How People Use ChatGPT*

Aaron Chatterji[1,2] Tom Cunningham[4] David Deming[3] Zoë Hitzig[1,3]
Christopher Ong[1,3] Carl Shan[1] Kevin Wadman[1]

[1]OpenAI
[2]Duke University
[3]Harvard University
[4]Model Evaluation and Threat Research

March 2026

Abstract

Three years after its launch, ChatGPT was used at least weekly by ten percent of the global adult population. This paper documents several important facts about how people use ChatGPT using a large random sample of ChatGPT messages linked to demographic data via an automated privacy-preserving process. ChatGPT usage was initially higher among men and users from high-income countries, but those gaps had closed substantially as of September 2025. Although non-work usage has grown relatively faster, users on consumer plans still send hundreds of millions of work-related messages each day. Work-related ChatGPT usage is highly concentrated in knowledge-seeking activities like decision support and analyzing and interpreting information, regardless of user occupation. We develop a principal-agent model of AI chatbot usage that is motivated by evidence from people using ChatGPT both to ask for advice and to delegate tasks directly. The model highlights the central role of users' demand for customization due to idiosyncratic preferences, which we call *context*. Using a novel classifier that measures conversational context, we find that users are more likely to ask AI chatbots for advice rather than work output when their queries have higher context requirements. Since context must be transmitted by the user, it may be a key bottleneck to unlocking economic value from AI.

*We acknowledge help and comments from Joshua Achiam, Hemanth Asirvatham, Ryan Beiermeister, Rachel Brown, Justin Bleuel, Cassandra Duchan Solis, Jason Kwon, Elliott Mokski, Kevin Rao, Harrison Satcher, Gawesha Weeratunga, Hannah Wong, seminar participants at NBER Labor Studies and the Analytics & Insights team. We especially thank Tyna Eloundou and Pamela Mishkin who laid the foundation for this work. This study was approved by Harvard IRB (IRB25-0983).

1 Introduction

ChatGPT launched in November 2022. By July 2025, 18 billion messages were being sent each week by 700 million users, about 10% of the global adult population. For a new technology, this speed of diffusion has no precedent (Bick et al., 2024). This paper studies consumer usage of ChatGPT, the first mass-market Large Language Model (LLM) chatbot.[1] The growth in LLM abilities and adoption has intensified interest in the impact of artificial intelligence on economic growth (Acemoglu, 2024; Korinek and Suh, 2024); employment (Eloundou et al., 2024); and society (Kulveit et al., 2025). However, despite rapid adoption of LLMs, there is limited public information on how they are used.[2]

We analyze a large random sample of ChatGPT messages and conversations from consumer plans (Free, Plus, Go, and Pro) between May 2024 and December 2025 using automated classifiers that categorize messages according to features like conversation topic, user intent, and job task.[3] This allows us to understand how people use ChatGPT without ever directly observing user messages.[4]

We first document several important facts about how people around the world use ChatGPT. As of December 2025 ChatGPT users were sending more than 3 billion messages per day, with a rapid growth trajectory since early 2025 that was driven by a mix of adding new users and increased intensity among existing users. While ChatGPT use was initially higher among users with male names and in high income countries, those gaps closed substantially between 2024 and 2025. Nearly 80 percent of ChatGPT messages fall into three conversation topics which we call *Practical Guidance*, *Seeking Information*, and *Writing*, while only about 4 percent of ChatGPT messages are related to technical help with activities like computer programming.[5]

Turning to work usage, we find that ChatGPT users on consumer plans send about 800 million work-related messages per day. However, non-work usage has grown relatively faster and comprised 72 percent of all ChatGPT messages as of late 2025. We also map ChatGPT messages to job tasks from the Occupational Information Network (O*NET) and find that work usage across all user occupations is highly concentrated in knowledge-seeking activities such as decision support and recording, analyzing, and interpreting information.[6]

Overall, we find that ChatGPT users are more likely to ask ChatGPT for information or advice to inform a decision rather than delegating tasks to AI directly (Deming, 2021; Caplin et al., 2023). This corresponds to problem-solving models of knowledge work (Garicano, 2000; Garicano and Rossi-

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

[2] 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.

[3] Our sample does not include data from business or enterprise users of ChatGPT, from API usage, or any other OpenAI product.

[4] Appendix B contains the actual text of our classifier prompts, and Appendix C includes more details about our validation procedure. The classification pipeline is protected by a series of privacy measures, detailed in Section 3.

[5] *Practical Guidance* includes activities like tutoring, how-to advice, and creative ideation that are highly customized to the user and can be adapted based on conversation follow-up, while *Seeking Information* 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*). More than half of all *Writing* messages ask ChatGPT to modify user text (editing, critiquing, translating, etc.) rather than creating new text from scratch.

[6] Handa et al. (2025) report that 37% of conversations are mapped to a "computer and mathematical" occupation category, and 30% or more of all imputed tasks are programming or IT-related. We believe the discrepancy between their results and ours is partly due to the difference in types of users between Claude and ChatGPT, and because Handa et al. (2025) only includes queries that can be mapped directly to job tasks.

Hansberg, 2006; Carnehl and Schneider, 2025; Ide and Talamas, 2025) and presents a challenge for economic models of AI that emphasize task automation (Acemoglu and Restrepo, 2022; Autor and Thompson, 2025).[7]

To understand the economic implications of ChatGPT usage, we develop a principal-agent model where users can delegate a job task to an AI chatbot, ask the chatbot for advice, or perform the task themselves.[8] A key determinant of user behavior is *context*, meaning how much detail the chatbot needs to produce output that is adequately customized to the user's needs. In our model, the ideal action depends on context that is observed by the principal but not by the chatbot agent. If the principal delegates, the agent produces "generic" output. For example, if the task is writing an email, the agent will produce a well-written draft in a standard style that reads like the average of all relevant emails in the chatbot's training data. This creates a *context gap* that is larger when the principal has idiosyncratic preferences or is seeking highly customized output.[9]

We test the implications of the model by studying user behavior across job tasks and over time. Using task-level measures of AI capabilities developed by Eloundou et al. (2024), we find that delegation is more likely when the AI chatbot is relatively more skilled than the human user. We also measure context directly using a conversation-level classifier that judges how much specific context previously unknown to the chatbot is required to adequately perform the user's task.[10] Controlling for user-by-task fixed effects, we find that users are more likely to ask AI chatbots for advice rather than work output when their queries have higher context requirements. We also find that ChatGPT conversations are becoming higher context over time, suggesting a shift into greater conversational complexity.

The growth of high-context conversations and increasing reliance on ChatGPT for advice rather than work output has important economic implications. Context must be transmitted by the user - even a superintelligent chatbot cannot read a person's mind. Thus context may be a key bottleneck for unlocking economic value through the automation of job tasks. Our findings may explain, for example, the growing popularity of LLMs that are embedded directly into company workflows so they can better observe context (Appel et al., 2025; Dillon et al., 2025).[11]

Our results are most consistent with Jahani et al. (2024a), who study how users dynamically adapt their prompts to model capabilities. They find that richer, more context-specific prompts improve model output in an image generation task, and also that automated prompt improvement is a poor substitute for human adaptation because the model lacks context. Lin (2025) finds that people prompt adaptively in response to model output, and that users respond to model improvements by providing

[7] Early evidence from the adoption of agentic AI also finds that planning, asking for explanations, and research are common use cases for AI agents (Sarkar (2025); Yang et al. (2025)).

[8] Our model is most similar to Ide and Talamas (2025), 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.

[9] A related paper is Gans (2026), who models jagged artificial intelligence as a function of the relatedness between the knowledge in a chatbot's training data.

[10] Specifically, we ask the model to infer the underlying real-world task that user is working on (e.g. draft a memo, analyze data, etc.) and then judge how much context is required. Examples of low context are common requests like summarize/rephrase, produce code snippets, or simple question and answer patterns. Examples of high context are internal company practices and norms, private or proprietary data analysis, or personal financial situations. See Appendix B for the actual text of the prompt, and Appendix C for details on our validation procedure and results.

[11] The growing value of context may also help explain the rapid growth of job listings for forward-deployed software engineers who are directly embedded in companies (Financial Times 2026 - https://www.ft.com/content/91002071-7874-4cb7-9245-08ca0571c408).

more context. Vendraminelli et al. (2025) find that LLMs help users close expertise gaps but only when they have "nearby" occupational expertise, which suggests that context is important for optimizing AI use at work.

Our work is also related to two recent papers that report statistics on chatbot conversations from Claude and Microsoft Copilot, 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 ChatGPT users is far larger, meaning our data may be a closer approximation to the average chatbot user.[12] Second, we report the diffusion of chatbot use across populations and the growth of different types of usage within cohorts. Third, we develop a simple economic framework to guide our interpretation of how people use ChatGPT. Fourth, we contribute to a growing literature on privacy preservation by developing a replicable protocol for secure analysis of user data and message content without observing either directly.

2 What is ChatGPT?

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 quality is measured in a variety of ways. 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.

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.[13] An LLM can be thought of as a function that maps a string of words to a probability distribution over the set of all possible words (more precisely, "tokens," which are word chunks). The functions are implemented with deep neural nets, typically with a transformer architecture (Vaswani et al., 2017), parameterized with billions of model "weights".

The weights in an LLM-based chatbot are trained in two stages, commonly called "pre-training" and "post-training". In the first stage ("pre-training"), the LLMs are trained to use an enormous corpus of text to predict the next word in a string, given the preceding words. 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 (2024), 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,

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

[13] 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). For a timeline of model launches, see Appendix D. We will refer to all of ChatGPT's models as language models, though most can additionally process tokens representing images, audio, or other media.

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 with human preferences between alternative responses to the same prompt (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, 2025b).

3.1 Growth Dataset

This dataset covers all usage on consumer (e.g. non-enterprise, non-education) ChatGPT Plans (Free, Go, Plus, Pro) since ChatGPT's launch in November 2022 through December 2025. This includes the total number of messages sent by each user on each day as well as de-identified metadata such as the timestamp of the user's 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

We randomly sampled approximately 2 million conversations[14] from logged-in ChatGPT users between May 2024 and December 2025.[15] We first randomly select a conversation between a user and ChatGPT, and then randomly select one user message from that conversation to annotate with automated classifiers (described in the next paragraph).[16] We applied several sampling restrictions to exclude conversations that were deleted or from users who opted out of sharing their messages for model training, users who self-report an age under 18, and users who have been deactivated or banned.[17]

[14] A conversation is a series of messages between the user and the chatbot that occur in sequence within a single context window.

[15] ChatGPT became available to logged-out users in April 2024. 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, who represent a minority share of ChatGPT users over the sample period.

[16] We use conversation-level sampling in the main text because it better reflects distinct user activities. Corresponding estimates from a sample constructed using message-level (rather than conversation-level) sampling—which draws randomly from all eligible messages—are reported in Appendix Section A.1; figure captions indicate where these robustness estimates are shown.

[17] Our sample is drawn from an internal OpenAI dataset that is itself sampled, with the sampling rate varying over time. Additionally, to simplify the process of extending the sample, we collect approximately 100,000 messages from each

We also construct a representative sample of approximately 130,000 ChatGPT *users*, again subject to the restrictions above. The sample contains about 1.58 million messages from these users, sampled at the conversation level. We apply the same LLM-based classifiers to all of these 1.58 million messages. The classifiers are introduced in more detail in Section 6, their exact text is reproduced in Appendix B, and our validation procedure is described in Appendix C.

Privacy via Automated Classifiers. We analyzed message content via automated LLM-based classifiers that we ran on de-identified and PII-scrubbed message data (see Appendix Figure A.1 for details).[18] Our classifications aim to discern the intent of a given message, and thus we include the prior 10 messages in a conversation as context. For example, a particular user message from a user to a chatbot might say "10 more", but the prior message was "give me 3 cultural activities to do with teens."[19]

We validated the model classification decisions against human-judged classifications of a sample of conversations from the publicly available WildChat dataset (Zhao et al., 2024), which users affirmatively assented to share publicly for research purposes.[20] Appendix C provides detail on our validation approach and model performance relative to human the human baseline.

3.3 Linked Dataset

We also match the classified message sample of 130,000 ChatGPT users to publicly available data on user characteristics such as industry, occupation, seniority level, company size, and educational attainment. These data were collected by a third party vendor and analyzed inside a secure Data Clean Room (DCR).

We never directly accessed the linked user-level records. All analysis occurred inside the DCR and was executed through pre-approved queries across independently held datasets, so that neither party could 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 Appendix Figure A.2). Our partner only approved code that returned cells meeting a threshold of 100 users, and any cells returning with a sample size of less than 100 were automatically suppressed.

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.

month. To produce counts and shares over time, we scale by daily aggregate message totals. In the main conversation-level series, this scaling uses raw daily message counts by plan; in the appendix message-level series, the daily totals first apply the same logged-in and eligibility filters as the analysis sample to reduce noise.

[18] The messages are first scrubbed of PII using an internal LLM-based tool and then classified 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).

[19] 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, using the prompts listed in Appendix B, with the exception of the interaction quality classifier, which used "gpt-5.2" to achieve better agreement with human labels.

[20] The dataset was collected from a third party chatbot using OpenAI's LLMs via their API. For additional transparency, we classify a sample of 100,000 public WildChat messages and provide those data in this paper's replication package.

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.[21]

3.4 Differential Privacy

In addition to the previously-described privacy protections, we add noise to several of the datasets we analyze to ensure that they are differentially private. This technique provides a mathematical guarantee that it is sufficiently difficult to use the aggregate statistics to discern information about any of the individual components included in the aggregate. The procedures we apply follow the definition of (ϵ, δ) differential privacy outlined in Dwork and Roth (2014).[22] The figures and tables produced using this data are based on separate differentially-private versions of each dataset.

4 The Growth of ChatGPT

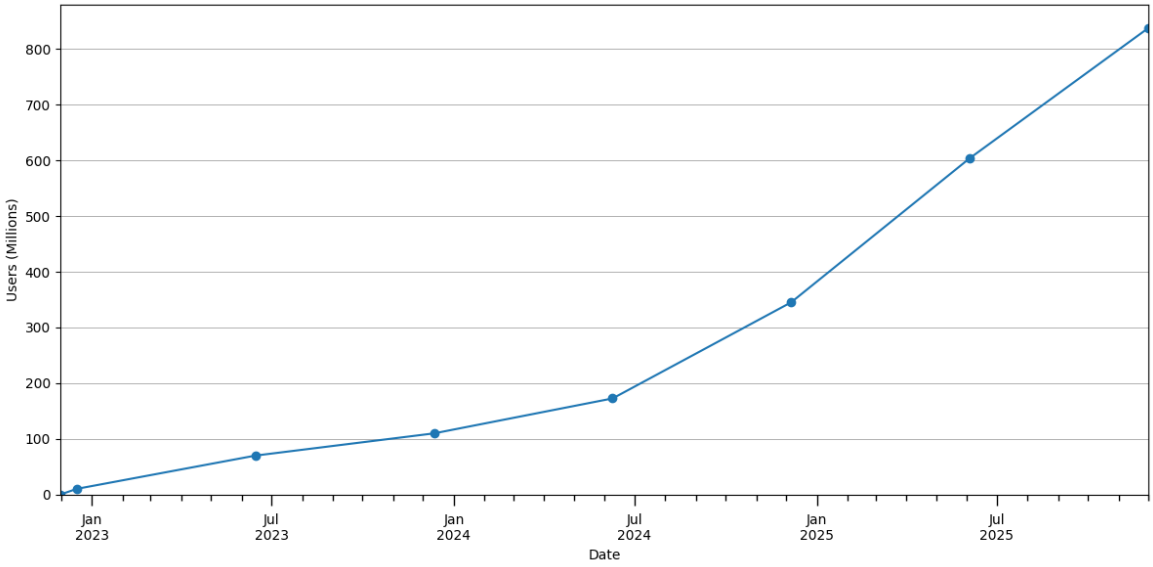


Figure 1: Weekly active ChatGPT users on consumer plans (Free, Go, Plus, Pro), shown as point-in-time snapshots every six months, November 2022–December 2025.

[21] 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).

[22] The amount of noise that is added to each cell is a function of several factors. ϵ defines the privacy budget, with lower values corresponding to a lower amount of noise, and correspondingly weaker privacy. The parameter δ corespondingly defines the probability that the differential privacy guarantee is violated, with smaller values providing more privacy. In Appendices G.1 and G.2, we outline how we use these two figures, as well as information about the aggregations we need to perform to create our desired datasets, to determine the amount of noise to add to each cell to ensure that our privacy guarantee is met.

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 1 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 2025, ChatGPT had more than 800 million total WAU, nearly 10% of the world's adult population.[23]

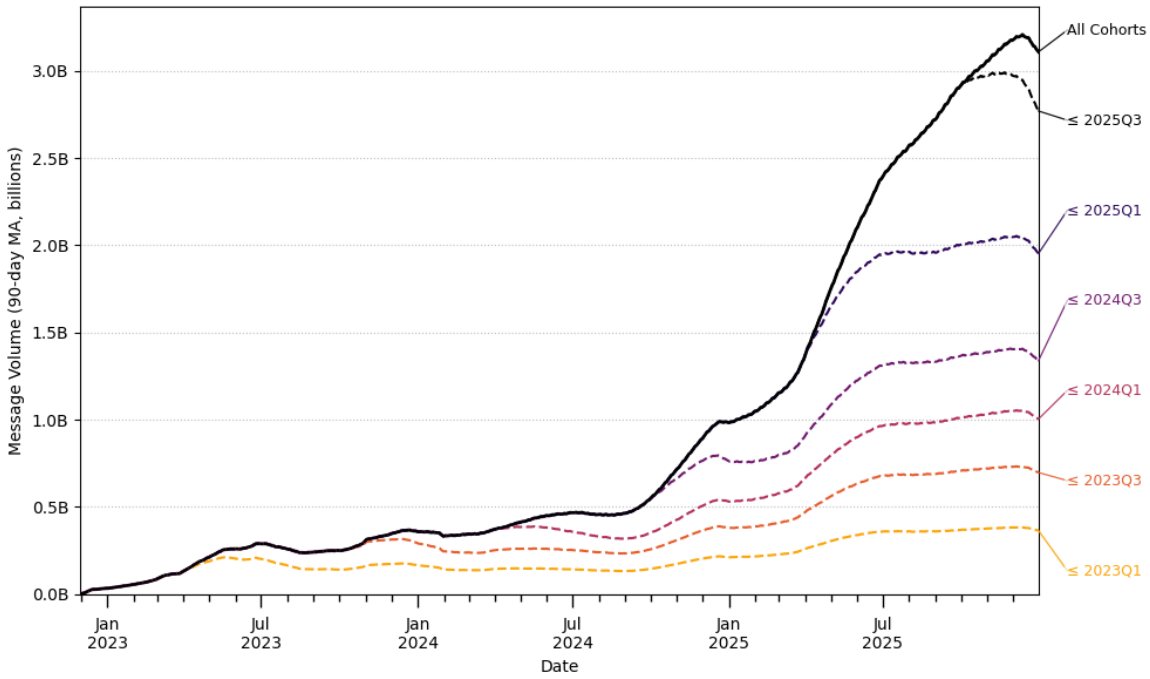


Figure 2: Daily message volumes from ChatGPT consumer plans (Free, Go, Plus, Pro), split by user sign-up date. Reported values are moving averages of the past 90 days for all cohorts up to and including the specified group. Y-axis is a 90 day moving average of message volume reported in billions.

Figure 2 presents total user message volume over time. By then end of 2025, ChatGPT users were sending more than 3 billion messages per day. Between July 2024 and December 2025, the number of daily messages sent grew more than sixfold.

Figure 2 also shows growth in message volume by the quarter in which the user first logged in to ChatGPT. For early adopters - those who started in the first three months of 2023 - usage declined somewhat over 2023, but started growing again in late 2024 and by late 2025 was higher than it had ever been.

Notably, Figure 2 shows that message volume grew substantially for all user cohorts beginning in early 2025. This suggests that the rapid growth of ChatGPT usage over the last year has been a time effect rather than a cohort effect, possibly because of improvements in the capabilities of the models or users discovering new use cases. Appendix Figure A.3 shows that daily messages per WAU have increased substantially since mid-2024, confirming growing usage intensity over time.

[23]The slight dip at the end of 2025 is consistent with a holiday dip in ChatGPT usage, which also occurs in prior years.

5 Who Uses 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, 2025). These papers typically find that generative AI is used more frequently by men, young people, and those with tertiary and/or graduate education.

We make two 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 usage gaps have changed over time.

5.1 Analyzing Gender Gaps using Names

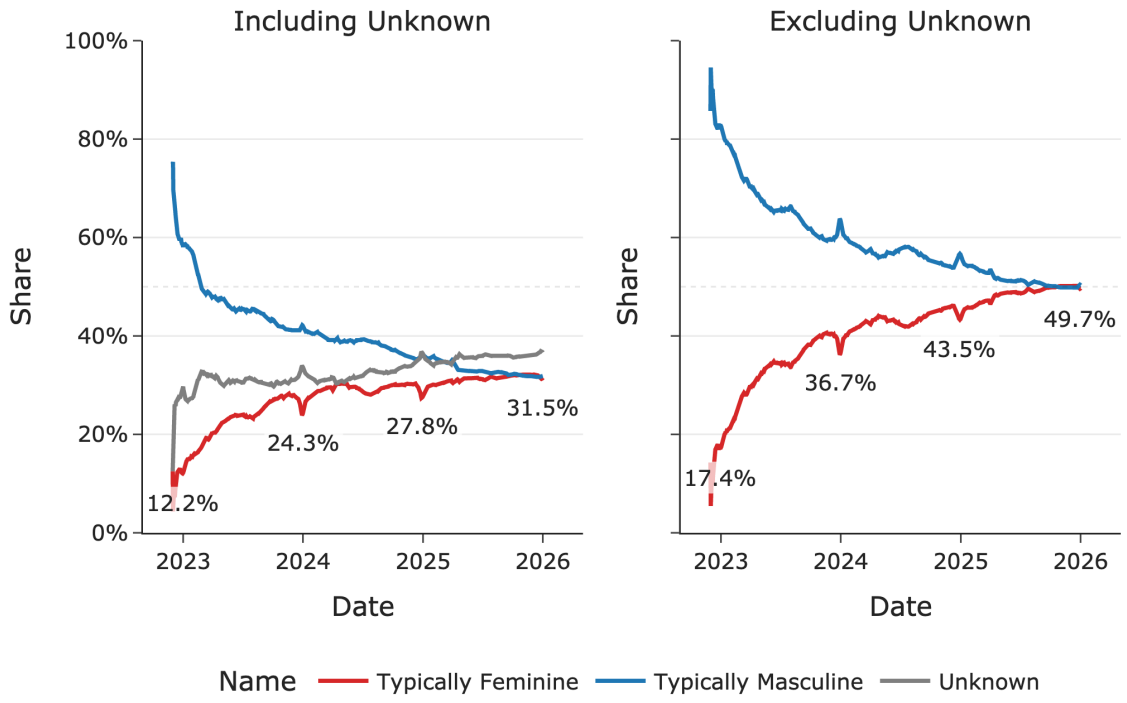


Figure 3: Breakdown of weekly active users by typically masculine and typically feminine first names. We draw on a uniform sample of 1.5M ChatGPT accounts, subject to the same user exclusion principles as other datasets we analyze. Note that this is a separate sample than those described in Section 3. First names are classified as typically masculine or typically feminine using public aggregated datasets of name-gender associations. See Appendix Figure A.28 for a complementary result based on a message-level sample rather than this main-text user-level series.

We investigate potential variation by gender by classifying 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 where there was ambiguity within them were classified as *Unknown*.

Excluding *Unknown*, more than 80% of the weekly active users (WAU) in the first few months after ChatGPT was released were by users with typically male first names. However, the share of active users with typically female and typically male names has gradually converged, and by late 2025 active users were slightly more likely to have typically female names. This suggests that gender gaps in ChatGPT usage have closed substantially over time.

Appendix Figure A.4 shows that users with typically female first names are relatively more likely to send messages related to writing, whereas users with typically male first names are more likely to use ChatGPT for technical tasks such as coding.

5.2 Variation by Age

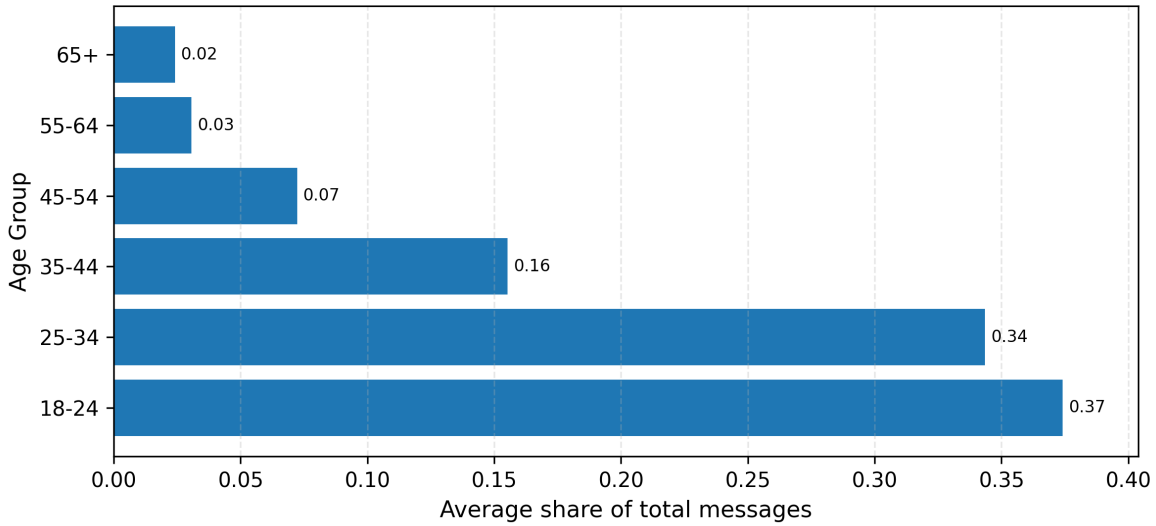


Figure 4: Average daily share of ChatGPT consumer-plan messages by self-reported age group in 2025. For each day, age-group shares are computed from sample-weighted message counts and then averaged across days with equal day weights. Estimates are based on approximately 1.2 million messages from the 2 million-message sample described in 3.2. See Appendix Figure A.26 for a complementary extension based on a message-level sample rather than this paper's conversation-level sampling design.

Figure 4 reports the average age distribution for the share of total messages sent in 2025. We observe that ChatGPT usage is heavily skewed toward young people. Among users who self-report their age, 37% of all messages are sent by users age 18-24 and 34% by users age 25-34. In contrast, only 5% of all messages are sent by users age 55 and above.

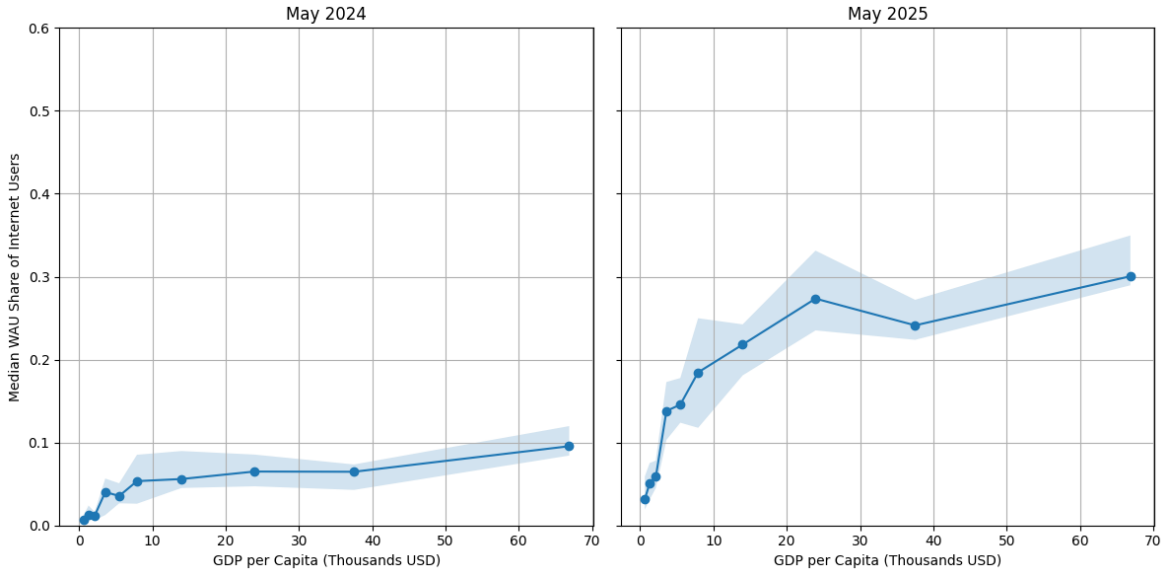


Figure 5: ChatGPT Weekly Active Users as Share of Internet Population vs GDP decile, May 2024 vs May 2025. Point estimates are medians within each decile. Internet Using Population uses 2023 estimates from the World Bank. Shaded regions indicate the interquartile range (25th–75th percentile) of country values within each GDP decile.

5.3 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. Figure 5 plots the weekly active user (WAU) share of internet users in each country against GDP per capita, with countries split into 10 percentile groups.[24]

Each dot in Figure 5 shows the median value within each percentile, and the shaded band is the interquartile range (25th–75th percentile) of each percentile. Between May 2024 and May 2025 the adoption of ChatGPT grew dramatically at all country income levels. However, it grew disproportionately in low to middle-income countries ($10,000–40,000 GDP-per-capita). As a result, the country income gradient in ChatGPT usage is relatively flat above the 25th percentile, suggesting substantial global income convergence in ChatGPT adoption.

6 How ChatGPT is Used

We now turn to the *content* of ChatGPT conversations, using a variety of taxonomies. Each taxonomy is associated with a "prompt" which defines a set of categories, so that an LLM can map each message to a category. Our categories often apply to the user's *intention* rather than the text of the conversation, so we do not directly observe the ground truth. Nevertheless the classifier results can

[24] GDP and population data are from the World Bank 2023 estimates, and nominal GDP is reported in thousands of U.S. dollars.

be interpreted as the best-guess inferences that a human would make. We show in Appendix C that the guesses from the LLM correlate highly with human guesses given the same prompt. We observed similar qualitative results when the prompt included an additional category for "uncertain."

6.1 What share of ChatGPT queries are work-related?

We label each user message in our dataset based on whether it appears to be related to work. The critical part of the prompt is as follows:[25]

> *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?")*

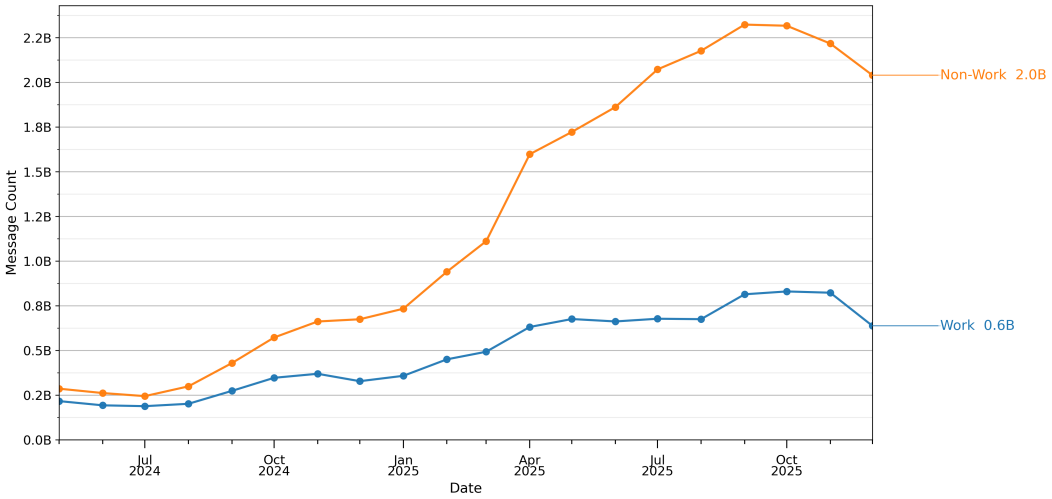


Figure 6: ChatGPT daily message counts (billions), broken down by likely work-related or non-work-related. Total daily counts are scaled estimates of message volume from all consumer plans. Daily counts of work and non-work related messages are estimated by classifying a random sample of conversations from each day, reweighted to reflect the total message counts for the corresponding 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 monthly averages from May 2024 to December 2025. Message-level robustness version: Appendix Figure A.10.

