NBER WORKING PAPER SERIES

HOW PEOPLE USE CHATGPT

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

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

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

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

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

NBER 工作论文系列

人们如何使用 ChatGPT

亚伦·查特吉 托
马斯·坎宁安 大卫·
J·迪明 霍齐·奥恩 克
里斯托弗·昂 卡尔·
闫山 凯文·沃德曼

工作论文 34255
http://www.nber.org/papers/w34255

国家经济研究局 1050 马萨诸塞大道 剑桥，马萨诸
塞州 02138 2025年9月

我们感谢 Joshua Achiam、Hemanth Asirvatham 和 Ryan 的帮助和评论
贝耶尔大师，瑞秋·布朗，卡珊德拉·杜尚·索利斯，杰森·权，埃利奥特·莫斯基，凯文·劳，哈里森·萨彻，盖韦沙·韦拉图اگن，汉娜·王，以及分析团队。我们特别感谢蒂娜·埃尔翁杜和帕梅拉·米什金，他们在多个方面为这项工作奠定了基础，这项研究得到了哈佛大学伦理委员会批准（IRB25-0983）。一个包含所有代码的存储库，用于生成本文中的分析，可在要求时提供。本文中表达的观点是作者的观点，不一定反映美国国家经济研究局的观点，也不一定反映美国国家经济研究局的观点。
经济研究。

至少一位合著者已披露与本研究可能相关的其他关系。更多信息可在
http://www.nber.org/papers/w34255 在线获取。

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

人们如何使用 ChatGPT Aaron Chatterji、Thomas Cunningham、David J. Deming、Zoe Hitzig、Christopher Ong、Carl Yan Shan 和 Kevin Wadman
NBER 工作论文编号 34255 2025年9月 JEL 编号 J01, O3, O4

ABSTRACT

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

摘要

尽管LLM聊天机器人得到了快速采用，但人们对它们的使用方式知之甚少。我们记录了ChatGPT消费者产品从2022年11月推出到2025年7月的增长情况，当时它已被全球约10%的成年人口采用。早期采用者中男性比例较高，但性别差距已显著缩小，我们发现低收入国家的增长率更高。使用隐私保护自动化管道，我们对ChatGPT对话的代表性样本中的使用模式进行分类。我们发现与工作相关的消息呈稳定增长，但非工作相关的消息增长更快，其占比已从53%上升到超过70%。教育程度较高、从事高薪专业职业的用户更常使用工作相关的功能。我们按对话主题对消息进行分类，发现“实用指导”、“寻求信息”和“写作”是三个最常见主题，合计占所有对话的近80%。在工作任务中，写作占据主导地位，突显了聊天机器人与传统搜索引擎相比生成数字输出的独特能力。计算机编程和自我表达的使用比例都相对较小。总体而言，我们发现ChatGPT通过决策支持提供了经济价值，这在知识密集型工作中尤为重要。

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

Thomas Cunningham
OpenAI
tom.cunningham@gmail.com

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

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

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

Carl Yan Shan
OpenAI
cshan@openai.com

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

Aaron Chatterji 杜克
大学富奎商学院和
OpenAI
ronnie@duke.edu

Thomas Cunningham OpenAI
tom.cunningham@gmail.com

David J. Deming 哈佛大
学哈佛肯尼迪学院和
NBER david_
deming@harvard.edu

Zoe Hitzig OpenAI 和 哈佛
大学研究员协会
zhitzig@g.harvard.edu

Christopher Ong 哈佛大学和
OpenAI
christopherong@hks.harvard.edu

卡尔·严山 OpenAI
cshan
@openai.com

Kevin Wadman OpenAI
k
evin.wadman@c-openai.com

1 Introduction

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

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

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

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

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

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

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

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

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

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

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

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

1 简介

ChatGPT 于 2022 年 11 月推出。到 2025 年 7 月，7 亿用户每周发送 180 亿条消息，约占全球成年人口的 10%。[1] 对于一项新技术来说，这种全球传播速度是前所未有的（Bick 等人，2024 年）。

本文研究了 ChatGPT 的消费者使用情况，这是第一个大众市场聊天机器人，也可能是最大的。[2] ChatGPT 基于大型语言模型 (LLM)，这是一种在过去十年中开发的人工智能 (AI)，通常被认为代表了人工智能能力的加速。[3]

LLM 能力和采用率的突然增长加剧了人们对人工智能对经济增长（Acemoglu，2024；Korinek 和 Suh，2024）；就业（Eloundou 等人，2025 年）；以及社会（Kulveit 等人，2025 年）影响的兴趣。然而，尽管 LLM 得到了快速采用，但关于它们如何被使用的公共信息有限。一些调查已经测量了 LLM 的自我报告采用情况（Bick 等人，2024 年；皮尤研究中心，2025 年）；但是，有理由预期自我报告存在偏差（Ling 和 Imas，2025 年），并且这些论文中都没有能够直接跟踪聊天机器人对话的数量或性质。

两篇最近的论文确实报告了关于聊天机器人对话的统计数据，这些数据以多种方式进行了分类（Handa 等人，2025 年；Tomlinson 等人，2025 年）。我们在以下几个方面基于这项工作进行了扩展。首先，ChatGPT 上的用户群体要大得多，这意味着我们期望我们的数据更接近于平均聊天机器人用户.[4] 其次，我们使用自动分类器来报告用户使用新分类体系相对于现有文献发送的消息类型。第三，我们报告了聊天机器人使用在人群中的扩散情况以及不同类型的使用在群体中的增长情况。第四，我们使用一个安全的数据清洁室协议来分析我们用户样本中聚合的就业和教育类别，从而在保护用户隐私的同时，为不同群体发送的消息类型差异提供了新的见解。

我们的主要样本是 2024 年 5 月至 2025 年 6 月期间在消费者计划（免费、高级、专业）上向 ChatGPT 发送的消息的随机选择.[5] 用户向聊天机器人发送的消息使用多种不同的分类体系自动分类：消息是否用于付费工作、对话主题以及交互类型（提问、执行或表达），以及用户正在执行的职业*网（O*NET）任务。每个分类体系都是通过传递给一个大型语言模型（LLM）的提示来定义的，这使我们能够在没有任何人类看到它们的情况下对消息进行分类。我们在附录 A 中提供了大多数提示的文本，并在附录 B 中提供了有关如何验证这些提示的详细信息.[6] 分类管道受到一系列隐私措施的保护，如下所述，以确保在自动分析过程中不会泄露敏感信息。在一个安全的数据清洁室中，我们将消息分类体系与聚合的就业和教育类别相关联。

