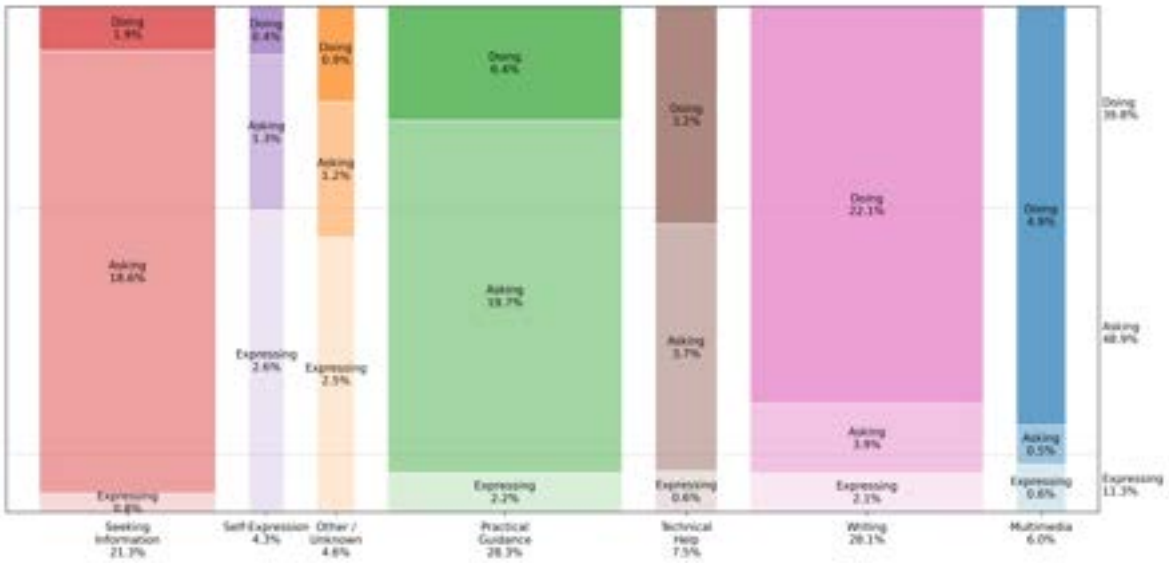


Figure 10: Breakdown of Conversation Topics by Asking/Doing/Expressing category, with topic columns sorted by relative share of "Doing" messages. Prompts for these automated classifiers are available in Appendix A. For a detailed breakdown of conversation topic contents, see Table 3. Each bin reports a percentage of the total population. Shares are calculated from a sample of approximately 1.1 million sampled conversations from May 15, 2024 through June 26, 2025. Observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3.

Figure 11: Breakdown of Conversation Topics by Asking/Doing/Expressing category for **only work-related messages**, with topic columns sorted by relative share of "Doing" messages. Prompts for these automated classifiers are available in Appendix A. For a detailed breakdown of conversation topic contents, see Table 3. Each bin reports a percentage of the total population. Shares are calculated from a sample of approximately 1.1 million sampled conversations from May 15, 2024 through June 26, 2025. Observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3.

Figure 12 presents changes over time in the composition of messages by user intent. In July 2024, usage was evenly split between *Asking* and *Doing*, with just under 8% of messages classified as *Expressing*. *Asking* and *Expressing* grew much faster than *Doing* over the next year, and by late June 2025 the split was 51.6% *Asking*, 34.6% *Doing*, and 13.8% *Expressing*.

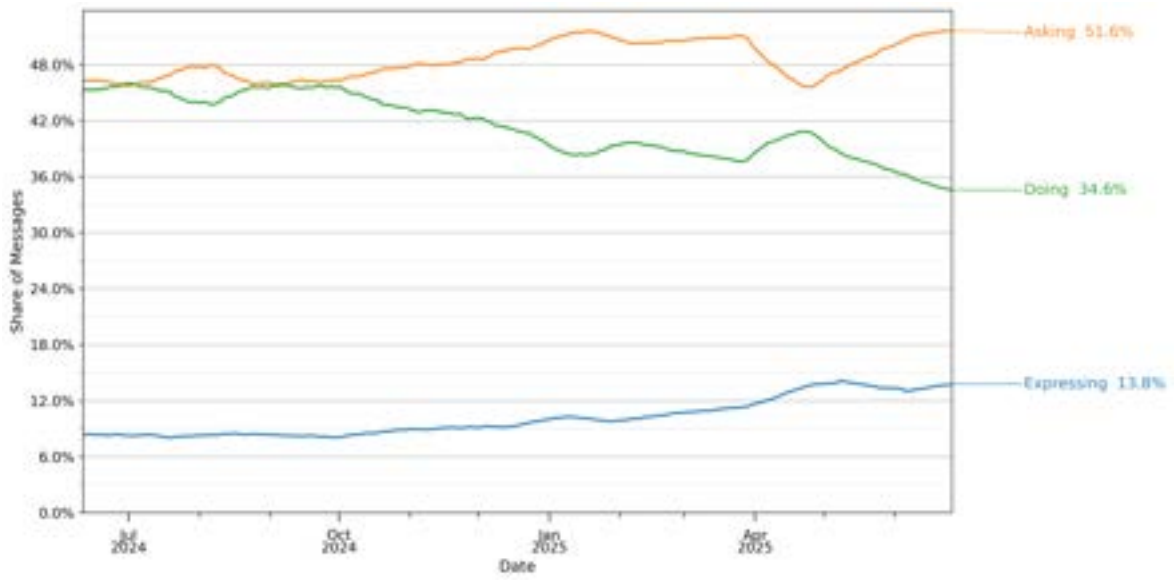


Figure 12: Shares of messages classified as Asking, Doing, or Expressing by an automated ternary classifier. Values are averaged over a 28 day lagging window. Shares are calculated from a sample of approximately 1.1 million sampled conversations from May 15, 2024 through June 26, 2025. Observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3.

Figure 13 presents the share of work-related messages by user intent. *Doing* messages, which account for approximately 40% of messages, have an even split of messages between work-related and non-work related.

5.4 O*NET Work Activities

We map message content to work activities using the Occupational Information Network (O*NET) Database Version 29.0, similar to Tomlinson et al (2025). O*NET was developed in partnership with the U.S. Department of Labor and systematically classifies jobs according to the skills, tasks, and work activities required to perform them. O*NET associates each occupation with a set of tasks that are performed at different levels of intensity. Each task is then aggregated up to three levels of detail - 2,087 detailed work activities (DWAs), 332 intermediate work activities (IWAs), and 41 generalized work activities (GWAs).

To understand the work activities associated with ChatGPT usage, we mapped messages to one of the 332 O*NET Intermediate Work Activities (IWA), with an additional option of *Ambiguous* to account for situations where the user message lacked sufficient context.[22] We then used the official

[22]We drew a sample of approximately 1.1 million conversations from May 2024 to June 2025, selected a random message within each, and classified it according to the prompt in A.

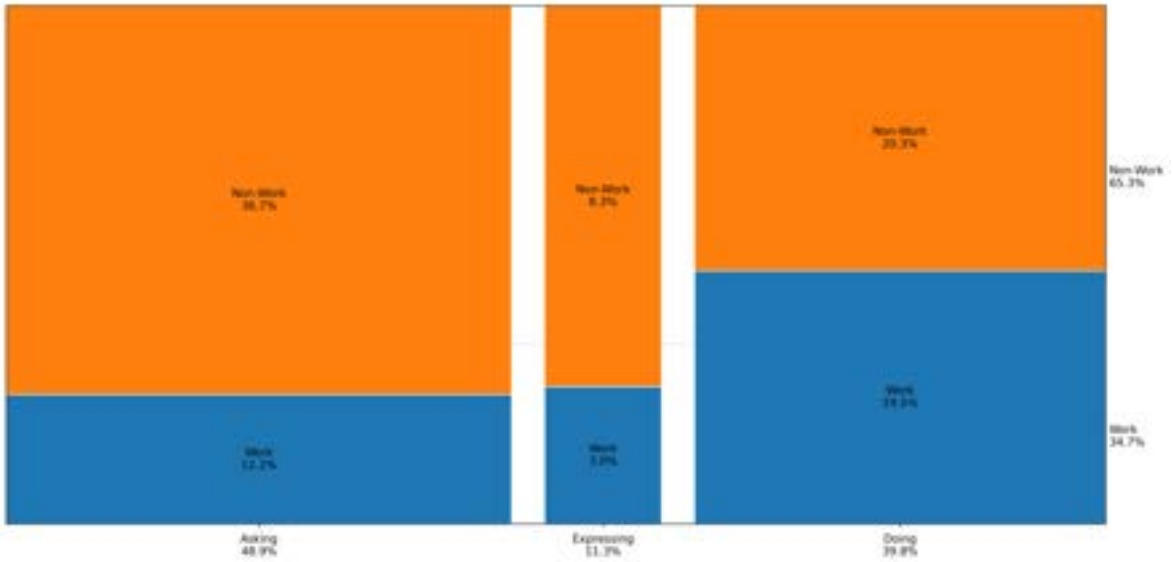


Figure 13: Shares of Asking, Doing, and Expressing messages split by work vs. non-work. See A to review the prompts used by the automated classifiers. The annotations on the right show the shares of work and non-work for the full sample. Each bin reports a percentage of the total population. Shares are calculated from a sample of approximately 1.1 million sampled conversations from May 15, 2024 through June 26, 2025. Observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3.

O*NET taxonomy to map these classified IWAs to one of the Generalized Work Activities (GWA). We do not show the shares for the following GWAs as there were fewer than 100 users sending messages for each category and group them into *Suppressed*.

Figure 14 presents the share of messages that belong to each GWA, in descending order. Nearly half of all messages (45.2%) fall under just three GWAs related to information use and manipulation: *Getting Information* (19.3%), *Interpreting the Meaning of Information for Others* (13.1%), and *Documenting/Recording Information* (12.8%). The next most common work activities are *Providing Consultation and Advice* (9.2%), *Thinking Creatively* (9.1%), *Making Decisions and Solving Problems* (8.5%), and *Working with Computers* (4.9%). These seven GWAs collectively account for 76.9% of all messages.

Figure 15 presents the distribution of GWAs for the subsample of messages we classify as work-related. Among work-related messages, the most common GWAs are *Documenting/Recording Information* (18.4%), *Making Decisions and Solving Problems* (14.9%), *Thinking Creatively* (13.0%), *Working with Computers* (10.8%), *Interpreting the Meaning of Information for Others* (10.1%), *Getting Information* (9.3%), and *Providing Consultation and Advice to Others* (4.4%). These seven GWAs collectively account for nearly 81% of work-related messages. Overall, the majority of ChatGPT usage at work appears to be focused on two broad functions: 1) obtaining, documenting, and interpreting information; and 2) making decisions, giving advice, solving problems, and thinking creatively.

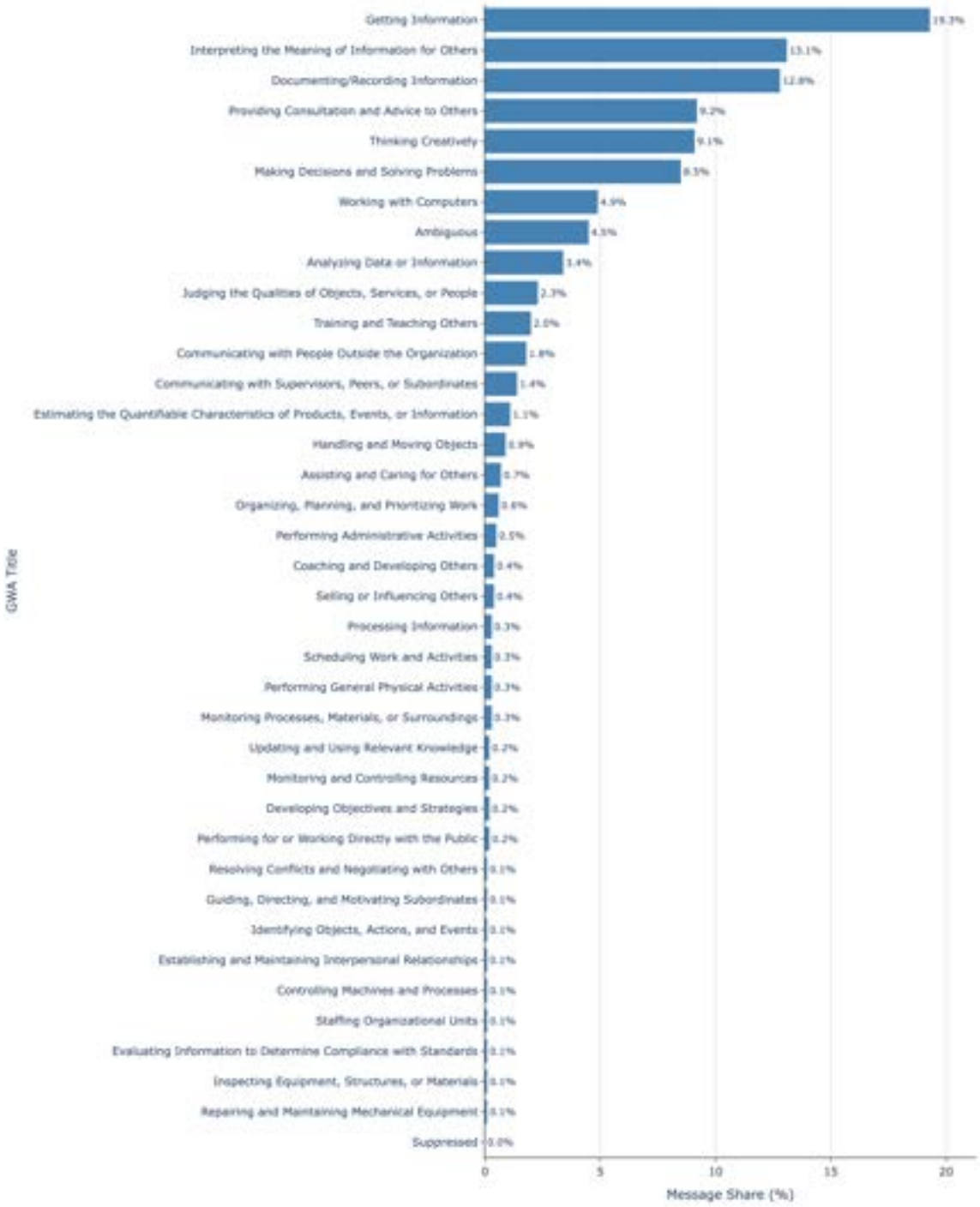


Figure 14: GWA Shares of 1.1M ChatGPT Messages. Messages are classified as pertaining to one of 332 O*NET IWAs, or *Ambiguous* using the prompt provided in the Appendix. IWAs were then aggregated to GWAs using the O*NET Work Activities taxonomy. Message sample from May 15, 2024 through June 26, 2025. We do not show the shares for the following GWAs as there were fewer than 100 users sending messages for each category and group them into *Suppressed*.