Figure 6 shows that both types of queries grew rapidly from May 2024 to December 2025. However non-work-related messages grew roughly tenfold by November 2025, compared to fivefold growth for work-related messages. As a result, the share of all messages that were work-related declined from 43% to 23% over this period. Appendix Figure A.5 shows that most of this change occurs within rather than across cohorts, and thus reflects a change in user behavior rather than a compositional

[25] See Appendix B for the full prompt, see Appendix C for validation.

shift.[26] 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.[27]

6.2 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, listed below. The full conversation categorization prompt is given in Appendix B.

Coarse Conversation Topic	**Detailed Conversation Topic**
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
	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 1: Mapping between Coarse and Detailed Conversation Topics

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

[26] We also built an alternative classifier that labels messages as *work*, *schoolwork*, or *other* rather than a binary work/non-work split. The corresponding time-series results are shown in Appendix Figure A.6.

[27] Users' revealed preference for non-work queries 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.

79% of all ChatGPT conversations. *Practical Guidance* grew modestly from 28% to 32% of overall usage. *Writing* declined from 34% of all usage in April 2024 to 23% at the end of 2025. *Seeking Information* grew from 18% to nearly 25% of all messages over the same period. *Multimedia* grew from 2% to just over 5%, with a large spike in April 2025 after ChatGPT released new image-generation capabilities. The share of *Technical Help* declined from 10% in May 2024 to 4% a year later. One possible explanation is that the use of LLMs for programming has grown very rapidly through other sources such as the API (outside of ChatGPT) and agentic AI assistance with code editing (e.g. Codex).

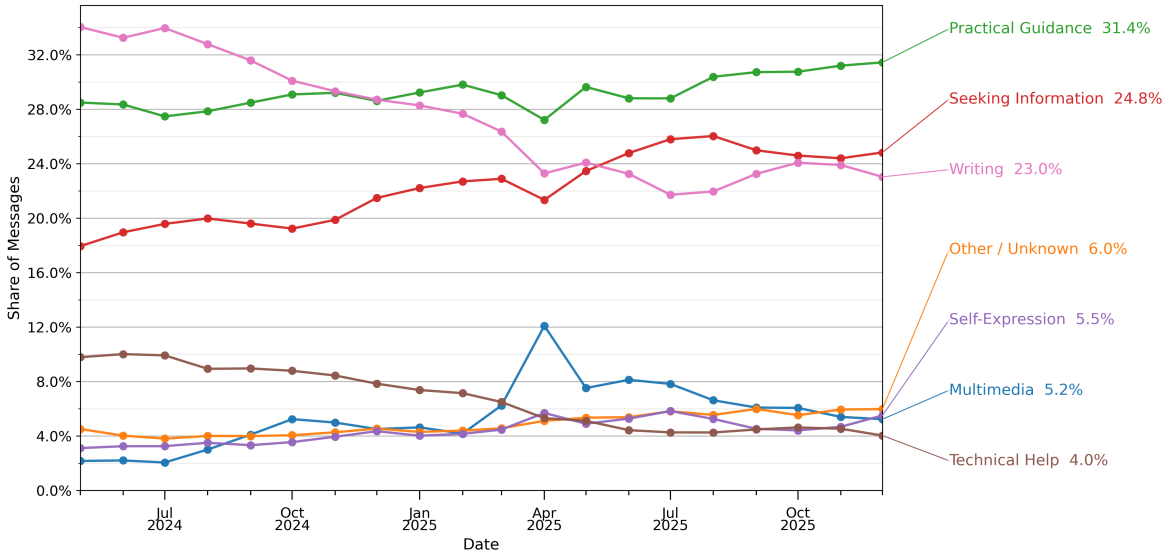


Figure 7: Share of consumer ChatGPT messages broken down by high level conversation topic, according to the mapping in Table 1. Values are bucketed by month. Shares are calculated from a sample of approximately 2 million sampled conversations from May 2024 through December 2025. Observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3. Message-level robustness version: Appendix Figure A.11.

Figure 8 shows Conversation Topics, restricting the sample to work-related messages. About 41% of all work-related messages by the end of 2025 are *Writing*, by far the most common Conversation Topic. *Practical Guidance* is the second most common use case at nearly 25%. *Technical Help* has declined from 15% of all work-related messages in May 2024 to just under 7% in December 2025. Figure 9 disaggregates the seven Conversation Topics into each of the 24 groups and sums up messages of each type over 2025. 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*.

Notably, two of those five categories (*Editing or Critiquing Provided Text* and *Translation*) are requests to modify text that has been provided to ChatGPT by the user, whereas the other three are requests to produce novel text. The former constitute more than half of all *Writing* conversations, which suggests that most *Writing* conversations with ChatGPT are requests to modify user inputs rather than to create something new.

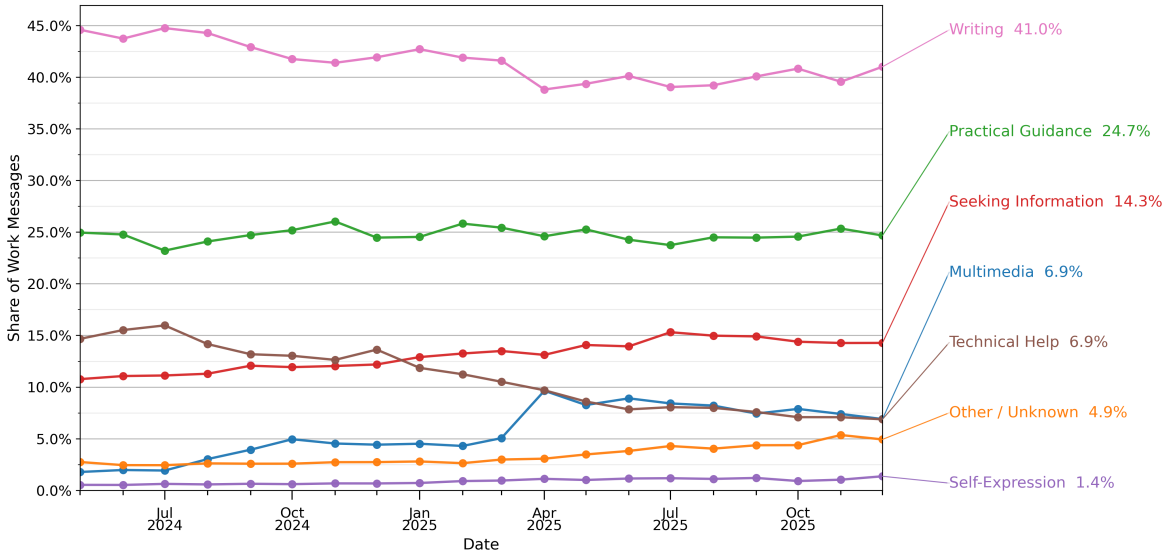


Figure 8: Share of **work related** consumer ChatGPT messages broken down by high level conversation topic, according to the mapping in Table 1. Values are bucketed by month. Shares are calculated from a sample of approximately 2 million sampled conversations from May 2024 through December 2025. Observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3. Message-level robustness version: Appendix Figure A.12.

Education is a major use case for ChatGPT, with 28% of *Practical Guidance* messages and 8.3% of all user messages being requests for *Tutoring or Teaching*. Another large share - 33% of *Practical Guidance* and 9.8% in total - is general how-to advice on a variety of topics. *Technical Help* includes *Computer Programming* (1.8% of messages), *Mathematical Calculations* (2.7%), and *Data Analysis* (0.3%). Looking at the topic of *Self-Expression*, 2.4% of all ChatGPT messages are on the topic of *Relationships and Personal Reflection*.

6.3 O*NET Work Activities

We map message content to work activities using the Occupational Information Network (O*NET) Database Version 29.0, following 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*. We then used the official O*NET taxonomy to map these classified IWAs to one of the Generalized Work Activities (GWA).[28]

[28] We do not show the shares for GWAs with fewer than 100 users sending messages for each category, grouping them instead into *Suppressed*.

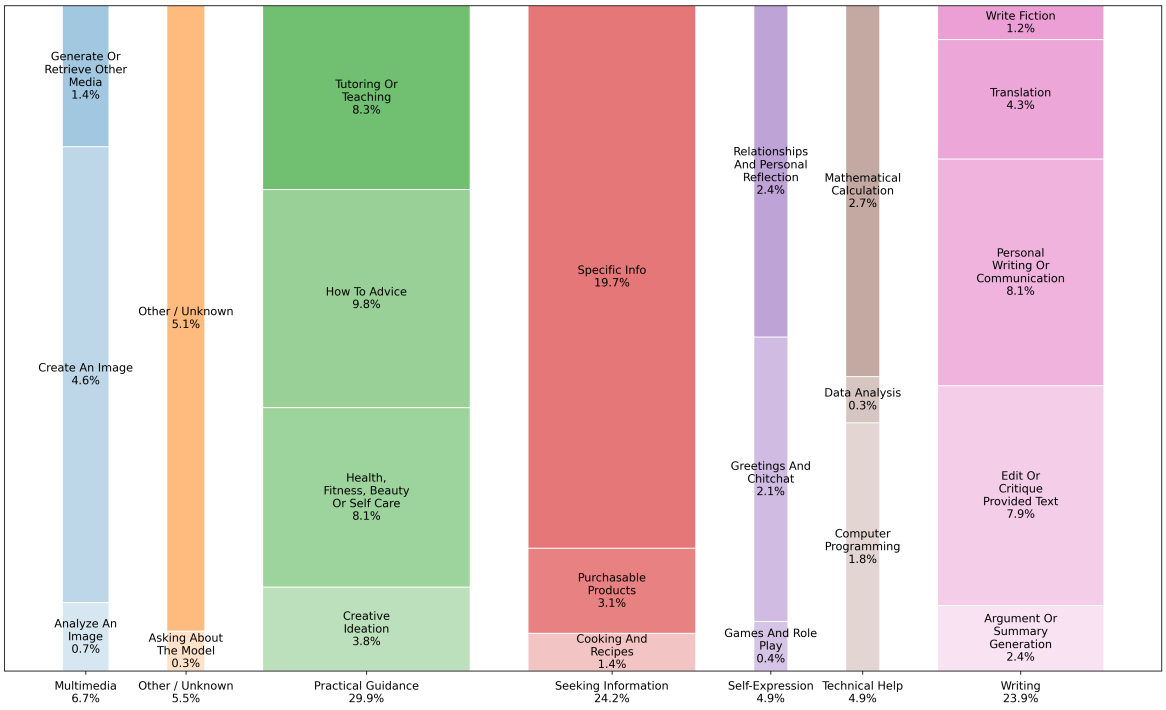


Figure 9: Breakdown of granular conversation topic shares within the coarse mapping defined in Table 1. The underlying classifier prompt is available in Appendix B. Each bin reports a percentage of the total population. Shares are calculated from a sample of approximately 1.2 million sampled conversations from January through December of 2025. Observations are reweighted to reflect total message volumes on a given day. This sample is a subset of the 2 million message sample described in Section 3.2. Message-level robustness version: Appendix Figure A.13.

Figure 10 presents the distribution of GWAs for the subsample of messages we classify as work-related.[29] The most common GWAs are *Documenting/Recording Information* (19.6%), *Thinking Creatively* (14.0%), *Making Decisions and Solving Problems* (12.6%), *Getting Information* (10.2%), *Interpreting the Meaning of Information for Others* (9.6%), *Interacting with Computers* (6.7%), and *Providing Consultation and Advice to Others* (5.0%). These seven GWAs collectively account for 78% of work-related messages.

Broadly speaking, we can group much of ChatGPT work usage into two functions - *information processing*, and *decision support*. The word "information" appears in three of the top five and five of the top fifteen tasks in Figure 10. Among the other common functions are tasks like making decisions, giving advice, solving problems and thinking creatively, all of which are meant to help ChatGPT users make better decisions at work.

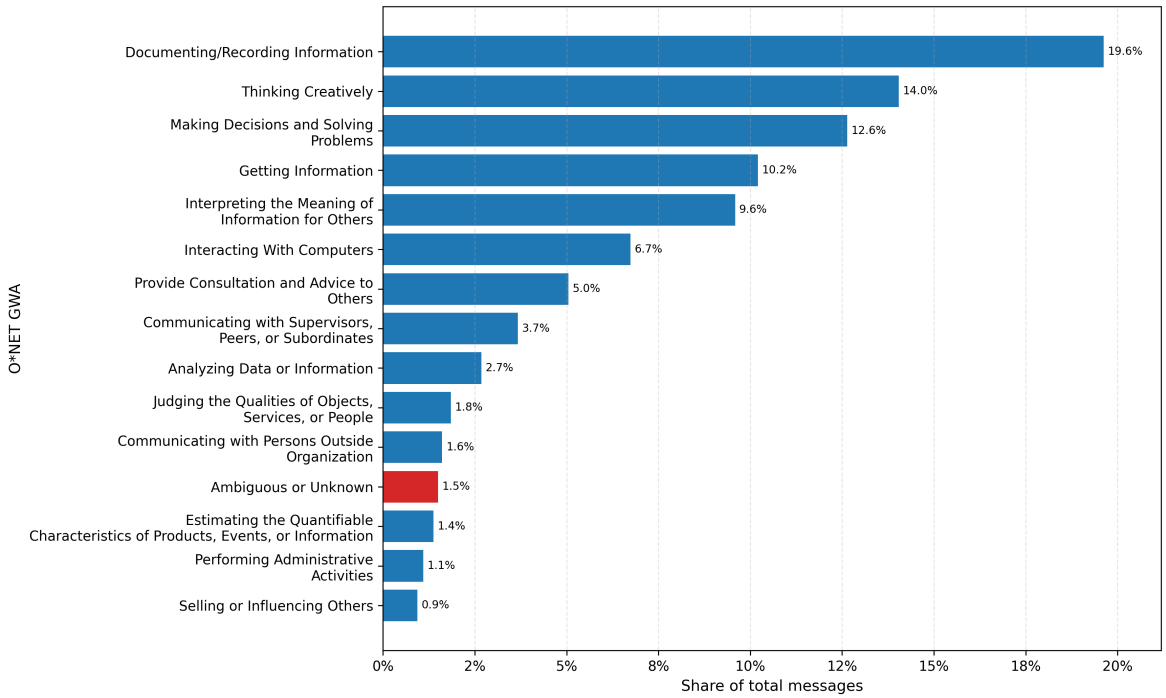


Figure 10: Top 15 GWA Shares of approximately 320,000 messages classified as work related from January to December 2025. Messages are classified as pertaining to one of 332 O*NET IWAs or *Ambiguous*. IWAs were then aggregated to GWAs using the O*NET Work Activities taxonomy. Observations are reweighted to reflect total message volumes on a given day. This sample is a subset of the 2 million message sample described in Section 3.2. The IWA prompt is provided in Appendix B. Message-level robustness version: Appendix Figure A.14.

We use the linked dataset described in Section 3.3 to study variation in ChatGPT usage by occupation. Appendix Figure A.7 presents raw and regression-adjusted differences by broad occupation groupings.[30] We find that users in highly paid professional and technical occupations are more likely

[29] Appendix Figure A.18 presents an analogous list for all messages, rather than restricting only to those that are work-related. Message-level robustness versions are shown in Appendix Figures A.14 and A.19.

[30] Appendix Figure A.7 shows results for five groups – (1) all nonprofessional occupations, including administrative,

to use ChatGPT for work, and that the topics of work-related messages are much more likely to be focused on core job tasks (e.g. coding for computer-related occupations).

clerical, service, and blue-collar occupations (SOC2 codes 31 to 53); (2) computer-related occupations (SOC2 code 15); (3) engineering and science occupations (SOC2 codes 17 and 19); (4) management and business occupations (SOC2 codes 11 and 13); and (5) all other professional occupations, including law, education, and health care (SOC2 codes 21 to 29).

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	1	2	3	4	5	6	8
Business	2	1	3	5	4	6	7
Computer/Math	2	3	5	1	4	6	7
Engineering	2	3	5	1	4	6	7
Science	2	1	5	3	4	6	8
Social Service	1	2	3	6	5	4	7
Legal	1	2	4	8	3	5	7
Education	1	2	3	4	6	5	7
Arts/Design/Media	1	2	3	4	5	6	7
Health Professionals	1	2	3	6	4	5	7
Health Support	1	2	3	-	5	4	6
Protective Service	1	2	3	-	4	5	7
Food Service	1	2	3	6	4	5	7
Personal Service	1	3	2	6	5	4	7
Sales	1	3	2	4	5	6	7
Administrative	1	2	3	6	4	5	7
Construction	1	2	4	5	3	6	7
Installation/Repair	2	1	3	6	4	5	9
Production	1	2	3	4	5	6	7
Transportation	2	3	4	1	5	6	7
Military	2	1	3	6	4	5	7

Figure 11: 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. Dashes indicate that the ranking is unavailable since fewer than 100 users from that occupation group requested that specific GWA within the sample. Omitted occupation groups are because no GWA was requested by more than 100 users from that occupation group. Not pictured are twelve other GWAs which are less frequently requested and are reported fully in Appendix E. See Appendix for full cross-tabulations between GWA and two-digit SOC2 codes.

We also examine the most common Generalized Work Activities (GWAs) associated with each broad occupation group as measured by 2-digit Standard Occupation Classification (SOC) codes, using the linked dataset described in Section 3.3. Table 11 presents the frequency ranking of work-related messages in each SOC code of the seven most common GWAs.

We find remarkable similarity across occupations in how ChatGPT is used at work. For example, *Making Decisions and Solving Problems* is one of the three most common GWAs in every single occupation group. Similarly, *Documenting and Recording Information* ranks in the top two and *Thinking Creatively* ranks in the top five of all 21 occupation groups. 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 Appendix E 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.

6.4 Asking, Doing and Expressing

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)). In economic models of AI at work, the replacement of job tasks has varying impacts depending on how important the tasks are in particular jobs (e.g. Acemoglu and Restrepo (2022), Eloundou et al. (2024)) or how the tasks are bundled together and whether replacing them substitutes or complements worker expertise (Autor and Thompson (2025)).

Our results show that many ChatGPT queries - including work-related queries - involve the user asking for information or advice rather than delegating tasks to the AI chatbot directly. Even so, AI chatbots are clearly used widely at work, and users report significant time savings (e.g. Bick et al. (2024)). This emphasis on advice-seeking and decision support is also evident in the early evidence from agentic AI: Yang et al. (2025) finds that "learning and research" is the second most common use case for AI agents, and Sarkar (2025) shows that planning and asking for explanations accounts for nearly 40% of agent usage.

We explore the distinction between asking for advice and delegating work with a novel classifier of user intent, which also captures a residual category for messages that involve neither. We call this classifier *Asking, Doing, or Expressing*. 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?").*

Intent	Prompt
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.

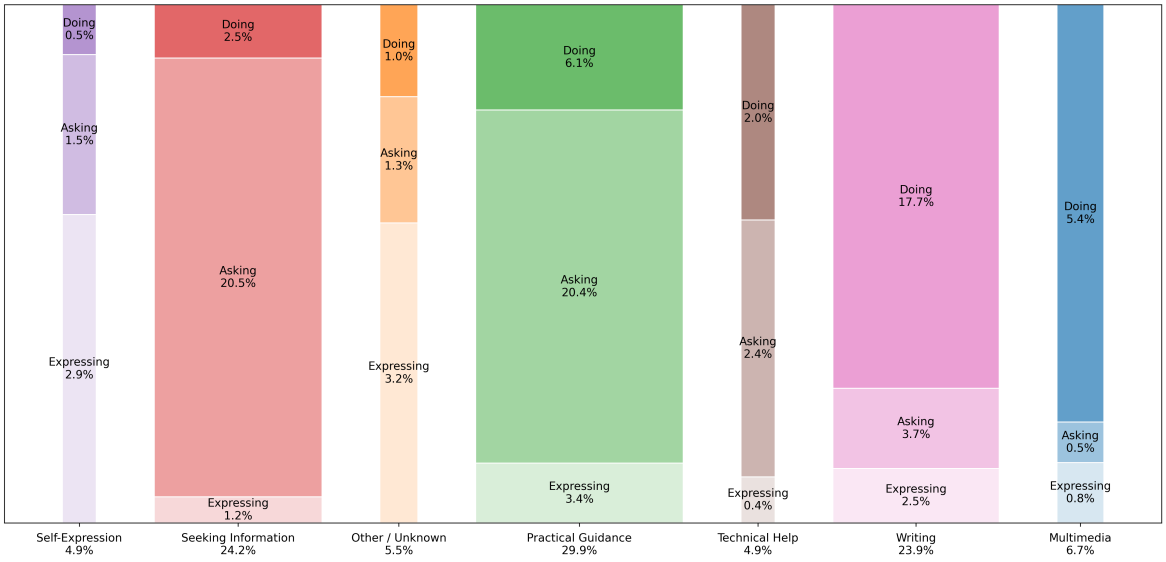


Figure 12: Breakdown of Conversation Topics by Asking/Doing/Expressing category, with topic columns sorted by relative share of "Doing" messages. Prompts for these automated classifiers are available in Appendix B. For a detailed breakdown of conversation topic contents, see Table 1. Each bin reports a percentage of the total population. Shares are calculated from a sample of approximately 1.2 million sampled conversations from January to December of 2025. Observations are reweighted to reflect total message volumes on a given day. This sample is a subset of the 2 million message sample described in Section 3.2. Message-level robustness version: Appendix Figure A.15.

Figure 12 shows the share of messages by each intent type in our sample. Overall, 50% of user messages are *Asking*, 35% are *Doing*, and 15% are *Expressing*. The figure also shows that conversation topic and user intent 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.[31]

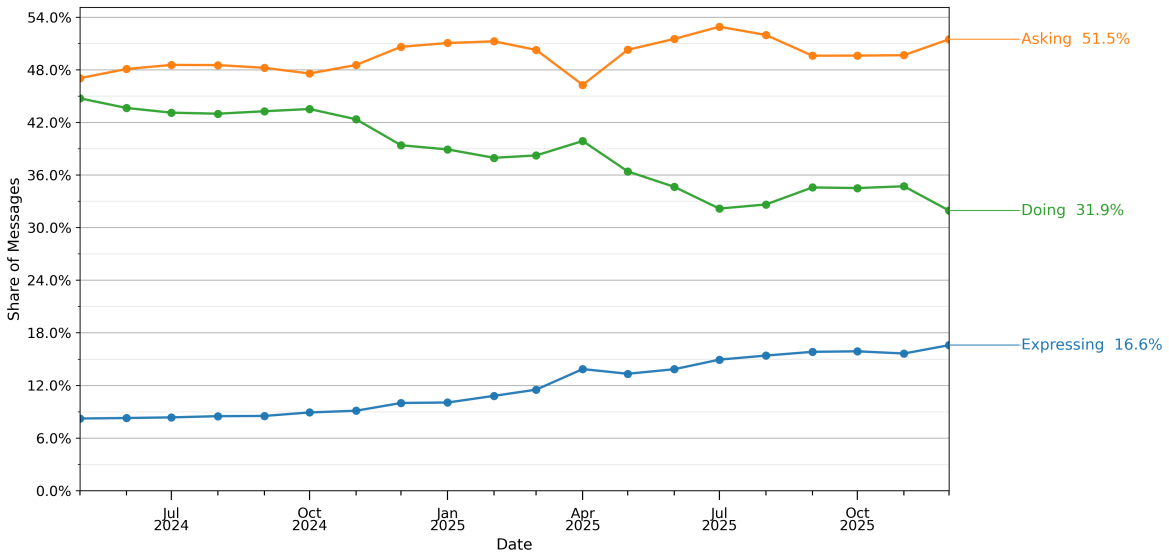


Figure 13: Shares of messages classified as Asking, Doing, or Expressing by an automated ternary classifier. Values bucketed by month. Shares are calculated from a sample of approximately 2 million sampled conversations from January 2024 through December 2025. Observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3.2. Message-level robustness version: Appendix Figure A.16.

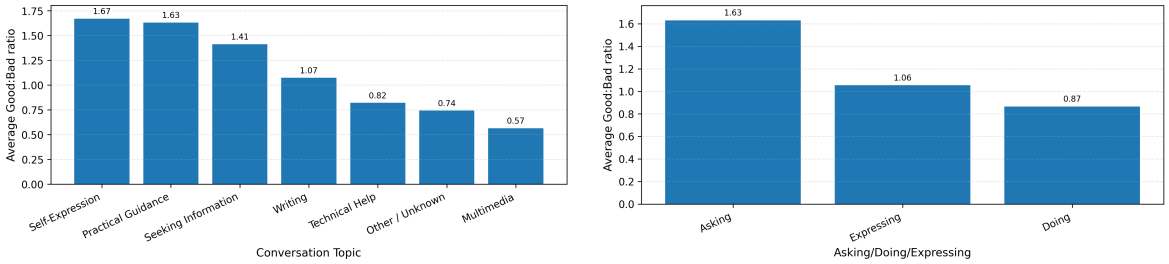
Figure 13 presents changes over time in the composition of messages by user intent. In May 2024, usage was nearly evenly split between *Asking* and *Doing*, with just under 10% of messages classified as *Expressing*. *Asking* and *Expressing* grew much faster than *Doing* over the next year, and by late 2025 the split was 52% *Asking*, 32% *Doing*, and 17% *Expressing*.[32]

We also study user satisfaction with an internal classifier called *Interaction Quality*, which looks for an expression of satisfaction or dissatisfaction in the user's subsequent message in the same conversation (if one exists).[33] Figure 14 shows the ratio of good-to-bad messages by user intent, as rated by the *Interaction Quality* classifier. *Asking* messages are 1.6 times more likely to receive a positive rating than a negative one, a much higher ratio than either *Doing* or *Expressing* messages. This is consistent with the stronger relative growth of *Asking* messages over time and suggests that users highly value the ability to ask for information and advice, even about work-related tasks.

[31] Appendix Figure A.8 presents shares of *Asking/Doing/Expressing* just for work-related messages. *Doing* constitutes 55% of work-related queries, compared to 34% for *Asking* and roughly 11% for *Expressing*. Nearly 33% of all work-related queries are *Doing* messages related to *Writing*. *Doing* and *Asking* comprise equal shares of *Technical Help* queries. The corresponding message-level robustness version is Appendix Figure A.20.

[32] Appendix Figure A.9 presents the share of work-related messages by user intent. *Doing* messages, which account for approximately 35% of messages, have the highest overall share of work related messages. The corresponding message-level robustness version is Appendix Figure A.21.

[33] 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 C.



Panel A **Panel B**

Figure 14: 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 B. Values represent the average ratio from January to December 2025, where observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3. Message-level robustness version: Appendix Figure A.17.

7 A Model of Chatbot Interaction

To understand the economic implications of *Asking* and *Doing*, we develop a simple principal-agent model of the interaction between people and chatbots. (For these analyses, we ignore the "expressing" category, which is relatively rare in work settings and has limited economic content.) A principal ("person") undertakes an action whose value depends on context. The principal can (i) work alone (DIY), (ii) delegate the task to the chatbot agent, or (iii) ask the agent for advice and then execute the task themselves. Delegation eliminates or reduces the principal's effort cost and may increase accuracy, but produces "generic" outputs that do not account for idiosyncratic preferences or specific context. Asking is costly but increases accuracy when the principal carries out the action themselves.[34]

7.1 Model Setup

A principal P faces a task $t \in \mathcal{T}$ and chooses an interaction mode

$$d \in \{\text{DIY}, \text{delegate}, \text{ask}\},$$

corresponding to (i) executing the task herself (DIY), (ii) delegating execution to an AI chatbot agent A (delegate), or (iii) asking the agent for help and then executing herself (ask). We suppress the task index t when it is not needed.

Each task instance has an idiosyncratic context $\omega \in \Omega$ drawn from a distribution F. The principal observes ω before choosing d. The chatbot agent does not observe ω - it knows only the prior F. Without additional information, the agent produces a generic output that is optimal for the population-wide context distribution F.

For example, if the task is "writing an email", the context ω is (for example) whether the email is being written to a boss, a coworker, or a friend, and more generally whether it is appropriate for

[34] Although we use "principal–agent" terminology, the principal's and agent's objectives are aligned in our model. The friction between the principal and the agent arises from information asymmetry and communication restrictions, rather than incentives or strategic behavior.

the situation. Delegating the email to the chatbot agent saves the principal's time and the email will be well-written, but the agent does not know the full context ω because the principal has not communicated it. As a result, the email will be written in a standard style that reads like the average of all relevant emails in the agent's training data.

The principal evaluates outcomes via a utility function $u : \mathcal{A} \times \Omega \to \mathbb{R}$ over an action $a \in \mathcal{A} \subseteq \mathbb{R}$ and context ω. Assume $\mathcal{A}$ is compact and $u(\cdot, \omega)$ is continuous and strictly concave in a for each ω, so the ideal action

$$a^*(\omega) \ \in \ \arg\max_{a \in \mathcal{A}} u(a, \omega) \tag{1}$$

is well-defined and unique. Let

$$\bar{a} \ \in \ \arg\max_{a \in \mathcal{A}} \mathbb{E}_{\omega \sim F}[u(a, \omega)] \tag{2}$$

denote the population-optimal generic action, indendent of any context (assume $\mathcal{A}$ is compact so that this maximizer exists).

For later use, define the *context gap* between the principal's ideal action and the generic action as

$$g(\omega) \ \equiv \ a^*(\omega) - \bar{a}, \qquad \tau(\omega) \ \equiv \ g(\omega)^2 = \big(a^*(\omega) - \bar{a}\big)^2.$$

Intuitively, $\tau(\omega)$ measures how idiosyncratic the context is in terms of the action it calls for: it is the squared difference between the ideal action to the generic action. The chatbot cannot observe the context ω, and for simplicity we assume the principal cannot communicate ω under delegation, although this can be easily relaxed.[35]

The party executing the task may implement the intended action imperfectly. If party $i \in \{P, A\}$ executes, the realized action equals the intended action plus mean-zero execution noise:

$$\tilde{a} = a + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i] = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma_i^2,$$

where ε_i is independent of ω. We interpret a smaller σ_i^2 as higher execution accuracy. If a party cannot execute the task (e.g. a human cannot solve a proof; an agent cannot walk up the stairs), we allow $\sigma_i^2 = +\infty$. Define the *AI advantage* as

$$\Delta \ \equiv \ \sigma_P^2 - \sigma_A^2,$$

so that $\Delta > 0$ indicates that the agent is more accurate in execution than the principal.

Performing the task oneself entails a time/effort cost $c \geq 0$. We normalize the direct cost of delegation to zero. If the principal asks for advice, they additionally choose an *asking intensity* $x \geq 0$ and pay a linear asking cost μx with $\mu > 0$.

Asking improves the principal's execution accuracy according to

$$\sigma_P^2(x) \ = \ \frac{\sigma_P^2}{1 + x}, . \tag{3}$$

[35] Appendix E extends the model by allowing the principal to transmit content at intensity $y \geq 0$ (e.g. by providing examples or documents) at linear cost κy, $\kappa > 0$, which reduces the effective context gap to $\tau(\omega; y) \equiv \frac{\tau(\omega)}{1+y}$. This modifies the cutoff rule so that the decision between DIY, delegating, or asking is a function of κ, but does not change the key comparative statics.