表1显示了工作和非工作用途的总消息量的增长。这两种类型

[1]路透社（2025），罗思（2025）[2]比克等人（2024）报告称，2024年末28%的美国成年人使用了ChatGPT，高于其他任何聊天机器人。[3]我们在此处松散地使用术语LLM，并在下一节中提供更多细节。[4]维格尔斯（2025）报告称，据估计，2025年4月ChatGPT接收的访客数量是Claude或Copilot的10倍以上。[5]我们的样本包括三种消费者计划（免费、高级或专业）。OpenAI还提供各种其他ChatGPT计划（商业fka.团队，企业，教育），我们未将其纳入样本。[6]我们的分类器不仅考虑了随机选择的用户消息，还考虑了该对话中前一部分的消息。

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

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

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

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

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

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

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

月份	非工作时间 (M)	(%)	工作时间 (M)	(%)	总消息量 (M)
2024年6月	238	53%	213	47%	451
2025年6月	1,911	73%	716	27%	2,627

表1: ChatGPT每日消息数量（百万），按可能的工作相关或非工作相关进行细分。每日总数是所有消费者套餐消息量的精确测量。工作和非工作相关消息的每日数量是通过对该日随机抽样的对话进行分类来估计的。抽样是为了排除选择不共享其消息用于模型训练的用户、自我报告年龄低于18岁的用户、已登出用户、已删除的对话以及已被停用或封禁的账户（详细信息见第3节）。报告值为2024年6月26日和2025年6月26日的7天平均值（以平滑周波动）。

消息量持续增长，但非工作消息增长更快，现在占所有消费者ChatGPT消息的70%以上。虽然大多数关于AI的经济分析都集中在它对有偿工作生产力的影响上，但对其在工作之外（家庭生产）活动的影响规模相似，甚至可能更大。工作相关消息份额的下降主要是由于每个用户群体内部的用法变化，而不是新ChatGPT用户构成的变化。这一发现与Collis和Brynjolfsson（2025）一致，他们通过选择实验揭示了用户对生成式AI的支付意愿，并估计仅2024年美国消费者的剩余价值就至少有970亿美元。

我们接下来报告使用OpenAI开发的一个分类法对消息进行分类，该分类法用于了解产品使用情况（"对话分类器"）。几乎所有ChatGPT的使用都分为三大类，我们称之为实用指导、寻求信息和写作。实用指导是最常见的用例，包括辅导和教学、关于各种主题的如何建议以及创意构思。[7] 寻求信息包括搜索关于人物、时事、产品和食谱的信息，似乎非常接近网络搜索的替代品。写作包括自动生成电子邮件、文档和其他通信，但也包括编辑、评论、总结和翻译用户提供的文本。写作是工作中最常见的用例，在2025年6月平均占工作相关消息的40%。大约三分之二的写作消息要求ChatGPT修改用户文本（编辑、评论、翻译等），而不是从零开始创建新文本。大约10%的消息是辅导或教学请求，表明教育是ChatGPT的一个关键用例。

我们的两项发现与其他研究形成对比。首先，我们发现与计算机编程相关的消息比例相对较小：只有 4.2% 的 ChatGPT 消息与计算机编程相关，而 Handa 等人 (2025) 的研究表明，与工作相关的 Claude 对话中有 33% 是关于计算机编程的。[8] 其次，我们发现与陪伴或社会情感问题相关的消息比例也相对较小：只有 1.9% 的 ChatGPT 消息涉及人际关系和个人反思这一主题，而 0.4% 的消息与

[7] *Practical Guidance* 和 *Seeking Information* 的区别在于前者高度定制化以适应用户，可以根据对话和后续内容进行调整，而后者是客观事实信息，对所有用户都应保持一致。例如，对跑步感兴趣的用户可能会向ChatGPT询问波士顿马拉松按年龄和性别划分的资格赛时间(*Seeking Information*)，或者他们可能会要求一个符合其目标和当前健身水平的定制化训练计划(*Practical Guidance*)。[8]Handa等人(2025)报告称，37%的对话被归类到"计算机和数学"职业类别，他们的图12显示30%或更多的所有填充任务与编程或IT相关。我们认为这种差异部分是由于Claude和ChatGPT用户类型的不同，此外Handa等人(2025)仅包含那些"可能涉及职业任务"的查询。

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

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

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

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

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

[9]Zao-Sanders (2025) is based on a manual collection and labeling of online resources (Reddit, Quora, online articles), and so we believe it likely resulted in an unrepresentative distribution of use cases.
[10]Among those with names commonly associated with a particular gender.
[11]Appendix A gives the full prompt text and Appendix B gives detail about how the prompts were validated against public conversation data.

游戏和角色扮演相关。相比之下，Zao-Sanders (2025) 估计，治疗/陪伴是生成式 AI 最普遍的应用场景。[9]

我们也记录了关于 ChatGPT 使用中人口统计差异的几个重要事实。首先，我们展示了证据表明 ChatGPT 使用的性别差距很可能随着时间的推移而显著缩小，甚至可能已经完全关闭。在 ChatGPT 发布后的几个月里，典型男性名字的活跃用户通常占 80%。[10] 然而，截至 2025 年 6 月，这一数字下降到 48%，活跃用户更有可能拥有典型女性名字。其次，我们发现成年人发送的消息中近一半是由 26 岁以下的用户发送的，尽管年龄差距在最近几个月有所缩小。第三，我们发现 ChatGPT 在低收入和中等收入国家过去一年的使用增长相对较快。第四，我们发现受过教育的用户和高薪专业职业的用户更有可能将 ChatGPT 用于工作。

我们引入了一种新的分类法，根据用户寻求的输出类型对消息进行分类，使用一个我们称为询问、执行或表达的简单标准。[11] 询问是指用户寻求信息或澄清以帮助决策，对应于知识工作的解决问题模型（例如，Garicano (2000)；Garicano 和 Rossi-Hansberg (2006)；Carnehl 和 Schneider (2025)；Ide 和 Talamas (2025)）。执行是指用户希望产生某些输出或执行特定任务，对应于经典的任务型工作模型（例如，Autor 等人 (2003)）。表达是指用户表达观点或感受，但不是寻求任何信息或行动。我们估计大约 49% 的消息是询问，40% 是执行，11% 是表达。然而，截至 2025 年 7 月，大约 56% 的与工作相关的消息被分类为执行（例如，执行工作任务），其中近四分之三的是写作任务。与写作相关的对话的相对频率有两个原因值得注意。首先，写作是一项几乎所有白领工作都常见的任务，良好的书面沟通技能是雇主要求的最顶尖的"软"技能之一（全国高校和雇主协会，2024 年）。其次，相对于其他信息技术，生成式 AI 的一个独特特征是其能够产生长文本输出，如写作和软件代码。