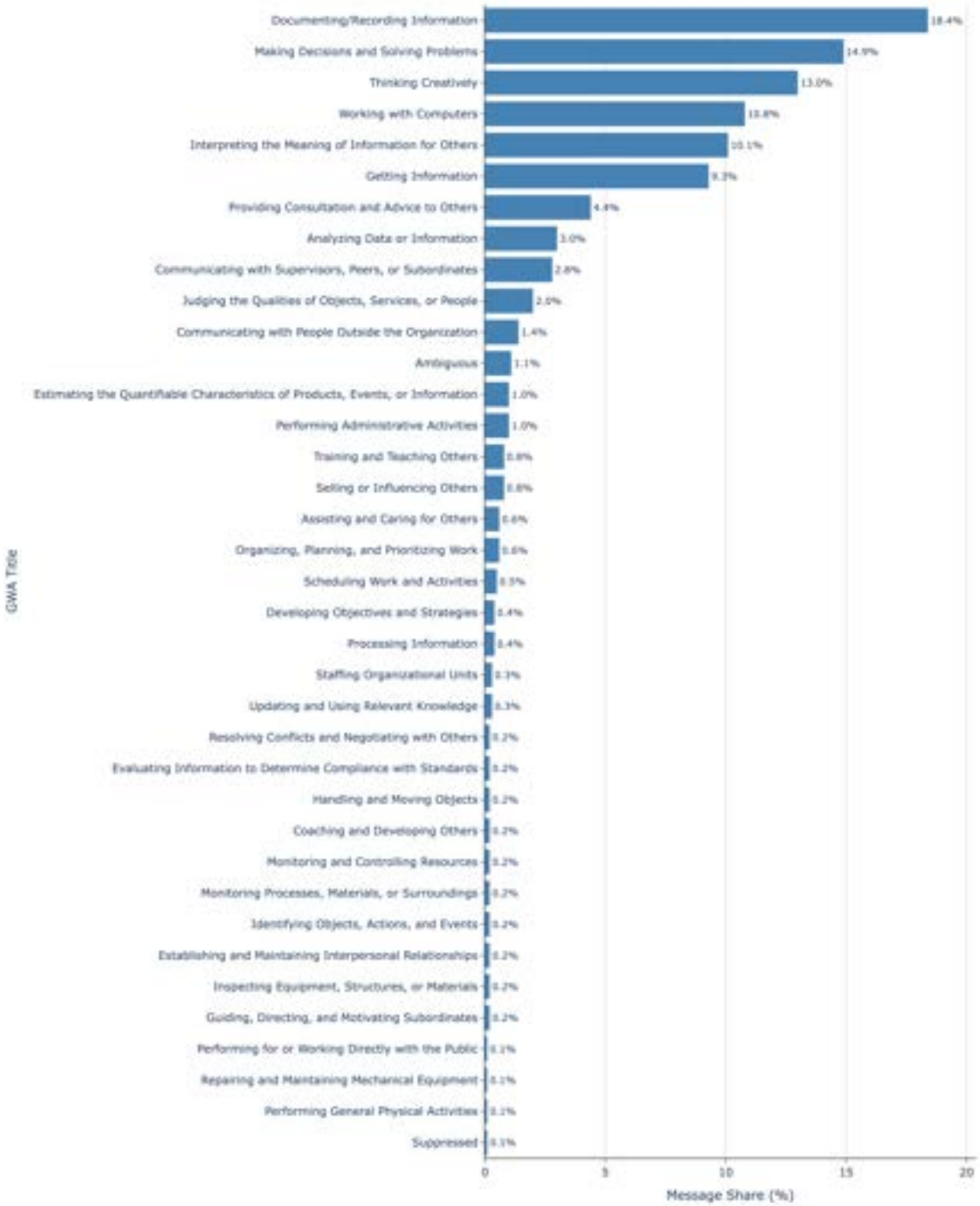


Figure 15: GWA Shares of approximately 366,000 Work-Classified Messages. Messages are classified as pertaining to one of 332 O*NET IWAs or *Ambiguous*. IWAs were then aggregated to GWAs using the O*NET Work Activities taxonomy. Messages were also additionally classified as pertaining to work or non-work. GWA shares are shown only for work-classified messages. Message sample from May 15, 2024 through June 26, 2025. We do not show the shares for the following GWAs as there were fewer than 100 users sending messages for each category and group them into *Suppressed*. Prompts are provided in the Appendix.

5.5 Quality of Interactions

We additionally used automated classifiers to study the user's apparent satisfaction with the chatbot's response to their request. Our *Interaction Quality* classifier looks for an expression of satisfaction or dissatisfaction in the user's subsequent message in the same conversation (if one exists), with three possible categories: *Good*, *Bad*, and *Unknown*. [23]

Figure 16 plots the overall growth of messages in these three buckets. In late 2024 *Good* interactions were about three times as common as *Bad* interactions, but *Good* interactions grew much more rapidly over the next nine months, and by July 2025 they were more than four times more common.

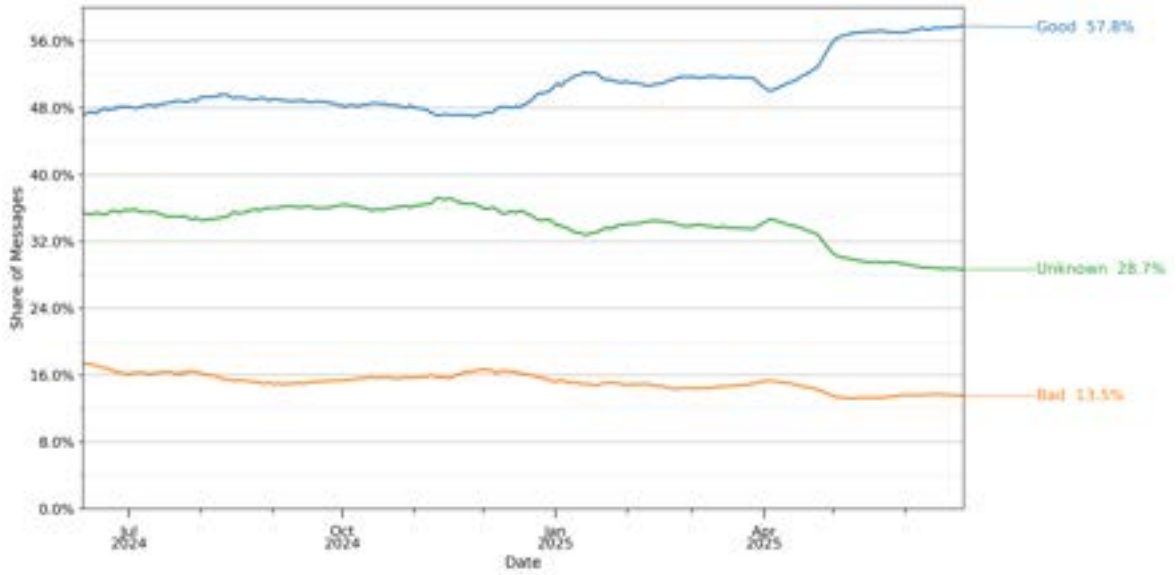


Figure 16: Interaction quality shares, based on automated sentiment analysis of the *next response* provided by the user. See Appendix B to understand how this classifier was validated. Values are averaged over a 28 day lagging window. Shares are calculated from a sample of approximately 1.1 million sampled conversations from May 15, 2024 through June 26, 2025. Observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3.

Details on the validation of this classifier, along with measurements of how it correlates with explicit thumbs up/thumbs down annotations from users, are included in Appendix B.

Figure 17 shows the ratio of good-to-bad messages by conversation topic and interaction type, as rated by Interaction Quality. Panel A shows that *Self-Expression* is the highest rated topic, with a good-to-bad ratio of more than seven, consistent with the growth in this category. *Multimedia* and *Technical Help* have the lowest good-to-bad ratios (1.7 and 2.7 respectively). Panel B shows that *Asking* messages are substantially more likely to receive a good rating than *Doing* or *Expressing* messages.

[23] For this classifier we do not disclose the prompt.

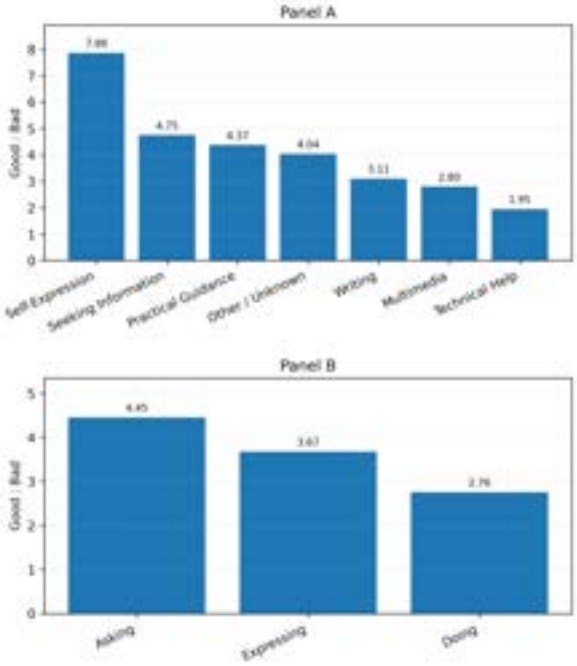


Figure 17: Average *Good* to *Bad* ratio for user interactions by Conversation Topic (Panel A) and Asking/Doing/Expressing classification (Panel B). The prompts for each of these automated classifiers (with the exception of interaction quality) are available in Appendix A. Values represent the average ratio from May 15, 2024 through June 26, 2025, where observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3.

6 Who Uses ChatGPT

In this section we report basic descriptive facts about who uses consumer ChatGPT. Existing work documents variation in generative AI use by demographic groups within representative samples in the U.S. (Bick et al. (2024), Hartley et al. (2025)) and within a subset of occupations in Denmark (Humlum and Vestergaard, 2025a). All of these papers find that generative AI is used more frequently by men, young people, and those with tertiary and/or graduate education.

We make three contributions relative to this prior literature. First, we confirm these broad demographic patterns in a global sample rather than a single country. Second, we provide more detail for selected demographics such as age, gender, and country of origin and study how gaps in each have changed over time. Third, we use a secure data clean room to analyze how ChatGPT usage varies by education and occupation.

6.1 Name Analysis

We investigate potential variation by gender by classifying a global random sample of over 1.1 million ChatGPT users' first names using public aggregated datasets of name-gender associations. We used the World Gender Name Dictionary, and Social Security popular names, as well as datasets of popular Brazilian and Latin American names. This methodology is similar to that in (Hofstra et al., 2020) and (West et al., 2013). Names that were not in these datasets, or were flagged as ambiguous in the datasets, or had significant disagreement amongst these datasets were classified as *Unknown*.

Excluding *Unknown*, a significant share (around 80%) of the weekly active users (WAU) in the first few months after ChatGPT was released were by users with typically masculine first names. However, in the first half of 2025, we see the share of active users with typically feminine and typically masculine names reach near-parity. By June 2025 we observe active users are more likely to have typically feminine names. This suggests that gender gaps in ChatGPT usage have closed substantially over time.

We also study differences in usage topics. Users with typically female first names are relatively more likely to send messages related to *Writing* and *Practical Guidance*. By contrast, users with typically male first names are more likely to use ChatGPT for *Technical Help*, *Seeking Out Information*, and *Multimedia* (e.g., modifying or creating images).

6.2 Variation by Age

A subset of users self-report their age when registering for OpenAI. Among those who self-report their age, around 46% of the messages in our dataset are accounted for by users 18-25.

A higher share of messages are work-related for older users. Work-related messages comprised approximately 23% of messages for users under age 26, with this share increasing with age. The one exception is users who self-attest to being 66 years-old or older, with only 16% of their classified messages being work-related. The plot below shows trends in the share of work-related messages by age group. ChatGPT usage has become less work-related over time for users of all ages.

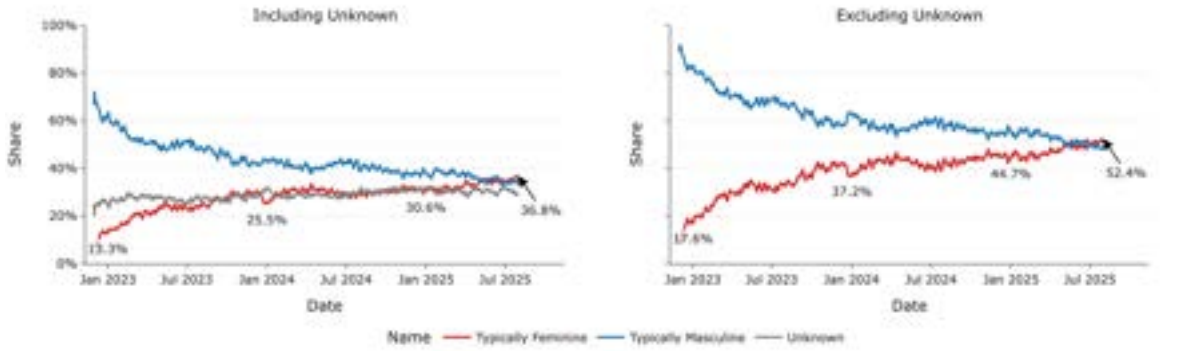


Figure 18: Breakdown of weekly active users by typically masculine and typically feminine first names. We draw on a uniform sample of 1.1M ChatGPT accounts, subject to the same user exclusion principles as other datasets we analyze. Note that this is a separate sample than those described in Section 3. First names are classified as typically masculine or typically feminine using public aggregated datasets of name-gender associations.