This captures the idea that asking (prompting, iterating, or requesting feedback) improves task performance but with diminishing returns. Other decreasing and convex variance functions deliver qualitatively similar predictions.

To summarize, the timing is as follows:

1. Nature draws $\omega \sim F$; the principal observes ω.
2. The principal chooses mode $d \in \{\text{DIY}, \text{delegate}, \text{ask}\}$. If $d = \text{ask}$, the principal also chooses intensity $x \geq 0$.
3. The intended action is selected and executed:
   - If $d = \text{DIY}$ or $d = \text{ask}$, the principal targets $a^*(\omega)$ and implements $\tilde{a} = a^*(\omega) + \varepsilon_P$, with $\varepsilon_P \sim \mathcal{N}(0, \sigma_P^2)$ under DIY and $\varepsilon_P \sim \mathcal{N}(0, \sigma_P^2(x))$ under Ask.
   - If $d = \text{delegate}$, the agent implements the generic action $\tilde{a} = \bar{a} + \varepsilon_A$, with $\varepsilon_A \sim \mathcal{N}(0, \sigma_A^2)$.
4. Payoffs are realized.

7.2 DIY, Delegation, and Asking

If $d = \text{DIY}$, the principal executes the ideal action and pays effort cost c:

$$V_{\text{DIY}}(\omega) \;=\; \mathbb{E}[u(a^*(\omega) + \varepsilon_P,\ \omega)] - c.$$

If $d = \text{delegate}$, the agent executes the generic action:

$$V_{\text{delegate}}(\omega) \;=\; \mathbb{E}[u(\bar{a} + \varepsilon_A,\ \omega)]\,.$$

If $d = \text{ask}$, the principal chooses an asking intensity $x \geq 0$, pays asking cost μx, and then executes the ideal action with reduced variance $\sigma_P^2(x)$:

$$V_{\text{ask}}(\omega; x) \;=\; \mathbb{E}[u(a^*(\omega) + \varepsilon_P(x),\ \omega)] - \mu x - c,$$

where $\varepsilon_P(x) \sim \mathcal{N}(0, \sigma_P^2(x))$. A solution to the model specifies an optimal mode $d^*(\omega)$ and, when $d^*(\omega) = \text{ask}$, an optimal asking intensity $x^*(\omega) \in \arg\max_{x \geq 0} V_{\text{ask}}(\omega; x)$.

7.3 Optimal human-chatbot interaction

We now solve for the principal's optimal interaction policy assuming a quadratic loss utility function.[36]

Assumption 1. *Payoffs are quadratic in the realized action:*

$$u(\tilde{a}, \omega) \;=\; -(\tilde{a} - \theta(\omega))^2,$$

where $\theta(\omega) = a^*(\omega)$.

[36] Appendix F extends the model beyond quadratic loss, to any environment in which the principal's value of DIY is context-invariant and the value of delegation is weakly decreasing in the context gap (τ) and the AI's execution variance (σ_A^2).

This functional form is convenient and can be interpreted as a second-order approximation to a more general payoff function around the ideal action. Using the execution-noise structure from above, the expected payoffs are:

$$V_{\text{DIY}}(\omega) = \mathbb{E}[u(\tilde{a}, \omega)] - c = -\sigma_P^2 - c, \tag{4}$$

$$V_{\text{ask}}(\omega; x) = \mathbb{E}[u(\tilde{a}, \omega)] - \mu x - c = -\sigma_P^2(x) - \mu x - c = -\frac{\sigma_P^2}{1+x} - \mu x - c, \tag{5}$$

$$V_{\text{delegate}}(\omega) = \mathbb{E}[u(\tilde{a}, \omega)] = -\Big(\tau(\omega) + \sigma_A^2\Big), \qquad \tau(\omega) \equiv \big(a^*(\omega) - \bar{a}\big)^2. \tag{6}$$

Two features are immediate. First, the value of "self-execution" (DIY or asking) does not depend on ω. When the principal chooses to execute herself, the context changes which action is ideal, but not the expected loss from executing it. Second, delegation is less valuable in more idiosyncratic contexts because the generic action is further from the ideal action.

Proposition 1. *Suppose utility is given by Assumption 1. Then:*

(i) Optimal asking intensity. *The optimal asking intensity is*

$$x^* \;=\; \max\left\{\sqrt{\frac{\sigma_P^2}{\mu}} - 1,\; 0\right\}. \tag{7}$$

(ii) Best self-execution payoff. *The best self-execution value is*

$$V_{\text{self}} \;\equiv\; \max\Big\{V_{\text{DIY}},\; \max_{x \geq 0} V_{\text{ask}}(\omega; x)\Big\} = \begin{cases} -\sigma_P^2 - c, & \sigma_P^2 \leq \mu, \\ -2\sqrt{\mu\sigma_P^2} + \mu - c, & \sigma_P^2 > \mu, \end{cases} \tag{8}$$

where the second line equals $V_{\text{ask}}(\omega; x^*)$.

(iii) Cutoff rule. *Delegation is optimal if and only if*

$$\tau(\omega) \;\leq\; \big(-V_{\text{self}}\big) - \sigma_A^2. \tag{9}$$

Equivalently, define

$$T_0 \;\equiv\; \sigma_P^2 - \sigma_A^2 + c, \qquad T_1 \;\equiv\; 2\sqrt{\mu\sigma_P^2} - \mu - \sigma_A^2 + c. \tag{10}$$

If $\sigma_P^2 \leq \mu$, then $d^*(\omega) =$ delegate if and only if $\tau(\omega) \leq T_0$; otherwise the principal uses DIY. If $\sigma_P^2 > \mu$, then $d^*(\omega) =$ delegate if and only if $\tau(\omega) \leq T_1$; otherwise the principal asks at x^* and executes herself. If the relevant cutoff (T_0 or T_1) is negative, delegation is never optimal.

The optimal interaction mode choice is governed by the context gap $\tau(\omega)$. Delegation is attractive when the generic output is close to the principal's ideal action; when the gap is large, the principal executes the action herself, either directly or after asking.

We next describe how the delegation region changes with the agent's execution accuracy, the cost

of self-execution, and the cost of asking. Throughout, recall that a larger cutoff means that delegation is optimal for a wider set of contexts.

Corollary 1 (Comparative statics in AI advantage)**.** *Under the conditions of Proposition 1, define the* AI advantage *is* $\Delta \equiv \sigma_P^2 - \sigma_A^2$*. Then:*

(i) AI advantage and self-execution costs. *Holding* σ_P^2 *fixed (so increases in* Δ *correspond to improvements in AI accuracy), the delegation region expands in* Δ *and in* c*:*

$$\frac{\partial T_0}{\partial \Delta} = 1, \quad \frac{\partial T_0}{\partial c} = 1, \qquad \textit{and} \qquad \frac{\partial T_1}{\partial \Delta} = 1, \quad \frac{\partial T_1}{\partial c} = 1.$$

Higher AI advantage raises the maximum context gap $\tau(\omega)$ *for which delegation is optimal.*

(ii) Context gap. *Holding* σ_P^2 *fixed, an increase in AI advantage* Δ *raises the cutoff in* $\tau(\omega)$*. As AI improves, users delegate in increasingly idiosyncratic contexts (higher* $\tau(\omega)$*).*

(iii) Asking costs. *In the asking regime (*$\sigma_P^2 > \mu$*), holding* (σ_P^2, Δ, c) *fixed (equivalently holding* $(\sigma_P^2, \sigma_A^2, c)$ *fixed),*

$$\frac{\partial T_1}{\partial \mu} = \sqrt{\frac{\sigma_P^2}{\mu}} - 1 \; > \; 0.$$

Lowering the marginal cost of asking (lower μ*) reduces* T_1 *and shrinks the delegation region. Cheaper asking improves self-execution, making delegation optimal in fewer high-context situations (e.g. smaller* $\tau(\omega)$*).*[37]

(iv) Heterogeneity by skill. *Holding AI accuracy fixed (fixed* σ_A^2*), principals with worse baseline execution (higher* σ_P^2*) delegate in a wider set of contexts. In particular,*

$$\frac{\partial T_0}{\partial \sigma_P^2} = 1 \quad (\sigma_P^2 \le \mu), \qquad \frac{\partial T_1}{\partial \sigma_P^2} = \sqrt{\frac{\mu}{\sigma_P^2}} > 0 \quad (\sigma_P^2 > \mu).$$

The lefthand panel of Figure 15 plots the optimal decision as a function of the AI advantage Δ and the context gap $\tau(\omega)$, holding other model parameters fixed. DIY dominates at low levels of δ but eventually disappears, leaving the choice between Asking and Delegation as a function of $\tau(\omega)$ and δ as described in parts (i) and (ii) of Corollary 1.

The righthand panel of Figure 15 plots the optimal decision as a function of Δ and the marginal cost of asking μ. As shown in part (iii) of Corollary 1, lowering μ shrinks the delegation region, implying that generic AI output must be a closer contextual fit for delegation to remain optimal. This is consisent with Jahani et al. (2024b), who show using a controlled experiment that automating part of the prompting process for a text-to-image model (e.g. lowering μ) increases the payoff to human-guided execution (asking) relative to accepting the model's default output (delegating). In a field experiment, Dell'Acqua et al. (2023) find that providing a prompt engineering overview reduces management consultants' effective cost of iterating with an AI model. They find that the most successful consultants adopted a "Cyborg" interaction style—continually prompting, evaluating, and re-prompting at a fine-grained level (e.g. asking)—rather than the "Centaur" style of cleanly delegating entire subtasks to

[37] Note that this derivative applies only within the asking regime. Changing μ can switch the optimal asking intensity to $x^* = 0$ (i.e. move into the DIY regime), in which case the relevant cutoff is T_0 rather than T_1.

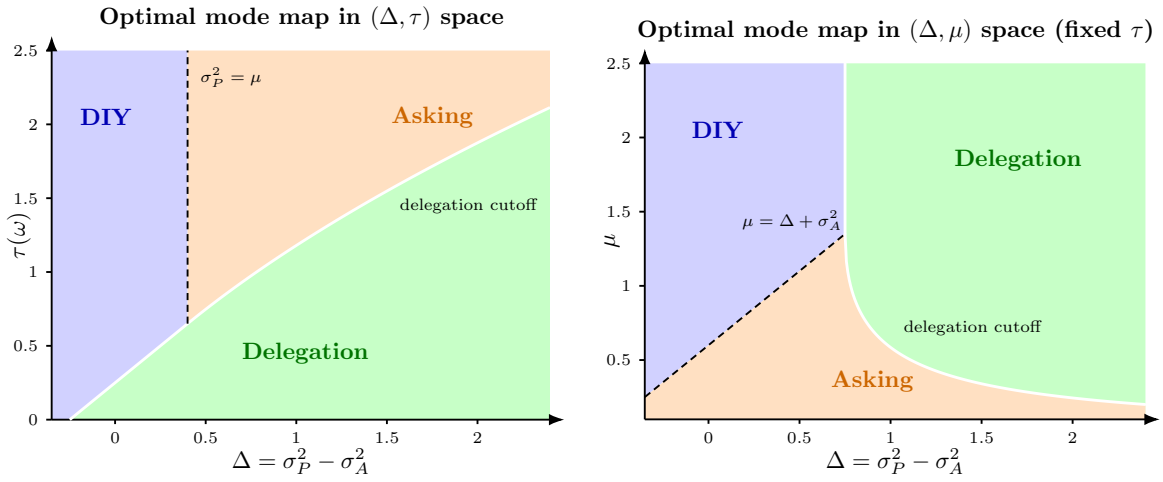


Figure 15: Optimal interaction modes under the model of human-AI interaction. In both panels, we plot the optimal decision between *DIY*, *Delegation*, and *Asking* using the cutoff rule from Proposition 1. Shared parameter values are $\sigma_A^2 = 0.6$ and $c = 0.25$. *Left:* context gap $\tau(\omega)$ varies on the y-axis and AI advantage Δ varies on the x-axis while holding asking cost fixed at $\mu = 1$. The dashed line marks the self-execution regime boundary $\sigma_P^2 = \mu$, and the white curve is the delegation cutoff $\tau = T(\Delta)$ separating delegation from self-execution. *Right:* asking cost μ varies on the y-axis and AI advantage Δ on the x-axis while holding context gap fixed at $\tau = 1$. The dashed segment marks $\mu = \Delta + \sigma_A^2$ (the DIY-vs-Asking threshold within self-execution), and the white boundary is the delegation cutoff in (Δ, μ) space.Green regions are delegation-optimal, blue and orange regions are self-execution (DIY and Asking, respectively).

the AI. Both of these interventions lower the asking cost μ, shifting the optimal mode of interaction with AI chatbots away from delegation and toward asking.

Part (iv) of Corollary 1 implies that the principal's accuracy and the agent's accuracy are substitutes, since $\Delta = \sigma_P^2 - \sigma_A^2$ rises with σ_P^2 when σ_A^2 is fixed. This suggests the asking value of AI is relatively greater for inaccurate users (e.g. higher σ_A^2) , which helps explain the greater benefits of generative AI for less-educated and lower-skilled workers (Noy and Zhang, 2023; Brynjolfsson et al., 2025; Cruces et al., 2026).

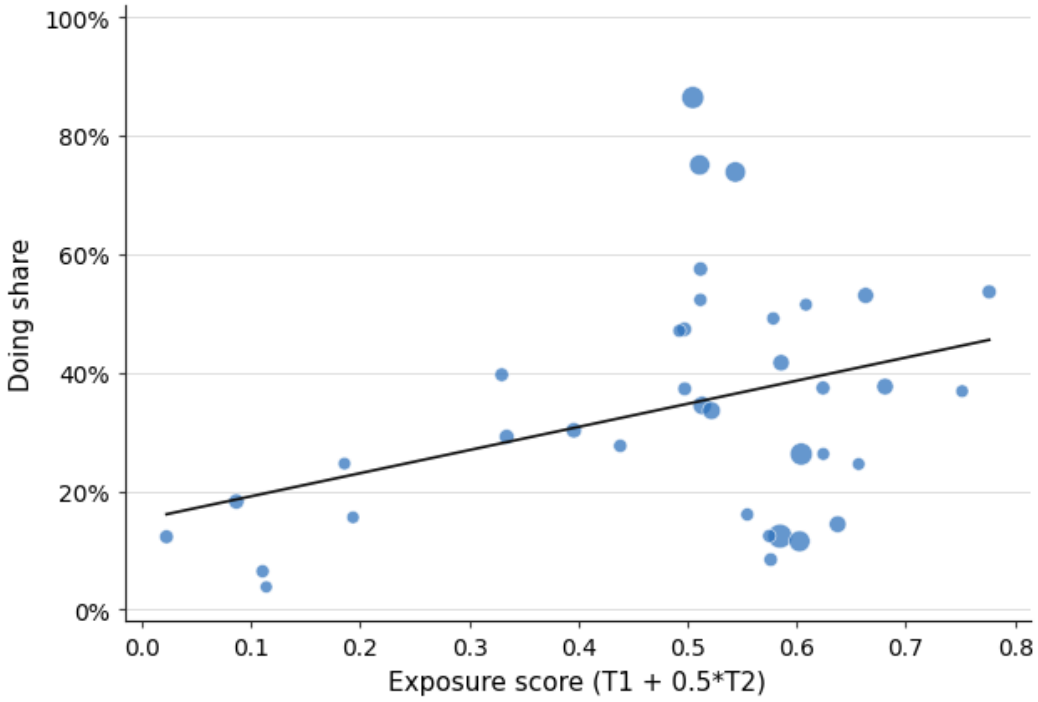
7.4 AI Advantage

Part (i) of Corollary 1 predicts that the share of interactions classified as *Doing* rather than *Asking* will be increasing in the AI advantage Δ. We proxy for task-level AI advantage using the exposure scores of Eloundou et al. (2024), which measure whether an LLM can substantially reduce completion time for O*NET tasks.[38] Assuming AI is more capable (e.g. lower σ_A^2) in these tasks, our model predicts that they will have a relatively higher *Doing* share.

Panel A of Figure 16 presents a scatterplot of the share of *Doing* messages for each task against the Eloundou et al. (2025) exposure score, with the 332 IWA-level tasks aggregated up to 38 GWAs.[39]

[38] Eloundou et al. (2025) asked GPT-4 and human validators to rate O*NET work tasks according to whether LLMs could decrease the time required to complete the task by at least 50 percent.

[39] We aggregate up because the Eloundou et al. (2025) exposure score is highly discrete, with values of either 0, 0.5, or 1 for each IWA. We use the β version of the Eloundou et al measure because it has the most variation, but our results hold with the other versions as well.



Panel A. Scatterplot: *Doing* share vs. Eloundou et al. (2025) exposure.

	(1)
	Doing share
Exposure score (β model)	0.0142***
	(0.0027)
Constant	0.3378***
	(0.0014)
Observations	1,284,647
R-Squared	0.140

Panel B. OLS regression: *Doing* share on exposure.

Figure 16: Panel A plots, for each O*NET Generalized Work Activity (GWA), the share of annotated ChatGPT messages classified as *Doing* against the GWA's AI exposure score under the β specification from Eloundou et al. (Eloundou et al., 2024). We construct exposure by mapping Eloundou et al.'s task-level automation labels to scores (`T1`=1, `T2`=0.5, `T0`=0), averaging human and GPT annotations when both are available, aggregating tasks to O*NET Intermediate Work Activities (IWAs) with greater weight on core than supplemental tasks, linking O*NET occupations to national BLS employment counts (U.S. Bureau of Labor Statistics, 2024), and then rolling BLS-employment-weighted IWA exposure measures up to the GWA level using the O*NET activity hierarchy. We construct the *Doing* share by assigning annotated messages to IWAs, aggregating them to the same GWA level, and computing the fraction labeled *Doing* among messages with expressing/asking/doing annotations. Each point is a GWA, with point size is proportional to the number of annotated messages in that GWA. Panel B presents the coefficient from a user-by-task level regression of the *Doing* share on the IWA-level exposure score, controlling for user fixed effects.

We find a strong positive correlation between task-level AI exposure and the *Doing* share, consistent with the model's predictions.

Panel B presents a multivariate regression version of the empirical test in Panel A by estimating a user-by-task level regression of the *Doing* share on the IWA-level Eloundou et al. (2025) β exposure score plus user fixed effects. We find that the same user is more likely to delegate tasks to ChatGPT (e.g. *Doing*) when those tasks are more AI-exposed, and the relationship is statistically significant at the less than one percent level. A user is about 1 percentage point more likely to delegate a task to AI rather than ask for advice when the task is in the 90th rather than 10th percentile of AI capability according to Eloundou et al. (2024). The magnitude of the relationship in Panel B is smaller than Panel A because it estimates a within-user slope, which controls for selection of users into different task shares based on occupation, education, and other individual characteristics.

7.5 The Context Gap

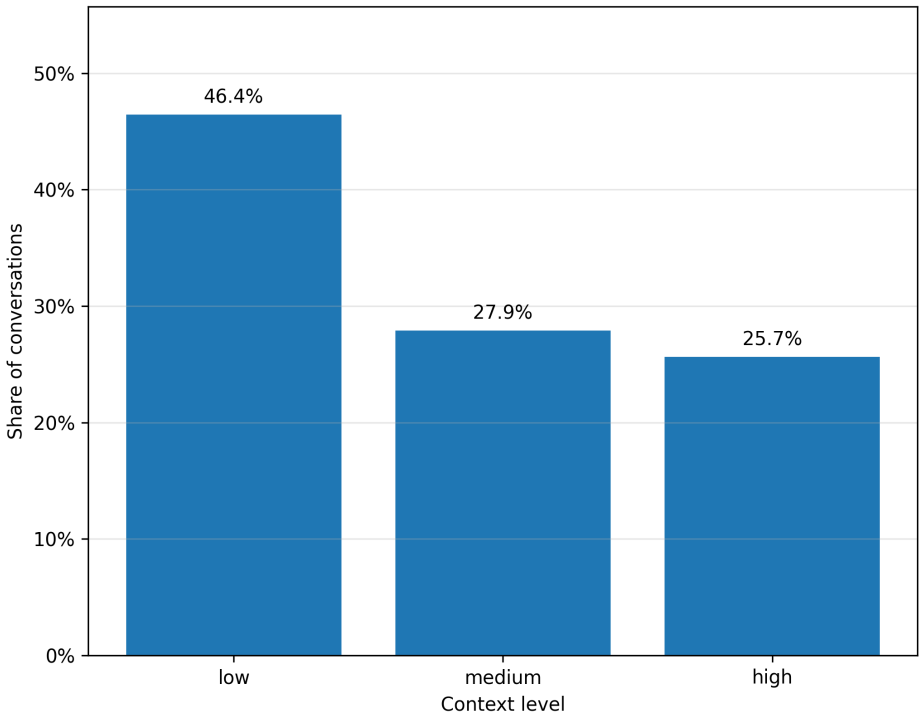
Part (ii) of Corollary 1 predicts that users are less likely to delegate and more likely to ask for advice when a task requires more *context*, meaning user-, organization-, or situation-specific information beyond generic world knowledge.

Context level	Definition
Low	The overall task does not depend at all on external knowledge known to the user but not to the assistant at the start of the conversation. The overall task is a common request we'd expect to see in a chatbot: summarize, rephrase, format, light edit, simple coding with a clear spec, straightforward Q&A.
Medium	The overall task does not depend strongly on external knowledge known to the user but not to the assistant at the start of the conversation. The conversation develops or revises something that is still fairly generic (e.g. a blog post, an essay, standard code on a common stack).
High	The overall task depends strongly on external knowledge known to the user but not to the assistant at the start of the conversation (for example: organization-specific norms, personal financial situation). A generic assistant without access to this external context would struggle to complete the task satisfactorily.

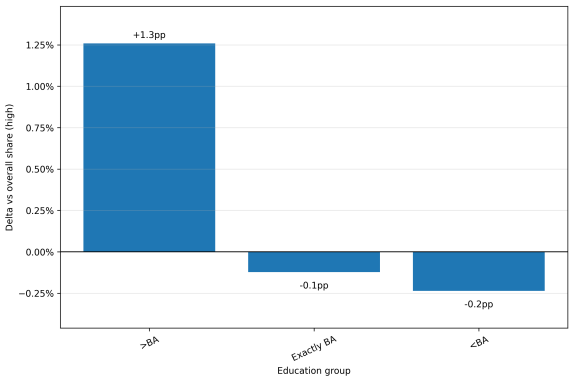
Table 3: Definitions for the conversation context classifier labels. The full classifier prompt text is provided in Appendix B.

Tasks where there is a well-established way of doing things will have a smaller $\tau(\omega)$ and thus are delegation-friendly. Writing boilerplate documents and software code are obvious examples of delegation-friendly tasks, because there is typically a standard, routine approach that most users take (e.g. Autor et al. (2003)). Context may be an important constraint for full delegation of work to AI agents. If tacit knowledge is important for completing a task successfully, the user must successfully transmit that knowledge to the chatbot agent to fully delegate their work.

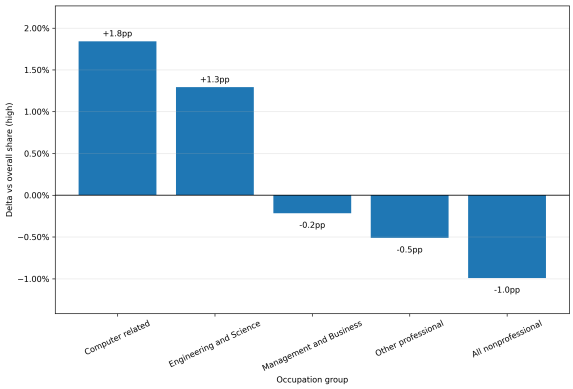
We attempt to measure the importance of context in a given task with a novel conversation-level classifier. The context classifier infers the single underlying real-world task the user is trying to accomplish and judges how much user-, organization-, or situation-specific information beyond general world knowledge would be required for the output of the conversation to be safely and effectively



Panel A. Context distribution (overall).



Panel B. Context distribution by education.



Panel C. Context distribution by occupation.

Figure 17: Conversation context distribution overall and by user characteristics. See Table 3 for the basic definitions and Appendix B for the full context classifier prompt. These results are generated using annotations for messages in the Linked Dataset described in section 3.3. Panel A captures overall context distribution among all users, while Panel B and C capture the percentage point difference, relative to the full sample, in high context message shares (as opposed to low or medium) for each subgroup of education and occupation. Not shown: context shares for the subset of users for whom education level was Ambiguous or Missing.

plugged into a real-world workflow. Table 3 presents the key text of the prompt, with the full text provided in Appendix B.[40] We ran the context classifier on all messages in our linked dataset.

Panel A of Figure 17 presents the distribution of context levels across conversations in the linked sample. 46 percent of conversations are low context, 28 percent are medium context, and 26 percent are high context. Panels B and C of Figure 17 show how conversation context varies by education and occupation respectively. Educated users and users in highly-paid professional and technical occupations have higher-context conversations. Users with graduate degrees are 1.5 percentage points more likely than users without a four-year college degree to have high-context conversations, and computer and engineering workers are 2.8 and 2.3 percentage points more likely than workers in non-professional occupations to have high-context conversations.

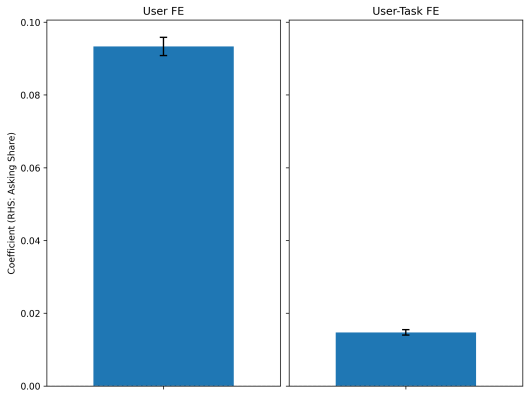


Figure 18: Presents the coefficients and 95% confidence intervals from a message-by-user level regression of the user's *Asking* intent on indicators for whether the conversation context is high (with low and medium combined as the left-out category), controlling for user and user-by-task fixed effects. Results are in Panels A and B respectively. Sample size is 6,180,984 conversations within the linked sample described in Section 3.3.

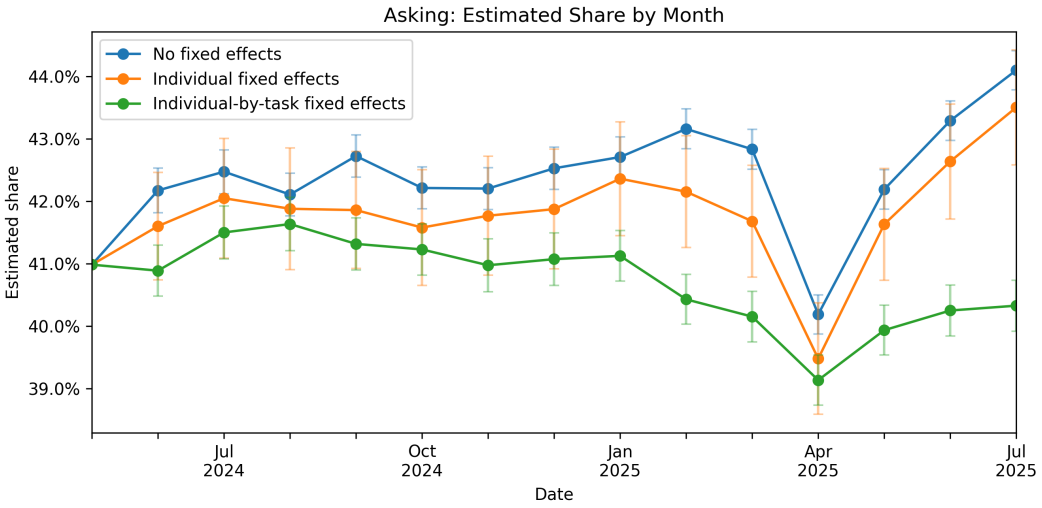
Part (ii) of Corollary 1 predicts that users performing high-context tasks are more likely to ask the chatbot for advice rather delegate tasks directly. Figure 18 tests this prediction with a message level regression of the user's *Asking* intent on indicators for whether the conversation is high context (with low or medium context as the left-out category), controlling for user and user-by-task fixed effects.[41]

The lefthand bar presents results from the user fixed effects specification. The same user is 9.3 percentage points more likely to ask for advice rather than delegate a task directly when the conversation is high context. The estimate is statistically significant at the less than one percent level.

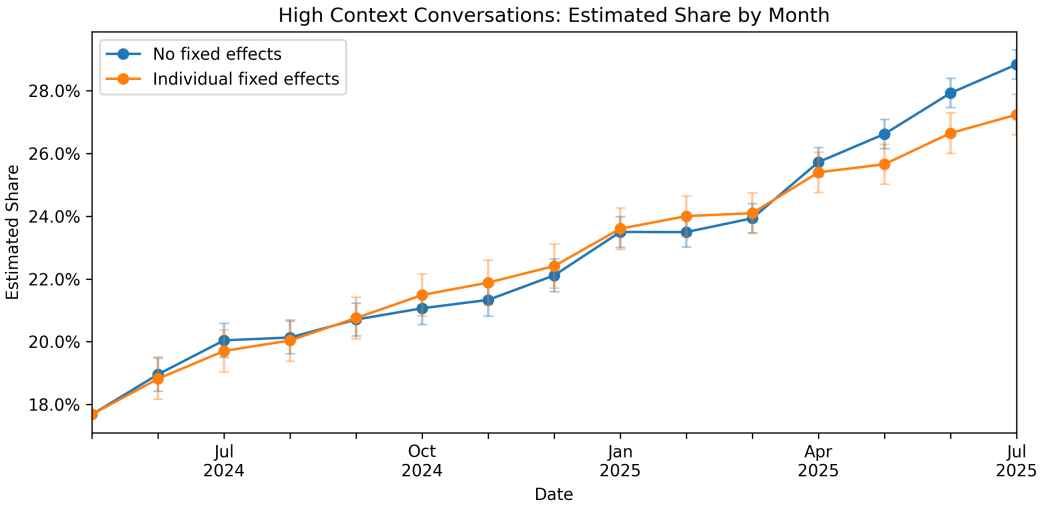
The righthand bar adds user-by-task fixed effects, where tasks are measured as one of the 338 IWAs. The same user performing the same very detailed work task is still about 1.6 percentage points more likely to ask rather than delegate when the conversation is high versus low or medium context, and the estimate is statistically significant at the less than one percent level. These results are consistent with the model and confirm the importance of *context* in driving user behavior. Users are much less willing to delegate high-context work assignments to an AI chatbot, which suggests that context may be a key bottleneck to full automation of job tasks.

[40] Reassuringly, we also find that higher context conversations have higher average word counts.

[41] we combine low and medium for simplicity, but our results are very similar when we compare low context to medium and high combined.



Panel A. Shares of messages classified as *asking*, by month.



Panel B. Share of conversations classified as high context, by month.

Figure 19: Time trends in the share of messages (conversations) that are classified as *Asking* (*High Context*), respectively. Figures were generated by regressing a binary indicator for whether the message (conversation) was *Asking*(*High Context*) against an intercept and month dummies (with May 2024 as the left-out category). Figures were generated by plotting the constant for the first month, and constant plus month indicators for subsequent months. The overall baseline was determined by the specification without fixed effects. Standard errors clustered at the level of the fixed effect.

How is user behavior changing over time? Panel A of Figure 19 plots the share of all messages that are *Asking* by month. The blue lines show the overall shares in our linked sample, while the orange and green lines control for user and user-by-task fixed effects respectively. The share of all user messages that are asking increased from 41% in May 2024 to more than 44% in July 2025. Adding user fixed effects has little impact on the results, which suggests that users are sending relatively more asking messages over time as they gain experience with ChatGPT. However, we find a slight *decrease* in the asking share when controlling for user-by-task fixed effects. This suggests that the overall rise in the asking share is due to a shift in users' task mix.