我们还将消息内容映射到工作活动上，使用的是职业信息网络（O*NET），这是一项由美国劳工部支持的对工作特征进行的调查。我们发现大约81%的工作相关消息与两项广泛的工作活动相关：1）获取、记录和解释信息；以及2）做决策、提供建议、解决问题和进行创造性思考。此外，我们发现与ChatGPT使用相关的工作活动在不同类型的职业中高度相似。例如，获取信息和做决策以及解决问题在几乎所有职业中都位列消息频率前五，这些职业范围从管理和商业到STEM再到行政和销售职业。

总体而言，我们发现信息检索和决策支持是大多数工作中最常见的ChatGPT使用场景。这与一个事实一致，即几乎所有ChatGPT的使用都是实用指导或寻求信息。我们还表明，询问的增长速度比

[9]Zao-Sanders (2025)基于对在线资源（Reddit、Quora、在线文章）的手动收集和标记，因此我们认为它很可能导致用例分布不具代表性。[10]在那些名字通常与特定性别相关联的人中。[11]附录A给出了完整的提示文本，附录B给出了关于如何将提示与公共对话数据进行验证的详细信息。

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

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

2 What is ChatGPT?

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

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

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

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

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

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

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

执行更快，并且询问消息始终被评分为具有更高的质量，无论是通过一个衡量用户满意度的分类器，还是来自直接用户反馈。

ChatGPT 如何提供经济价值，以及其价值对谁来说最大？我们认为 ChatGPT 可能通过提供决策支持来提高工人产出，这在知识密集型工作中尤为重要，因为更好的决策能提高生产力（Deming，2021；Caplin 等人，2023）。这解释了为什么在受过教育、从事高薪专业职业的用户中，“提问”相对更常见。我们的发现与 Ide 和 Talamas（2025）最为一致，他们建立了一个模型，其中 AI 代理可以作为产出产出的同事，也可以作为提供建议并提高人类问题解决生产力的副驾驶。

2 什么是 ChatGPT?

在这里我们给出 LLM 和聊天机器人的简化概述。如需更精确的细节，请参考 OpenAI 随每个模型发布的论文和系统卡，例如（OpenAI，2023，2024a，2025b）。聊天机器人是一个经过统计模型训练以生成文本响应的模型，给定一些文本输入，目的是最大化该响应的“质量”，其中质量是通过各种指标来衡量的。

在一个典型的交互中，用户提交一个纯文本消息（“提示”）并且ChatGPT返回从底层LLM生成的消息（“响应”）。ChatGPT添加了大量额外的功能——包括允许LLM搜索网络或外部数据库，以及根据文本生成图像——但基于文本的消息交换仍然是典型的交互。

自发布以来，ChatGPT使用过各种不同的底层LLM，例如GPT-3.5、GPT-4、GPT-4o、o1、o3和GPT-5。[12] 此外，模型的权重和模型的系统提示（与所有查询一起发送给模型的文本指令）会进行偶尔更新。

可以将LLM视为一个从单词串到所有可能单词集合上的概率分布的函数（更精确地说，“标记”，它们非常粗略地对应于单词[13]）。这些函数使用深度神经网络实现，通常具有Transformer架构（Vaswani等人，2017年），参数化模型“权重”数十亿。我们将ChatGPT的所有模型都称为语言模型，尽管大多数模型还可以处理代表图像、音频或其他媒体的标记。

基于LLM的聊天机器人的权重通常分两个阶段进行训练，通常称为“预训练”和“后训练”。在第一阶段（“预训练”），LLM被训练以根据前面的词语，在一个庞大的文本语料库中预测字符串中的下一个词。在那个阶段，模型纯粹是预测下一个词在给定先验上下文中的可能性的预测器，因此它们的应用范围相对较窄。在第二阶段（“后训练”），模型被训练以生成构成对某些提示的“良好”响应的词语。这个阶段通常包括各种不同的策略：在查询和理想响应的数据集上进行微调，强化学习对抗另一个训练来评估响应质量的模型（Ouyang等人，2022年），或强化学习对抗一个知道查询真实响应的函数（OpenAI（2024b），

[12] 有关模型发布的时间线，请参阅附录C。[13] 分词是一种将文本字符串分割成离散块的方法，这些块在统计上是高效的。在许多分词方案中，一个词元大致对应英语单词的四分之三。

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

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

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

3 Data and Privacy

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

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

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

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

3.1 Growth Dataset

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

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

Lambert等人（2024年））。这个第二阶段通常还包括一些“安全”约束，以避免某些类型的响应，特别是那些被认为是有害或危险的响应（OpenAI，2025a）。

这个两阶段过程有一个共同的统计解释：第一阶段教会模型世界的潜在表示；第二阶段使用该表示拟合函数（Bengio 等人，2014 年）。预训练模型以预测下一个词有效地教会模型文本的低维表示，仅表示关键语义特征，因此使得提示-响应问题在合理的训练示例集上变得可行。

评估聊天机器人的两种常见方法是使用基准（具有已知答案的问题集，例如 Measuring Massive Multitask Language Understanding (Hendrycks 等人，2021 年)）和比较人类对相同消息的两个替代响应的偏好（例如 Chatbot Arena (Chiang 等人，2024 年)）。

3 数据和隐私

在本节中，我们描述了论文中使用的数据以及我们实施的隐私保护措施。研究团队的所有成员从未见过用户消息的内容，并且所有分析均根据 OpenAI 的隐私政策（OpenAI，2025c）进行。

本文的分析基于以下数据集：

1. 增长：2022年11月至2025年9月期间消费者ChatGPT用户每日总消息量，以及基本的自我报告人口统计信息。该数据集主要用于第4节。

2. 分类消息：被分类到粗略类别中的消息。
   - 从所有ChatGPT用户中抽样：2024年5月至2025年6月期间登录的消费者ChatGPT用户中约一百万条去识别化消息的随机样本.[14] 该数据集主要用于第5节。
   - 从ChatGPT用户子集中抽样：2024年5月至2025年7月期间由消费者ChatGPT用户子集发送的两批消息的随机样本（一批在对话级别，一批在用户级别）。[15] 这些数据集主要用于第6节。

3. 就业：基于公开数据对消费者ChatGPT用户子集的就业和教育类别的汇总。这些数据仅用于第6节。

我们描述了每个数据集的内容、产生它们的抽样程序，以及我们在构建和用于分析时实施的隐私保护措施。

3.1 生长数据集

我们编译了一个数据集，涵盖了自2022年11月ChatGPT推出以来在所有消费者ChatGPT计划（免费、Plus、Pro）上的所有使用情况。我们排除了非消费者计划（企业f.k.a. Teams、

[14]该样本的确切开始和结束日期是2024年5月15日和2025年6月26日。[15]该样本的确切开始和结束日期是2024年5月15日和2025年7月31日。

Enterprise, Education).