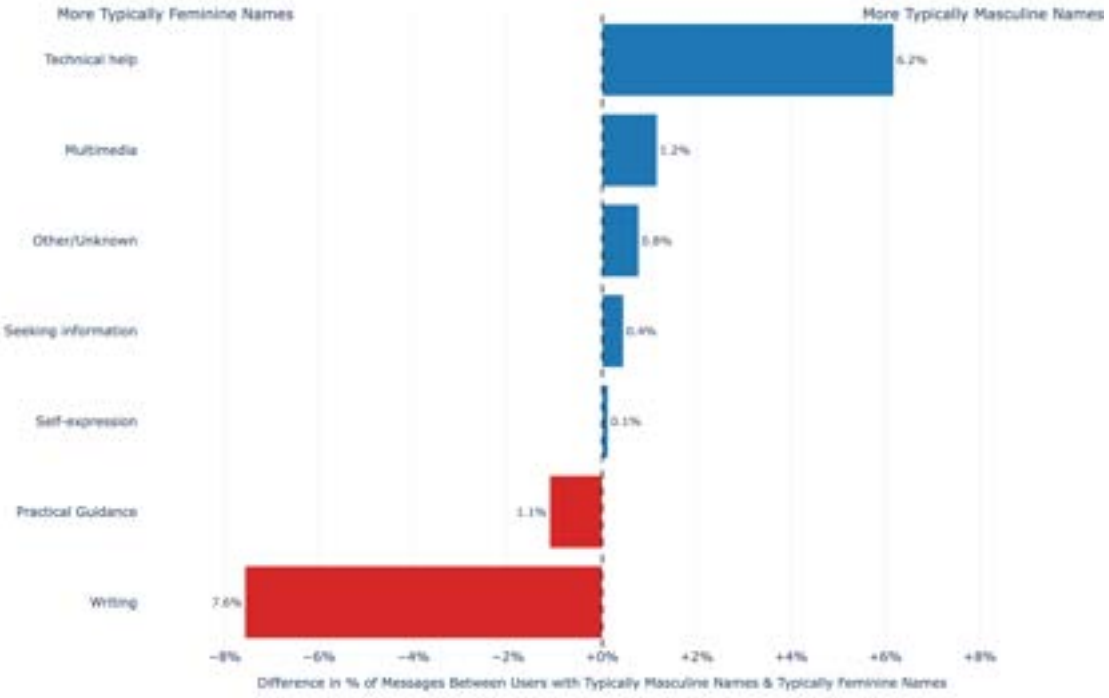


Figure 19: Difference in share of topic prevalence in messages by users with typically masculine/feminine first name. We draw on a uniform sample of 1.1M ChatGPT accounts, subject to the same user exclusion principles as other datasets we analyze. Note that this is a separate sample than those described in Section 3. First names are classified as typically masculine or typically feminine using public aggregated datasets of name-gender associations. Topics are aggregated groupings from a classifier whose prompt we provide in Appendix A.

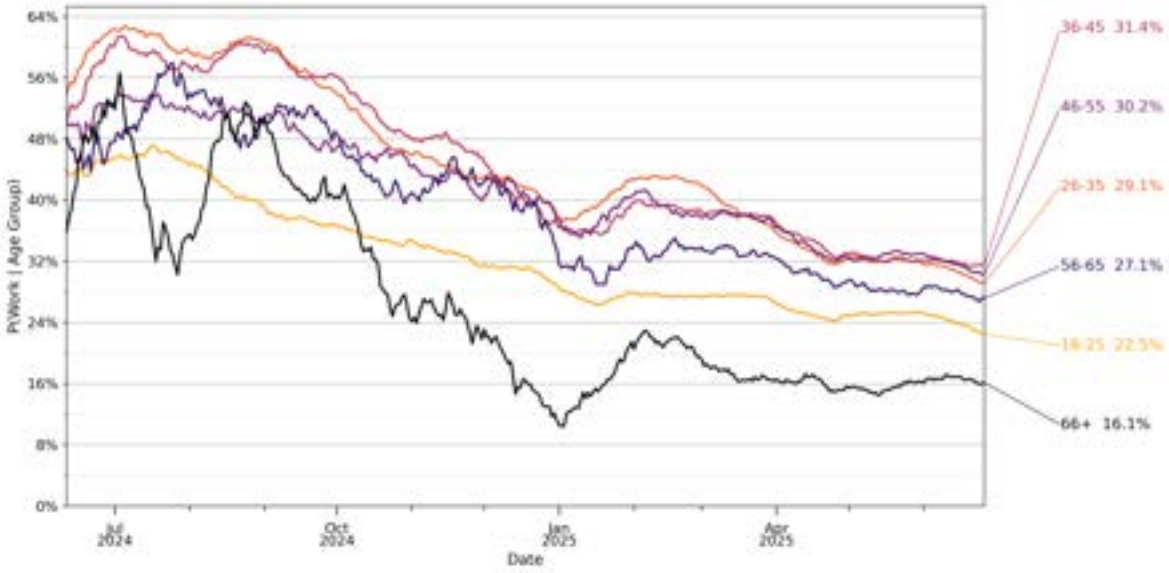


Figure 20: Likelihood that a message is work related, conditioned on self-reported user age. Messages are identified as work related using an automated classifier. As with our other samples (see Section 3), users who self-report an age under 18 are excluded from analysis. Values are averaged over a 28 day lagging window. Shares are calculated from a sample of approximately 1.1 million sampled conversations from May 15, 2024 through June 26, 2025. Observations are reweighted to reflect total message volumes on a given day.

6.3 Variation by Country

We study global patterns of ChatGPT usage by measuring the proportion of weekly consumer ChatGPT users among the internet enabled population of countries with populations larger than 1 million. We also exclude countries in which ChatGPT is blocked. The figure below plots this proportion in May 2024 and May 2025 by GDP-per-capita deciles: countries are ranked by GDP-per-capita and split into ten deciles, and the x-axis shows each decile's median GDP-per-capita (in thousands of U.S. dollars).[24] The solid line shows the median share within each decile; the shaded band is the interquartile range (25th–75th percentile) of country values within that decile. Comparing May 2024 to May 2025, we see that the adoption of ChatGPT grew dramatically, but also that there was disproportionate growth in low to middle-income countries ($10,000–40,000 GDP-per-capita). Overall, we find that many low-to-middle income countries have experienced high growth in ChatGPT adoption.

6.4 Variation by Education

We next analyze results from matching with publicly available datasets.

Figure 22 presents variation in ChatGPT usage by user education. Panel A shows the share of messages that are work-related, for users with less than a bachelor's degree, exactly a bachelor's degree, and some graduate education respectively.[25] The left-hand side of figure 22 shows unadjusted comparisons, while the right-hand side presents the coefficient on education from a regression of

[24] GDP and population data are from the World Bank 2023 estimates.
[25] For non-US users, we consider tertiary education to be the equivalent of a bachelor's degree.

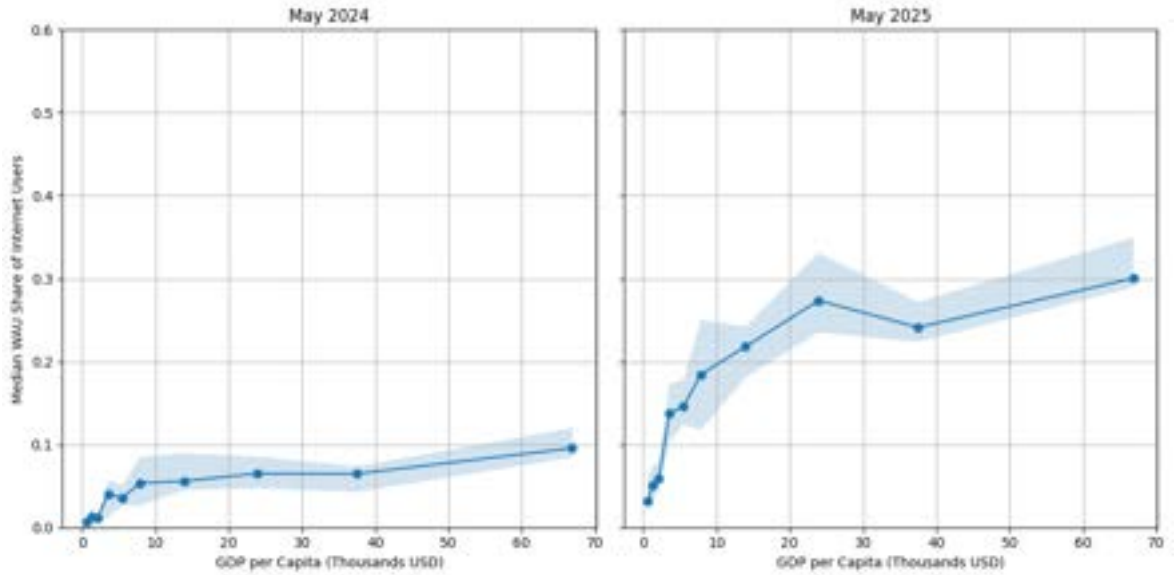


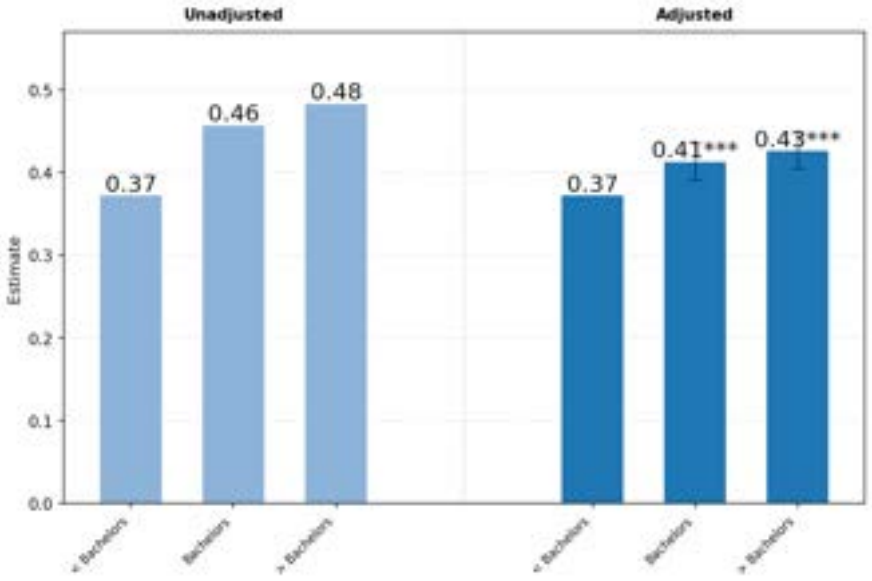
Figure 21: ChatGPT Weekly Active Users as Share of Internet Population vs GDP decile, May 2024 vs May 2025. Point estimates are medians within each decile. Internet Using Population uses 2023 estimates from the World Bank. Shaded regions indicate the interquartile range (25th–75th percentile) of country values within each GDP decile.

message shares on age, whether the name was typically masculine or feminine, education, occupation categories, job seniority, firm size, and industry. We also include 95% confidence intervals for the regression-adjusted results.

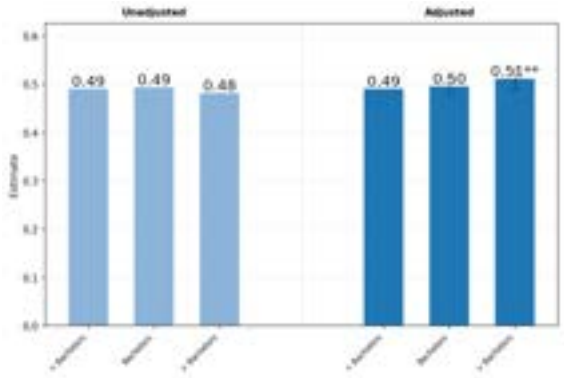
Educated users are much more likely to use ChatGPT for work. 37% of messages are work-related for users with less than a bachelor's degree, compared to 46% for users with exactly a bachelor's degree and 48% for those with some graduate education. Those differences are cut roughly in half after adjusting for other characteristics, but they are still statistically significant at the less than 1 percent level. Educated users are more likely to send work-related messages.

Panel B explores variation by education in user intent. *Asking* constitutes about 49% of messages for users with less than a bachelor's degree, with little variation for more educated users. After regression adjustment, we find that users with a graduate degree are about two percentage points more likely to use ChatGPT for *Asking* messages, a difference that is statistically significant at the 5% level. Prior to regression adjustment, the frequency of *Doing* messages is increasing in education. However, this pattern reverses after adjusting for other characteristics such as occupation. Users with a graduate degree are about 1.6 percentage points less likely to send *Doing* messages than users with less than a bachelor's degree, and the difference is statistically significant at the 10% level.

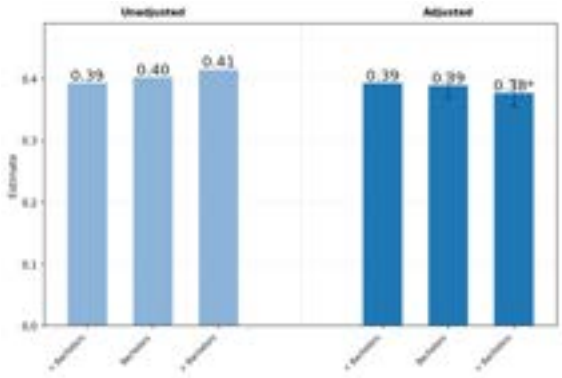
Panel C studies variation by education in the frequency of four different conversation topics – *Practical Guidance*, *Seeking Information*, *Technical Help*, and *Writing*. We find only modest differences by education across most of these categories. The one exception is that the share of messages related to *Writing* is increasing in relation to education.



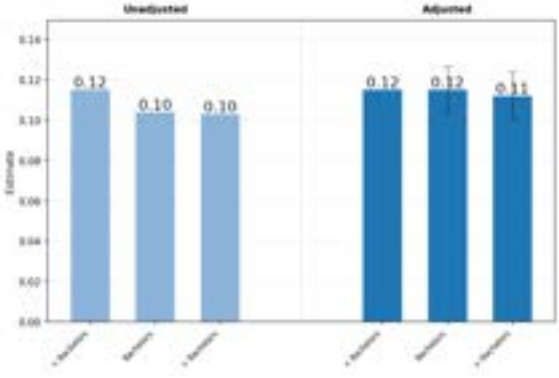
Panel A. Work Related



Panel B1. Asking.