Panel B of Figure 19 plots the monthly share of user conversations that are high context. Between May 2024 and July 2025 the share of conversations that are high context increased from less than 18% to more than 28%, a marked increase. This fixed effects specification suggests that about 80% of the increase in context was within-user rather than across users.

Overall, Figure 19 shows that ChatGPT user behavior has shifted toward higher-context conversations and tasks over time, which helps explain why delegation hasn't increased even as model capabilities improve. Users have responded to model improvement by shifting toward higher-context tasks where *Asking* is relatively more beneficial. Increasing demand for high-context tasks may also explain the growing popularity of AI chatbots that are embedded into organization workflows, where context is more easily observed (Appel et al., 2025; Dillon et al., 2025).

8 Conclusion

ChatGPT has grown rapidly since its launch in November 2022, and by December 2025 it had more than 800 million weekly users who were collectively sending more than 3 billion messages per day. Yet despite the rapid adoption of ChatGPT and Generative AI more broadly, there is little evidence on how this new technology is used and who is using it.

This paper studies how people use ChatGPT by classifying a large random sample of ChatGPT messages according to features like conversation topic and job task, and linking them to user demographic data with privacy-preserving methods. No user messages were never directly observed by the researchers.

We first show that ChatGPT usage has grown rapidly since early 2025, both by adding new users and by increasing usage intensity among existing cohorts of users. ChatGPT usage was initially concentrated among users with male names and in high income countries, but those gaps closed substantially by mid-2025. Nearly 75% of all ChatGPT messages around the world are sent by users under the age of 35.

Nearly 80% of ChatGPT messages fall into three conversation topics that we call *Practical Guidance*, *Seeking Information*, and *Writing*. The most common sub-category within *Practical Guidance* is *Tutoring and Teaching* at 10% of all messages. Less than 4% of messages are related to computer programming. The most common work-related topic is *Writing* at 40% of work messages. More than half of *Writing* messages are requests to edit, summarize, translate, or otherwise modify user text rather than generating new text from scratch.

Non-work usage has grown faster than work usage and accounted for nearly three-quarters of all

messages as of late 2025, although ChatGPT users on consumer plans still sent about 800 million work-related messages each day. We map messages to O*NET work tasks and find that work usage is highly concentrated in a small number of tasks related to information processing and decision support. The most common work tasks vary surprisingly little by user occupation. Overall, we find striking evidence that workers ask for information and advice just as often as they delegate work output to ChatGPT directly.

To understand these patterns, we develop a principal-agent model where users can delegate a job task to an AI chatbot, ask the chatbot for advice, or perform the task themselves. A key predictor of user behavior in the model is *context*, meaning how much information must be transmitted to the chatbot so that it can adequately customize its output to the user's needs. Consistent with the model's predictions, we find that users are more likely to ask for advice rather than delegate tasks when the underlying task they are performing has higher context requirements. We also find that users are more likely to delegate to the chatbot in tasks where it is more capable.

Importantly, we find that ChatGPT conversations are becoming higher-context over time, suggesting a shift into more difficult tasks and greater conversational complexity. Thus, achieving greater economic benefits from AI may require not only improvements in model capabilities, but also more efficient and deeper transmission of human context.

References

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

_ **and Pascual Restrepo**, "Tasks, automation, and the rise in US wage inequality," *Econometrica*, 2022, *90* (5), 1973–2016.

Appel, Ruth, Peter McCrory, Alex Tamkin, Miles McCain, Tyler Neylon, and Michael Stern, "Anthropic economic index report: Uneven geographic and enterprise ai adoption," *arXiv preprint arXiv:2511.15080*, 2025.

Autor, David and Neil Thompson, "Expertise," *Journal of the European Economic Association*, 2025, p. jvaf023.

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.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond, "Generative AI at work," *The Quarterly Journal of Economics*, 2025, *140* (2), 889–942.

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.

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.

Cruces, Guillermo, Diego Fernández Meijide, Sebastian Galiani, Ramiro H Gálvez, and María Lombardi, "Does Generative AI Narrow Education-Based Productivity Gaps? Evidence from a Randomized Experiment," Technical Report, National Bureau of Economic Research 2026.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," Technology & Operations Management Unit Working Paper No. 24-013, Harvard Business School September 2023.

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.

Dillon, Eleanor W, Sonia Jaffe, Nicole Immorlica, and Christopher T Stanton, "Shifting work patterns with generative ai," Technical Report, National Bureau of Economic Research 2025.

Dwork, Cynthia and Aaron Roth, "The algorithmic foundations of differential privacy," *Foundations and trends® in theoretical computer science*, 2014, *9* (3-4), 211–487.

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.

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

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.

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

Jahani, Eaman, Benjamin S Manning, Joe Zhang, Hong-Yi TuYe, Mohammed Alsobay, Christos Nicolaides, Siddharth Suri, and David Holtz, "Prompt Adaptation as a Dynamic Complement in Generative AI Systems," *arXiv preprint arXiv:2407.14333*, 2024.

_ **, Benjamin S. Manning, Joe Zhang, Hong-Yi TuYe, Mohammed Alsobay, Christos Nicolaides, Siddharth Suri, and David Holtz**, "Prompt Adaptation as a Dynamic Complement in Generative AI Systems," 2024. Last revised January 2026 (v7).

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.

Li, Ninghui, Wahbeh Qardaji, and Dong Su, "On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy," in "Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security" 2012, pp. 32–33.

Lin, Sijie, "Learning to prompt: Human adaptation in production with generative ai," *Available at SSRN 5788402*, 2025.

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.

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

OpenAI, "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.

_ , "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.

Sarkar, Suproteem K, "Ai agents, productivity, and higher-order thinking: Early evidence from software development," *Available at SSRN 5713646*, 2025.

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

U.S. Bureau of Labor Statistics, "Occupational Employment and Wage Statistics, May 2023 National Occupational Employment and Wage Estimates," `https://www.bls.gov/oes/current/oes_nat.htm` 2024. Accessed December 17, 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.

Vendraminelli, Luca, Matthew DosSantos DiSorbo, Annika Hildebrandt, Edward McFowland III, Arvind Karunakaran, and Iavor Bojinov, "The GenAI Wall Effect: Examining the Limits to Horizontal Expertise Transfer Between Occupational Insiders and Outsiders," *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, 2025, (26-011), 26–011.

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.

Yang, Jeremy, Noah Yonack, Kate Zyskowski, Denis Yarats, Johnny Ho, and Jerry Ma, "The Adoption and Usage of AI Agents: Early Evidence from Perplexity," *arXiv preprint arXiv:2512.07828*, 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: Additional Figures

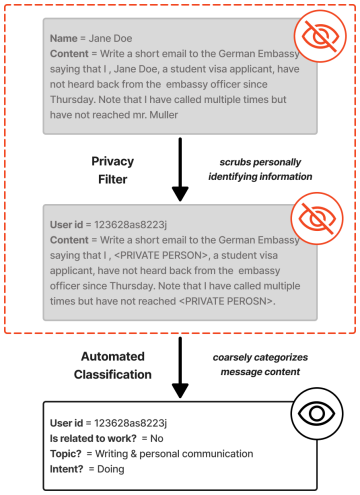


Figure A.1: Illustration of Privacy-Preserving Automated Classification Pipeline (Synthetic Example). Messages are first stripped of PII via an internal LLM-based tool called *Privacy Filter*. Then they are classified by LLM-based automated classifiers, described in detail in Appendices B and C. Humans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

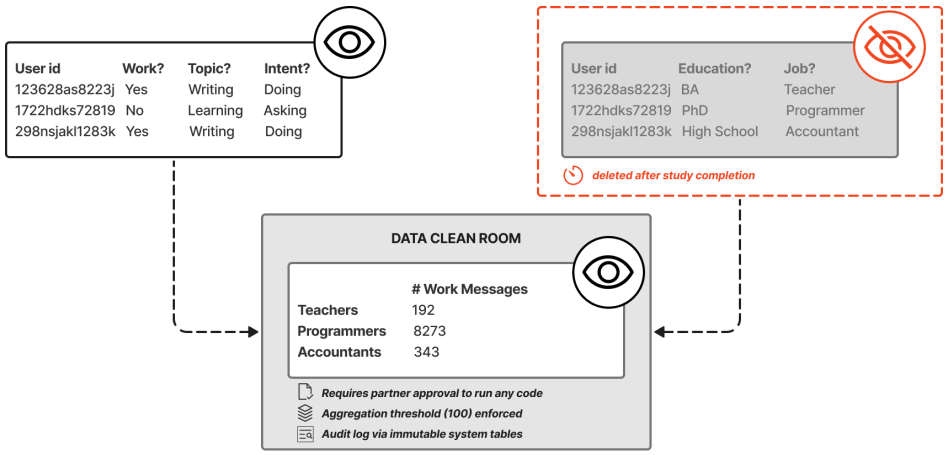


Figure A.2: Illustration of Aggregated Employment Category Analysis via a Data Clean Room. All queries run in the Data Clean Room must be approved by our data partner, enforcing a strict aggregation threshold (100 observations). As a result, researchers cannot access user-level employment data, only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

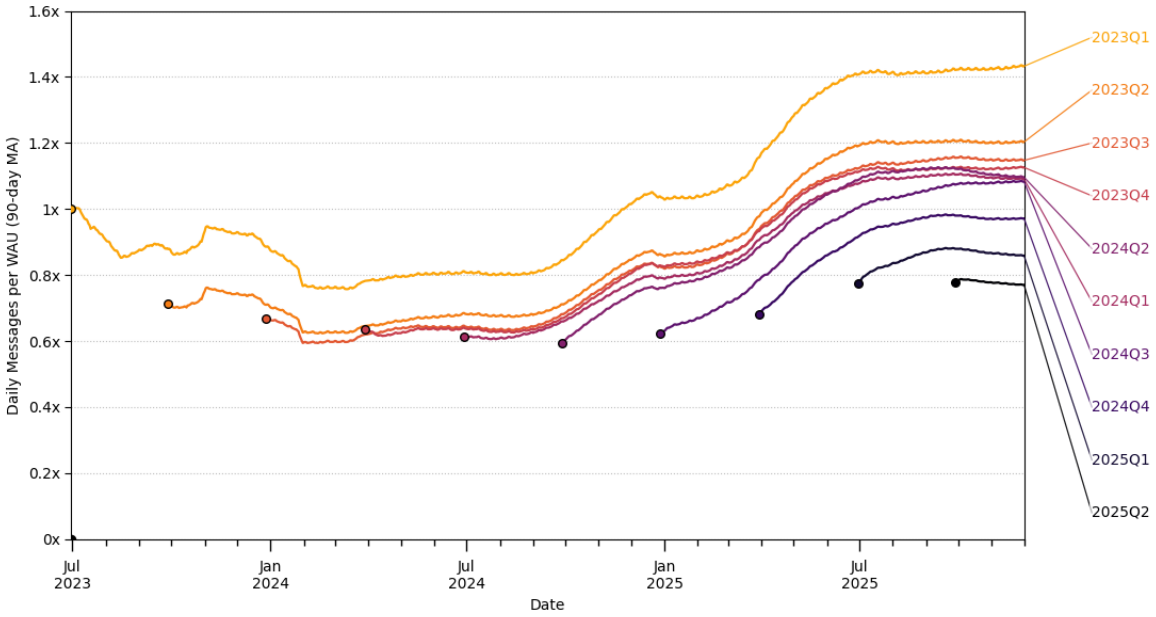


Figure A.3: Daily messages sent per weekly active user, split by sign-up cohort. Sample only considers users of ChatGPT consumer plans (Free, Plus, Pro). Reported values are moving averages of the past 90 days and are reported starting 90 days after the cohort is fully formed. Y-axis is an index normalized to the first reported value for the Q1 2023 cohort.

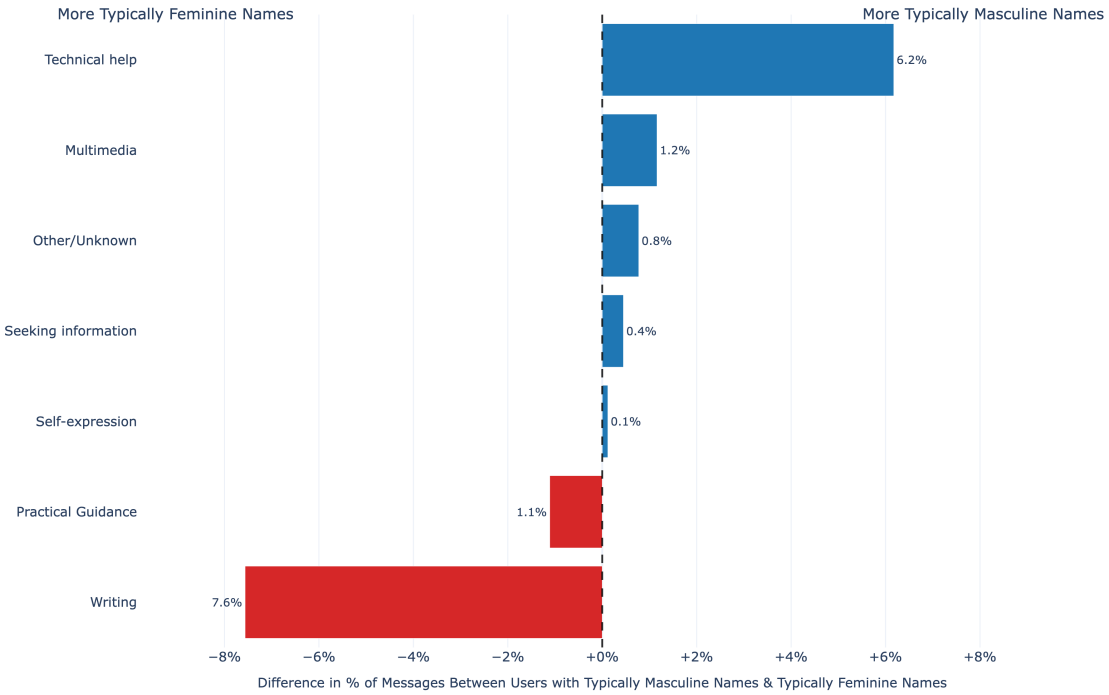


Figure A.4: Difference in share of topic prevalence in messages by users with typically masculine/feminine first name. We draw on a uniform sample of 1.1M ChatGPT accounts, subject to the same user exclusion principles as other datasets we analyze. Note that this is a separate sample than those described 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 B.

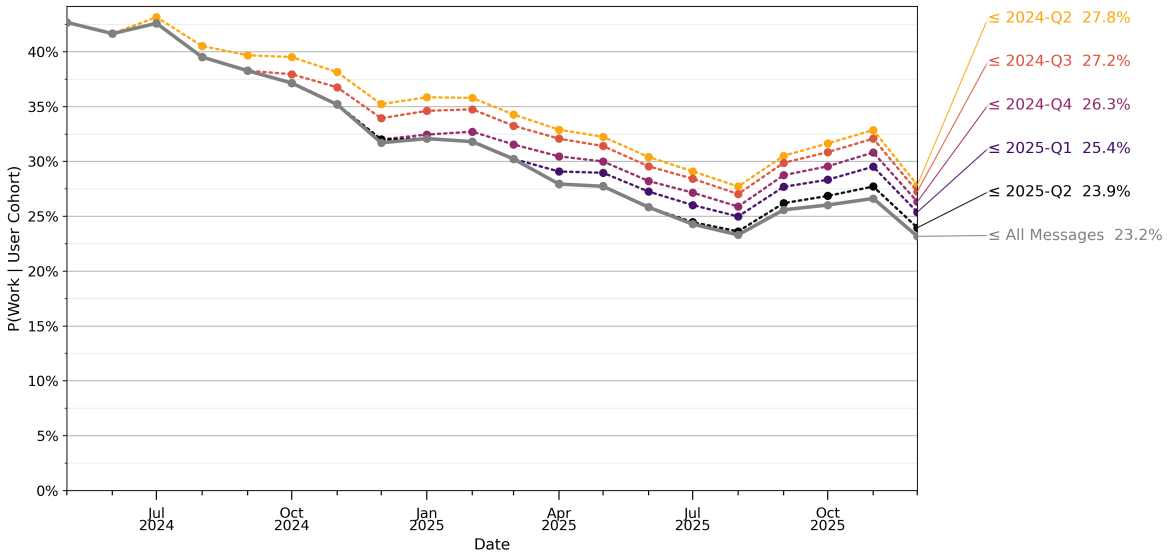


Figure A.5: The solid gray line represents the probability that a messages on a given day is related to work, as determined by an automated classifier. Values are bucketed by month. The dotted orange line shows the same calculation, but conditioned on messages being from users who first used ChatGPT 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 2 million sampled conversations from May 2024 through December 2025. Observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3.

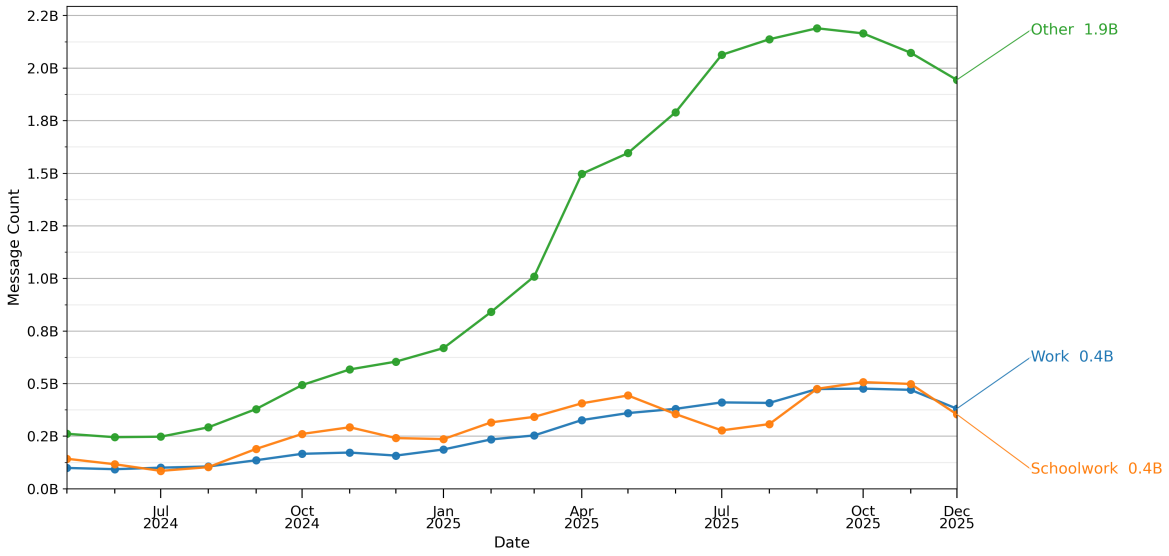
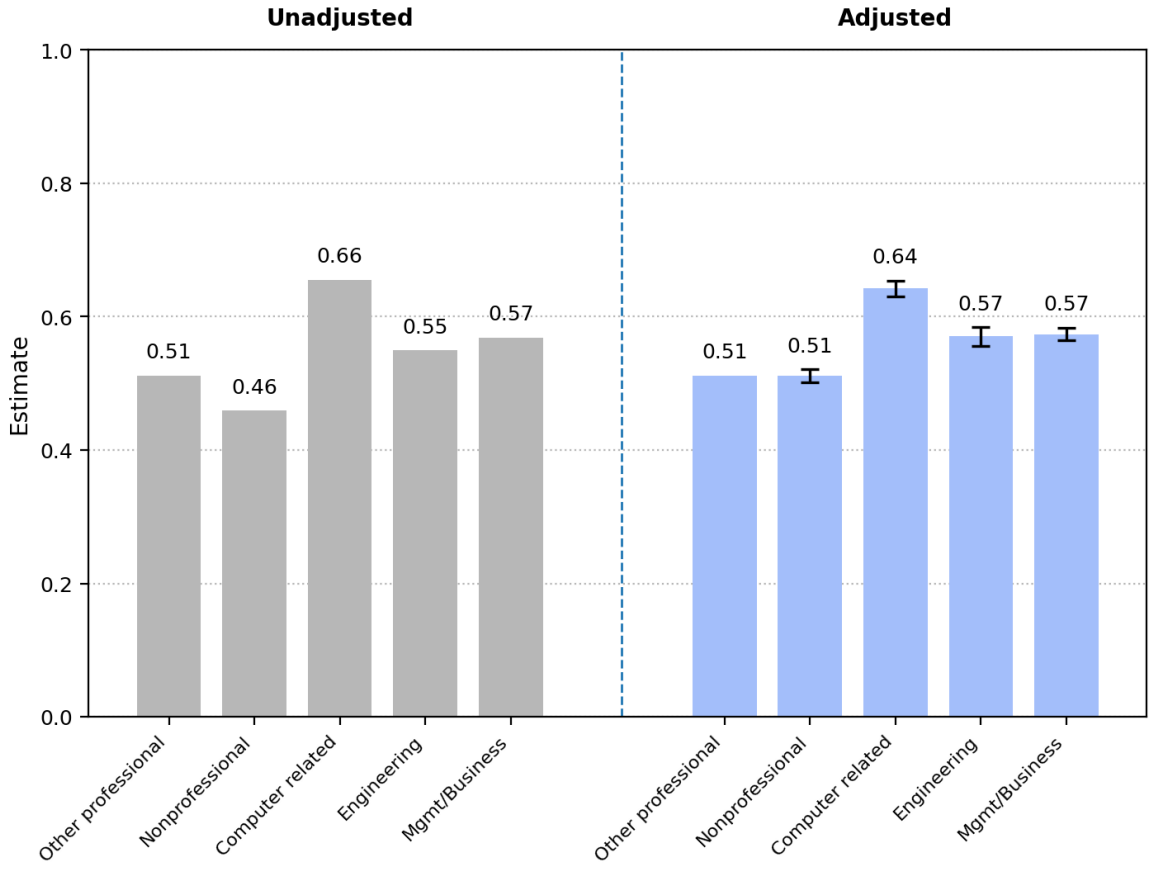
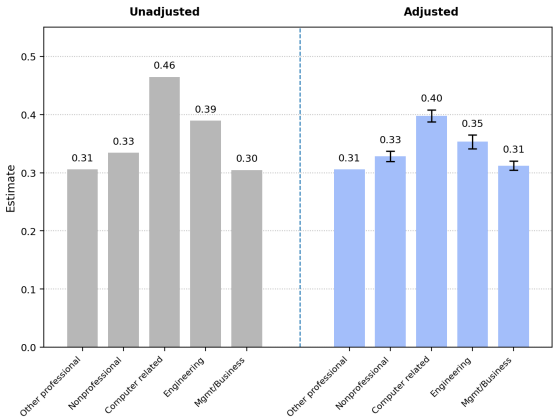


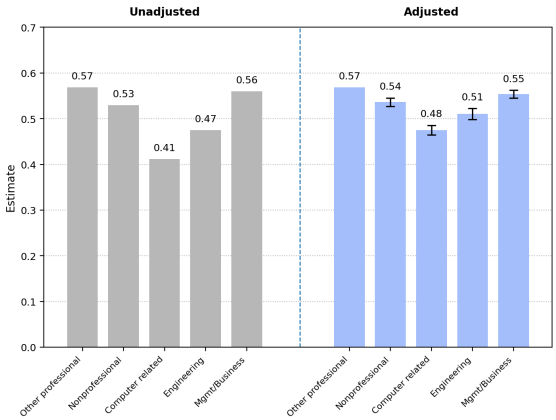
Figure A.6: Daily message-share trends under an alternative three-way classifier that labels messages as work, schoolwork, or other. The verbatim classifier prompt is provided in Appendix B.6. Reported values are monthly averages from May 2024 to December 2025, estimated from sampled conversations and reweighted to match daily total message volumes.



Panel A. Work Related

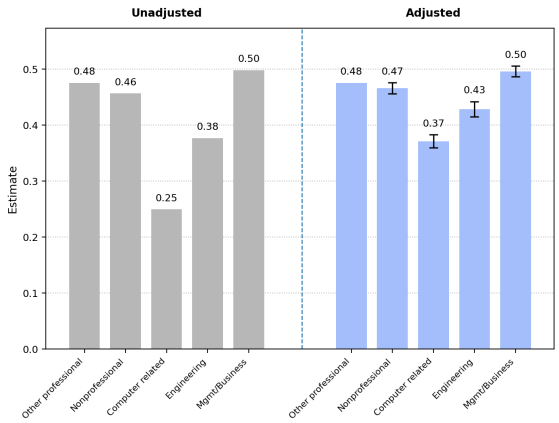


Panel B1. Asking.

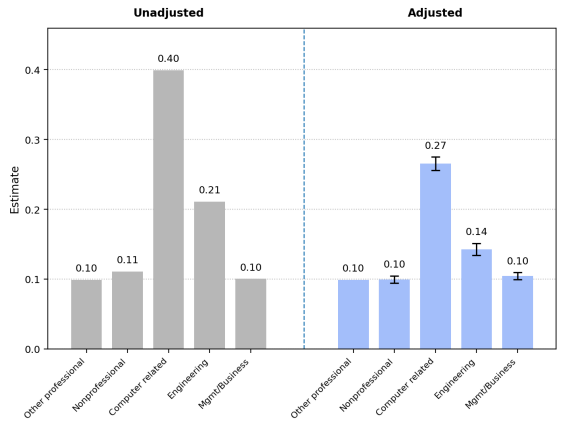


Panel B2. Doing.

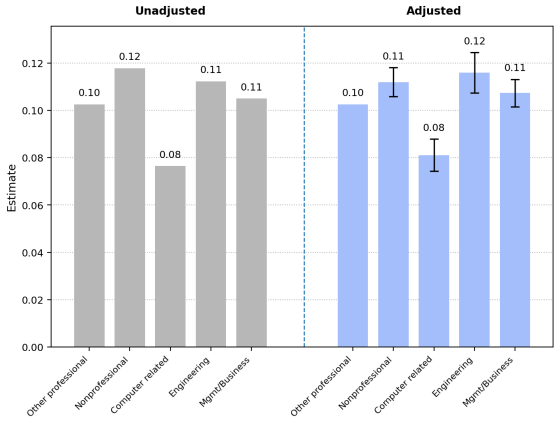
Figure A.7: *(continued on next page)*



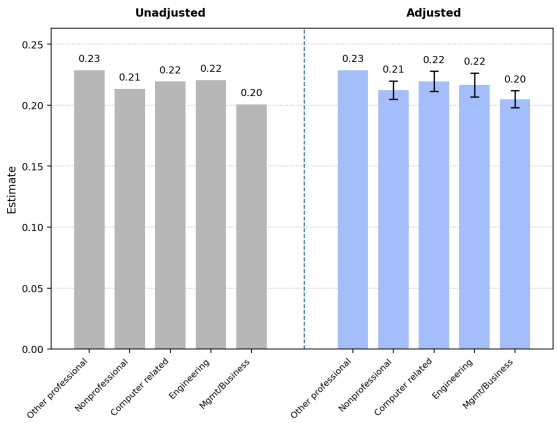
Panel C1. Writing.



Panel C2. Technical Help.



Panel C3. Seeking Information.



Panel C4. Practical Guidance.

Figure A.7: 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 A.7.

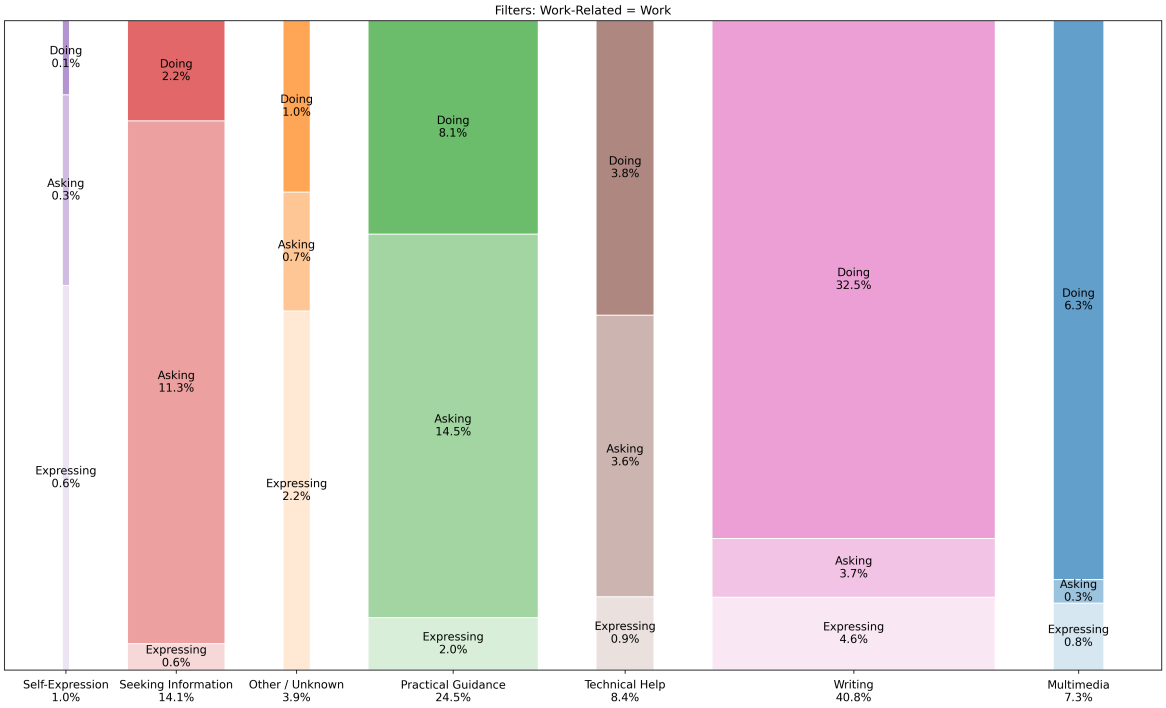


Figure A.8: Breakdown of Conversation Topics by Asking/Doing/Expressing category for **only work-related messages**, with topic columns sorted by relative share of "Doing" messages. Prompts for these automated classifiers are available in Appendix B. For a detailed breakdown of conversation topic contents, see Table 1. Each bin reports a percentage of the total population. Shares are calculated from a sample of approximately 1.2 million sampled conversations from January to December 2025. Observations are reweighted to reflect total message volumes on a given day, correcting for uneven sampling effects. This sample is a subset of the 2 million message sample described in Section 3.2. Message-level robustness version: Figure A.20.