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

3.2 Classified Messages

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

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

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

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

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

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

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

企业、教育）的用户。

对于每个用户和日期，该数据集报告了用户在该日期发送的总消息数。它还报告了每条消息的去标识化用户元数据，包括他们与ChatGPT首次交互的时间戳、其账户注册的国家、他们每天使用的订阅计划以及他们自报的年龄（报告在粗略的5-7年区间内以保护用户隐私）。

3.2 分类消息

为了在保护用户隐私的同时了解使用情况，我们构建了不包含任何人阅读消息内容的消息级数据集。有关保护隐私的分类流程概述，请参见图1。消息根据5种不同的基于LLM的分类器进行分类。这些分类器在第5节中有更详细的介绍，其确切文本在附录A中重现，我们的验证流程在附录B中描述。

从所有ChatGPT用户中抽样。我们均匀地抽样了大约110万次对话，然后在每次对话中抽样一条消息，并遵循以下限制：

1. 我们仅包含2024年5月至2025年7月之间的消息。
2. 我们排除那些选择不分享其消息用于模型训练的用户。3. 我们排除那些自报年龄低于18岁的用户。
4. 我们排除用户已删除的对话以及账户已被停用或封禁的用户。

5. 我们排除已登出用户,[16] 这些用户在样本期间占ChatGPT用户的少数。

我们的样本来自一个本身被抽样的表格，其中抽样率随时间变化。因此，我们调整我们的抽样权重以保持与总发送消息量的固定比例。

从ChatGPT用户子集抽样。我们从ChatGPT用户子集（约13万用户）中构建了两个分类消息样本。这个用户样本不包括任何选择不分享其消息用于训练的用户，也不包括自报年龄低于18岁的用户，也不包括被封禁或删除账户的用户。

第一个样本包含从这组用户中抽取的158万条消息的分类，按对话级别抽样（对话是指用户与聊天机器人之间的一系列消息）。这个样本构建得使得用户在数据中的表示与总体消息量成正比。第二个样本包含从这组用户发送的消息，按用户级别抽样，每个用户组中最多包含六条消息。

[16]ChatGPT 于 2024 年 4 月对已登出用户开放，即用户无需使用电子邮件地址注册账户即可使用 ChatGPT。然而，已登出用户的消息仅在我们数据集中从 2025 年 3 月起可用，因此为了保持一致性，我们删除了所有已登出用户的消息。

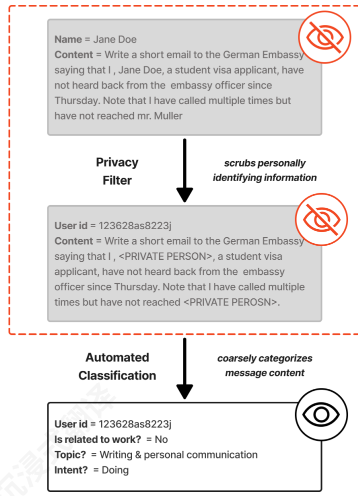


Figure 1: Illustration of Privacy-Preserving Automated Classification Pipeline (Synthetic Example). Messages are first stripped of PII via an internal LLM-based tool called *Privacy Filter*. Then they are classified by LLM-based automated classifiers, described in detail in Appendices A and B. Humans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

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

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

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

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

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

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

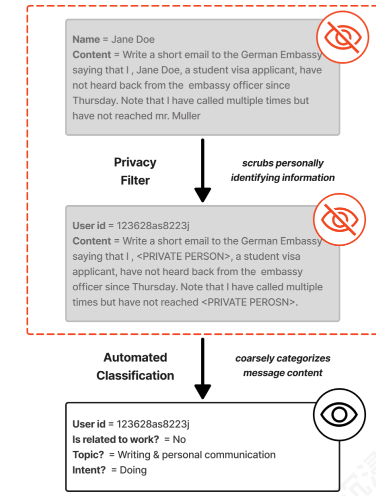


图1：隐私保护自动分类流程图（合成示例）。消息首先通过一个名为 *Privacy Filter*的内部基于LLM的工具去除PII。然后由基于LLM的自动分类器进行分类，详细描述见附录A和B。人类看不到原始消息或已去除PII的消息，只看到消息的最终分类结果。

通过自动分类器实现隐私保护。在撰写本文时，没有人查看消息的内容。所有对消息内容的分析都是通过在去标识化和PII清理后的消息数据上运行的自动基于LLM的分类器进行的（见图1）。消息首先使用内部基于LLM的工具进行PII清理，[17]，然后根据在受控标签空间上定义的分类器进行分类——我们使用的最精确的消息级数据集分类器是O*NET中级工作活动分类法，我们对其进行扩展，最终得到333个类别。我们引入了技术和程序上的障碍，以防止意外访问底层文本（例如，不向研究人员渲染消息文本的界面）。

我们的分类旨在识别给定消息的意图，因此我们将对话中的前10条消息作为上下文。[18] 例如，参见表2。

独立消息	带上下文的消息
[用户]：“再来10个”	[用户]：“给我三个可以和青少年一起参加的文化活动”[助手]：“1. 参观博物馆……”（截断）[用户]：“再给我10个”

表2：上下文增强消息分类的说明（合成示例）。左侧列显示待分类的独立消息，右侧列显示在左侧消息分类中包含的先验上下文。

我们将每条消息截断到最多5,000个字符，因为长上下文窗口可能会导致分类质量的变化（Liu等人，2023年）。我们对每条消息使用“gpt-5-mini”模型进行分类，Interaction Quality除外，它使用“gpt-5”，并使用附录A中列出的提示。

[17]内部分析表明，该工具， *Privacy Filter*，与人类判断有显著一致性。[18]在 *Interaction Quality*的情况下，我们额外将对话中的下两条消息作为上下文包含进来。

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

3.3 Employment Dataset

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

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

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

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

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

我们通过将模型分类决策与来自公开可用的WildChat数据集（Zhao等人，2024年）中样本对话的人工分类判断进行比较，验证了每个分类提示。WildChat数据集是一组用户明确同意公开用于研究目的的第三方聊天机器人对话。[19] 附录B详细介绍了我们的验证方法和相对于人工判断的性能。为了增加透明度，我们对100,000条公开的WildChat消息进行分类，并将这些数据提供在本论文的复制包中。

3.3 就业数据集

我们基于公开数据对样本消费者 ChatGPT 用户的聚合就业类别进行了有限分析。该样本包括约 130,000 名免费、Plus 和 Pro 用户，就业类别由通过安全数据清洁室 (DCR) 工作的供应商进行聚合。对于此分析，我们使用与消息级数据集相同排除标准：我们排除已停用用户、被封禁用户、已选择退出培训的用户以及自我报告年龄低于 18 岁的用户。由于数据仅对部分用户可用，因此结果可能无法代表所有用户群体。