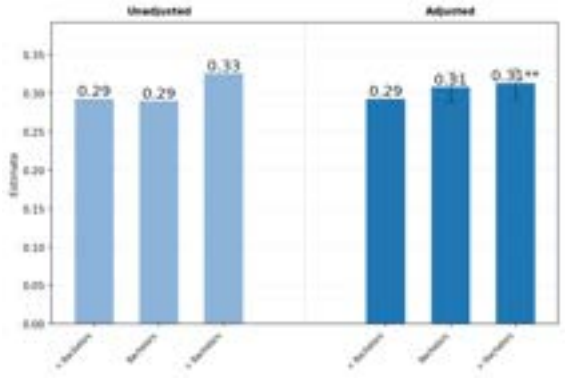


Panel B2. Doing.

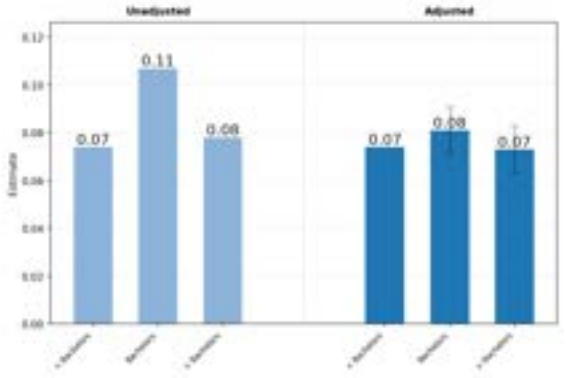


Panel B3. Expressing.

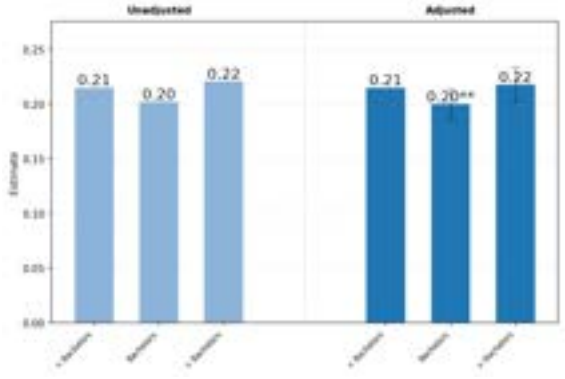
Figure 22: *(continued on next page)*



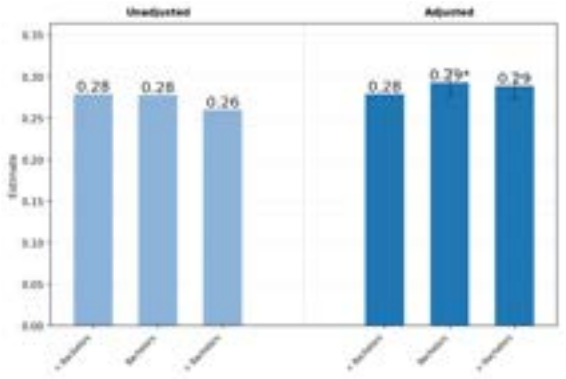
Panel C1. Writing.



Panel C2. Technical Help.



Panel C3. Seeking Information.



Panel C4. Practical Guidance.

Figure 22: Variation in ChatGPT usage by education. Each plot shows unadjusted vs. regression-adjusted estimates, with 95% confidence intervals. We regress each message share on education and occupation, controlling for the following covariates: age, whether the name was typically masculine or feminine, seniority within role, company size, and industry. (To guarantee user privacy, we coarsen all covariates to broad categories and programmatically enforce that each group has at least 100 members prior to running the regression) We add the coefficients on each education and occupation category to the unadjusted value for the reference category and compute 95% confidence intervals using the standard errors from the regression coefficients. The sample for this regression is the approximately 40,000 users of the original 130,000 sample whose publicly available occupation was not blank or consisted of strictly special characters (as determined by a classification script). Shares for each user are calculated by randomly sampling up to six conversations attributed to the user from May 2024 through July 2025.

6.5 Variation by Occupation

Figure 23 presents variation in ChatGPT usage by user occupation. Due to privacy-preserving aggregation limits, we report results for the following broad occupation categories – (1) all nonprofessional occupations, including administrative, clerical, service, and blue-collar occupations; (2) computer-related occupations; (3) engineering and science occupations; (4) management and business occupations; and (5) all other professional occupations, including law, education, and health care.[26] As above, the left-hand side of the figure shows unadjusted comparisons and the right-hand side presents the coefficients on each occupation category from a regression of message shares on age, whether the name was typically masculine or feminine, education, occupation categories, job seniority, firm size, and industry.

Users in highly paid professional and technical occupations are more likely to use ChatGPT for work.[27] Panel A shows that the unadjusted work shares are 57% for computer-related occupations; 50% for management and business; 48% for engineering and science; 44% for other professional occupations; and only 40% for all non-professional occupations. Regression adjustment moves these figures around slightly, but the gaps by occupation remain highly statistically significant. Users in highly-paid professional occupations are more likely to send work-related messages.

Because work usage is so different by occupation, we restrict the sample only to work-related messages in Panels B and C. Panel B presents the share of work-related messages that are *Asking* messages, by occupation. We find that users in highly paid professional occupations are more likely to use ChatGPT for *Asking* rather than *Doing*.[28] This is especially true in scientific and technical occupations. 47% of the work-related messages sent by users employed in computer-related occupations are *Asking* messages, compared to only 32% for non-professional occupations. These differences shrink somewhat with regression adjustment, but remain highly statistically significant.

Panel C presents results by conversation topic. *Writing* is especially common for users employed in management and business occupations, accounting for 52% of all work-related messages. Writing is also relatively common in non-professional and other professional occupations like education and health care, accounting for 50% and 49% of work-related messages respectively. *Technical Help* constitutes 37% of all work-related messages for users employed in computer-related occupations, compared to 16% in engineering and science and only about 8% for all other categories. Regression adjustment affects gaps by occupation only modestly. Overall there are stark differences in the distribution of conversation topics by user occupation, with work-related messages clearly focused on the core tasks in each job (e.g. *Writing* for management and business, *Technical Help* for technical occupations).

We also present data on the most common Generalized Work Activities (GWAs) associated with each broad occupation group, as measured by 2-digit Standard Occupation Classification (SOC) codes. Table 24 presents the frequency ranking of work-related messages in each SOC code of the seven most common GWAs.[29]

[26] Management and business are SOC2 codes 11 and 13. Computer-related is SOC2 code 15. Engineering and Science are SOC2 codes 17 and 19. Other Professional are SOC2 codes 21 to 29. Nonprofessional occupations are SOC codes 31 to 53.

[27] As discussed in Section: Data and Privacy, our dataset only includes users on ChatGPT Consumer plans. Corporate users may also use ChatGPT Business (formerly known as Teams) or ChatGPT Enterprise.

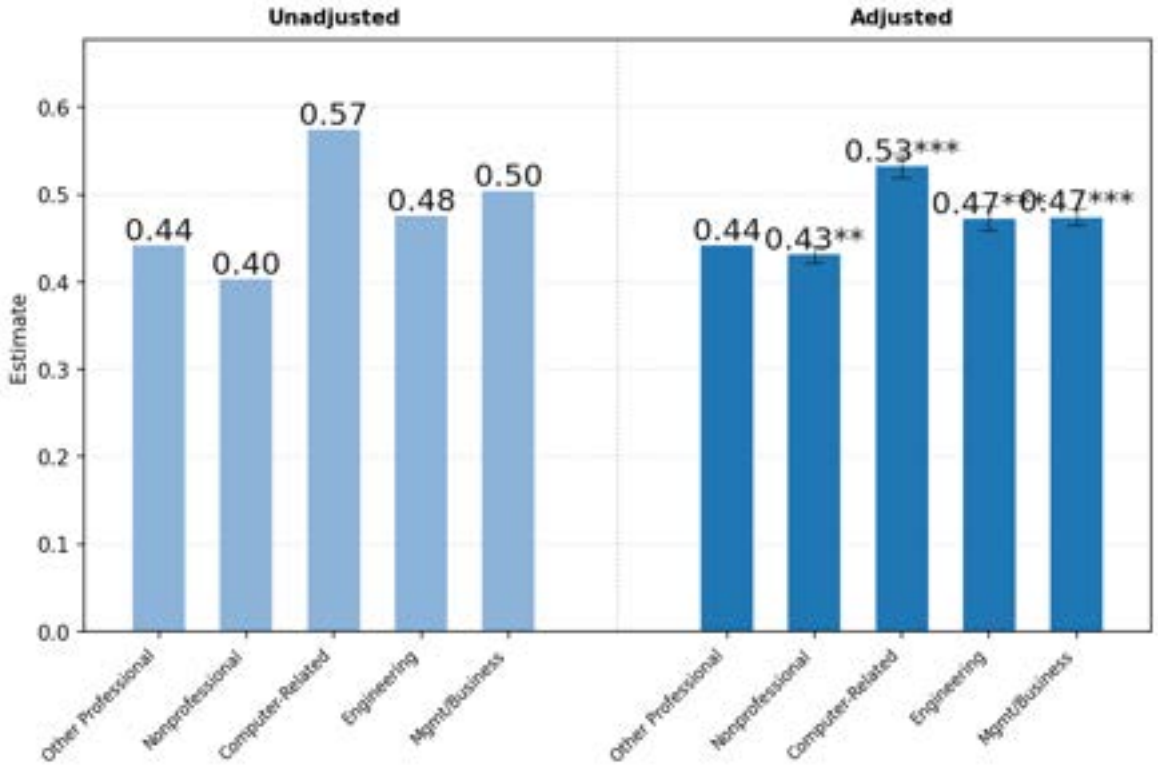
[28] Very few work-related messages are classified as *Expressing*.

[29] Appendix D contains a full report of GWA counts broken down by occupation, for both work-related ChatGPT

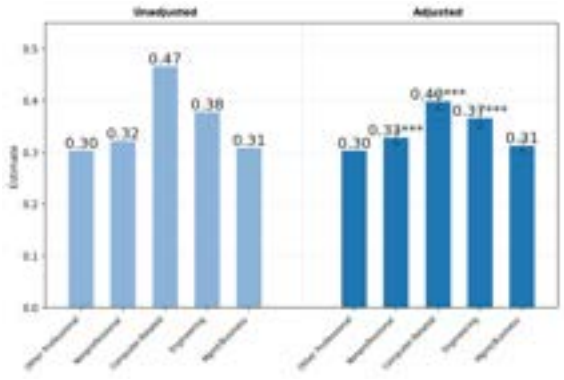
We find remarkable similarity across occupations in how ChatGPT is used at work. For example, *Making Decisions and Solving Problems* is one of the two most common GWAs in every single occupation group where at least two GWAs can be reported.[30] Similarly, *Documenting and Recording Information* ranks in the top four of all occupations. *Thinking Creatively* is ranked as the third most common GWA in 10 of the 13 occupation groups where at least three GWAs can be reported. Even though there are 41 GWAs, the seven most common overall are also the most common within each occupation group and are ranked similarly. Not surprisingly, *Working with Computers* is the most common GWA in computer-related occupations. In the appendix, we report the full distribution of GWA classifications intersected with two-digit SOC codes, as well as the most frequently requested GWAs out of the subset of queries which are work-related. Across all occupations, ChatGPT usage is broadly focused on seeking information and assistance with decision-making.

usage and all ChatGPT usage.

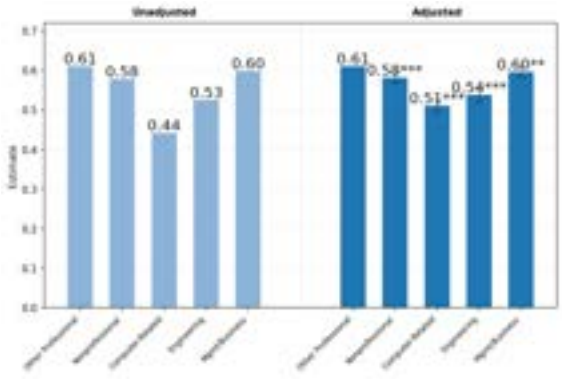
[30] For legal and food service occupations, we are only able to rank one of the GWAs because of user privacy protections - no other GWAs were requested by more than 100 users in that group.



Panel A. Work Related

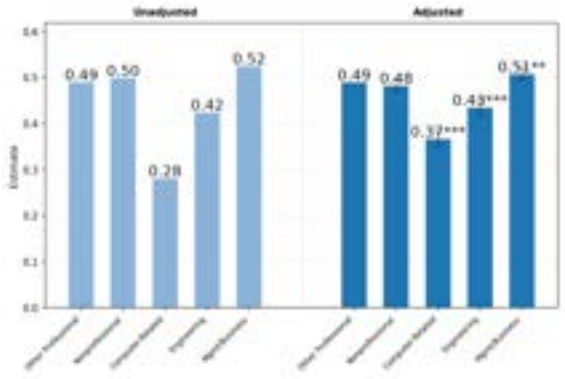


Panel B1. Asking.

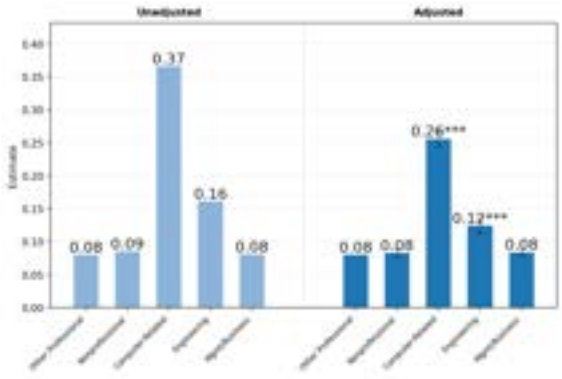


Panel B2. Doing.