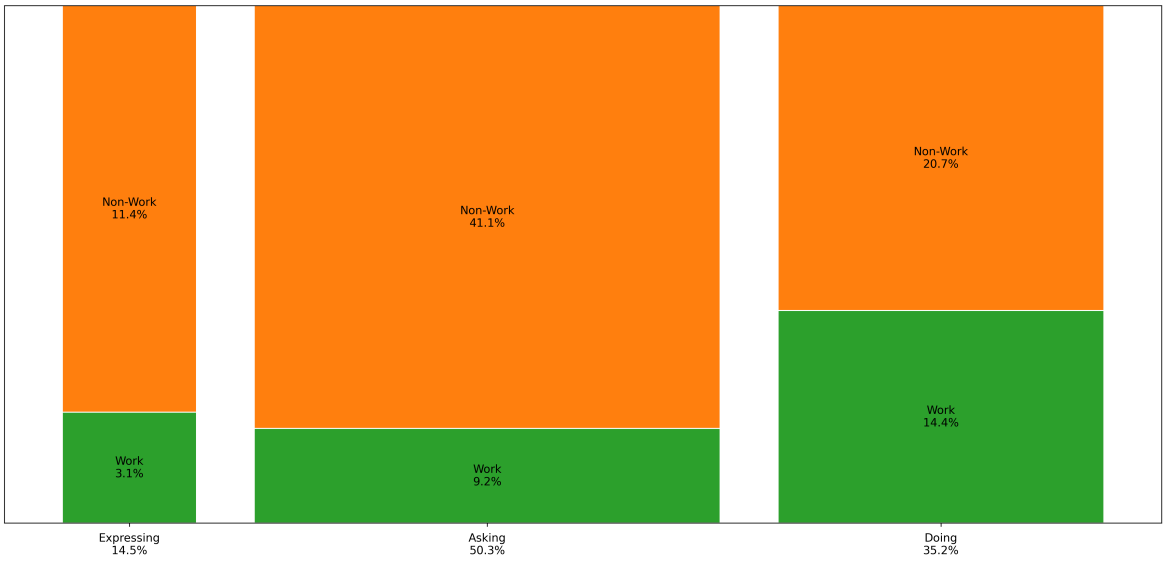


Figure A.9: Shares of Asking, Doing, and Expressing messages split by work vs. non-work. See B to review the prompts used by the automated classifiers. The annotations on the right show the shares of work and non-work for the full sample. Each bin reports a percentage of the total population. Shares are calculated from a sample of approximately 1.2 million sampled conversations from January to December 2025. Observations are reweighted to reflect total message volumes on a given day, correcting for uneven sampling effects. This sample is a subset of the 2 million message sample described in Section 3.2. Message-level robustness version: Figure A.21.

A.1 Results using Message-Level Sampling

The principal results in this paper are based on a sample of messages selected randomly at the conversation level, using a procedure outlined in Section 3. We highlight that, by sampling at this level, we can reflect the set of distinct user activities performed on the platform.

In this section, we present alternative figures based on a sample of performed randomly at the message level. By sampling in this way, we are able to better approximate the intensity of usage within each task, increasing the importance of tasks that require more messages relative to those that are shorter in duration.

As in the main sample, we continue to include context about the conversation outside of the message, by including the nine messages preceding each sampled message as context when classifying messages.

The time frame captured in this data is narrower than that used in the conversation-level sample, beginning in July 2024. For each month in the resulting sample, we include 100,000 messages in the sample that we classify, as in the conversation-level sample. Aside from this difference, the restrictions we impose on the sample used in this appendix otherwise match those described in Section 3.

We present two sets of analyses in this Appendix. In Section A.1.1, we present robustness versions of figures presented in the main text, which facilitate comparisons between the conversation-level and message-level samples. In Section A.1.2, we present additional analyses based on this message-level sample, which do not correspond to conversation-level figures.

The statistics underlying these figures are available for download as a part of the accompanying OpenAI Signals release. To protect users' privacy, these statistics were infused with noise to provide a mathematical guarantee that it is sufficiently hard to reveal information about the messages used to construct them. We describe the process used to calibrate the amount of noise added in Appendix G.2.

A.1.1 Message-Level Sampling Robustness Figures

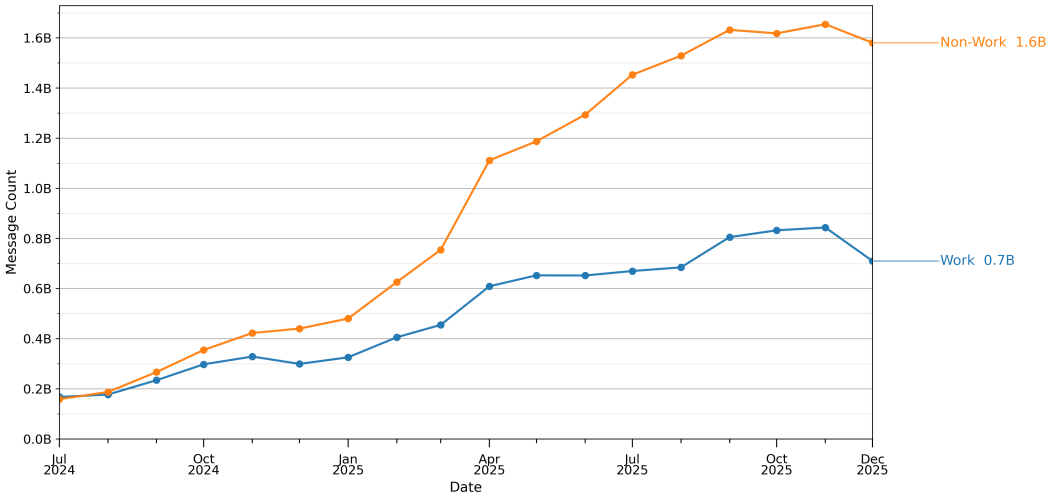


Figure A.10: ChatGPT daily message counts (billions), broken down by likely work-related or non-work-related. Total daily counts are exact measurements of message volume from all consumer plans. Daily counts of work and non-work related messages are estimated by classifying a random sample of messages from that day. Sampling excludes users who opt out of sharing messages for model training, users who self-report an age under 18, logged-out users, deleted conversations, and deactivated or banned accounts. Reported values are monthly averages from July 2024 to December 2025. Estimates are based on approximately 1.8 million sampled messages. Daily totals used for scaling apply the same logged-in and eligibility filters before computing shares.

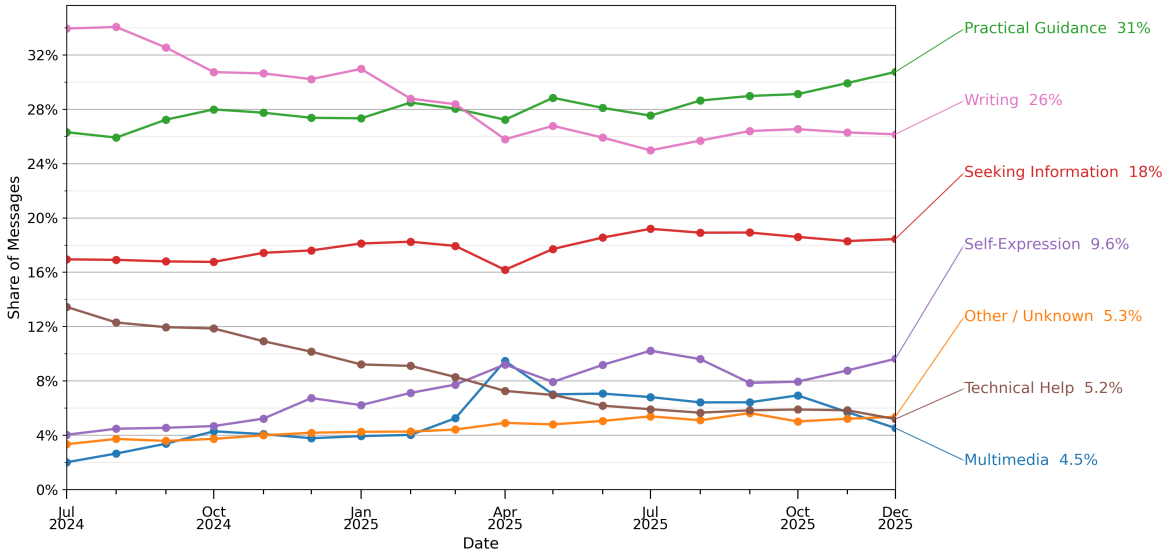


Figure A.11: Share of consumer ChatGPT messages broken down by high-level topic, according to the mapping in Table 1. Values are bucketed by month. Shares are calculated from approximately 1.8 million sampled messages from July 2024 through December 2025. Observations are reweighted using daily totals that apply logged-in and eligibility filters before computing shares.

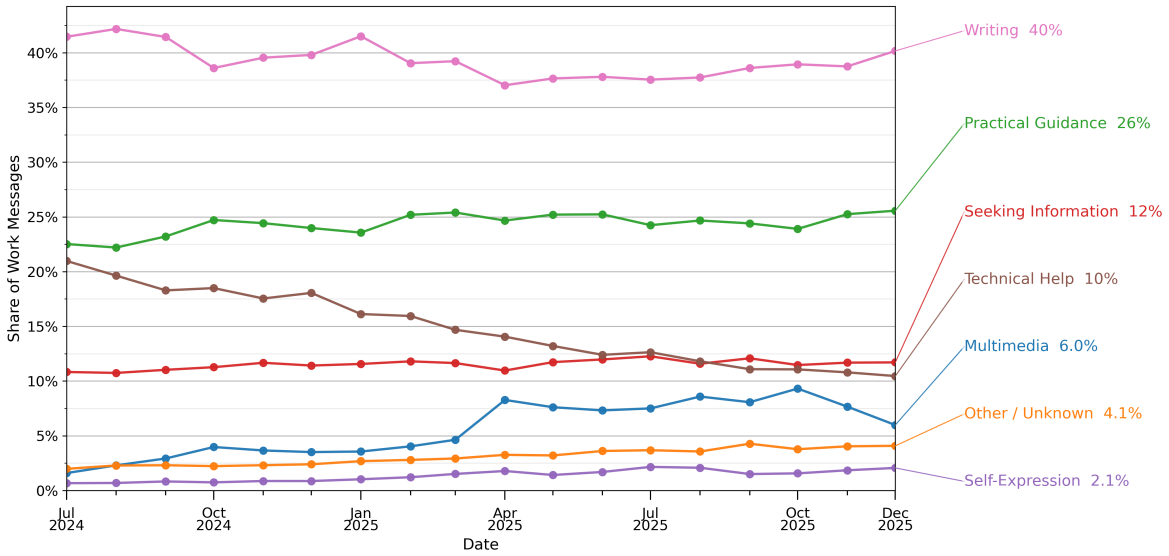


Figure A.12: Share of **work-related** consumer ChatGPT messages broken down by high-level topic, according to the mapping in Table 1. Values are bucketed by month. Shares are calculated from approximately 1.8 million sampled messages from July 2024 through December 2025. Observations are reweighted using daily totals that apply logged-in and eligibility filters before computing shares.

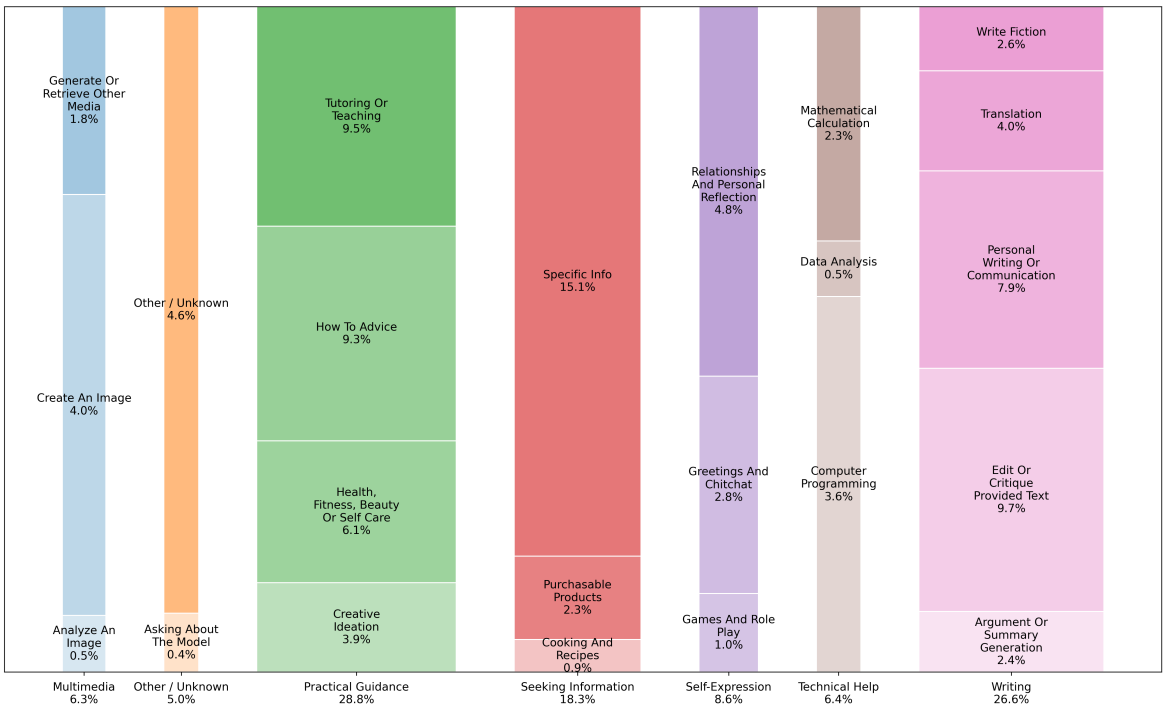


Figure A.13: Breakdown of granular message topic shares within the coarse mapping defined in Table 1. Each bin reports a percentage of the total population. Shares are calculated from approximately 1.2 million sampled messages from January to December 2025, and daily scaling totals apply logged-in and eligibility filters before computing shares.

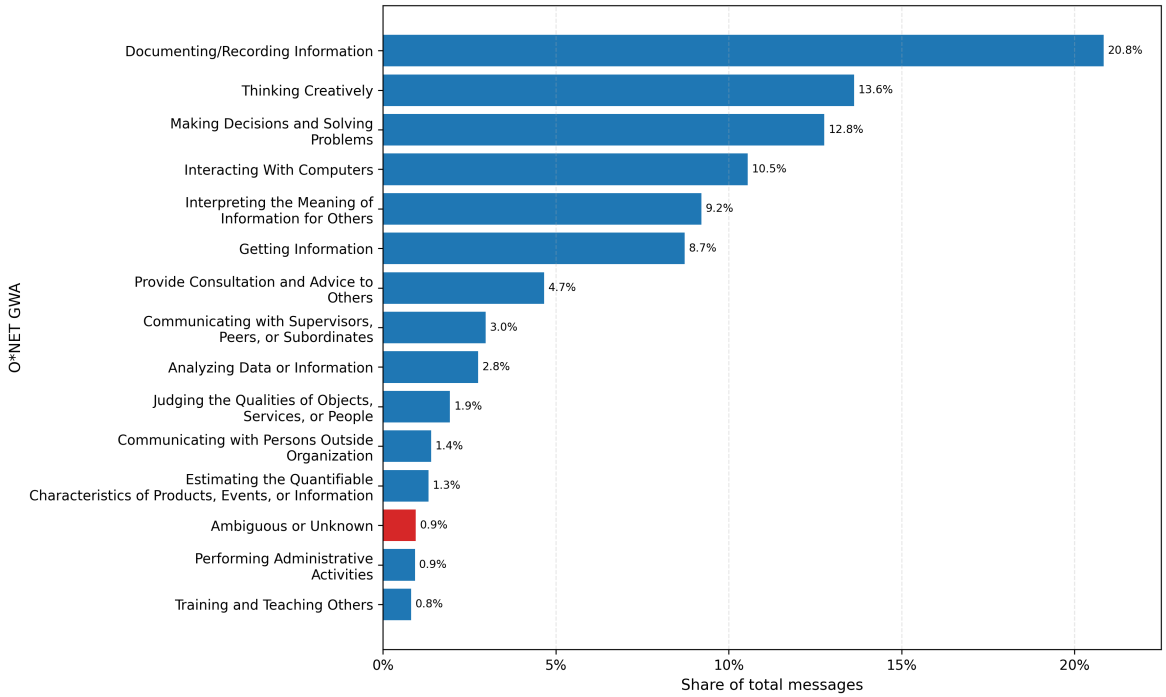


Figure A.14: GWA shares for messages classified as work-related. Messages are classified as pertaining to one of 332 O*NET IWAs or *Ambiguous*, then aggregated to GWAs using the O*NET Work Activities taxonomy. Estimates use the message-level sample of approximately 1.2 million messages from January through December 2025, with daily scaling totals applying logged-in and eligibility filters before computing shares.

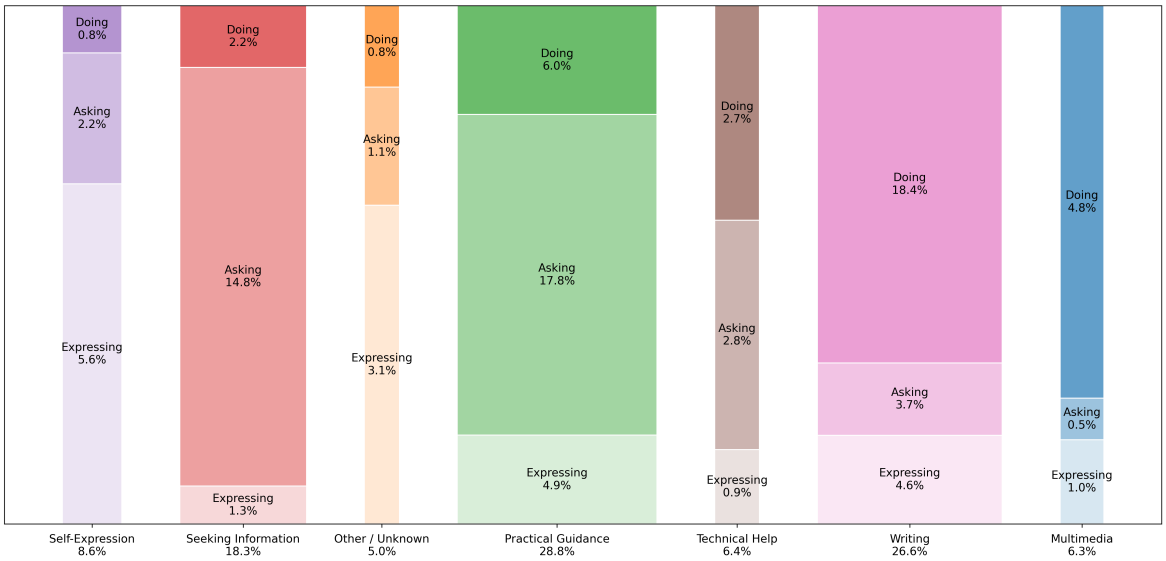


Figure A.15: Breakdown of Message Topics by Asking/Doing/Expressing category, with topic columns sorted by relative share of "Doing" messages. Each bin reports a percentage of the total population. Shares are calculated from approximately 1.2 million sampled messages from January through December 2025, and daily scaling totals apply logged-in and eligibility filters before computing shares.

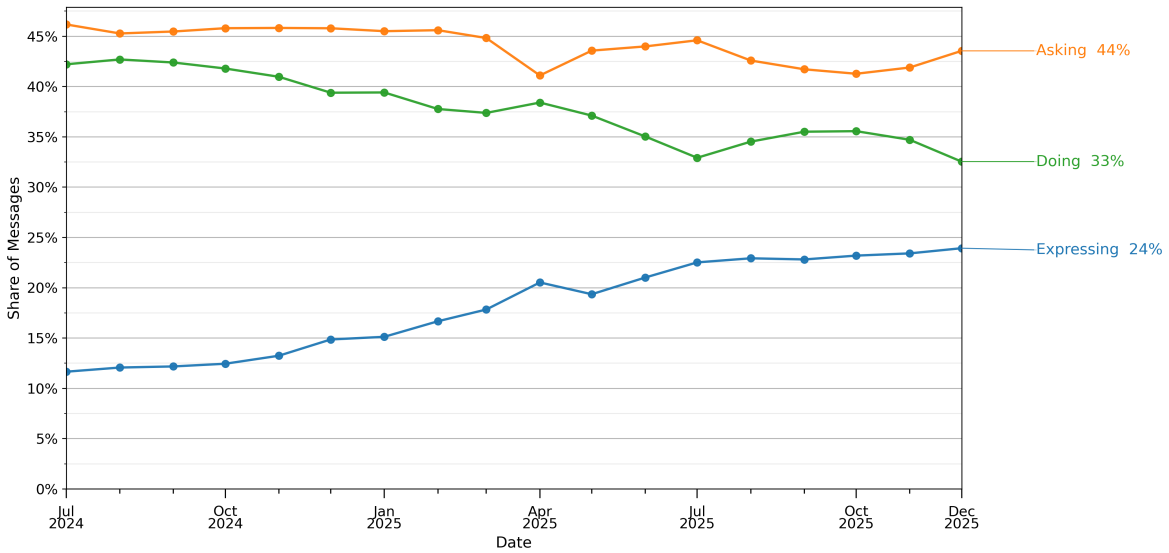


Figure A.16: Shares of messages classified as Asking, Doing, or Expressing by an automated ternary classifier. Values are bucketed by month. Shares are calculated from approximately 1.8 million sampled messages from July 2024 through December 2025. Observations are reweighted using daily totals that apply logged-in and eligibility filters before computing shares.

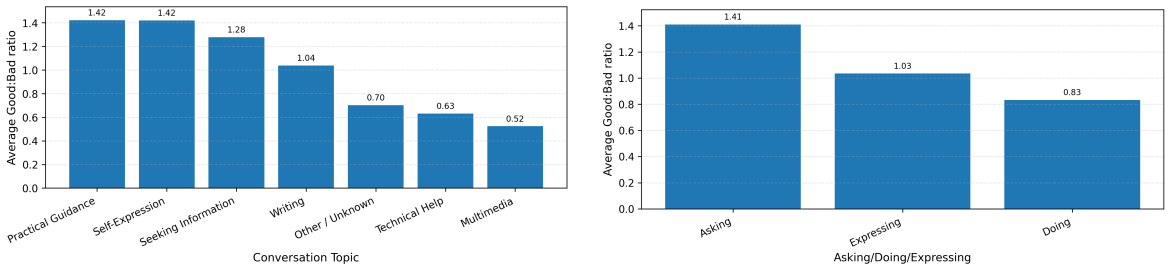


Figure A.17: Average *Good* to *Bad* ratio for user interactions by Message Topic (Panel A) and Asking/Doing/Expressing classification (Panel B). Values represent averages from January to December 2025 using approximately 1.2 million sampled messages and daily scaling totals that apply logged-in and eligibility filters before computing shares.

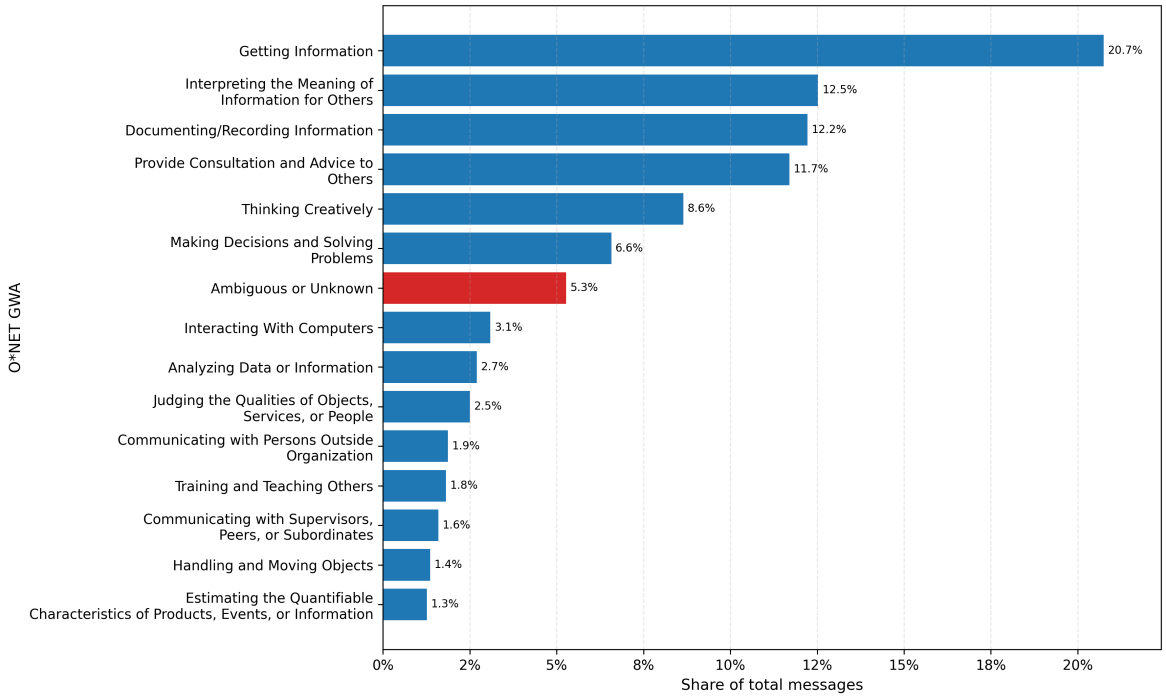


Figure A.18: GWA Shares of 1.2M ChatGPT Messages. Messages are classified as pertaining to one of 332 O*NET IWAs, or *Ambiguous* using the prompt provided in the Appendix. IWAs were then aggregated to GWAs using the O*NET Work Activities taxonomy. Message sample from January through December 2025. GWAs representing a small portion of overall use were combined into a "Minor Categories" group for each column.

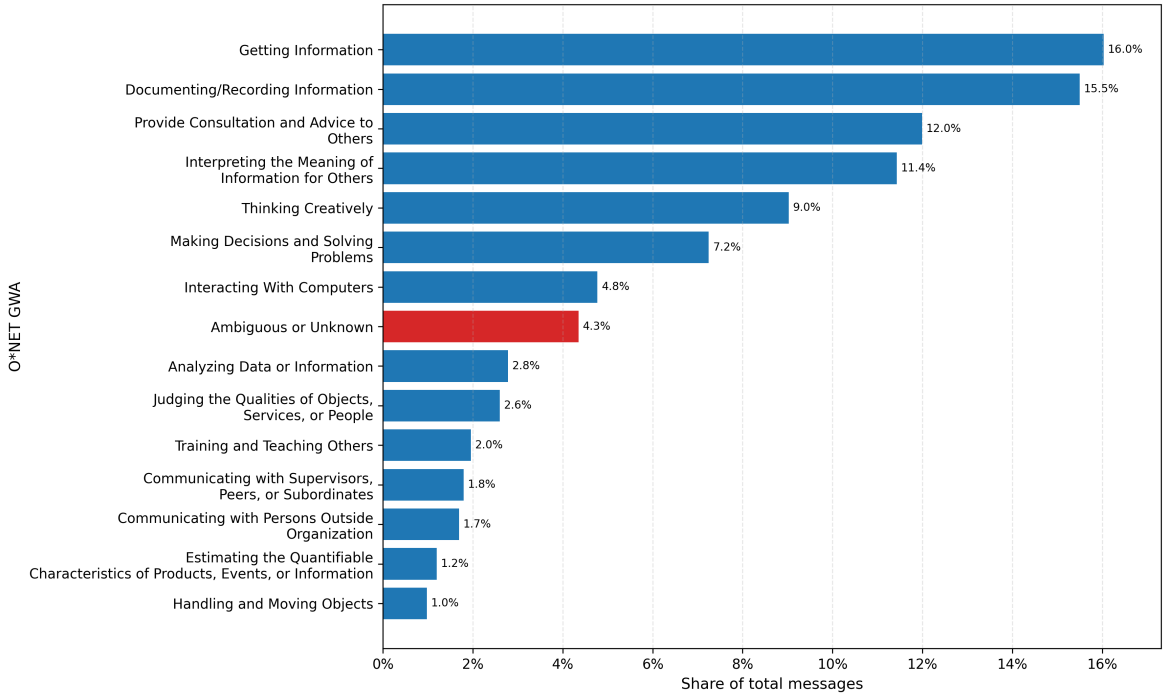


Figure A.19: GWA shares across all messages. Messages are classified as pertaining to one of 332 O*NET IWAs or *Ambiguous*, then aggregated to GWAs using the O*NET Work Activities taxonomy. Estimates use approximately 1.2 million sampled messages from January through December 2025, and daily scaling totals that apply logged-in and eligibility filters before computing shares.

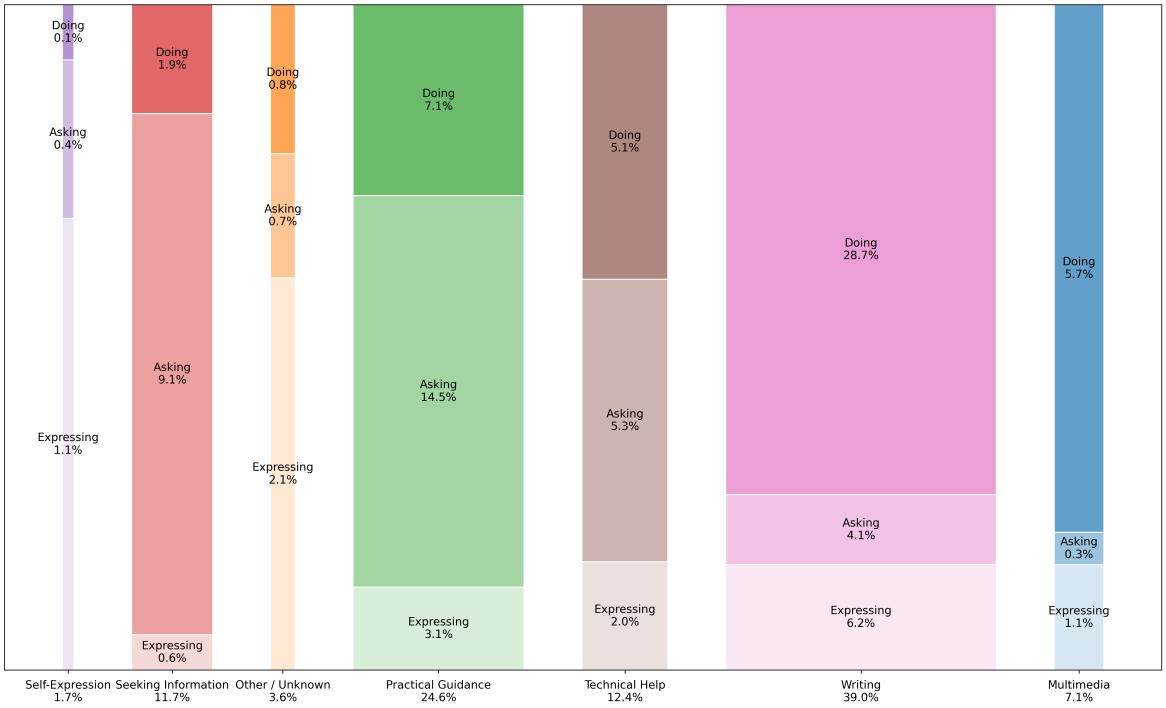


Figure A.20: Breakdown of Message Topics by Asking/Doing/Expressing category for **only work-related messages**, with topic columns sorted by relative share of "Doing" messages. Each bin reports a percentage of the total population. Shares are calculated from approximately 1.8 million sampled messages from July 2024 through December 2025, and daily scaling totals apply logged-in and eligibility filters before computing shares.

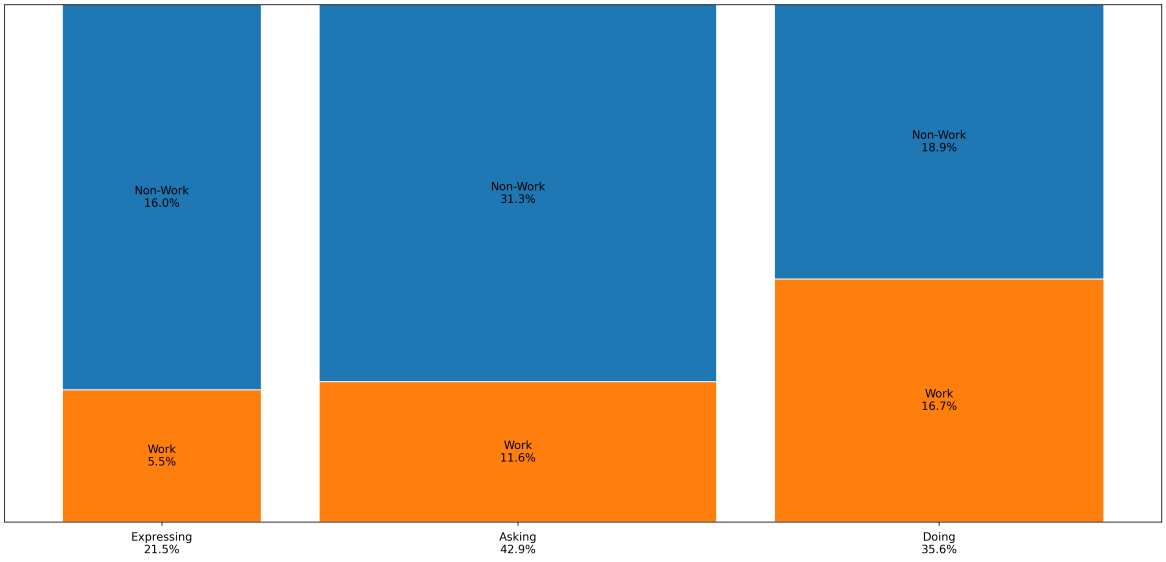


Figure A.21: Shares of Asking, Doing, and Expressing messages split by work vs. non-work. The annotations show work and non-work shares for the full sample. Estimates are based on approximately 1.8 million sampled messages from July 2024 through December 2025, with daily scaling totals that apply logged-in and eligibility filters before computing shares.