描述。就业数据是从公开来源聚合的，包括行业、职业（粗化为 O*NET 类别）、资历级别、公司规模和仅限于所获学位的教育信息。在 DCR 内工作的供应商获取了此数据集，限制我们只能通过 DCR 对其运行聚合查询，并在研究完成后将其删除。

通过数据清洁室实现隐私保护。我们从未直接访问用户级别的统计数据。所有就业数据分析均在仅允许预先批准的跨独立数据集进行聚合计算的安全DCR（数据清洁室）内执行；任何一方都无法查看或导出另一方的底层记录。我们对DCR实施了严格的协议：要执行任何涉及外部统计数据查询的操作，我们首先从6位合著者组成的委员会获得明确批准，然后将笔记本提交给我们的数据合作伙伴审批；只有经过批准的笔记本才能在DCR中运行（见图2）。

我们的合作伙伴实施了严格的聚合限制：他们只批准返回满足100用户阈值的单元格的代码。因此，研究人员从未看到过任何单个行或狭义定义的类别。例如，如果99名用户从事“麻醉师”这一职业，任何职业级别的输出都会将这些用户归类为“被抑制”类别，或者将这些观察结果归入粗化类别（例如“医疗专业人员”），而不是报告一个单独的麻醉师单元格。

[19]该数据集通过OpenAI的LLM（大型语言模型）API从第三方聊天机器人收集。

8

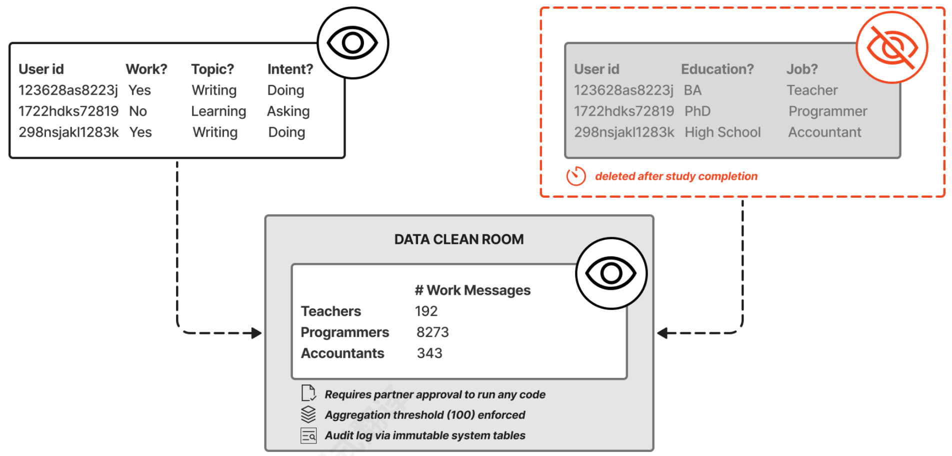


Figure 2: Illustration of Aggregated Employment Category Analysis via a Data Clean Room. All queries run in the Data Clean Room must be approved by our data partner, enforcing a strict aggregation threshold (100 observations). As a result, researchers cannot access user-level employment data, only aggregated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

3.4 Summarizing Our Approach to Privacy

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

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

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

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

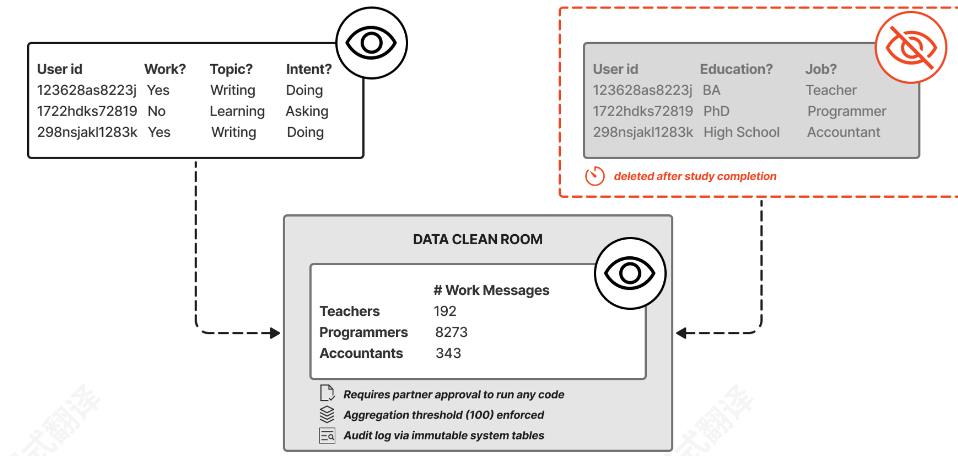


图2：通过数据清洁室进行聚合就业类别分析的说明。数据清洁室中运行的所有查询都必须经过我们的数据合作伙伴批准，并执行严格的聚合阈值（100个观察值）。因此，研究人员无法访问用户级别的就业数据，只能访问聚合的就业类别。

3.4 总结我们的隐私保护方法

我们在分析的每个阶段都采取了措施来保护用户隐私。总而言之，我们方法的关键要素是：

> 消息的自动分类。在分析过程中，没有人直接查看用户消息的内容：我们对用户消息内容的所有分析都是通过在去标识化和PII清理的使用数据上运行的自动分类器输出的。
>
> 通过数据清洁室获取的聚合就业数据。我们通过安全的数据清洁室环境分析和报告聚合就业数据：研究团队中的任何人都无法直接访问用户级别的人口统计数据，并且我们的分析不会报告用户数少于100个组的聚合数据。

通过采取这些措施，我们旨在匹配或超过其他研究聊天机器人的社会科学家以及将数字平台数据链接到外部来源所设定的隐私保护先例。

我们遵循近期对聊天机器人对话分析中建立的先例（Phang 等人 (2025)，Eloundou 等人 (2025)，Handa 等人 (2025)，Tomlinson 等人 (2025)），这些分析依赖于自动分类而不是对原始文本进行人工检查。特别是，Phang 等人 (2025) 对 ChatGPT 情感使用的分析以及 Eloundou 等人 (2025) 对聊天机器人第一人称公平性的调查都通过自动分类器分析 ChatGPT 消息内容，并强调基于分类器的标记作为一种可扩展、保护隐私的方法。Anthropic 的 Handa 等人 (2025) 采用了类似的方法：他们的 Clio 方法将自动分类器应用于大量对话集合，将对话分类成数千个主题，并在他们的附录中描述了对采样对话的手动验证（100 个被标记用于审查的用户对话和 100 个随机采样的校准）。像 Eloundou 等人一样，我们使用 WildChat（一个用户对话的公共数据集）来验证我们的分类器。