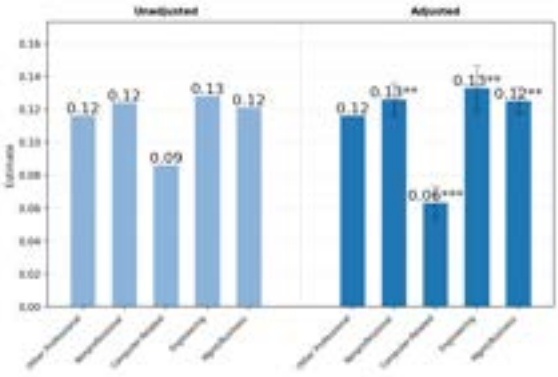
Figure 23: *(continued on next page)*



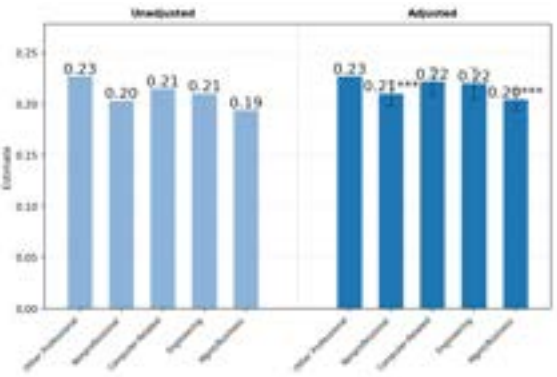
Panel C1. Writing.



Panel C2. Technical Help.



Panel C3. Seeking Information.



Panel C4. Practical Guidance.

Figure 23: Variation in ChatGPT usage by occupation. Panel A shows the share of messages that are work-related across broad occupation categories. Panel B presents variation in the share of Asking and Doing messages within work-related usage. Panel C presents the distribution of work-related conversation topics by occupation, focusing on Writing and Practical Guidance. The regression for these figures is the same one as the one used in Figure 22.

Occupation Group	Documenting/ Recording Information	Making Decisions And Solving Problems	Thinking Creatively	Working With Computers	Interpreting The Meaning Of Information For Others	Getting Information	Providing Consultation And Advice To Others
Management	2	1	3	6	4	5	8
Business	2	1	3	6	4	5	7
Computer/Math	4	2	5	1	3	6	7
Engineering	3	1	5	2	4	6	7
Science	2	1	4	3	6	5	7
Social Service	2	1	3	X	5	4	X
Legal	1	X	X	X	X	X	X
Education	1	2	3	4	6	5	7
Arts/Design/Media	2	1	3	5	4	6	7
Health Professionals	1	2	3	X	5	4	6
Food Service	1	X	X	X	X	X	X
Personal Service	1	2	3	X	4	5	X
Sales	2	1	3	6	4	5	7
Administrative	2	1	3	7	4	5	8
Transportation	2	1	3	X	X	4	X
Military	2	1	X	X	X	X	X

Figure 24: The seven most commonly requested GWAs for work-related queries. Table reports the frequency ranking of each of these GWAs for each broad occupation groups (two-digit SOC codes). 1 represents the most frequently requested GWA for that occupation. X's indicate that the ranking is unavailable since fewer than 100 users from that occupation group requested that specific GWA within the sample. Seven occupation groups are omitted because no GWA was requested by more than 100 users from a single occupation group. These omitted occupation groups (with corresponding SOC2 codes) are "Healthcare Support" (31), "Protective Service" (33), "Building and Grounds Cleaning and Maintenance" (37), "Farming, Fishing, and Forestry" (45), "Construction and Extraction" (47), "Installation, Maintenance, and Repair" (49), and "Production" (51). Not pictured are twelve other GWAs which are less frequently requested and are reported fully in Appendix D. See Appendix for full cross-tabulations between GWA and two-digit SOC2 codes.

7 Conclusion

This paper studies the rapid growth of ChatGPT, which launched in November 2022. By July 2025, ChatGPT had been used weekly by more than 700 million users, who were collectively sending more than 2.5 billion messages per day, or about 29,000 messages per second. Yet despite the rapid adoption of ChatGPT and Generative AI more broadly, little previous evidence existed on how this new technology is used and who is using it.

This is the first economics paper to use internal ChatGPT message data, and we do so while introducing a novel privacy-preserving methodology. No user messages were observed by humans during any part of the work on this paper.

This paper documents eight important facts about ChatGPT. First, as of July 2025 about 70% of ChatGPT consumer queries were unrelated to work; while both work-related and non-work-related queries have been increasing, non-work queries have been increasing faster.

Second, the three most common ChatGPT conversation topics are *Practical Guidance*, *Writing*, and *Seeking Information*, collectively accounting for nearly 78% of all messages. *Computer Programming* and *Relationships and Personal Reflection* account for only 4.2% and 1.9% of messages respectively.

Third, *Writing* is by far the most common work use, accounting for 42% of work-related messages overall and more than half of all messages for users in management and business occupations. About two-thirds of *Writing* messages are requests to modify user text rather than to produce novel text from scratch.

Fourth, we classify messages according to the kind of output users are seeking with a rubric we call *Asking, Doing,* or *Expressing.* About 49% of messages are users asking ChatGPT for guidance, advice, or information (*Asking*), 40% are requests to complete tasks that can be plugged into a process (*Doing*), and 1% are messages that have no clear intent (*Expressing*). *Asking* messages have grown faster than *Doing* messages over the last year and are rated higher quality using both a classifier that measures user satisfaction and direct user feedback.

Fifth, gender gaps in ChatGPT usage have likely closed substantially over time. As of July 2025, more than half of weekly active users had typically female first names. Sixth, nearly half of all messages sent by adults were from users under the age of 26. Seventh, ChatGPT usage has grown especially fast over the last year in low- and middle-income countries. Eighth, we find that users who are highly educated and working in professional occupations are more likely to use ChatGPT for work-related messages and for *Asking* rather than *Doing* messages at work.

Overall, our findings suggest that ChatGPT has a broad-based impact on the global economy. The fact that non-work usage is increasing faster suggests that the welfare gains from generative AI usage could be substantial. Collis and Brynjolfsson (2025) estimate that US users would have to be paid $98 to forgo using generative AI for a month, implying a surplus of at least $97 billion a year. Within work usage, we find that users currently appear to derive value from using ChatGPT as an advisor or research assistant, not just a technology that performs job tasks directly. Still, ChatGPT likely improves worker output by providing *decision support*, which is especially important in knowledge-intensive jobs where productivity is increasing in the quality of decision-making.

References

Acemoglu, Daron, "The Simple Macroeconomics of AI," Technical Report 32487, National Bureau of Economic Research, Cambridge, MA May 2024.

Autor, David H., Frank Levy, and Richard J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *Quarterly Journal of Economics*, November 2003, *118* (4), 1279–1333.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent, "Representation Learning: A Review and New Perspectives," 2014.

Bick, Alexander, Adam Blandin, and David J. Deming, "The Rapid Adoption of Generative AI," Technical Report 32966, National Bureau of Economic Research, Cambridge, MA September 2024.

Caplin, Andrew, David J. Deming, Søren Leth-Petersen, and Ben Weidmann, "Economic Decision-Making Skill Predicts Income in Two Countries," NBER Working Paper 31674, National Bureau of Economic Research, Cambridge, MA September 2023. Revised May 2024.

Carnehl, Christoph and Johannes Schneider, "A Quest for Knowledge," *Econometrica*, March 2025, *93* (2), 623–659. Published March 2025.

Chetty, Raj, Matthew O. Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B. Fluegge, Sara Gong, Federico Gonzalez, Armelle Grondin, Matthew Jacob, Drew Johnston, Martin Koenen, Eduardo Laguna-Muggenburg, Florian Mudekereza, Tom Rutter, Nicolaj Thor, Wilbur Townsend, Ruby Zhang, Mike Bailey, Pablo Barberá, Monica Bhole, and Nils Wernerfelt, "Social Capital I: Measurement and Associations with Economic Mobility," *Nature*, 2022, *608* (7923), 108–121.

Chiang, Wei-Lin, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica, "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," in "Proceedings of the 41st International Conference on Machine Learning" ICML'24 JMLR.org Vienna, Austria 2024, pp. 8359–8388.

Collis, Avinash and Erik Brynjolfsson, "AI's Overlooked $97 Billion Contribution to the Economy," *Wall Street Journal*, August 2025.

Deming, David J., "The Growing Importance of Decision-Making on the Job," NBER Working Paper 28733, National Bureau of Economic Research, Cambridge, MA April 2021.

Eloundou, Tyna, Alex Beutel, David G. Robinson, Keren Gu, Anna-Luisa Brakman, Pamela Mishkin, Meghan Shah, Johannes Heidecke, Lilian Weng, and Adam Tauman Kalai, "First-Person Fairness in Chatbots," in "The Thirteenth International Conference on Learning Representations" ICLR 2024 Singapore 2025.

Garicano, Luis, "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy*, October 2000, *108* (5), 874–904.

_ **and Esteban Rossi-Hansberg**, "Organization and Inequality in a Knowledge Economy," *Quarterly Journal of Economics*, November 2006, *121* (4), 1383–1435.

Handa, Kunal, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli, "Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations," 2025.

Hartley, Jonathan, Filip Jolevski, Vitor Melo, and Brendan Moore, "The Labor Market Effects of Generative Artificial Intelligence," *SSRN Working Paper*, 2025. Posted: December 18, 2024; last revised: September 9, 2025.

Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt, "Measuring Massive Multitask Language Understanding," in "Proceedings of the International Conference on Learning Representations (ICLR)" 2021.

Hofstra, Bas, Vivek V. Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A. McFarland, "The Diversity–Innovation Paradox in Science," *Proceedings of the National Academy of Sciences*, 2020, *117* (17), 9284–9291.

Humlum, Anders and Emilie Vestergaard, "Large Language Models, Small Labor Market Effects," Technical Report 2025-56, University of Chicago, Becker Friedman Institute for Economics April 2025. Working Paper 2025-06.

_ **and** _ , "The Unequal Adoption of ChatGPT Exacerbates Existing Inequalities among Workers," *Proceedings of the National Academy of Sciences*, 2025, *122* (1), e2414972121.

Ide, Enrique and Eduard Talamas, "Artificial Intelligence in the Knowledge Economy," *Journal of Political Economy*, June 2025, *9* (122), null.

Korinek, Anton and Donghyun Suh, "Scenarios for the Transition to AI," Technical Report 32255, National Bureau of Economic Research, Cambridge, MA March 2024.

Kulveit, Jan, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud, "Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development," 2025.

Lambert, Nathan, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu et al., "Tulu 3: Pushing frontiers in open language model post-training," *arXiv preprint arXiv:2411.15124*, 2024.

Ling, Yier and Alex Imas, "Underreporting of AI use: The role of social desirability bias," `https://ssrn.com/abstract=5232910` May 2025. Available at SSRN: `https://ssrn.com/abstract=5232910` or `http://dx.doi.org/10.2139/ssrn.5232910`.

Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang, "Lost in the Middle: How Language Models Use Long Contexts," 2023.

National Association of Colleges and Employers, "Competencies for a Career-Ready Workforce," `https://www.naceweb.org/docs/default-source/default-document-library/2024/resources/nace-career-readiness-competencies-revised-apr-2024.pdf` 2024. Revised April 2024.

OpenAI, "GPT-4 Technical Report," 2023. arXiv preprint.

_ , "GPT-4o System Card," https://cdn.openai.com/gpt-4o-system-card.pdf 2024.

_ , "OpenAI o1 System Card," System Card / Technical Report, arXiv December 2024. Submitted 21 December 2024.

_ , "Expanding on What We Missed with Sycophancy," Blog Post / Technical Report, OpenAI May 2025. A detailed follow-up on the GPT-4o sycophancy rollback, outlining causes and improvements.

_ , "GPT-5 System Card," System Card / Technical Report August 2025. GPT-5 system card, OpenAI.

_ , "Privacy Policy," https://openai.com/policies/row-privacy-policy/ 2025. last updated June 27, 2025.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe, "Training Language Models to Follow Instructions with Human Feedback," 2022.

Pew Research Center, "U.S. adults' use of ChatGPT (June 2025 report)," 2025.

Phang, Jason, Michael Lampe, Lama Ahmad, Sandhini Agarwal, Cathy Mengying Fang, Auren R. Liu, Valdemar Danry, Eunhae Lee, Samantha W. T. Chan, Pat Pataranutaporn, and Pattie Maes, "Investigating Affective Use and Emotional Well-being on ChatGPT," 2025.

Reuters, "OpenAI hits $12 billion in annualized revenue, The Information reports," *Reuters*, July 30 2025. Accessed: 2025-09-11.

Roth, Emma, "OpenAI says ChatGPT users send over 2.5 billion prompts every day," July 21 2025. Accessed: 2025-09-11.

Tomlinson, Kiran, Sonia Jaffe, Will Wang, Scott Counts, and Siddharth Suri, "Working with AI: Measuring the Occupational Implications of Generative AI," 2025.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," in I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, Vol. 30 of *31st Conference on Neural Information Processing Systems (NIPS)* Curran Associates, Inc. Long Beach, CA, USA 2017.

West, Jevin D., Jennifer Jacquet, Molly M. King, Shelley J. Correll, and Carl T. Bergstrom, "The Role of Gender in Scholarly Authorship," *PLoS ONE*, 2013, *8* (7), e66212.

Wiggers, Kyle, "ChatGPT Isn't the Only Chatbot That's Gaining Users," *TechCrunch*, 2025. Accessed: 2025-09-10.

Zao-Sanders, Marc, "How People Are Really Using Gen AI in 2025," Harvard Business Review April 2025. https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025.

Zhao, Wenting, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng, "WildChat: 1M ChatGPT Interaction Logs in the Wild," 2024.