A.1.2 Additional Figures using the Message-Level Sample

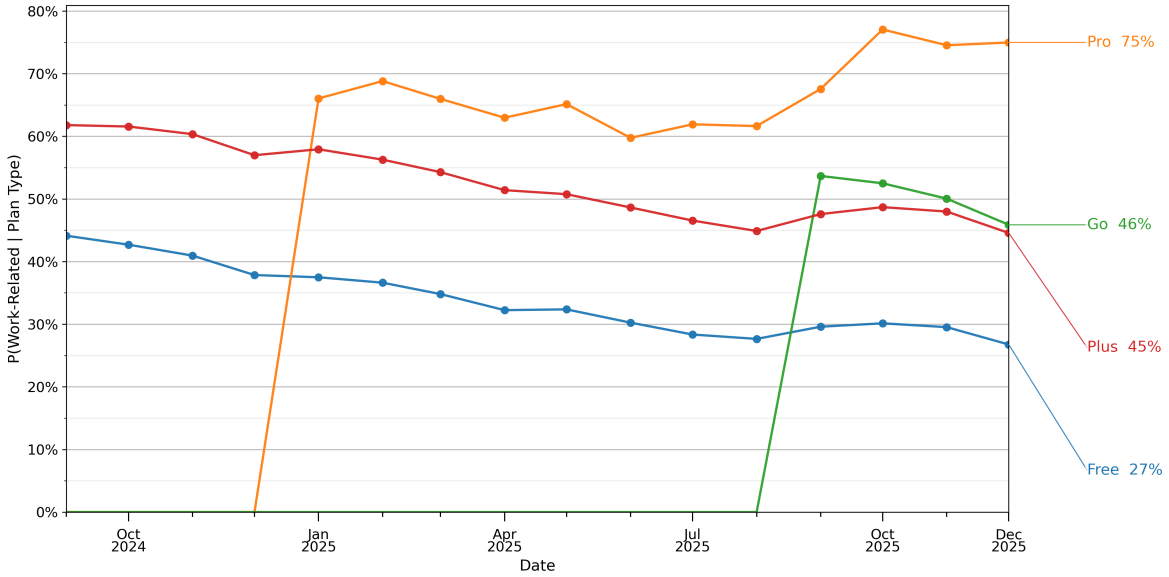


Figure A.22: Conditional probability that a message is work-related by the sender's current ChatGPT plan type, shown in monthly buckets. The series are reproduced from the global OpenAI Signals DP aggregation of work-related message classifications.

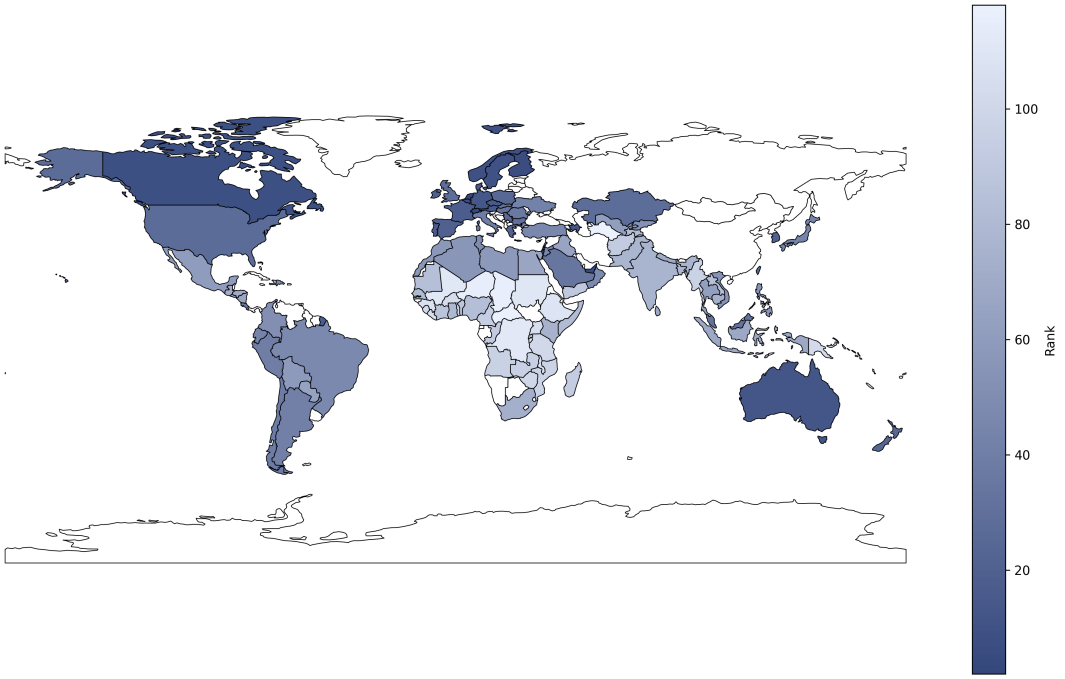


Figure A.23: Rank of countries by ChatGPT messages sent per capita in 2025, where rank 1 denotes the highest per-capita message volume. This figure restyles the corresponding OpenAI Signals country-level DP rank output for use in the paper appendix.

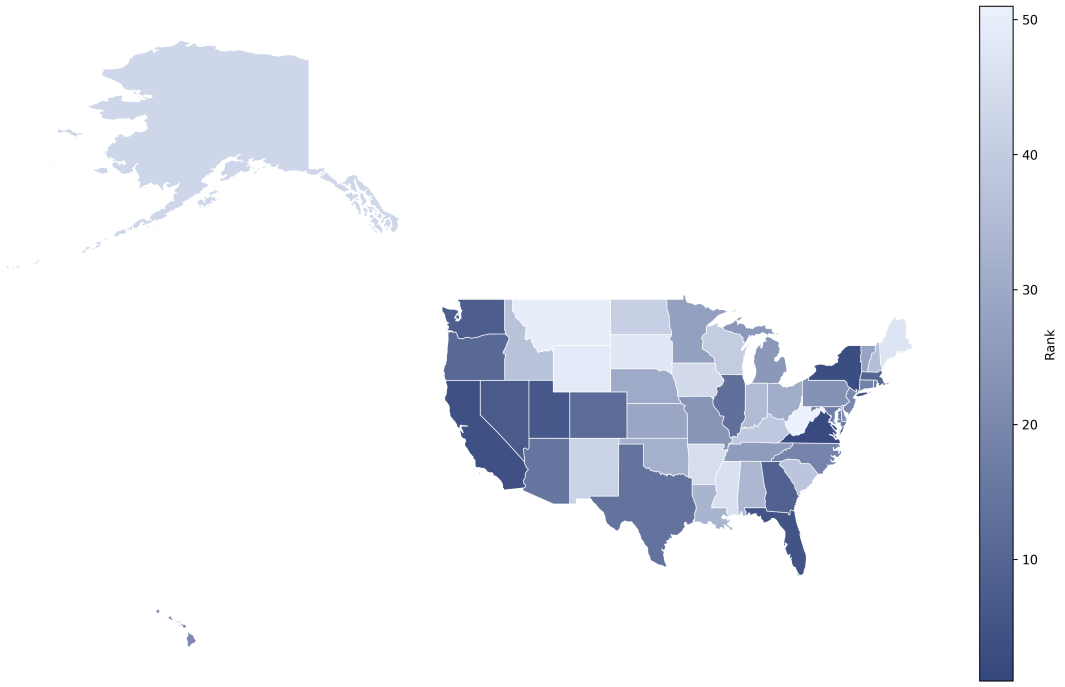
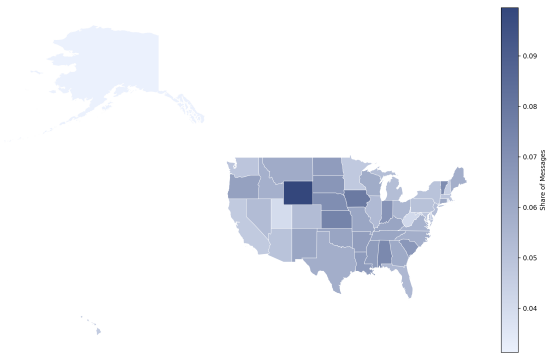
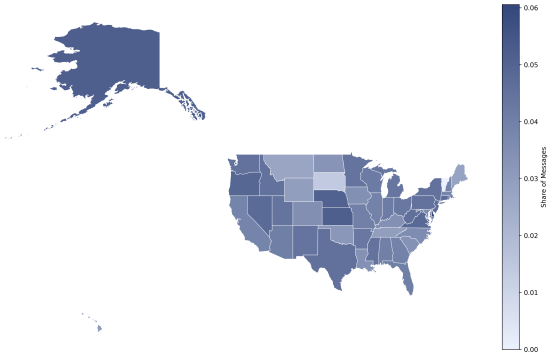


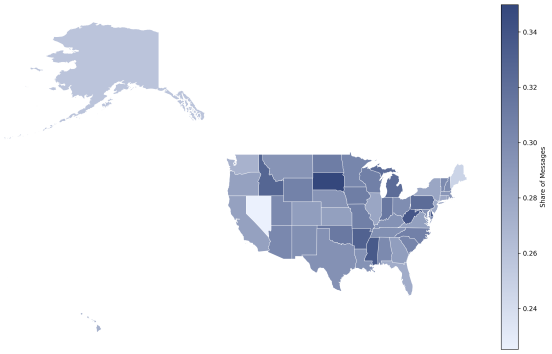
Figure A.24: Rank of U.S. states by ChatGPT messages sent per capita in 2025, where rank 1 denotes the highest per-capita message volume. This figure restyles the corresponding OpenAI Signals U.S. state-level DP rank output for use in the paper appendix.



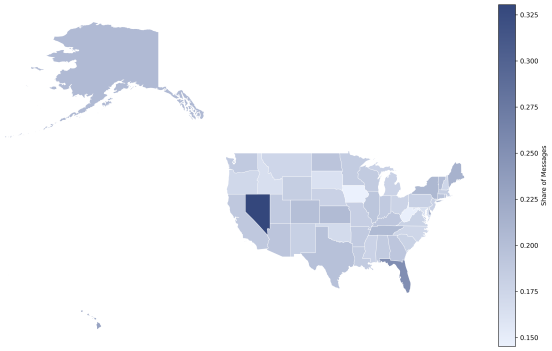
Panel A. Multimedia



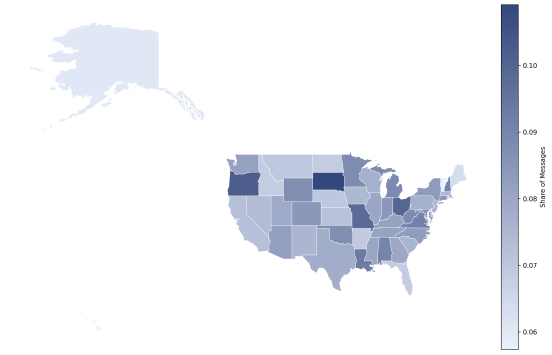
Panel B. Other/Unknown



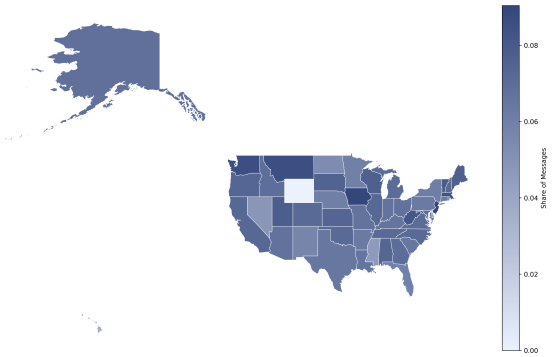
Panel C. Practical Guidance



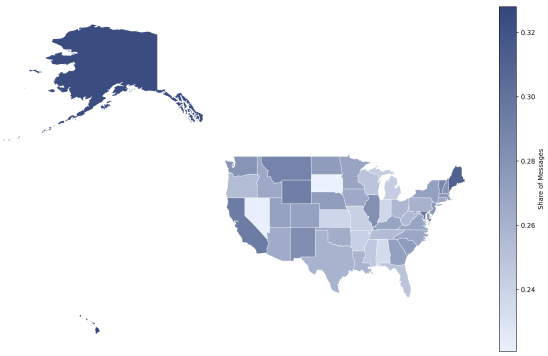
Panel D. Seeking information



Panel E. Self-expression



Panel F. Technical help



Panel G. Writing

Figure A.25: Geographic variation in topic mix across U.S. states in 2025. Each panel shows one OpenAI Signals topic's share within a state's messages, so darker values indicate that the topic accounts for a larger share of messages in that state.

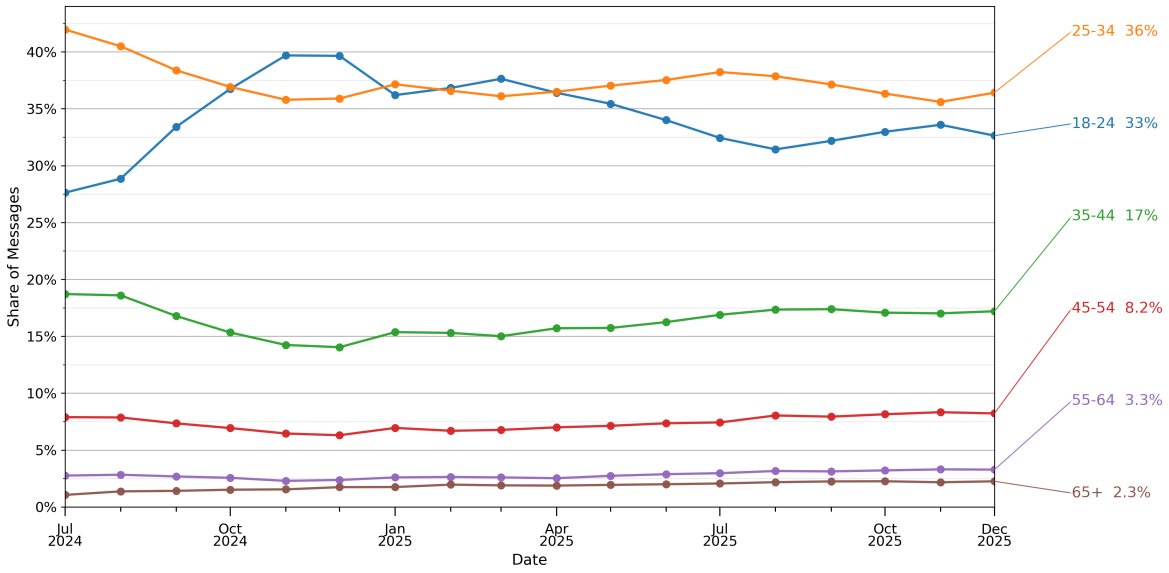


Figure A.26: Share of messages by self-reported age group over time, shown in monthly buckets for the global OpenAI Signals sample. This complements Figure 4 by shifting the focus from the age composition of users to the age composition of messages.

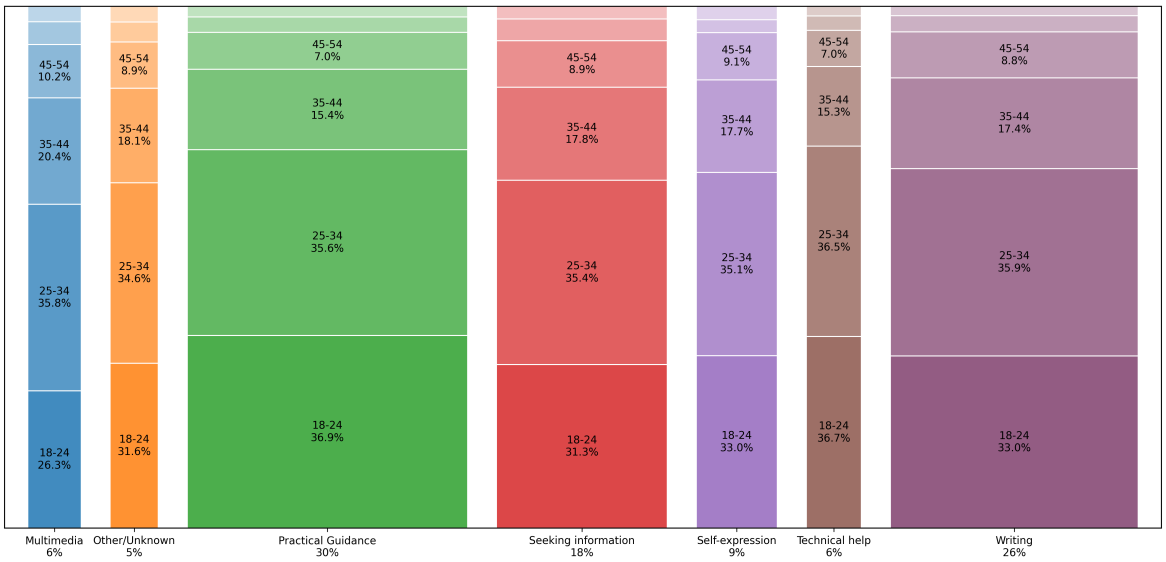


Figure A.27: Share of messages by self-reported age group within major topic category in the November 2025 OpenAI Signals snapshot. Column widths represent each topic's overall message share, so the figure should be read as the age composition of messages within topic rather than the topic mix within age group; it complements Figures 7 and 9.

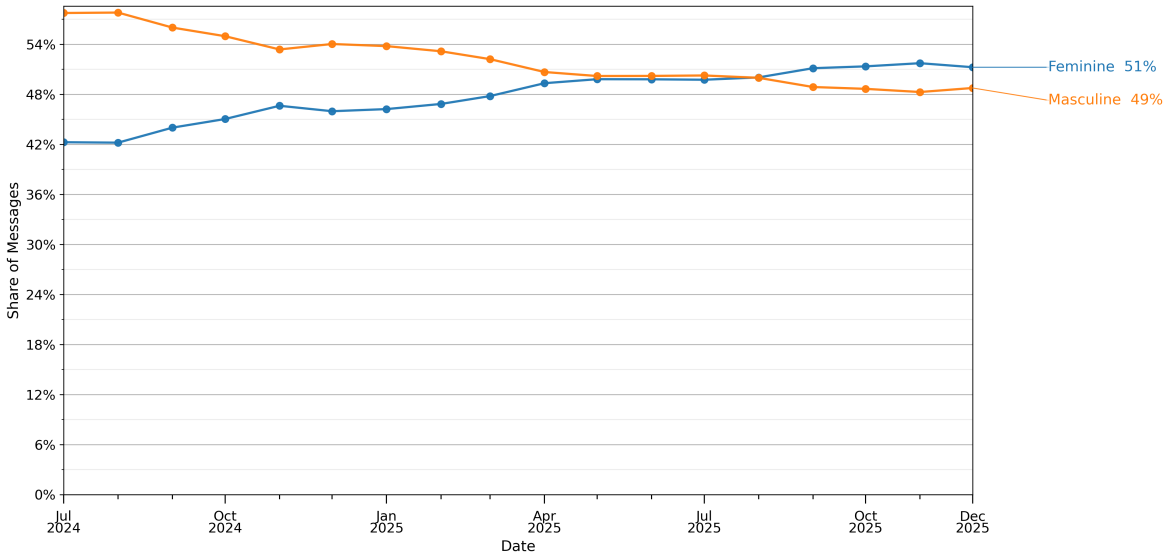


Figure A.28: Share of messages by users with typically feminine or masculine names over time, shown in monthly buckets for the global OpenAI Signals sample. Because these groups are inferred from names rather than observed gender, the figure should be interpreted as descriptive evidence on inferred name groups; see also Figure A.29 and Figure 3.

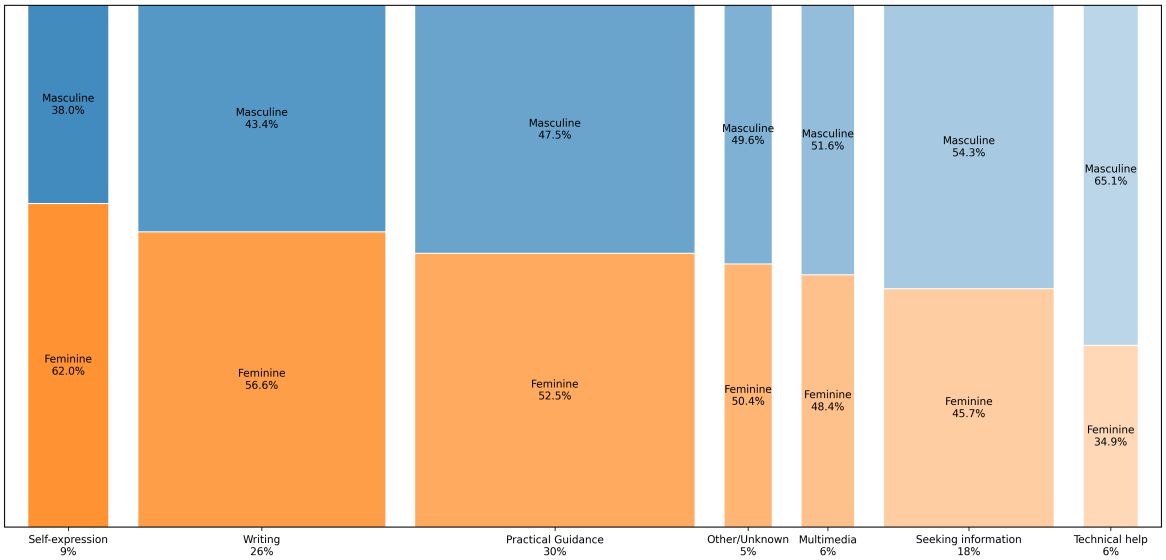


Figure A.29: Share of messages by topic and inferred name group in the November 2025 OpenAI Signals snapshot. Column widths represent each topic's overall message share, and topics are ordered by the feminine-name share of messages; this supplements Appendix Figure A.4 with message-share evidence.

B Appendix: Classifier Prompts

B.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.
```

B.2 Asking/Doing/Expressing

```
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.
```

B.3 Conversation Context Level (Low/Medium/High)

```
You are given the full transcript of a conversation between a user and an assistant
↪ (like ChatGPT).

Your job is to infer the single underlying real-world task the user is trying to
↪ accomplish across the entire conversation (ignoring small digressions), and how
↪ much context is required for the output of this conversation to be safely and
↪ effectively plugged into a real-world workflow or process. Here, 'context'
↪ means user-, organization-, or situation-specific information beyond general
↪ world knowledge.

Important:
- First, infer the underlying real-world task the user is ultimately working on
↪ (for example, "draft a memo for leadership," "analyze survey data for a
↪ research paper," "write an email to a client," "prepare talking points for a
↪ meeting").
- Then, based solely on the previously inferred underlying real-world task,
↪ classify how much context that underlying task requires to be done well.
- By context that the underlying task requires, we mean the user-, organization-,
↪ or situation-specific information required to successfully accomplish a task.
↪ For instance, a task that involves solving a simple math question or looking up
↪ a factual piece of information does require much or any additional critical
↪ information to successfully complete. In contrast, a task such as writing an
↪ internal report about a topic requires more critical information to
↪ successfully accomplish, such as the intended audience, tone, length, previous
↪ reports the audience may have seen etc.
- The user may be asking the assistant to directly do the work product that plugs
↪ into the real-world process, or they may be asking for information, advice, or
↪ partial help so that they can do the work product themselves. In both cases,
↪ label the same underlying task and its context level.
- If the conversation is mostly about learning concepts or asking general
↪ questions, make your best educated guess about what real-world task the user is
↪ preparing for (e.g. writing a report, making an investment decision, preparing
↪ for an exam), and base the context_level on that inferred task.

You must output a single strict JSON object with the following fields:
- "task_summary": A concise statement (<40 words) capturing what the user is
↪ ultimately trying to accomplish while using the assistant. Focus on the likely
↪ end goal, even if the end goal is not represented in conversation itself. Do
↪ not focus on local sub-steps.
```

```
- "task_category": A short noun phrase describing the main type of work (e.g. "data
↪  analysis", "coding", "creative writing", "project planning", "policy drafting",
↪  "email editing").
- "context_level": A categorical label indicating how much context (background,
↪  prior turns, external artifacts) is required for the user or the assistant to
↪  complete the overall task well. It must be one of:
    - "low"
    - "medium"
    - "high"
- "context_rationale": A short explanation citing concrete cues from the transcript
↪  that justify the chosen context_level (e.g. references to prior work, multiple
↪  artifacts, long-term projects, custom constraints, or the need for
↪  user-specific background).

Guidance on "context_level" (for the overall task, not individual messages):

- "low" context:
    - The overall task does not depend at all on external knowledge known to the
    ↪  user but not to the assistant at the start of the conversation.
    - The overall task is a common request we'd expect to see in a chatbot:
    ↪  summarize, rephrase, format, light edit, simple coding with a clear spec,
    ↪  straightforward Q&A.

- "medium" context:
    - The overall task does not depend strongly on external knowledge known to the
    ↪  user but not to the assistant at the start of the conversation.
    - The conversation develops or revises something that is still fairly generic
    ↪  (e.g. a blog post, an essay, standard code on a common stack).

- "high" context:
    - The overall task depends strongly on external knowledge known to the user but
    ↪  not to the assistant at the start of the conversation (for example:
    ↪  internal policies, proprietary code, private data, organization-specific
    ↪  norms, personal financial situation).
    - A generic assistant without access to this external context would struggle to
    ↪  complete the task satisfactorily.

Always return a single JSON object, with no comments or extra text.

Examples:

Example 1 (low context):
```

```
TRANSCRIPT:
"""
User: Here's the text of my talk. Please turn it into a clear, 6-bullet summary for
↪  a slide.
[User pastes full talk text]
Assistant: Sure, here is a 6-bullet summary...
"""

OUTPUT:
{
  "task_summary": "Summarize a provided talk into concise bullet points for a
  ↪  slide.",
  "task_category": "summarization",
  "context_level": "low",
  "context_rationale": "The overall task is to compress a single provided text into
  ↪  slide bullets; all necessary information is in the transcript and no
  ↪  user-specific or external background is required."
}

Example 2 (medium context):

TRANSCRIPT:
"""
User: I'm writing a blog post about the benefits of remote work. What are the most
↪  common reasons people cite for why they want to work remotely?
Assistant: Here are some common reasons....
User: Great, could you outline the blog post for me?
Assistant: [drafts an outline]

User: Thanks. Please critique this draft of the piece....
"""

OUTPUT:
{
  "task_summary": "Develop and refine a multi-section blog post about the benefits
  ↪  of remote work.",
  "task_category": "content writing",
  "context_level": "medium",
  "context_rationale": "The overall task is to produce a generic blog post for a
  ↪  broad audience; the assistant relies on general knowledge and the evolving
  ↪  draft, not on user-specific constraints or private documents."
```

```
}

Example 3 (high context):

TRANSCRIPT:
""
User: I am writing a memo for our leadership team about whether to expand our pilot
↪  with Client X.
Constraints:
- Legal is worried about data retention.
- Sales wants us to move before a competitor does.
- We've had reliability issues in the last two rollouts.