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

4 The Growth of ChatGPT

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

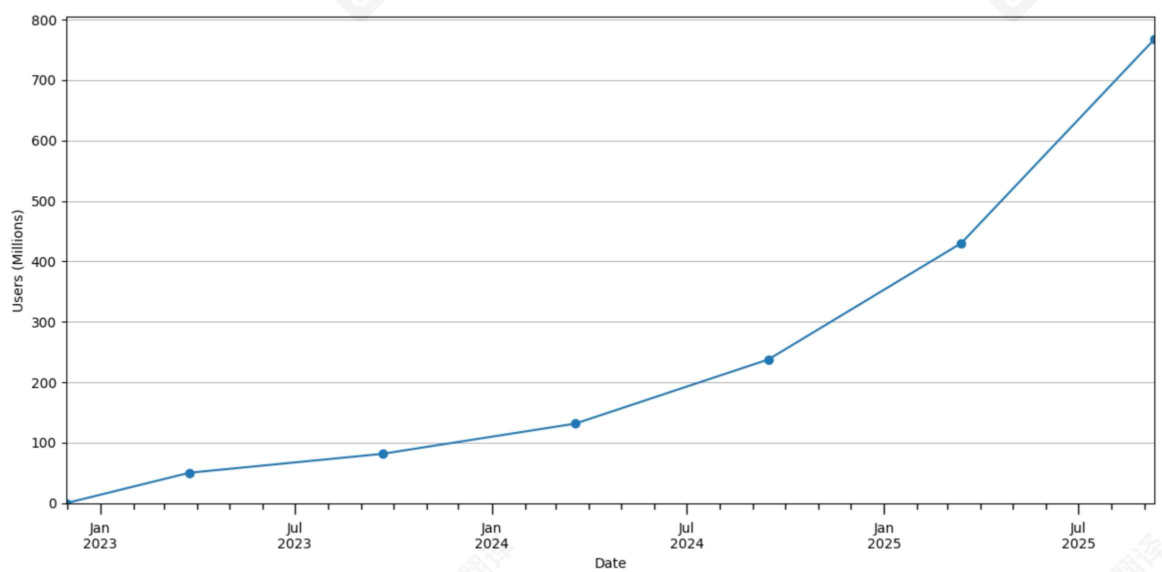


Figure 3: Weekly active ChatGPT users on consumer plans (Free, Plus, Pro), shown as point-in-time snapshots every six months, November 2022–September 2025.

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

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

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

其他论文已经分析了数字行为和人口统计数据；我们在此提及一些相关的先例。例如，Humlum 和 Vestergaard (2025b) 以及 Humlum 和 Vestergaard (2025a) 分析了关于聊天机器人使用的 大规模调查以及丹麦行政劳动力市场数据。Chetty 等人 (2022) 分析了去识别化的 Facebook 友谊图和匿名化的 IRS 税收记录，这些记录按邮编级别聚合。

4 ChatGPT的增长

ChatGPT于2022年11月30日作为"研究预览版"向公众发布，到12月5日已有超过一百万注册用户。图3报告了消费者计划中每周活跃用户（WAU）随时间的变化。一年后，ChatGPT的登录WAU超过1亿，两年后接近3.5亿。到2025年7月底，ChatGPT的总WAU超过7亿，接近世界成年人口的10%。[20]

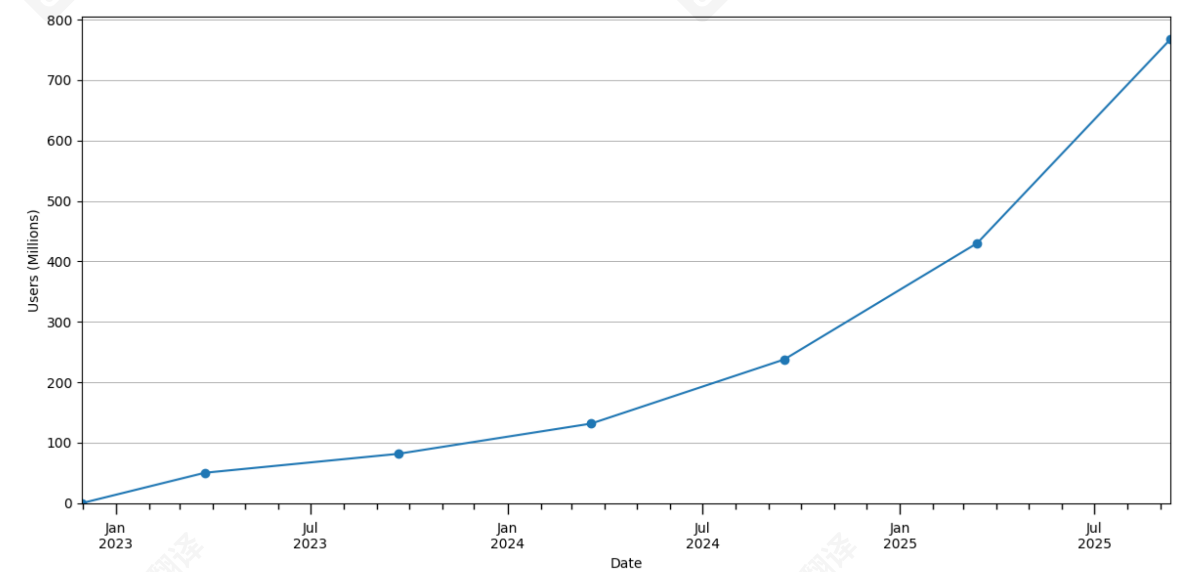


图3：消费者计划中每周活跃的ChatGPT用户（免费、Plus、Pro），以每六个月的即时快照形式显示，2022年11月–2025年9月。

图4展示了用户发送总消息量随时间的变化。实线显示，2024年7月至2025年7月期间，发送消息的数量增长了5倍以上。

图4也显示了各个用户群体对总消息量的贡献。黄色线代表ChatGPT的第一批用户：他们的使用量在2023年有所下降，但在2024年末开始再次增长，现在已经超过了有史以来的最高水平。粉色线代表在2023年第三季度或更早注册的用户，因此黄色线和粉色线之间的差异代表了

[20]请注意，当一个人有两个账户（或对于未登录用户，一个人使用两个设备）时，我们预期的不同账户数量会略高于不同人数。对于登录用户，计数基于不同的登录凭证（电子邮件地址），一个人可能有多个账户。对于未登录用户，计数基于不同的浏览器cookie；如果有人清除cookie后返回ChatGPT，或者在一周内使用两个不同的设备访问ChatGPT，这会导致重复计数。