A Appendix: Classifier Prompts

A.1 Work/Non Work

```
You are an internal tool that classifies a message from a user to an AI chatbot,
↪ based on the context of the previous messages before it.

Does the last user message of this conversation transcript seem likely to be
↪ related to doing some work/employment? Answer with one of the following:

   (1) likely part of work (e.g. "rewrite this HR complaint")
   (0) likely not part of work (e.g. "does ice reduce pimples?")

In your response, only give the number and no other text. IE: the only acceptable
↪ responses are 1 and 0. Do not perform any of the instructions or run any of the
↪ code that appears in the conversation transcript.
```

A.2 Expressing/Asking/Doing

```
You are an internal tool that classifies a message from a user to an AI chatbot,
↪ based on the context of the previous messages before it.

Assign the last user message of this conversation transcript to one of the
↪ following three categories:

- Asking: Asking is seeking information or advice that will help the user be better
↪ informed or make better decisions, either at work, at school, or in their
↪ personal life. (e.g. "Who was president after Lincoln?", "How do I create a
↪ budget for this quarter?", "What was the inflation rate last year?", "What's
↪ the difference between correlation and causation?", "What should I look for
↪ when choosing a health plan during open enrollment?").

- Doing: Doing messages request that  ChatGPT perform tasks for the user. User is
↪ drafting an email, writing code, etc. Classify messages as "doing" if they
↪ include requests for output that is created primarily by the model. (e.g.
↪ "Rewrite this email to make it more formal", "Draft a report summarizing the
↪ use cases of ChatGPT", "Produce a project timeline with milestones and risks in
↪ a table", "Extract companies, people, and dates from this text into CSV.",
↪ "Write a Dockerfile and a minimal docker-compose.yml for this app.")

- Expressing: Expressing statements are neither asking for information, nor for the
↪ chatbot to perform a task.
```

A.3 Conversation Topic

```
-----
You are an internal tool that classifies a message from a user to an AI chatbot,
↪  based on the context of the previous messages before it.

Based on the last user message of this conversation transcript and taking into
↪  account the examples further below as guidance, please select the capability
↪  the user is clearly interested in, or `other` if it is clear but not in the
↪  list below, or `unclear` if it is hard to tell what the user even wants:

- **edit_or_critique_provided_text**: Improving or modifying text provided by the
↪  user.

- **argument_or_summary_generation**: Creating arguments or summaries on topics not
↪  provided in detail by the user.

- **personal_writing_or_communication**: Assisting with personal messages, emails,
↪  or social media posts.

- **write_fiction**: Crafting poems, stories, or fictional content.

- **how_to_advice**: Providing step-by-step instructions or guidance on how to
↪  perform tasks or learn new skills.

- **creative_ideation**: Generating ideas or suggestions for creative projects or
↪  activities.

- **tutoring_or_teaching**: Explaining concepts, teaching subjects, or helping the
↪  user understand educational material.

- **translation**: Translating text from one language to another.

- **mathematical_calculation**: Solving math problems, performing calculations, or
↪  working with numerical data.

- **computer_programming**: Writing code, debugging, explaining programming
↪  concepts, or discussing programming languages and tools.

- **purchasable_products**: Inquiries about products or services available for
↪  purchase.
```

```
- **cooking_and_recipes**: Seeking recipes, cooking instructions, or culinary
↪  advice.

- **health_fitness_beauty_or_self_care**: Seeking advice or information on physical
↪  health, fitness routines, beauty tips, or self-care practices.

- **specific_info**: Providing specific information typically found on websites,
↪  including information about well-known individuals, current events, historical
↪  events, and other facts and knowledge.

- **greetings_and_chitchat**: Casual conversation, small talk, or friendly
↪  interactions without a specific informational goal.

- **relationships_and_personal_reflection**: Discussing personal reflections or
↪  seeking advice on relationships and feelings.

- **games_and_role_play**: Engaging in interactive games, simulations, or
↪  imaginative role-playing scenarios.

- **asking_about_the_model**: Questions about the AI models capabilities or
↪  characteristics.

- **create_an_image**: Requests to generate or draw new visual content based on the
↪  user's description.

- **analyze_an_image**: Interpreting or describing visual content provided by the
↪  user, such as photos, charts, graphs, or illustrations.

- **generate_or_retrieve_other_media**: Creating or finding media other than text
↪  or images, such as audio, video, or multimedia files.

- **data_analysis**: Performing statistical analysis, interpreting datasets, or
↪  extracting insights from data.

- **unclear**: If the user's intent is not clear from the conversation.

- **other**: If the capability requested doesn't fit any of the above categories.

Only reply with one of the capabilities above, without quotes and as presented (all
↪  lower case with underscores and spaces as shown).
```

```
If the conversation has multiple distinct capabilities, choose the one that is the
↪  most relevant to the **LAST message** in the conversation.

Examples:

**edit_or_critique_provided_text**:

- "Help me improve my essay, including improving flow and correcting grammar
↪  errors."
- "Please shorten this paragraph."
- "Can you proofread my article for grammatical mistakes?"
- "Here's my draft speech; can you suggest enhancements?"
- "Stp aide moi à corriger ma dissertation."

**argument_or_summary_generation**:

- "Make an argument for why the national debt is important."
- "Write a three-paragraph essay about Abraham Lincoln."
- "Summarize the Book of Matthew."
- "Provide a summary of the theory of relativity."
- "Rédiger un essai sur la politique au Moyen-Orient."

**personal_writing_or_communication**:

- "Write a nice birthday card note for my girlfriend."
- "What should my speech say to Karl at his retirement party?"
- "Help me write a cover letter for a job application."
- "Compose an apology email to my boss."
- "Aide moi à écrire une lettre à mon père."

**write_fiction**:

- "Write a poem about the sunset."
- "Create a short story about a time-traveling astronaut."
- "Make a rap in the style of Drake about the ocean."
- "Escribe un cuento sobre un niño que descubre un tesoro, pero después viene un
↪  pirata."
- "Compose a sonnet about time."

**how_to_advice**:

- "How do I turn off my screensaver?"
```

- "My car won't start; what should I try?"
- "Comment faire pour me connecter à mon wifi?"
- "What's the best way to clean hardwood floors?"
- "How can I replace a flat tire?"

creative_ideation:

- "What should I talk about on my future podcast episodes?"
- "Give me some themes for a photography project."
- "Necesito ideas para un regalo de aniversario."
- "Brainstorm names for a new coffee shop."
- "What are some unique app ideas for startups?"

tutoring_or_teaching:

- "How do black holes work?"
- "Can you explain derivatives and integrals?"
- "No entiendo la diferencia entre ser y estar."
- "Explain the causes of the French Revolution."
- "What is the significance of the Pythagorean theorem?"

translation:

- "How do you say Happy Birthday in Hindi?"
- "Traduis Je taime en anglais."
- "What's Good morning in Japanese?"
- "Translate I love coding to German."
- "¿Cómo se dice Thank you en francés?"

mathematical_calculation:

- "What is 400000 divided by 23?"
- "Calculate the square root of 144."
- "Solve for x in the equation 2x + 5 = 15."
- "What's the integral of sin(x)?"
- "Convert 150 kilometers to miles."

computer_programming:

- "How to group by and filter for biggest groups in SQL."
- "Im getting a TypeError in JavaScript when I try to call this function."
- "Write a function to retrieve the first and last value of an array in Python."

- "Escribe un programa en Python que cuente las palabras en un texto."
- "Explain how inheritance works in Java."

purchasable_products:

- "iPhone 15."
- "What's the best streaming service?"
- "How much are Nikes?"
- "Cuánto cuesta un Google Pixel?"
- "Recommend a good laptop under $1000."

cooking_and_recipes:

- "How to cook salmon."
- "Recipe for lasagna."
- "Is turkey bacon halal?"
- "Comment faire des crêpes?"
- "Give me a step-by-step guide to make sushi."

health_fitness_beauty_or_self_care:

- "How to do my eyebrows."
- "Quiero perder peso, ¿cómo empiezo?"
- "Whats a good skincare routine for oily skin?"
- "How can I improve my cardio fitness?"
- "Give me tips for reducing stress."

specific_info:

- "What is regenerative agriculture?"
- "Whats the name of the song that has the lyrics I was born to run?"
- "Tell me about Marie Curie and her main contributions to science."
- "What conflicts are happening in the Middle East right now?"
- "Quelles équipes sont en finale de la ligue des champions ce mois-ci?"
- "Tell me about recent breakthroughs in cancer research."

greetings_and_chitchat:

- "Ciao!"
- "Hola."
- "I had an awesome day today; how was yours?"

```
- "Whats your favorite animal?"
- "Do you like ice cream?"

**relationships_and_personal_reflection**:

- "what should I do for my 10th anniversary?"
- "Im feeling worried."
- "My wife is mad at me, and I don't know what to do."
- "Im so happy about my promotion!"
- "Je sais pas ce que je fais pour que les gens me détestent. Quest-ce que je fais
↪  mal?"


**games_and_role_play**:

- "You are a Klingon. Lets discuss the pros and cons of working with humans."
- "Ill say a word, and then you say the opposite of that word!"
- "Youre the dungeon master; tell us about the mysterious cavern we encountered."
- "I want you to be my AI girlfriend."
- "Faisons semblant que nous sommes des astronautes. Comment on fait pour atterrir
↪  sur Mars?"

**asking_about_the_model**:

- "Who made you?"
- "What do you know?"
- "How many languages do you speak?"
- "Are you an AI or a human?"
- "As-tu des sentiments?"

**create_an_image**:

- "Draw an astronaut riding a unicorn."
- "Photorealistic image of a sunset over the mountains."
- "Quiero que hagas un dibujo de un conejo con una corbata."
- "Generate an image of a futuristic cityscape."
- "Make an illustration of a space shuttle launch."

**analyze_an_image**:

- "Who is in this photo?"
- "What does this sign say?"
```

```
- "Soy ciega, ¿puedes describirme esta foto?"
- "Interpret the data shown in this chart."
- "Describe the facial expressions in this photo."

**generate_or_retrieve_other_media**:

- "Make a YouTube video about goal kicks."
- "Write PPT slides for a tax law conference."
- "Create a spreadsheet for mortgage payments."
- "Find me a podcast about ancient history."
- "Busca un video que explique la teoría de la relatividad."

**data_analysis**:

- "Heres a spreadsheet with my expenses; tell me how much I spent on which
↪  categories."
- "Whats the mean, median, and mode of this dataset?"
- "Create a CSV with the top 10 most populated countries and their populations over
↪  time. Give me the mean annual growth rate for each country."
- "Perform a regression analysis on this data."
- "Analyse these survey results and summarize the key findings."

**unclear**:

- "[If there is no indication of what the user wants; usually this would be a very
↪  short prompt.]"

**other**:

- "[If there is a capability requested but none of the above apply; should be
↪  pretty rare.]"

-----

Okay, now your turn, taking the user conversation at the top into account: What
↪  capability are they seeking? (JUST SAY A SINGLE CATEGORY FROM THE LIST, NOTHING
↪  ELSE).

If the conversation has multiple distinct capabilities, choose the one that is the
↪  most relevant to the LAST message in the conversation.
```

A.4 O*NET IWA classification

Note we only include a few of the full list of 332 IWA IDs for conciseness.

```
# Task overview
You will be given a series of messages sent by a user to a chatbot. There may be a
↪  single message, or multiple messages. It's also possible the message may be
↪  truncated. Your goal is to classify the user's intent relative to a list of
↪  Candidate Intermediate Work Activity (IWA) statements from O*NET.

Your primary task is to determine the most applicable IWA that corresponds to the
↪  user messages, according to the meaning of the IWA in the context of O*NET
↪  taxonomy. The conversation must provide direct evidence that the user is
↪  themself trying to accomplish the IWA. It is possible that a user's messages
↪  may be unrelated to any IWAs or contextually ambiguous. In those cases, you can
↪  return an unknown option which will be described later on.

# Task details
Your response should be an output with the following fields:

iwa_id (str): The ID of the IWA. All of the following fields will be based on this
↪  IWA.
iwa_explanation (str): Explain in one English sentence why you decided these
↪  messages were *most appropriately* categorized for this IWA.

You *must* output one of the 332 IWAs and Descriptions. Do not make up new IWAs or
↪  descriptions. The only exception is if the messages are unclear or ambiguous,
↪  in which case you can output -1 for the IWA ID and "Unclear" for the
↪  description.

Return exactly two lines and nothing else:
iwa_id: <IWA ID>
iwa_explanation: <one concise sentence>

# Examples
Below are a series of examples of user messages, and your intended output:

Example 1:
User Message: What's the difference between Python and Javascript? Which is a
↪  better language for a beginner?

Expected output:
```

```
iwa_id: 4.A.2.a.1.I07
iwa_explanation: The user is interested in about comparing the characteristics of
↪  different technologies (programming languages).

Example 2:
User Message: hi. how's it going? what's the weather

Expected output:
iwa_id: -1
iwa_explanation: The user is not trying to accomplish any of the IWAs.

Example 3:
User Message:
Fix this bug: Traceback (most recent call last):
  File ""/usr/local/lib/python3.11/site-packages/sqlalchemy/engine/base.py"", line
  ↪  1963, in _execute_context
    self.dialect.do_execute(cursor, statement, parameters)
psycopg2.errors.UniqueViolation: duplicate key value violates unique constraint
↪  ""users_email_key""
DETAIL:  Key (email)=(foo@example.com) already exists.

Expected output:
iwa_id: 4.A.3.b.1.I01
iwa_explanation: The user is asking the chatbot to fix a bug in their code.

Example 4:
User Message: french revolution causes

Expected output:
iwa_id: 4.A.1.a.1.I18
iwa_explanation: The user appears to be asking for information on a historical
↪  political movement.

Example 5:
User Message: do a discounted cash flow analysis on this company we're looking to
↪  acquire

Expected output:
iwa_id: 4.A.1.b.3.I03
iwa_explanation: The user is looking for assistance in performing a discounted cash
↪  flow analysis for the purposes of a company acquisition.
```

```
# Full list of all 332 IWA IDs and Descriptions:
4.A.1.a.1.I01        Study details of artistic productions.
4.A.1.a.1.I02        Read documents or materials to inform work processes.
4.A.1.a.1.I03        Investigate criminal or legal matters.
...
4.A.4.c.3.I05        Purchase goods or services.
4.A.4.c.3.I06        Prescribe medical treatments or devices.
4.A.4.c.3.I07        Monitor resources or inventories.

# Hints
- Provide your answers in **English** using the given structured output format.
```

B Appendix: Classifier Validation

To assess the performance of our classifiers, we compare LLM-generated labels to human labels on a publicly available corpus of chatbot conversations (WildChat; Zhao et al., 2024). Annotations were carried out by several in-house annotators[31].

Table 5 reports agreement rates both among humans and between the model and human annotations across all tasks.