Please use the concerns from the \Client X { Pilot Notes" doc we discussed last
↪  week and keep the tone consistent with the Q2 strategy memo to leadership.
Assistant: Understood. I'll propose a structure and sample language for the memo.
User: Do the recommendations fit with our existing data retention policy and our
↪  current infrastructure limitations?
""

OUTPUT:
{
  "task_summary": "Draft a leadership memo on expanding a client pilot, balancing
  ↪  legal, sales, and reliability concerns using prior internal documents and
  ↪  policies.",
  "task_category": "strategic writing",
  "context_level": "high",
  "context_rationale": "The memo will be used in an internal leadership decision
  ↪  and must align with prior internal notes, strategy memos, data retention
  ↪  policy, and infrastructure constraints that are not visible in the
  ↪  transcript."
}

Example 4 (high context):

TRANSCRIPT:
""
User: I keep hearing about bond yields going up. Can you explain what a bond yield
↪  actually is in simple terms?
Assistant: [explains bond yields]
User: How does the yield curve work, and what does it mean when it inverts?
Assistant: [explains yield curves]
```

```
User: If interest rates rise, what happens to the price of existing bonds?
Assistant: [explains price-yield relationship]
User: This is helpful. Suppose I have savings I won't need for 5{10 years. What
↪  kinds of risks should I think about if I'm considering government bonds versus
↪  a high-yield savings account?
Assistant: [discusses risk and horizon]
""

OUTPUT:
{
  "task_summary": "Understand bond yields and interest rate risk in order to make a
  ↪  medium-term savings or investment decision.",
  "task_category": "personal finance planning",
  "context_level": "high",
  "context_rationale": "Although the user mainly asks general questions, the focus
  ↪  on bond yields, yield curves, and a 5{10 year horizon suggests an investment
  ↪  decision that depends on their personal finances and risk tolerance, which
  ↪  are not in the transcript."
}
```

B.4 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.
```

B.5 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.6 Work/Schoolwork/Other

```
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. The message may
↪ contain a truncated series of back-and-forths between the user and AI chatbot.

Your job is to classify the message according to whether it is likely to be related
↪ to doing a task or activity for the user's professional work or employment,
↪ schoolwork, or something else? Answer with one of the following:

2: related to schoolwork or homework.
1: related to a task or activity occurring during employment
0: other

In your response, only give the number and no other text. The only acceptable
↪ responses are 2, 1 and 0.

---

### General Guidance

Use the **content and context** of the message to infer whether the user's goal is
↪ tied to:
- **Professional work 1:** Tasks that are part of the user's job or professional
↪ responsibilities.
- **Schoolwork 2:** Tasks for a class, course, or degree program.
- **Other 0:** Hobby projects, self-learning, or casual experimentation with no
↪ clear school or job context.

If the message is ambiguous, choose the most probable category based on context and
↪ phrasing style. For instance, if the message mentions their classmate,
↪ instructor, or professor, this suggests they are enrolled in a school and their
↪ message is likely related to schoolwork. In contrast, if a message mentions
↪ their boss, manager, customer, vendor, interview candidate, teammates, etc. it
↪ is likely related to employment.

---

### Examples

User: ""hola, cómo estás""
Expected response: 0
```

```
Explanation: ""The user greeted the chatbot with no other information being
↪  shared.""

User: ""i have a lot of acne on my back. what should I do""
Expected response: 0
Explanation: ""The user asked about their acne, which is not related to schoolwork
↪  nor professional work""

User: ""write a funny scene from The Office where michael is arguing with Pam about
↪  whether star wars is better than star trek. make it goofy and have a lot of
↪  jokes that fans of The Office will like""
Expected response: 0
Explanation: ""The user asked a creative writing prompt, likely for personal
↪  entertainment and therefore not likely to be related to schoolwork nor
↪  professional work.""

User: ""make this cover letter more clear. right now my friend told me it's too
↪  wordy and repetitive""
Expected response: 0
Explanation: ""The user is asking for edits to a cover letter, which is related to
↪  job seeking rather than schoolwork nor their actual current employment.""

User: ""Why does this JavaScript promise not resolve when I run it? I'm just trying
↪  to learn how async works.""
Expected response: 0
Explanation: ""The user says they are 'just trying to learn,' suggesting a personal
↪  learning activity.""

User: ""equation for volume of pipe?""
Expected response: 0
Explanation: ""The user is asking about a math equation, but there is not
↪  sufficient contextual clues to determine whether this is for school, work or
↪  another purpose.""

User: ""i don't have a part time job. i want to start a business""
Expected response: 0
Explanation: ""The user is expressing an interest in starting a business, but their
↪  question is not related to an actual task done for work.""

User: ""write an essay on the Foucault's analysis of the panopticon""
Expected response: 2
```

```
Explanation: ""The user requested an essay about an academic topic, which is likely
↪  related to schoolwork""

User: ""Find the vector components for a = (4, 0, -3)""
Expected response: 2
Explanation: ""The user requested a solution to a math problem which is likely
↪  related to schoolwork""

User: ""What is the result of multiplying 15 by 20? Answer Choices: (a) 200 (b) 300
↪  (c) 315 (d) None of the above""
Expected response: 2
Explanation: ""The user is asking a math question, structured in a way that
↪  indicates it is schoolwork""

User: ""Write three large paragraphs to analyse and explain the effectiveness of
↪  the strategies and resources, such as capital costs, revenue costs, expertise
↪  and intellectual capital, using the structure points for each paragraph below I
↪  will now provide ...""
Expected response: 2
Explanation: ""The user is asking the chatbot to write several paragraphs economics
↪  and management strategy, following a specific outline. The structure of the
↪  outline suggests this may be for an essay related to a school assignment.""

User: ""I'm getting a syntax error in my Python homework, can you help me find
↪  where it's wrong?""
Expected response: 2
Explanation: ""The user explicitly mentions 'homework,' indicating this is
↪  schoolwork.""

User: ""Please translate to german: The course of one particular revolt makes
↪  immediately clear why a protracted scholarly dispute could break out.""
Expected response: 2
Explanation: ""The user is asking for a translation of a passage that seems to be
↪  for an academic topic.""

User: ""Fix the indentation in this code snippet from my senior project report.""
Expected response: 2
Explanation: ""Mentions 'senior project report,' which likely refers to an academic
↪  capstone or thesis assignment.""
```

```
User: ""Sarah owns a small bakery that sells cupcakes and cookies. Her fixed costs
↪ are $1,200 per month, and each cupcake costs $2 to make and sells for $5. How
↪ many cupcakes does she need to sell to break even?
(a) 200
(b) 300
(c) 400
(d) 600""
Expected response: 2
Explanation: ""Although the question describes a small business owner, the
↪ structure of the question - multiple-choice, specific numerical data, and a
↪ clear problem-solving format - indicates it is a test or quiz question from a
↪ business or economics class, not a real-world work task.""

User: ""Can you optimize this SQL query? It's running too slowly in production.""
Expected response: 1
Explanation: ""The user is asking for optimization of a SQL query. SQL is commonly
↪ used in business functions but less common in schoolwork and other contexts.
↪ They also specify using it in 'production' which hints that it is for their
↪ professional work.""

User: ""make a math question for my students. it should involve geometry""
Expected response: 1
Explanation: ""The user is asking for the creation of a math question for their
↪ students. From their question it suggests they are employed as an educator.""

User: ""Generate a project update summary for the team. Make it clear and
↪ professional.""
Expected response: 1
Explanation: ""The user is asking for a project update summary. They indicate the
↪ audience is their 'team', suggesting it is for their professional work. They
↪ also ask for the style to be 'professional' which contextually also suggests it
↪ is work related.""

User: ""how should i ask for a raise? what if i haven't been performing that well
↪ on my team?""
Expected response: 1
Explanation: ""The user is asking for advice on how to ask for a raise, which is
↪ likely related to employment and professional work.

User: ""Rewrite this email to our client so it sounds more professional.""
Expected response: 1
```

```
Explanation: ""The user is asking for a rewrite of an email to a client, suggesting
↪  it is related to professional work.""

---

Here is the user conversation with a chatbot you will now classify:
```

C 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 in-house annotators (three per task).

Table 4 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 (humans+model)	Mean pairwise Cohen's κ (human–human)	Mean pairwise Cohen's κ (model–human)	Cohen's κ (model vs. plurality)
Work Related (binary)	147	0.67 [0.55, 0.77]	0.69 [0.60, 0.78]	0.67	0.72 [0.64, 0.79]	0.83 [0.72, 0.92]
Asking / Doing / Expressing (3-class)	146	0.62 [0.52, 0.70]	0.64 [0.56, 0.71]	0.62	0.67 [0.61, 0.72]	0.74 [0.64, 0.83]
Conversation Topic (coarse)	145	0.46 [0.38, 0.53]	0.48 [0.42, 0.55]	0.48	0.51 [0.46, 0.56]	0.56 [0.46, 0.65]
IWA Classification	100	0.35 [0.27, 0.43]	0.44 [0.38, 0.51]	0.36	0.54 [0.48, 0.59]	0.75 [0.65, 0.83]
GWA Classification	100	0.38 [0.29, 0.46]	0.47 [0.40, 0.54]	0.39	0.56 [0.51, 0.61]	0.76 [0.67, 0.85]
Context Level (Low/Medium/High)	149	0.27 [0.18, 0.35]	0.26 [0.18, 0.33]	0.28	0.25 [0.18, 0.32]	0.29 [0.16, 0.42]
Context Level (binary: Low vs Medium+High)	149	0.37 [0.27, 0.48]	0.36 [0.26, 0.45]	0.38	0.34 [0.26, 0.42]	0.45 [0.30, 0.58]
Interaction Quality (3-class incl. unknown)	145	0.14 [0.04, 0.23]	0.19 [0.11, 0.26]	0.21	0.24 [0.19, 0.30]	0.33 [0.21, 0.44]

Table 4: Validation topline results. Confidence intervals are 95% percentile intervals (2.5th and 97.5th percentiles) from a nonparametric bootstrap with 2,000 resamples.

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; (iv) the mean pairwise Cohen's κ between the model and each human annotator; and (v) 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. The reported n_{labels} is the minimum number of items contributing to any agreement statistic for that task. 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.'
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[42] 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-5.2 was selected based on development-set performance.

C.1 Results

C.1.1 Work-Related Classifier

As shown in Table 4, model–plurality agreement is high (Cohen's $\kappa = 0.83$), exceeding the mean human–human agreement ($\kappa = 0.67$). The mean model–human agreement is likewise substantial ($\kappa = 0.72$; 95% CI $[0.64, 0.79]$). The heatmap in Figure C.1 indicates close alignment with the human plurality and limited systematic bias.

[42] Ties were broken by a senior annotator.

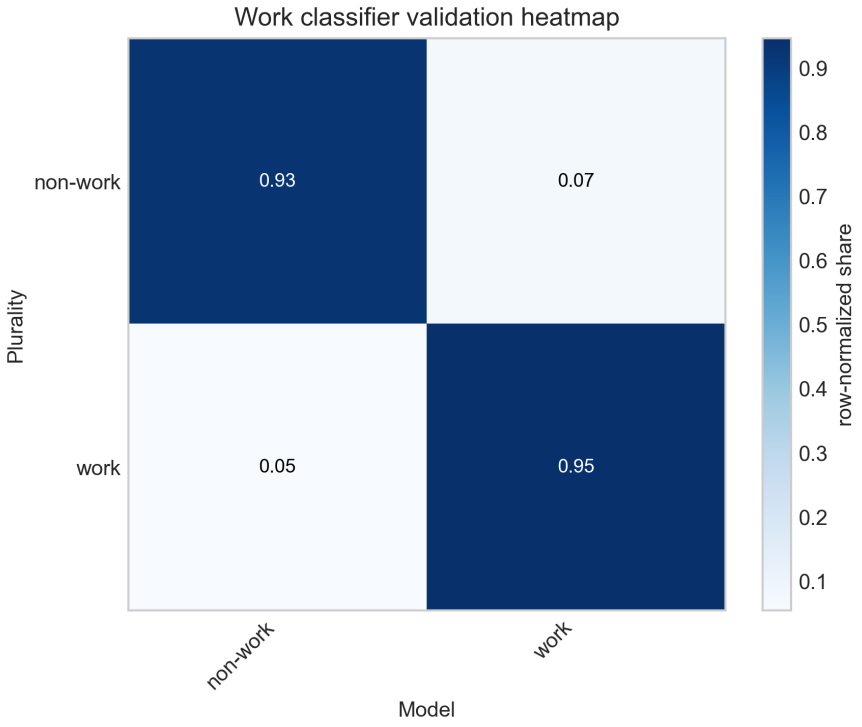


Figure C.1: Agreement between Model and Human Plurality for Work/Non-Work Classifications.

As an additional robustness check, Figures C.2 and C.3 plot the density of the work-classifier share by local time and day of week. The same work-signal is concentrated during standard business hours, with the highest density on weekdays between roughly 9:00 and 17:00 local time and a visible midday dip consistent with a lunch break. This temporal pattern aligns with a canonical work schedule and supports the interpretation that the classifier is identifying genuine work-related usage.

C.1.2 Asking/Doing/Expressing Classifier

Human annotations exhibit substantial agreement (mean human–human Cohen's $\kappa = 0.62$), and the classifier improves on this benchmark with mean model-human agreement of $\kappa = 0.67$ (95% CI $[0.61, 0.72]$) and $\kappa = 0.74$ against the human plurality (Table 4). Figures C.4 and C.5 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.

C.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 ($\kappa = 0.48$). The mean model–human agreement is $\kappa = 0.51$ (95% CI $[0.46, 0.56]$). Misclassifications are concentrated between *Seeking Information* and *Practical Guidance* (Figure C.6), which are conceptually adjacent categories. Relative to human

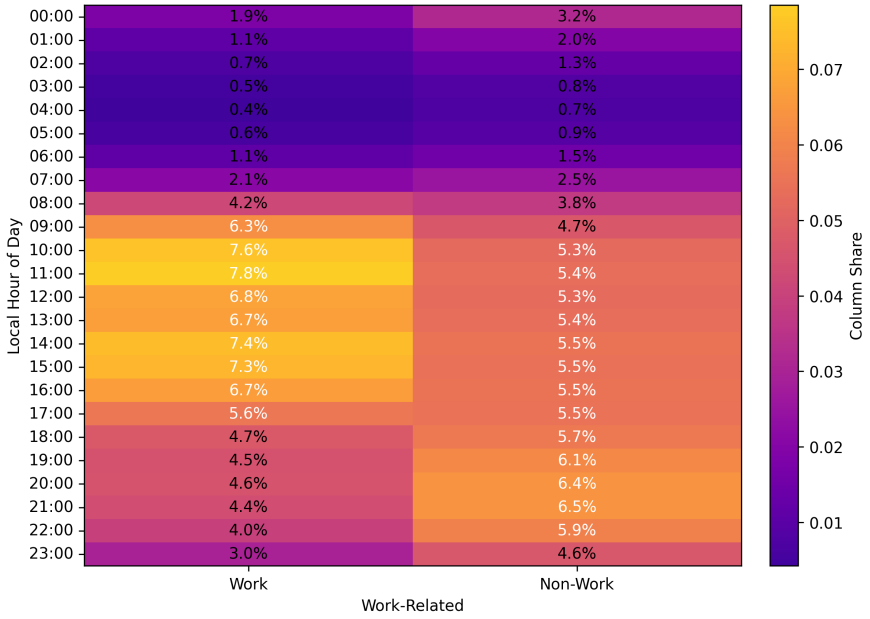


Figure C.2: Robustness check: density of work-classifier share by local hour of day (weekdays only).

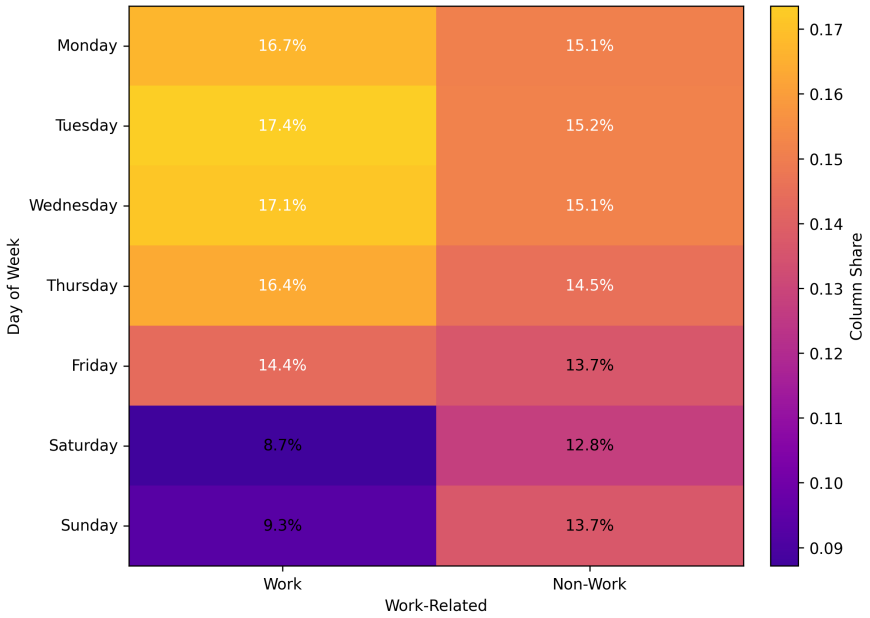


Figure C.3: Robustness check: density of work-classifier share by day of week.

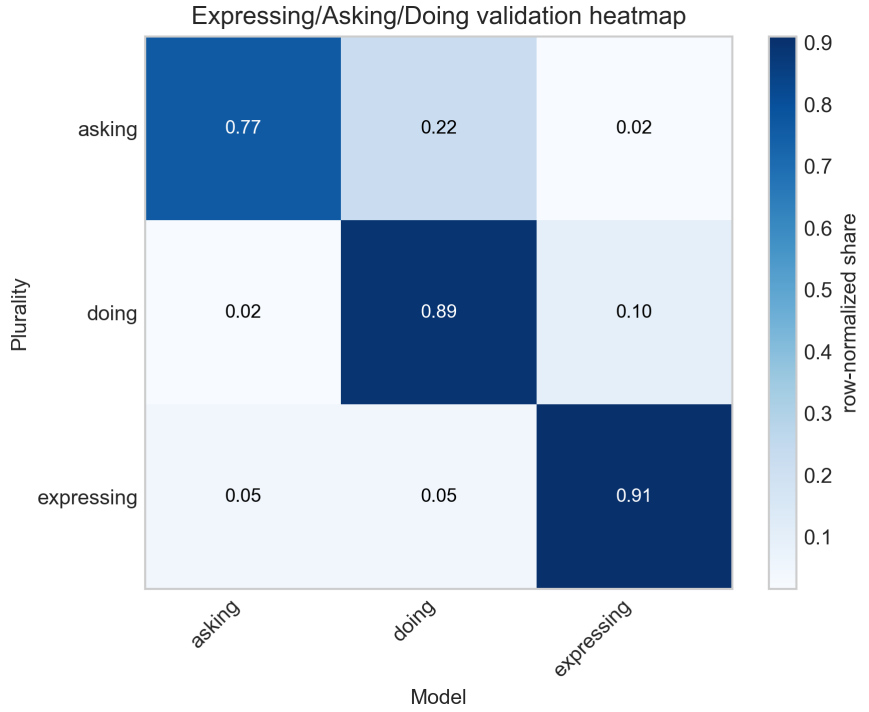


Figure C.4: Agreement Between Model and Plurality for Asking/Doing/Expressing

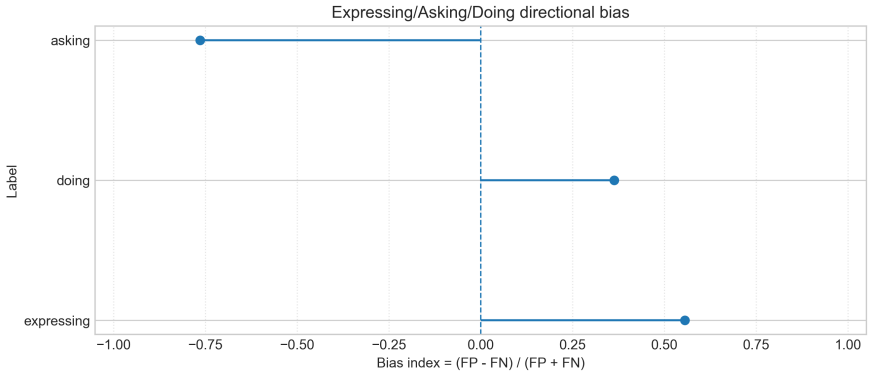


Figure C.5: Per-label Bias, Model vs Plurality for Asking/Doing/Expressing

annotators, the model under-labels *Seeking Information*, *Technical Help*, and *Self-Expression*, and over-labels *Practical Guidance*, *Multimedia*, and *Other* (Figure C.7).

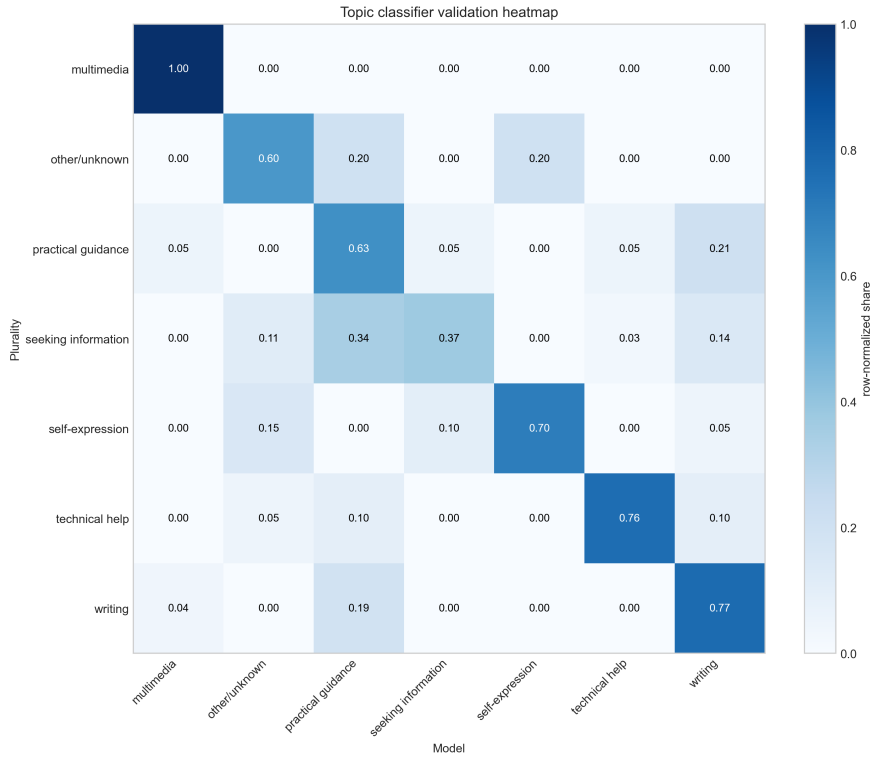


Figure C.6: Agreement Between Model and Plurality for Topic Classification

C.1.4 O*NET Intermediate Work Activity

Three human labelers annotated 100 WildChat messages over the 332 O*NET Intermediate Work Activities (IWAs), with an additional *none* category for ambiguous or non-informative prompts (e.g., greetings or empty messages). We evaluate agreement both at the IWA level and after aggregating to Generalized Work Activities (GWAs), defined as the IWA identifier truncated to its nine-character prefix.

Given the large label space and conceptual overlap across IWAs, human agreement is moderate (mean human–human Cohen's $\kappa = 0.36$ for IWA; $\kappa = 0.39$ for GWA; Table 4). Despite this ambiguity, the model aligns closely with the human plurality (IWA: $\kappa = 0.75$; GWA: $\kappa = 0.76$), and its mean agreement with individual human raters is $\kappa = 0.54$ (95% CI $[0.48, 0.59]$) for IWA and $\kappa = 0.56$ (95% CI $[0.51, 0.61]$) for GWA. From review, we observe that disagreements often reflect reasonable alternative mappings among semantically adjacent IWAs (e.g., different writing-related activities) rather than clear-cut errors.

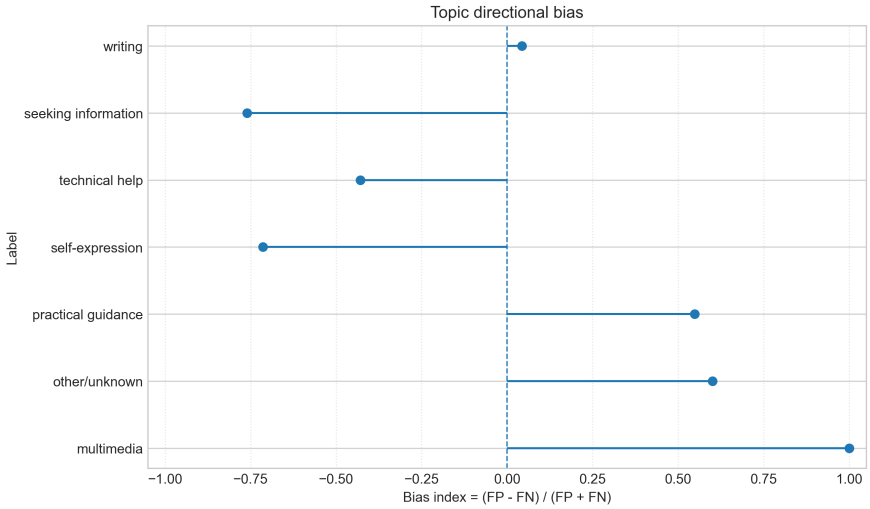


Figure C.7: Bias Between Model and Plurality for Topic Classification

C.1.5 Interaction Quality Classifier

Human and model annotations of interaction quality remain noisy, and required the use of the stronger GPT 5.2 model to generate reasonable results. The classifier attains fair agreement with the human plurality (Cohen's $\kappa = 0.33$; 95% CI $[0.21, 0.44]$), above the mean human–human agreement ($\kappa = 0.21$; Table 4). Mean model–human agreement also rises to $\kappa = 0.24$ (95% CI $[0.19, 0.30]$), and Fleiss' κ increases to 0.19 (95% CI $[0.11, 0.26]$) when treating the model as an additional rater. Figure C.8 indicates meaningful overlap, though disagreement remains substantial. Figure C.9 shows that the model is strongly biased towards rating items as *Bad* relative to humans, indicating that results that show a high rate of good responses are more likely to be genuine signals of user satisfactions. We interpret this classifier cautiously: these κ statistics underscore the difficulty of inferring latent user satisfaction from text alone, but the additional evidence below suggests that the classifier captures meaningful user-satisfaction signal.

While the latent prior of user sentiment is unobserved by our human validators, it is partially observable for a subset of messages that include explicit user feedback in the form of thumbs-up/down responses to assistant messages. To assess whether the classifier captures a signal aligned with user experience, we link model predictions to voluntary feedback on assistant messages. We draw a sample of roughly 100,000 conversations from June 2025 where (i) the assistant message received explicit feedback from the user in the form of a thumbs up or thumbs down and (ii) the user sent a subsequent message that our classifier can score. 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 C.10 shows that *Unknown* classifications are split roughly evenly between thumbs-down and thumbs-up feedback. Thumbs-up comprises more than 90% of all feedback. Conversations with thumbs-down feedback are about 2.5 times as likely to be classified as *Bad* than *Good*, a similar ratio is observed for *Good* messages that received a thumbs-up.

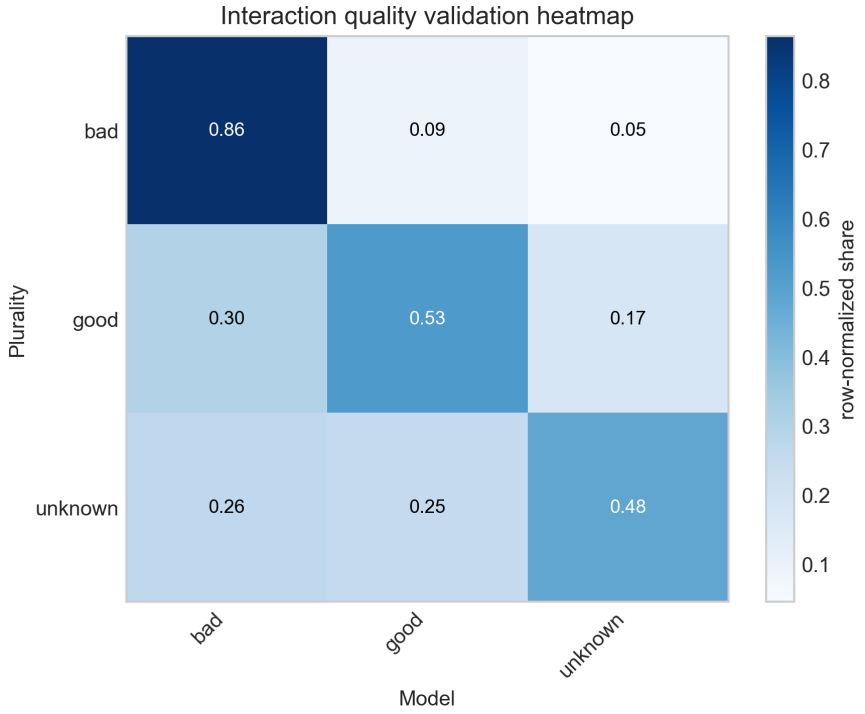


Figure C.8: Agreement Between Model and Plurality for Interaction Quality

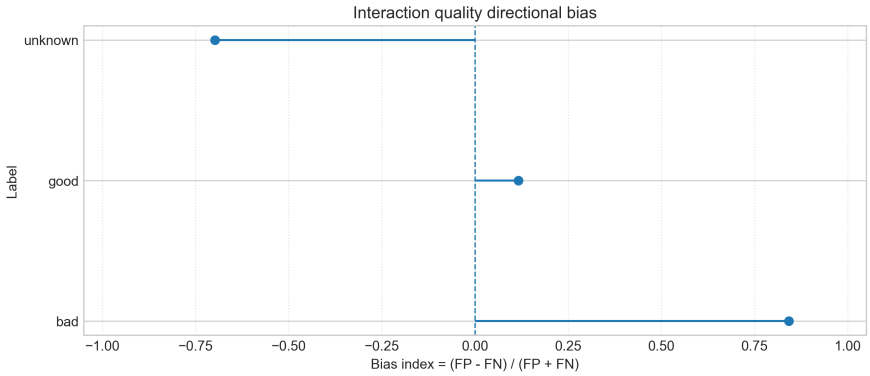


Figure C.9: Bias Between Model and Plurality for Interaction Quality

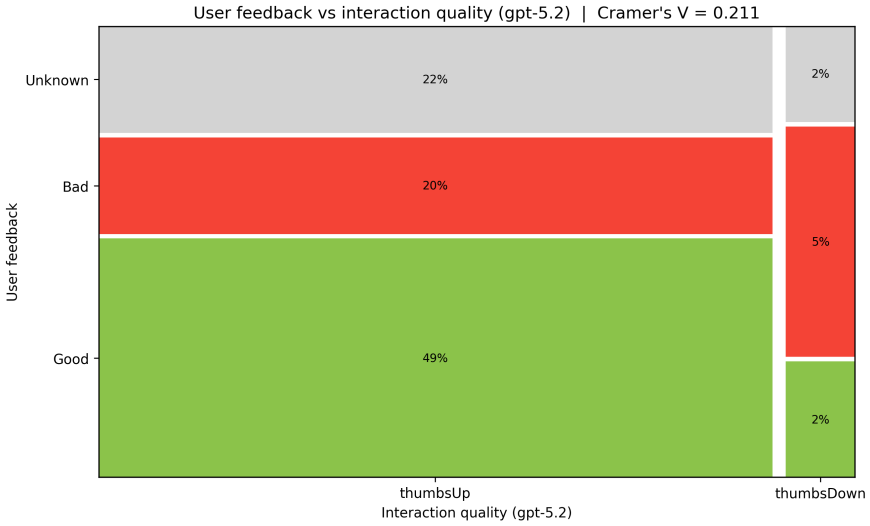


Figure C.10: Correlation of User Rating and Interaction Quality Annotation

C.2 Conversation Context Level (Low/Medium/High)

C.2.1 Conversation Context Level Classifier

The *Context Level* task asks whether classifying the final user message requires little prior conversational context (*Low*), some prior context (*Medium*), or substantial prior context (*High*).

For the ternary task, human agreement is fair (Fleiss' $\kappa = 0.27$ [0.18, 0.35]; mean pairwise human–human Cohen's $\kappa = 0.28$; Table 4). The model aligns similarly with human annotators (mean model–human $\kappa = 0.25$ [0.18, 0.32]) and with the human plurality (Cohen's $\kappa = 0.29$ [0.16, 0.42]).

We also evaluate a binary *Low vs. Medium+High* variant, combining all non low context results. In this setting, human agreement is moderate (Fleiss' $\kappa = 0.37$ [0.27, 0.48]; mean pairwise human–human $\kappa = 0.38$), and model agreement improves meaningfully relative to the ternary task (mean model–human $\kappa = 0.34$ [0.26, 0.42]; model vs. plurality $\kappa = 0.45$ [0.30, 0.58]). Including the model as an additional rater yields Fleiss' $\kappa = 0.36$ [0.26, 0.45].

Figures C.11 and C.12 summarize the model–plurality confusion patterns and per-level labeling bias, showing that the model has a bias towards identifying activities as low context. This is in contrast to share prevalence in the actual data, and suggests that the types of messages in the WildChat reference set have a much higher propensity to be low context than the actual distribution of messages sent to ChatGPT.

Figures C.13 and C.14 provide the corresponding confusion matrix and per-label bias for the *High vs. Medium* version of the context classifier.

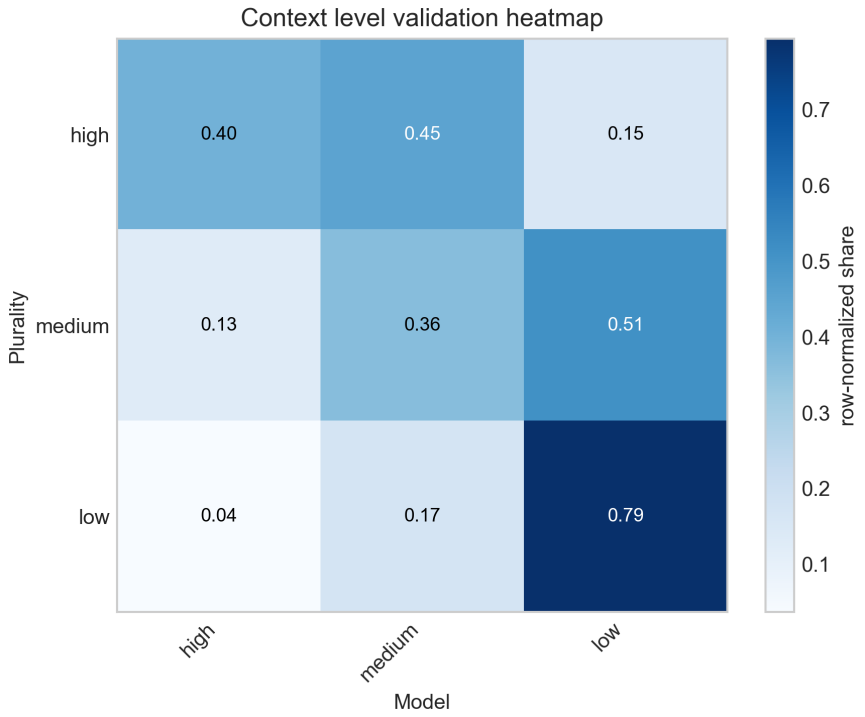


Figure C.11: Agreement Between Model and Human Plurality for Context Level (Low/Medium/High).