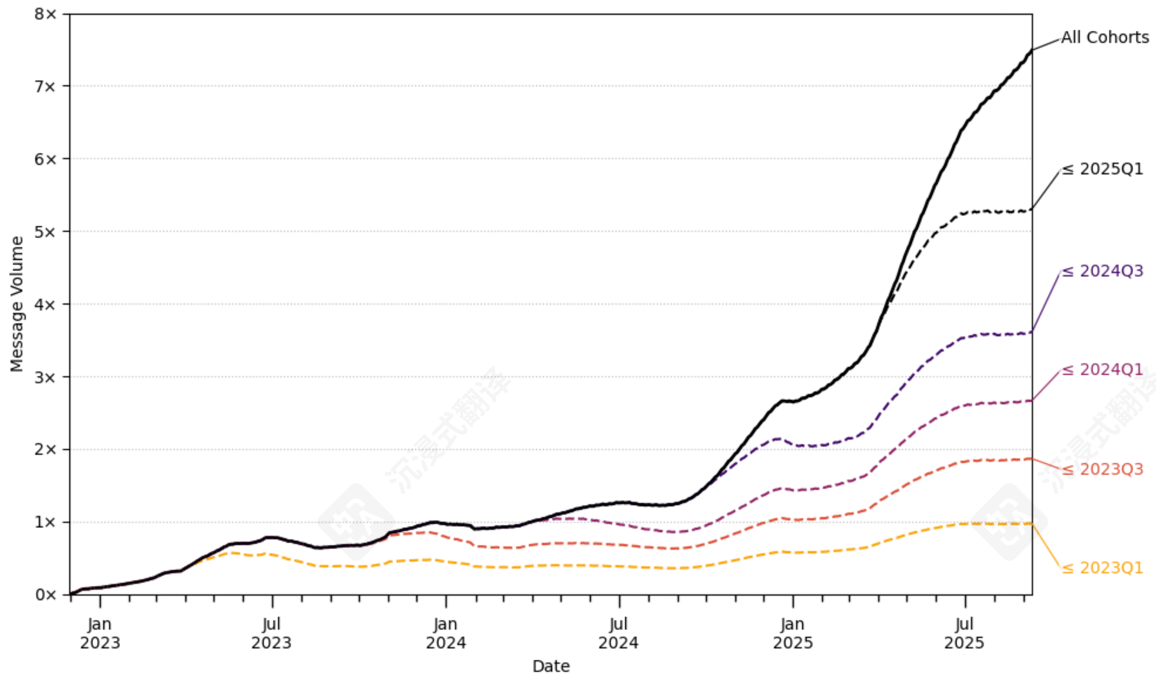


Figure 4: Daily message volumes from ChatGPT consumer plans (Free, Plus, Pro), split by sign-up date of the requesting user. Reported values are moving averages of the past 90 days. Y-axis is an index normalized to the reported value for "All Cohorts" at the end of Q1 2024 (April 1, 2024).

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

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

5 How ChatGPT is Used

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

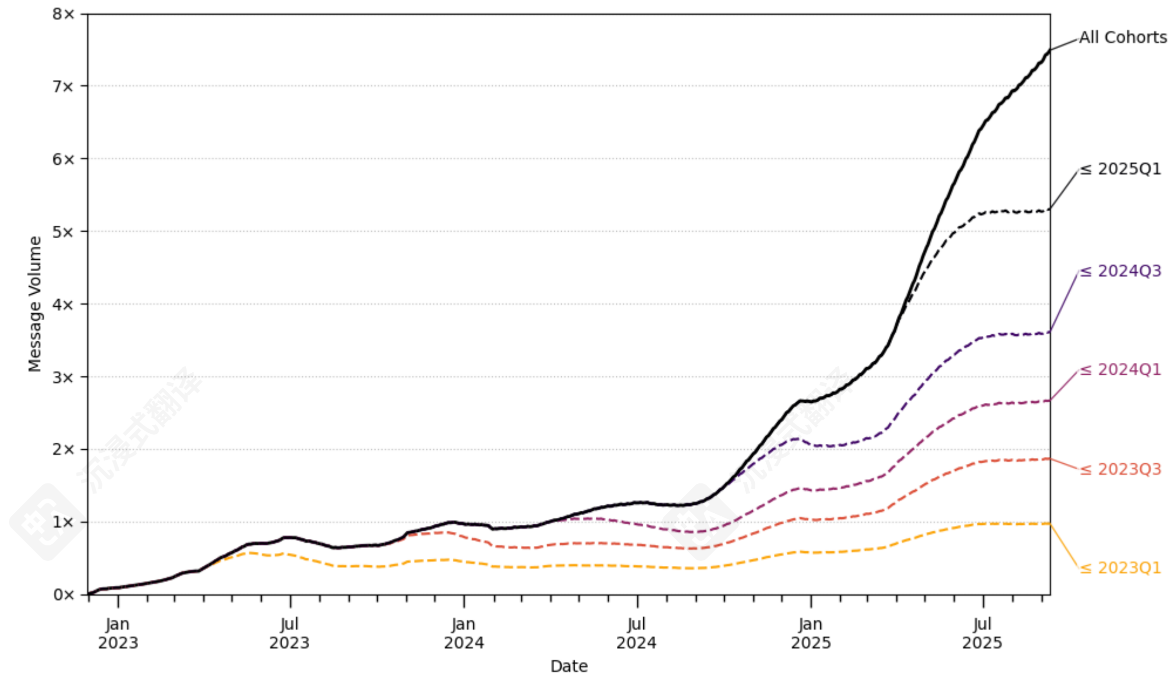


图4：来自ChatGPT消费者套餐（免费、高级、专业）的每日消息量，按请求用户的注册日期划分。报告值为过去90天的移动平均值。Y轴是一个指数，已归一化到2024年第一季度末（2024年4月1日）的“所有群体”报告值。

在2023年第二季度和第三季度注册的用户发送的消息。无论是新用户群体还是现有用户群体的消息量都出现了显著增长。

图5对每个群体进行标准化，绘制每周活跃用户每日消息量。每条线代表一个独立的群体（而不是图4中的累积群体）。该图显示，较早注册的用户始终具有更高的使用量，但每个群体内的使用量也始终在增长，我们将其解释为由于（1）模型能力的改进和（2）用户逐渐发现现有功能的新用途。

5 ChatGPT的使用方式

我们接下来报告了使用多种不同分类体系对ChatGPT对话内容进行分析的结果。对于每种分类体系，我们描述了一个“提示”，该提示定义了一组类别，然后应用一个大型语言模型（LLM）将每条消息映射到一个类别。我们的类别通常适用于用户的意图，而不是对话文本本身，因此我们从未直接观察到真实情况。尽管如此，分类器的结果可以被解释为人类可能做出的最佳猜测：LLM的猜测与相同提示下人类的猜测高度相关，当提示包含一个“不确定”的第三类别时，我们也会得到类似的定性结果。