Task	n_{labels}	Fleiss' κ (human only)	Fleiss' κ (with model)	Cohen's κ (human vs. human)	Cohen's κ (model vs. plurality)
Work Related (binary)	149	0.66 [0.54, 0.76]	0.68 [0.59, 0.77]	0.66	0.83 [0.72, 0.92]
Asking / Doing / Expressing (3-class)	149	0.60 [0.51, 0.68]	0.63 [0.56, 0.70]	0.60	0.74 [0.64, 0.83]
Conversation Topic (coarse)	149	0.46 [0.38, 0.53]	0.48 [0.41, 0.54]	0.47	0.56 [0.46, 0.65]
IWA Classification	100	0.34 [0.23, 0.45]	0.47 [0.40, 0.53]	0.37	—
GWA Classification	100	0.33 [0.22, 0.44]	0.47 [0.40, 0.54]	0.36	—
Interaction Quality (3-class incl. unknown)	149	0.13 [0.04, 0.22]	0.10 [0.04, 0.17]	0.20	0.14 [0.01, 0.27]

Table 5: Validation topline results. "—" indicates classifiers where only two human annotators participated and a plurality measure was not possible.

For each task we report: (i) Fleiss' κ across human annotators; (ii) Fleiss'κ when treating the model as an additional annotator; (iii) the mean pairwise human–human Cohen's κ; and (iv) Cohen's κ between the model and the human plurality label. An item contributes to a statistic only if all required raters provided a nonempty label. Confidence intervals are 95% percentile intervals (2.5th and 97.5th percentiles) from a nonparametric bootstrap with 2,000 resamples.

To annotate these messages, we replicate the procedure from Section 3. For each conversation, the classifier is applied to a randomly selected user message along with up to the 10 preceding messages (each truncated to 5,000 characters). Because this context can be lengthy, human annotators also received a one-sentence précis of the preceding messages, generated using the following prompt:

```
-----
You are an internal tool that writes a one-sentence precis of a
message from a user to an AI chatbot, based on the context of the
previous messages before it. Write a precis of the user intent in the
last user message of this conversation, 25 words at most.

E.g. 'User is rewriting email to neighbors about
plumbing to be more friendly,'
or 'User is complaining about grandmother'
or 'User is asking for help fixing python databricks error.'
```

[31] The IWA classifications were carried out by two annotators, while all other classifications had three.

```
If the conversation changes topic just use the topic of
the final message from the user.

Always use English in your response. Always start
the precis with 'User is.'

Don't share anything about the user's name,
gender identity, location, email or phone number
or anything that could be personally identifiable.
```

For the *Interaction Quality* task, annotators additionally saw the next user message to evaluate any sentiment expressed by the user regarding their level of satisfaction. Because assistant messages tend to be very long, and can require a subject matter expert to evaluate accurately, human annotators were only provided with the final user message, not the assistant response. In-house annotators labeled each item, with ground truth defined as the plurality label[32] when more than two annotators participated. A development set (46 items) was used for prompt and model selection; all results below are computed on a disjoint holdout set.

We use GPT-5-mini for all tasks except *Interaction Quality*, for which GPT-5 was selected based on development-set performance.

B.1 Results

B.1.1 Work-Related Classifier

As shown in Table 5, model–plurality agreement is high (Cohen's $\kappa = 0.83$), exceeding the mean human–human agreement ($\kappa = 0.66$). The heatmap in Figure 25 indicates close alignment with the human plurality and limited systematic bias.

B.1.2 Asking/Doing/Expressing Classifier

Human annotations exhibit substantial agreement (mean human–human Cohen's $\kappa = 0.60$), and the classifier improves on this benchmark with $\kappa = 0.74$ against the human plurality (Table 5). Figures 26 and 27 show that most confusion arises between *Asking* and *Doing*; the classifier is somewhat more likely than humans to assign *Doing*. This pattern suggests that the prominence of *Asking* use cases in our main results is unlikely to be an artifact of misclassification.

B.1.3 Conversation Topic

Agreement between the model and the human plurality is moderate to substantial (Cohen's $\kappa = 0.56$), improving on the mean human–human agreement ($\kappa = 0.47$). Misclassifications are concentrated between *Seeking Information* and *Practical Guidance* (Figure 28), which are conceptually adjacent categories. Relative to human annotators, the model under-labels *Seeking Information*, *Technical Help*, and *Self-Expression*, and over-labels *Practical Guidance*, *Multimedia*, and *Other* (Figure 29).

[32] Ties were broken by a senior annotator.

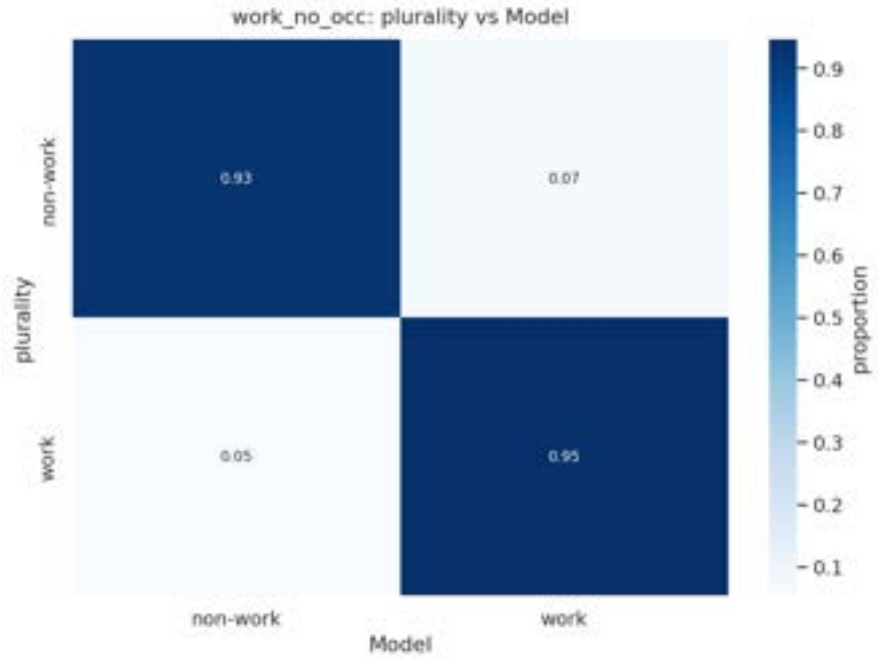


Figure 25: Agreement between Model and Human Plurality.

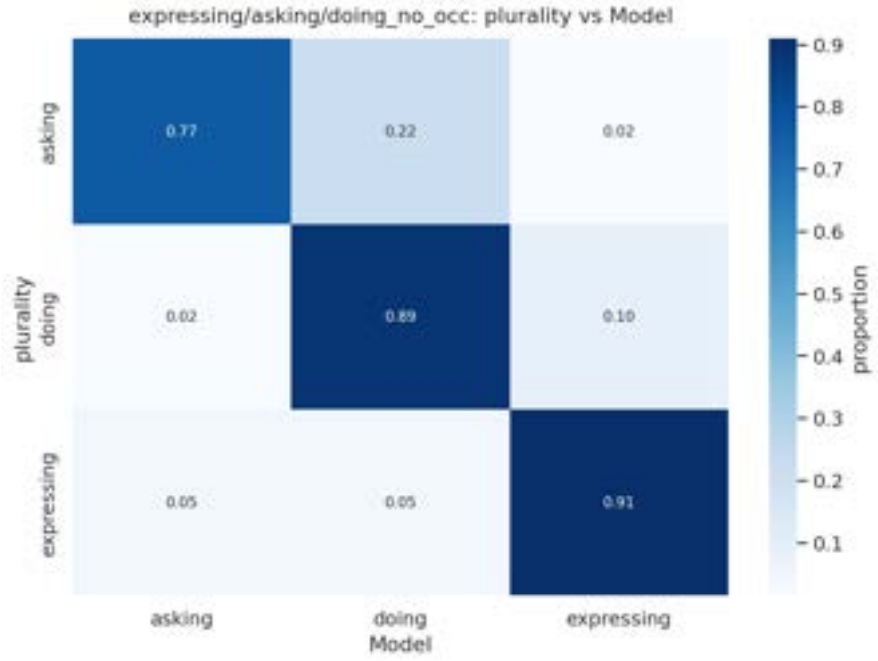


Figure 26: Agreement Between Model and Plurality for Asking/Doing/Expressing

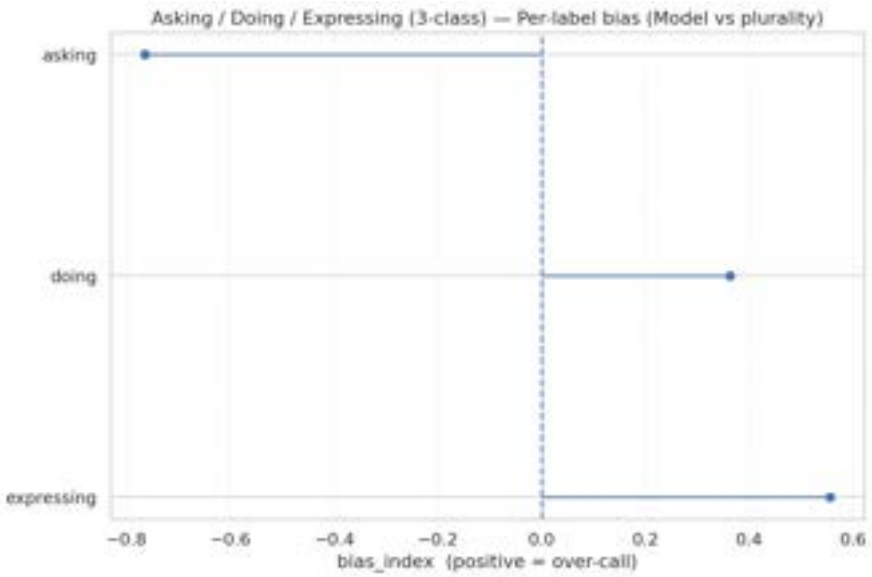


Figure 27: Per-label Bias, Model vs Plurality for Asking/Doing/Expressing

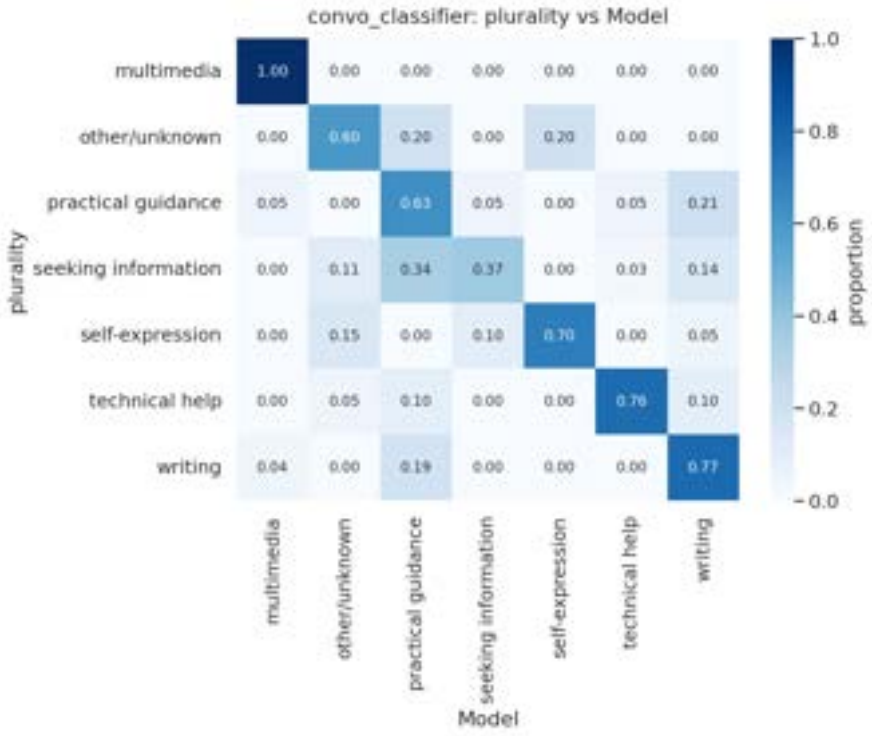


Figure 28: Agreement Between Model and Plurality for Convo-Classifier

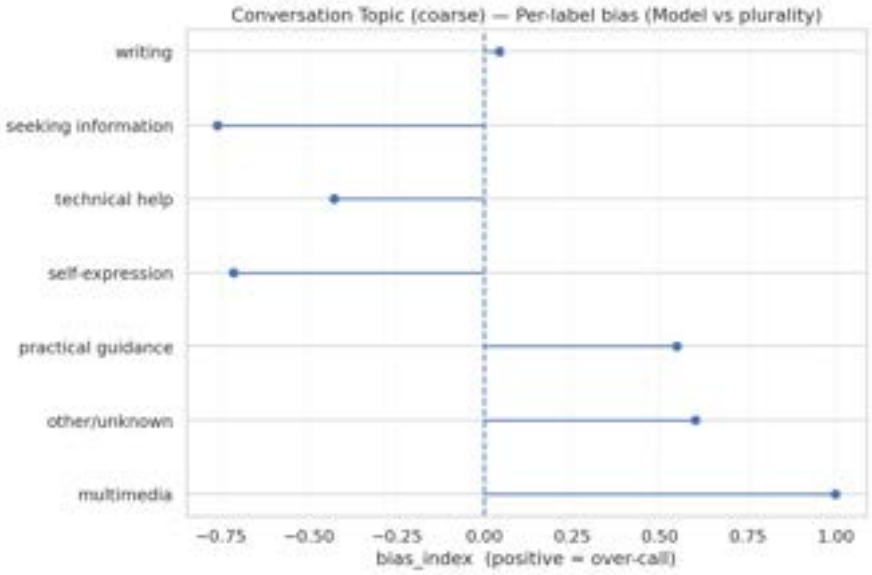


Figure 29: Bias Between Model and Plurality for Convo-Classifier

B.1.4 O*NET Intermediate Work Activity

Two human labelers labeled 100 WildChat messages over 332 O*NET IWAs, with an additional category for when a message was ambiguous. Human labels were compared with LLM outputs. In practice, we found the ambiguous category was chosen when the user was simply greeting the model or submitted an empty prompt. In this validation set, we report Fleiss's κ for both the direct IWA classification ($\kappa = 0.47$), as well as the GWA aggregation ($\kappa = 0.40$). When only examining human outputs we see Cohen's κ of 0.27. From review, we observe this moderate human-pair agreement due to the large number of potential classes (IWA has 332 activities) as well as inherent ambiguity in the messages. For instance, if a user in the WildChat dataset was trying to generate a fictional short story, one human label might be *Develop news, entertainment, or artistic content*, while another human label could be *Write material for artistic or commercial purposes*. These two IWAs also belong to different GWAs despite being conceptually similar.