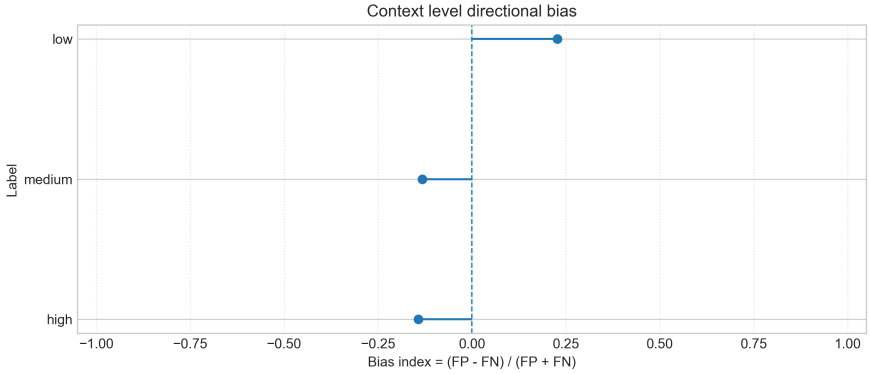


Figure C.12: Per-label Bias, Model vs. Plurality for Context Level (Low/Medium/High).

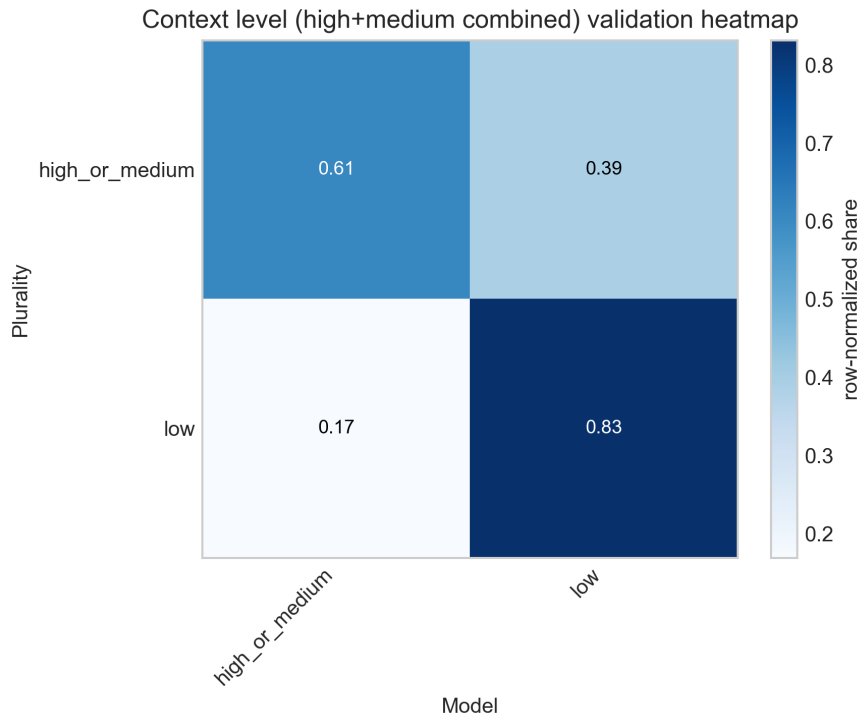


Figure C.13: Agreement Between Model and Human Plurality for Context Level (High/Medium).

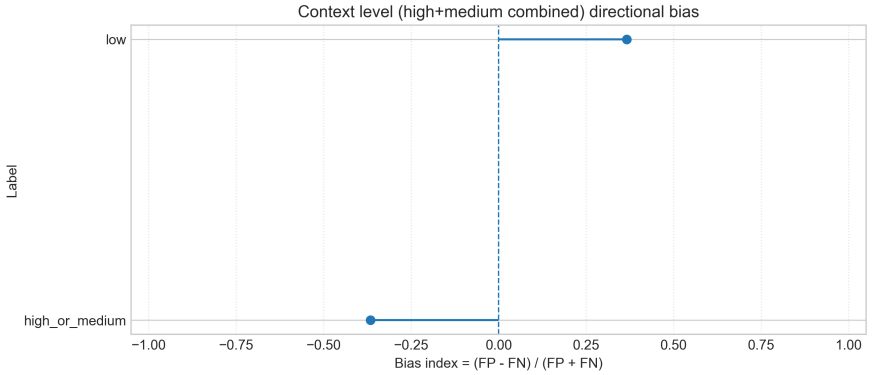


Figure C.14: Per-label Bias, Model vs. Plurality for Context Level (High/Medium).

D 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

E Appendix: Occupational Results

E.0.1 GWA Breakdowns by Occupation

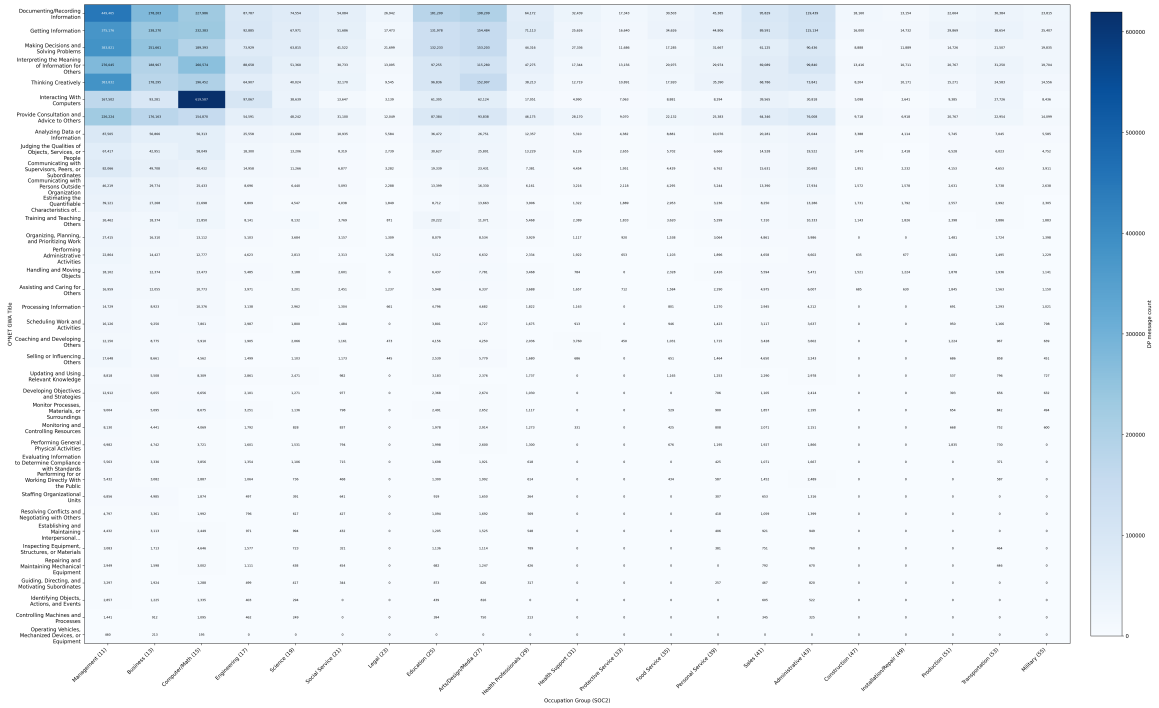


Figure E.1: Classified Queries, Organized by Generalized Work Activity (of the query) and Occupation (of the user). Queries are from approximately 40,000 ChatGPT users, from May 2024 through July 2025. Cells with contributions from fewer than 100 users are suppressed to zero. The title of one GWA is not fully shown due to space constraints: "Estimating the Quantifiable Characteristics of Products, Events, or Information."

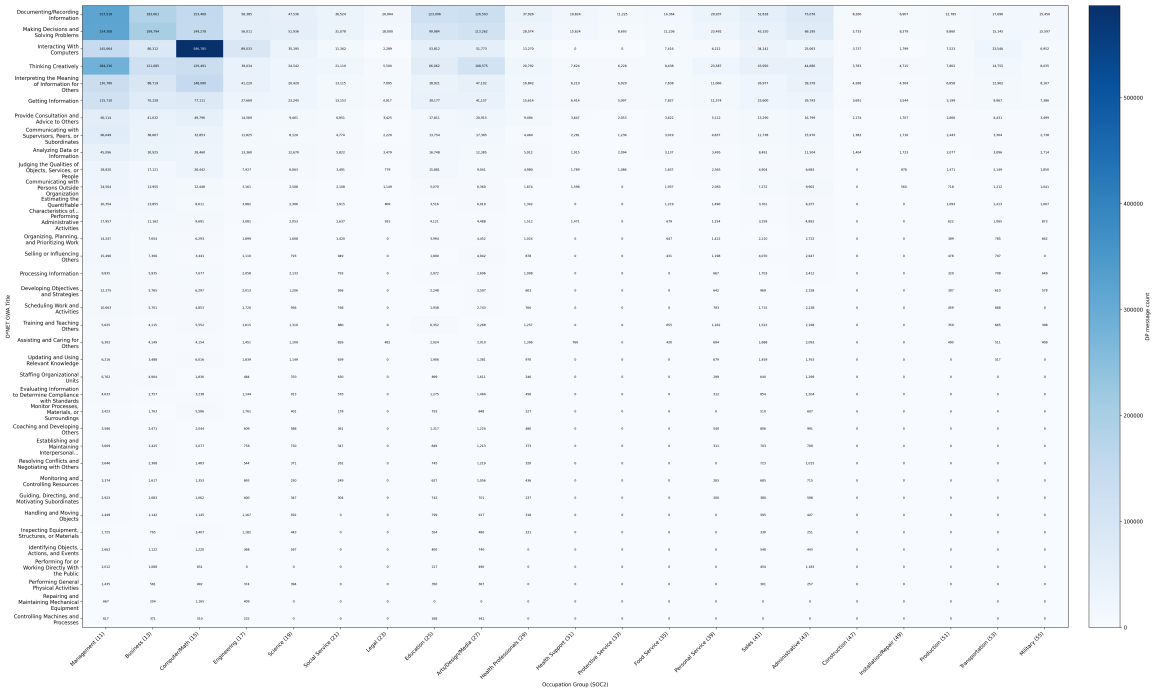


Figure E.2: Classified Work-Related Queries, Organized by Generalized Work Activity (of the query) and Occupation (of the user). Queries are from approximately 40,000 ChatGPT users, from May 2024 through July 2025. Cells with contributions from fewer than 100 users are suppressed to zero. The title of 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	1	3	2	7	5	4	6
Business	1	2	5	7	4	3	6
Computer/Math	4	6	5	1	2	3	7
Engineering	4	5	6	1	3	2	7
Science	1	3	6	7	4	2	5
Social Service	1	3	4	7	6	2	5
Legal	1	2	6	10	4	3	5
Education	1	2	5	7	4	3	6
Arts/Design/Media	1	3	4	7	5	2	6
Health Professionals	2	5	6	7	3	1	4
Health Support	1	3	6	9	5	4	2
Protective Service	1	4	5	7	3	2	6
Food Service	2	6	5	7	4	1	3
Personal Service	1	4	3	9	5	2	6
Sales	1	6	4	7	3	2	5
Administrative	1	4	6	7	3	2	5
Construction	1	5	6	7	3	2	4
Installation/Repair	2	3	5	9	4	1	6
Production	2	6	5	8	4	1	3
Transportation	3	7	5	4	2	1	6
Military	2	3	5	7	4	1	6

Figure E.3: 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. Occupation groups which are omitted are because no GWA was requested by more than 100 users from that occupation group.

F Model Extensions

This section states a characterization of the optimal interaction mode under two assumptions: (i) the value of self-execution (DIY or asking) does not depend on the realized context ω, whereas (ii) the value of delegation varies with ω only through how far the principal's ideal action is from the generic action. With these two assumptions, the optimal interaction mode follows a cutoff rule in the *context gap*.

The first assumption is that DIY and asking yield the same expected payoff across all contexts. That is, context affects only which action is ideal, not the expected value of self-execution (and asking's value depends only on intensity).

This section gives a simple description of the optimal interaction mode. We use two reduced-form assumptions. First, the value of self-execution (DIY or asking) does not depend on the realized context ω; context changes which action is ideal, but not the expected value of carrying it out. Second, the value of delegation depends on ω only through the context gap $\tau(\omega) = (a^*(\omega) - \bar{a})^2$, and it is (weakly) lower when the gap is larger and when the AI executes less accurately.

Assumption 2 (Self-execution payoffs are context-invariant)**.** *There exist a constant $V_{\text{DIY}} \in \mathbb{R}$ and a function $V_{\text{ask}} : \mathbb{R}_+ \to \mathbb{R}$ such that for all $\omega \in \Omega$ and all $x \geq 0$,*

$$V_{\text{DIY}}(\omega) = V_{\text{DIY}}, \qquad V_{\text{ask}}(\omega; x) = V_{\text{ask}}(x).$$

Lemma 1 (Optimal asking intensity is context-invariant)**.** *Under Assumption 2, the set of optimal asking intensities*

$$x^* \in \arg\max_{x \geq 0} V_{\text{ask}}(x)$$

is independent of ω. Define the best self-execution value

$$V_{\text{self}} \;\equiv\; \max\Big\{V_{\text{DIY}},\; \max_{x \geq 0} V_{\text{ask}}(x)\Big\}.$$

Then for every ω, the principal's optimal self-execution payoff equals V_{self}. Within self-execution, the principal uses DIY iff $V_{\text{DIY}} \geq \max_{x \geq 0} V_{\text{ask}}(x)$; otherwise she asks at some x^*.

Proof. Assumption 2 implies that $\max_{x \geq 0} V_{\text{ask}}(\omega; x) = \max_{x \geq 0} V_{\text{ask}}(x)$ for every ω, so the maximizers do not depend on context. The remaining claims follow by comparing the constant values V_{DIY} and $\max_{x \geq 0} V_{\text{ask}}(x)$. □

Lemma 1 reduces the choice between DIY and asking to a single constant payoff V_{self}. The remaining decision is whether to delegate or self-execute, which depends on how well a generic output fits the current context and on the AI's execution accuracy.

Assumption 3 (Delegation depends on context gap and AI accuracy)**.** *There exists a function $\Phi : \mathbb{R}_+ \times [0, \infty) \to \mathbb{R}$ such that for all $\omega \in \Omega$ and all $\sigma_A^2 \geq 0$,*

$$V_{\text{delegate}}(\omega; \sigma_A^2) \;\equiv\; \mathbb{E}[u(\bar{a} + \varepsilon_A, \omega)] \;=\; \Phi\big(\tau(\omega), \sigma_A^2\big), \qquad \varepsilon_A \sim \mathcal{N}(0, \sigma_A^2),$$

with ε_A independent of ω. Moreover:

1. *for each σ_A^2, the map $\tau \mapsto \Phi(\tau, \sigma_A^2)$ is weakly decreasing;*
2. *for each τ, the map $\sigma_A^2 \mapsto \Phi(\tau, \sigma_A^2)$ is weakly decreasing.*

Proposition 2. *Suppose Assumptions 2 and 3 hold. Fix σ_A^2 and let V_{self} be as in Lemma 1. Define*

$$\tau^*(\sigma_A^2) \;\equiv\; \sup\Big\{\tau \geq 0 :\; \Phi(\tau, \sigma_A^2) \geq V_{\text{self}}\Big\} \in [-\infty, \infty],$$

with the conventions $\sup(\emptyset) = -\infty$ and $\sup(\mathbb{R}_+) = +\infty$. Then:

1. *If $\tau(\omega) < \tau^*(\sigma_A^2)$, delegation is optimal at ω.*
2. *If $\tau(\omega) > \tau^*(\sigma_A^2)$, self-execution is optimal at ω (DIY vs. asking as in Lemma 1).*

The cutoff $\tau^*(\sigma_A^2)$ summarizes when delegation beats the best self-execution option. When $\tau^*(\sigma_A^2) = -\infty$, delegation is never optimal; when $\tau^*(\sigma_A^2) = +\infty$, delegation is always optimal. At the boundary $\tau(\omega) = \tau^*(\sigma_A^2)$, the optimal choice depends on ties.

Proposition 3. *Maintain Assumptions 2 and 3. Then:*

1. (Monotonicity in the context gap.) *Fix σ_A^2. If $\tau(\omega_1) \leq \tau(\omega_2)$ and delegation is optimal at ω_2, then delegation is also optimal at ω_1.*
2. (Monotonicity in AI accuracy.) *If $\sigma_A^{2\prime} < \sigma_A^2$, then $\tau^*(\sigma_A^{2\prime}) \geq \tau^*(\sigma_A^2)$. Equivalently, holding σ_P^2 fixed, a larger AI advantage $\Delta = \sigma_P^2 - \sigma_A^2$ weakly expands the delegation region.*

A convenient sufficient condition for Assumptions 2 and 3 is that payoffs admit a "location-family" representation in which context enters only through the ideal action and utility depends on deviations from that ideal in a context-invariant way. Specifically, suppose there exist a constant $\bar{u} \in \mathbb{R}$, a mapping $a^* : \Omega \to \mathbb{R}$, and a loss function $\ell : \mathbb{R} \to \mathbb{R}_+$ such that

$$u(a, \omega) = \bar{u} - \ell\big(a - a^*(\omega)\big),$$

where ℓ is even ($\ell(z) = \ell(-z)$) and strictly convex, and execution noise is mean-zero and symmetric (e.g. Gaussian) with variance σ_i^2 independent of ω. Then self-execution is context-invariant because targeting $a^*(\omega)$ yields

$$\mathbb{E}[u(a^*(\omega) + \varepsilon, \omega)] = \bar{u} - \mathbb{E}[\ell(\varepsilon)]\,,$$

which does not depend on ω; likewise, delegation depends on ω only through the deviation $g(\omega) = a^*(\omega) - \bar{a}$, and by symmetry this dependence is through $|g(\omega)| = \sqrt{\tau(\omega)}$, so one can write $V_{\text{delegate}}(\omega; \sigma_A^2) = \Phi(\tau(\omega), \sigma_A^2)$ with Φ weakly decreasing in τ and (by concavity) weakly decreasing in σ_A^2. This class is broad—quadratic loss is a special case, but power losses $\ell(z) = |z|^p$ for any $p > 1$ and smooth Huber-type losses also qualify—yet the reduced-form assumptions are not limited to it: for example, they are also satisfied by the state-dependent quadratic family

$$u(a, \omega) = \kappa(\omega) - \lambda(\omega)\,\big(a - a^*(\omega)\big)^2$$

provided $\kappa(\omega)$ and $\lambda(\omega)$ are constant across ω (or normalized), even though this specification is not of the form $\bar{u} - \ell(a - a^*(\omega))$ with a single common ℓ unless $\lambda(\omega)$ is constant.

F.1 Extension: Costly context transmission

In the baseline, delegation implements the generic action $\bar{a}$, so the principal incurs a loss based on the full context gap $\tau(\omega) = (a^*(\omega) - \bar{a})^2$. Suppose instead that when delegating the principal can transmit context at intensity $y \geq 0$ (e.g. by providing examples or documents) at linear cost κy, $\kappa > 0$. We assume transmission reduces the effective context gap to

$$\tau(\omega; y) \equiv \frac{\tau(\omega)}{1+y}. \tag{11}$$

Under Assumption 1 (quadratic loss), delegation with transmission yields

$$V_{\text{delegate}}(\omega; y) = -\Big(\tau(\omega; y) + \sigma_A^2\Big) - \kappa y = -\frac{\tau(\omega)}{1+y} - \kappa y - \sigma_A^2. \tag{12}$$

Proposition 4 (Costly context transmission under delegation)**.** *Maintain Assumption 1 and the asking technology from Proposition 1. Then:*

Let V_{self} and the baseline cutoff (T_0, T_1) be as in Proposition 1: if $\sigma_P^2 \leq \mu$, $V_{\text{self}} = -\sigma_P^2 - c$ and $T_0 \equiv \sigma_P^2 - \sigma_A^2 + c$; if $\sigma_P^2 > \mu$, $V_{\text{self}} = -2\sqrt{\mu\sigma_P^2} + \mu - c$ and $T_1 \equiv 2\sqrt{\mu\sigma_P^2} - \mu - \sigma_A^2 + c$. Define the transmission-adjusted cutoff

$$\widetilde{T}_j \equiv \begin{cases} T_j, & T_j \leq \kappa, \\ \dfrac{(T_j+\kappa)^2}{4\kappa}, & T_j > \kappa, \end{cases} \qquad j \in \{0, 1\}.$$

Then the principal delegates (with optimal transmission $y^*(\omega)$) if and only if $\tau(\omega) \leq \widetilde{T}_j$, where $j = 0$ in the DIY regime ($\sigma_P^2 \leq \mu$) and $j = 1$ in the asking regime ($\sigma_P^2 > \mu$). If $\tau(\omega) > \widetilde{T}_j$, the principal self-executes: she uses DIY when $\sigma_P^2 \leq \mu$, and asks at the optimal intensity x^* from Proposition 1 when $\sigma_P^2 > \mu$.

Proof. Fix ω and write $\tau \equiv \tau(\omega)$. Conditional on delegation, the optimal transmission intensity is

$$y^*(\omega) = \max\left\{\sqrt{\frac{\tau(\omega)}{\kappa}} - 1,\ 0\right\}.$$

The optimal value of delegation is

$$\max_{y \geq 0} V delegate(\omega; y) = \begin{cases} -(\tau(\omega) + \sigma_A^2), & \tau(\omega) \leq \kappa, \\ -2\sqrt{\kappa\,\tau(\omega)} + \kappa - \sigma_A^2, & \tau(\omega) > \kappa. \end{cases}$$

Proposition 1 implies that the best self-execution value is V_{self} and the baseline delegation condition is $\tau(\omega) \leq -V_{\text{self}} - \sigma_A^2$, equivalently $\tau(\omega) \leq T_j$ with $j \in \{0, 1\}$ depending on the regime. With transmission, delegation is optimal iff $\max_{y \geq 0} V delegate(\omega; y) \geq V_{\text{self}}$. Let $T \equiv -V_{\text{self}} - \sigma_A^2 (= T_j)$. If

$\tau \le \kappa$, the inequality becomes $\tau \le T$. If $\tau > \kappa$, it becomes

$$-2\sqrt{\kappa\tau} + \kappa - \sigma_A^2 \ge V_{\text{self}} \quad \Longleftrightarrow \quad 2\sqrt{\kappa\tau} \le T + \kappa \quad \Longleftrightarrow \quad \tau \le \frac{(T+\kappa)^2}{4\kappa}.$$

Combining cases yields the cutoff $\widetilde{T}_j$ and the stated policy, with DIY vs. asking within self-execution as in Proposition 1. □

G Differential Privacy Implementations

G.1 Linked Dataset

We add noise to the data derived from the employment-linked dataset described in Section 3.3 in order to protect user privacy in the figures presenting disaggregated views on this dataset, including Figure 11 and Appendix Figures E.1, E.2, and E.3.

We define a message-level privacy budget for the data underlying these analyses such that $\epsilon = 1$ and $\delta = \frac{1}{N+1}$, where N is defined to be the total number of messages sent by the users who were linked to their employer in Section 3.3.

The figures using this dataset rely on ten aggregations:

1. Number of queries by Generalized Work Activity and SOC2 Occupation
2. Number of users sending a query by Generalized Work Activity and SOC2 Occupation
3. Number of work-related queries by Generalized Work Activity and SOC2 Occupation
4. Number of users sending a work-related query by Generalized Work Activity and SOC2 Occupation
5. Number of users in each education cell
6. Number of users in each job family cell
7. Share of low-context messages in each education cell × number of users in that cell
8. Share of medium-context messages in each education cell × number of users in that cell
9. Share of low-context messages in each job family cell × number of users in that cell
10. Share of medium-context messages in each job family cell × number of users in that cell

Since the cells in each aggregation are mutually exclusive, a given message can affect at most one cell. In each granularity, a given message can affect the value of the cell by at most 1. In the aggregations that are count the number of queries in a given cell, this is trivially true. In the cells that count the number of distinct users, the metric also has a maximum message-level of sensitivity of 1, which occurs when the inclusion of a given message causes a user to be included in the aggregation at hand. In the aggregations calculating the share of messages in each context level within each context

grouping times the number of users in the corresponding cell, the maximum sensitivity is 1, which occurs when we classify only a single message from the given user[43].

Using this information, we can calculate a measure called the L2 sensitivity, which allows us to summarize the maximum influence a single message can have across all of the aggregations present in the data we present. This can be calculated using the following formula:

$$\Delta = \sqrt{\sum_{i=1}^{k} w_i^2} \tag{13}$$

where k is the number of aggregations we perform (here, 10) and w_i is the maximum sensitivity of aggregation i to the presence of a given message. In the context of the counts used in the linked dataset, this is equal to 1 across all aggregations, allowing us to see that, in this context, $\Delta \approx 3.16$.

We can use Δ, ϵ, and δ to calculate the standard deviation of the normally-distributed noise that we add, by finding a value that fulfills the following inequality:

$$\Phi(\frac{\Delta}{2\sigma} - \frac{\epsilon\sigma}{\Delta}) - e^{\epsilon}\Phi(-\frac{\Delta}{2\sigma} - \frac{\sigma\epsilon}{\Delta}) \leq \delta \tag{14}$$

where Φ represents the cumulative distribution function of the standard normal distribution.

In practice, we solve the inequality above using a slightly different set of parameters, ϵ' and δ', in place of ϵ and δ. These parameters allow us to add a smaller amount of noise than would otherwise be possible, while preserving the same privacy guarantee, in recognition of the fact that, since our statistics are computed on a sample of all messages sent on ChatGPT, they preserve privacy better than statistics calculated on a similarly-sized, unsampled dataset. To determine the appropriate values of ϵ' and δ', we follow theorem 1 of Li et al. (2012) and calculate:

$$\epsilon' = \ln\left(1 + \frac{e^{\epsilon} - 1}{s}\right),\ \delta' = \frac{\delta}{s} \tag{15}$$

where s corresponds to the share of users that were sampled for inclusion in the data linkage, following the procedure described in Section 3.3.

Using Δ, δ', and ϵ', we find that our privacy can be guaranteed by setting $\sigma \approx 2.23$. We then add $\mathcal{N}(0, \sigma^2)$ noise to the unprotected value in each cell and round the resulting value to the nearest integer.

G.2 Signals Dataset

The results in Appendix A.1 are built using a dataset constructed using message-level sampling, which was released publicly as a part of the February 2026 OpenAI Signals data release.

The approach we use to add differential privacy noise to that dataset is broadly similar to the approach described in Appendix G.1[44].

[43] We do not directly add calculate the DP-protected terms for the high-context cells for each job family and education cell. Instead, we subtract the low- and medium-context shares from 1 to calculate this term. We divide each of the corresponding values by the DP-protected number of users in each cell to calculate the measures plotted in Figure 17.

[44] Further details about this procedure can be found in the data dictionary accompanying the OpenAI Signals release.

In this dataset, we provide a message-level privacy guarantee using $\epsilon = 2$ and $\delta = \frac{1}{N}$, where N is the number of messages sent in a given period by users in our sample. The OpenAI Signals Dataset comprises 24 distinct aggregations:

1. Share of messages by gender within each month (global)
2. Share of messages by gender within each month (India)[45]
3. Share of messages by gender within each topic and month (global)
4. Share of messages by gender within each topic and month (India)[45]
5. Share of messages by age group within each month (global)
6. Share of messages by age group within each month (India)[45]
7. Share of messages by age group within each topic and month (global)
8. Share of messages by age group within each topic and month (India)[45]
9. Share of messages by topic within each month (global)
10. Share of messages by topic within each month (India)[45]
11. Share of messages by topic within US state (2025 average)
12. Share of messages by work-related status within each topic and month (global)
13. Share of messages by work-related status within each topic and month (India)[45]
14. Share of messages by topic within each work-related status and month (global)
15. Share of messages by topic within each work-related status and month (India)[45]
16. Share of messages by work-related status within each month (global)
17. Share of messages by work-related status within each month (India)[45]
18. Share of messages by work-related status within each plan type and month (global)
19. Share of messages by O*NET IWA within each month (USA)
20. Share of work-related messages by O*NET IWA within each month (USA)
21. Share of messages by Ask/Do/Express classification within each work-related status and month (global)
22. Share of messages by Ask/Do/Express classification within each work-related status and month (India)[45]
23. Rank for messages sent per capita (USA, states)
24. Rank for messages sent per capita (by country)

The aggregations used in the Signals dataset are somewhat more complicated than the raw counts used in the aggregations described in Appendix G.1, with several being constructed using several underlying aggregations. In these cases, we add noise satisfying our privacy guarantee to each of the base aggregations.

[45]This data is not shared in CSV format and is only used to construct a visualization.

To understand this process, consider the function $G_{2025-06}(x)$, which calculates the (weighted) share of messages in June 2025 (represented here by x) that were made by users with predominantly masculine names. We implement this function using two simpler functions, $M_{2025-06}(x)$ and $F_{2025-06}(x)$, which counts the (weighted) number of messages in June 2025 by users with names inferred to be male and female, respectively. We can then write:

$$G_{2025-06}(x) = \frac{M_{2025-06}(x)}{M_{2025-06}(x) + F_{2025-06}(x)}$$

We can ensure that a composite function (here, $G_{2025-06}$) provides a guarantee of (ϵ, δ) differential privacy if each of the functions underlying it (in this case, $M_{2025-06}$ and $F_{2025-06}$) are replaced with functions that themselves each provide a guarantee of (ϵ, δ) differential privacy, which we achieve here by adding calibrated Gaussian noise to the true value of the two functions. Concretely, we define:

$$M'_{2025-06}(x) = M_{2025-06}(x) + Z_1, \;\; F'_{2025-06} = F_{2025-06} + Z_2$$

where Z_1, Z_2 $\mathcal{N}(0, \sigma^2)$, and σ is sufficiently large to ensure that $M'_{2025-06}$ and $F'_{2025-06}$ are each (ϵ, δ) differentially private. In this case, we can ensure that our overarchinng calculation of the shares is (ϵ, δ) differentially private by writing it as:

$$G'_{2025-06}(x) = \frac{M'_{2025-06}(x)}{M'_{2025-06}(x) + F'_{2025-06}(x)}$$

.

In each of the aggregations defined above, the underlying cells are mutually exclusive, such that each message can influence at most a single cell in each aggregation. The aggregations performed are based on weighted counts, where each message is given a weight, which is calculated at the daily level. These weights ensure that, within a given month, the (weighted) share of messages in our sample from each day matches the true proportion of messages from that month that were sent on that day. This procedure accounts for changes over time in the logging infrastructure we use to sample messages. The day-level weights are never higher than 1.5.

We can then use this information on the maximum weight applied to any message (which is bounded from above by 1.5) with information on the maximum number of granularities that a message can affect (21^{46}):

$$\Delta = \sqrt{\sum_{i=1}^{k} w_i} \approx 6.87$$

.

As in Appendix G.1, when calculating the amount of noise to add to the data, we account for the privacy-amplifying effects of sampling, following Equation 15. Since we sample a constant 100,000 messages in each month, which represents a varying fraction of the overall data, we calculate a separate value for ϵ' and δ' in each month, and therefore calculate a different value for σ each month using Equation 14. We then add the corresponding amount of noise to the (weighted) counts performed at

[46] This is fewer than the 24 aggregations calculated, as some are mutually exclusive—a message cannot be sent from both India and the United States

each step of each aggregation.