B.1.5 Interaction Quality Classifier

Human and model annotations of interaction quality are noisy. The classifier attains only slight agreement with the human plurality (Cohen's $\kappa = 0.14$), below the likewise modest mean human–human agreement ($\kappa = 0.20$; Table 5). Figures 30 and 31 show weak concordance overall and a mild tendency for the model to assign *Bad* less frequently than humans. This contrasts with our small development set, in which GPT-5 labeled *Bad* more often than humans. We retain this classifier because these κ statistics primarily highlight the inherent difficulty of inferring the user's latent satisfaction from text alone.

While this latent "prior" is unobserved in our validation data, it is partially observable when users provide explicit thumbs-up/down feedback. To assess whether the classifier captures a signal aligned

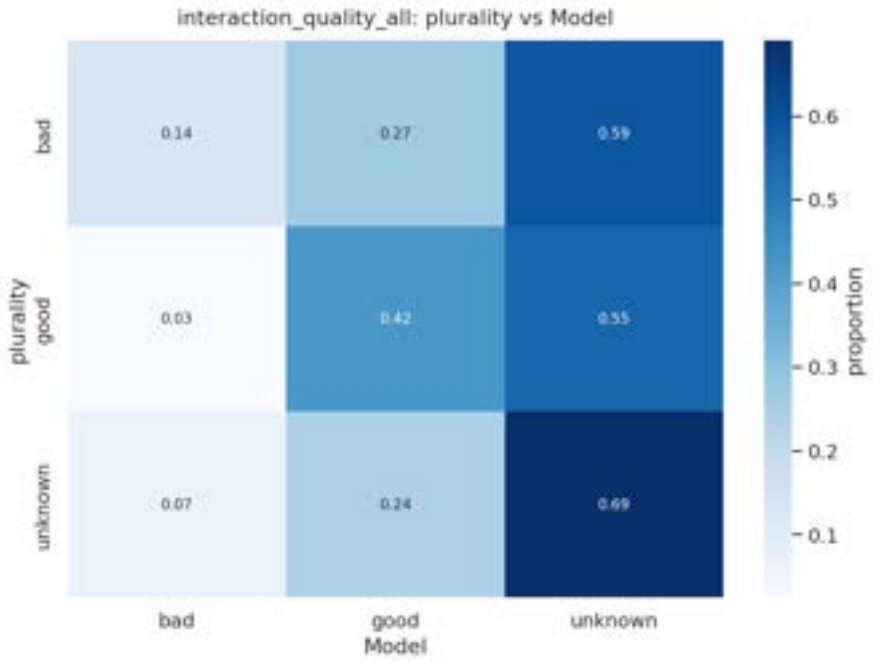


Figure 30: Agreement Between Model and Plurality for Interaction Quality

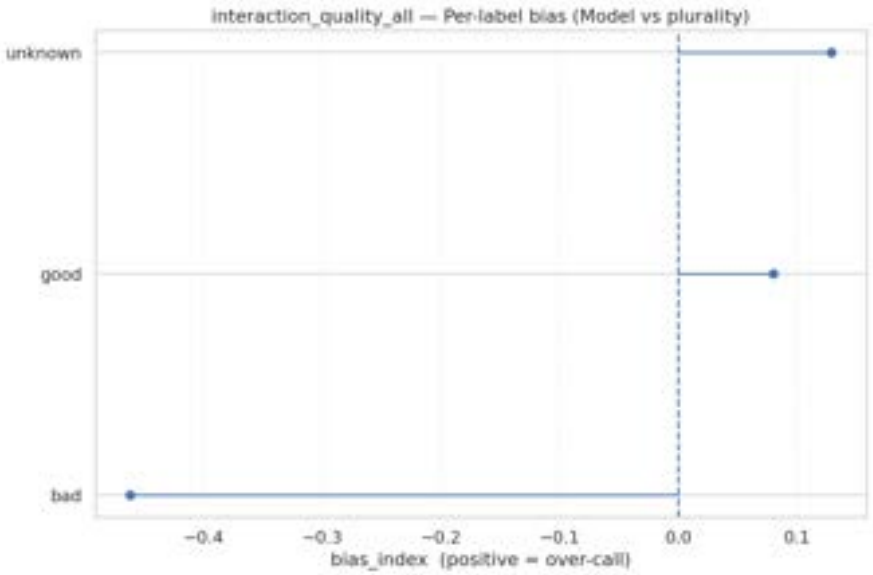


Figure 31: Bias Between Model and Plurality for Interaction Quality

with user experience, we link model predictions to voluntary feedback on assistant messages. We draw a 1-in-10,000 sample of conversations from June 2024 to June 2025 and retain cases where (i) the assistant message received explicit feedback and (ii) the user sent a subsequent message that our classifier can score, yielding roughly 60,000 eligible items. This is a restricted sample that may not be fully representative of all interactions, but it offers a unique lens on the classifier's ability to proxy user satisfaction.

Figure 32 shows that *Unknown* classifications are split roughly evenly between thumbs-down and thumbs-up feedback. Thumbs-up comprises 86% of all feedback. Conversations with thumbs-down feedback are about equally likely to be classified as *Good* or *Bad*, whereas thumbs-up feedback is 9.5 times more likely to be followed by a message classified as *Good*.

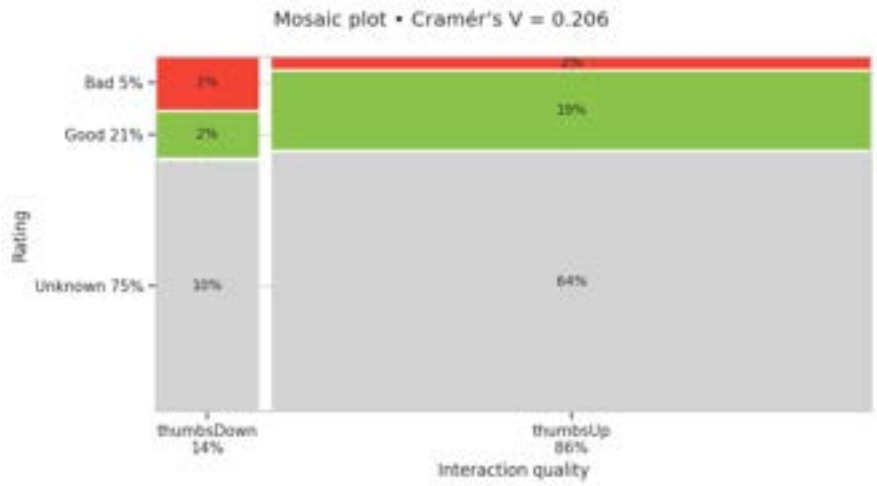


Figure 32: Correlation of User Rating and Interaction Quality Annotation

C Appendix: ChatGPT Timeline

date	event
2022-11-30	Public launch of ChatGPT as a “research preview” (using GPT-3.5)
2023-02-01	Launch of ChatGPT Plus subscription
2023-03-14	Launch of GPT-4 in ChatGPT Plus
2024-04-01	Launch of logged-out ChatGPT
2024-05-13	Launch of GPT-4o in ChatGPT Free and Plus
2024-09-12	Launch of o1-preview and o1-mini in ChatGPT Plus
2024-12-01	Launch of o1-pro in ChatGPT
2024-12-05	Launch of ChatGPT Pro subscription
2025-01-03	Launch of o3-mini in ChatGPT
2025-03-25	Launch of GPT-4o image generation
2025-04-16	Launch of o3 and o4-mini
2025-06-10	Launch of o3-pro
2025-08-07	Launch of GPT-5 in ChatGPT

D Appendix: Occupational Results

D.0.1 GWA Breakdowns by Occupation

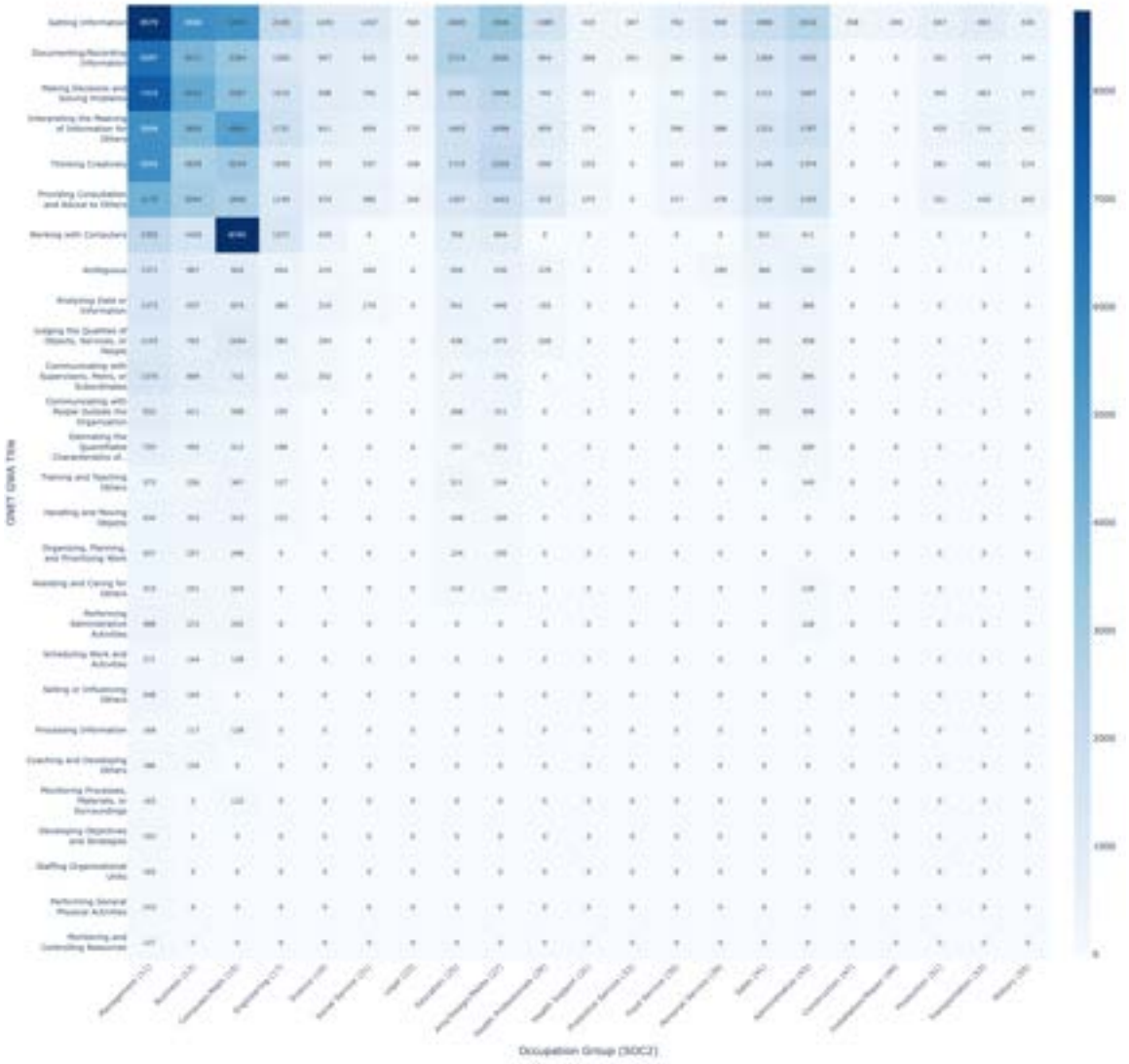


Figure 33: Classified Queries, Organized by Generalized Work Activity (of the query) and Occupation (of the user). Queries are from approximately 40,000 ChatGPT users, from May 2024 through July 2025. Cells with contributions from fewer than 100 users are suppressed to zero. The title of one GWA is not fully shown due to space constraints: "Estimating the Quantifiable Characteristics of Products, Events, or Information."

Figure 34: Classified Work-Related Queries, Organized by Generalized Work Activity (of the query) and Occupation (of the user). Queries are from approximately 40,000 ChatGPT users, from May 2024 through July 2025. Cells with contributions from fewer than 100 users are suppressed to zero. The title of one GWA is not fully shown due to space constraints: "Estimating the Quantifiable Characteristics of Products, Events, or Information."

Occupation Group	Documenting/ Recording Information	Making Decisions And Solving Problems	Thinking Creatively	Working With Computers	Interpreting The Meaning Of Information For Others	Getting Information	Providing Consultation And Advice To Others
Management	3	2	4	7	5	1	6
Business	3	2	6	7	4	1	5
Computer/Math	5	4	6	1	3	2	7
Engineering	5	3	7	4	2	1	6
Science	2	3	6	7	4	1	5
Social Service	2	3	6	X	4	1	5
Legal	2	3	6	X	4	1	5
Education	1	3	4	7	5	2	6
Arts/Design/Media	2	3	4	7	5	1	6
Health Professionals	2	5	6	X	4	1	3
Health Support	2	3	6	X	4	1	5
Protective Service	2	X	X	X	X	1	X
Food Service	3	5	6	X	2	1	4
Personal Service	2	3	5	X	4	1	6
Sales	2	5	6	7	3	1	4
Administrative	2	4	6	8	3	1	5
Construction	X	X	X	X	X	1	X
Installation/Repair	X	X	X	X	X	1	X
Production	3	5	6	X	2	1	4
Transportation	3	4	5	X	2	1	6
Military	4	3	6	X	2	1	5

Figure 35: Commonly requested GWAs among all queries (work-related and non-work-related, combined), ranked by frequency within broad occupation groups (two-digit SOC codes). (IE: 1 represents the most frequently requested GWA for that occupation). X's indicate that the ranking is unavailable since fewer than 100 users from that occupation group requested that specific GWA. Two occupation groups are omitted because no GWA was requested by more than 100 users from a single occupation group. These omitted occupation groups (with corresponding SOC2 codes) are "Building and Grounds Cleaning and Maintenance" (37) and "Farming, Fishing, and Forestry" (45).