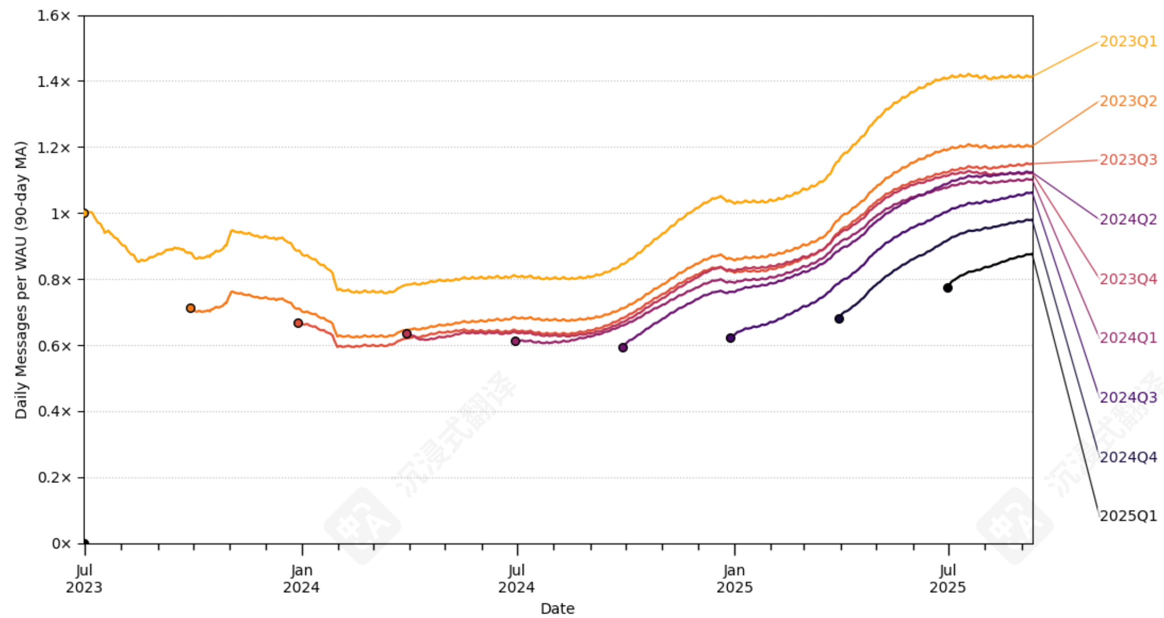


Figure 5: Daily messages sent per weekly active user, split by sign-up cohort. Sample only considers users of ChatGPT consumer plans (Free, Plus, Pro). Reported values are moving averages of the past 90 days and are reported starting 90 days after the cohort is fully formed. Y-axis is an index normalized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

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

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

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

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

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

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

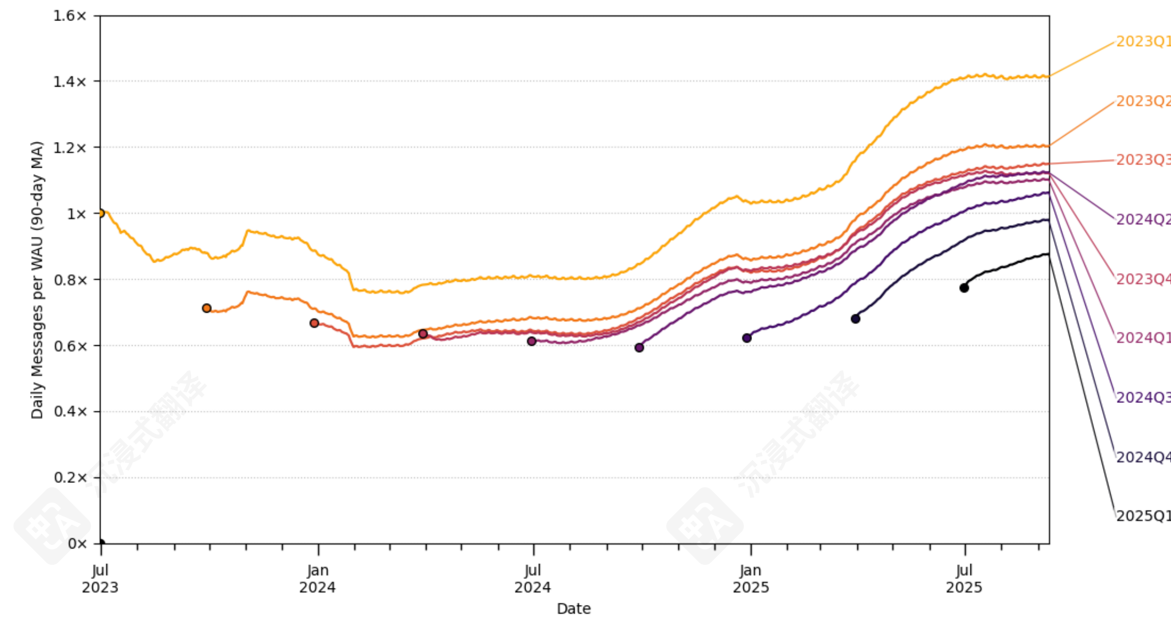


图5：每周活跃用户每日发送的消息量，按注册组分层。样本仅考虑使用ChatGPT消费者套餐（免费、Plus、Pro）的用户。报告的值为过去90天的移动平均值，并在组完全形成后90天开始报告。Y轴为Q1 2023组首次报告值的标准化指数。

5.1 ChatGPT查询中有多少比例与付费工作相关?

我们根据用户消息是否似乎与工作相关，使用一个LLM分类器对数据集中的每条用户消息进行标记。提示的关键部分如下：[21]

> 这次对话转录的最后一个用户消息似乎可能与做些工作/就业有关吗？用以下选项之一回答：
>
> *(1)* 可能是工作的一部分（例如，“重写这个*HR*投诉”），
>
> *(0)* 可能不是工作的一部分（例如，“冰能减少粉刺吗？”），

表1显示，这两种类型的查询在2024年6月至2025年6月期间迅速增长，但与工作无关的消息增长更快：2024年6月有53%的消息与工作无关，到2025年6月这一比例上升至73%。

图6展示了按累计注册群体分解的非工作消息份额。后续群体中非工作消息的份额更高，但在每个群体内部，其非工作使用量也有所增加。将所有用户的份额（黑线）与最早用户群体的份额（黄线）进行比较，我们可以看到它们非常紧密地跟踪。

[21]请参见附录A获取完整提示，请参见附录B进行验证。

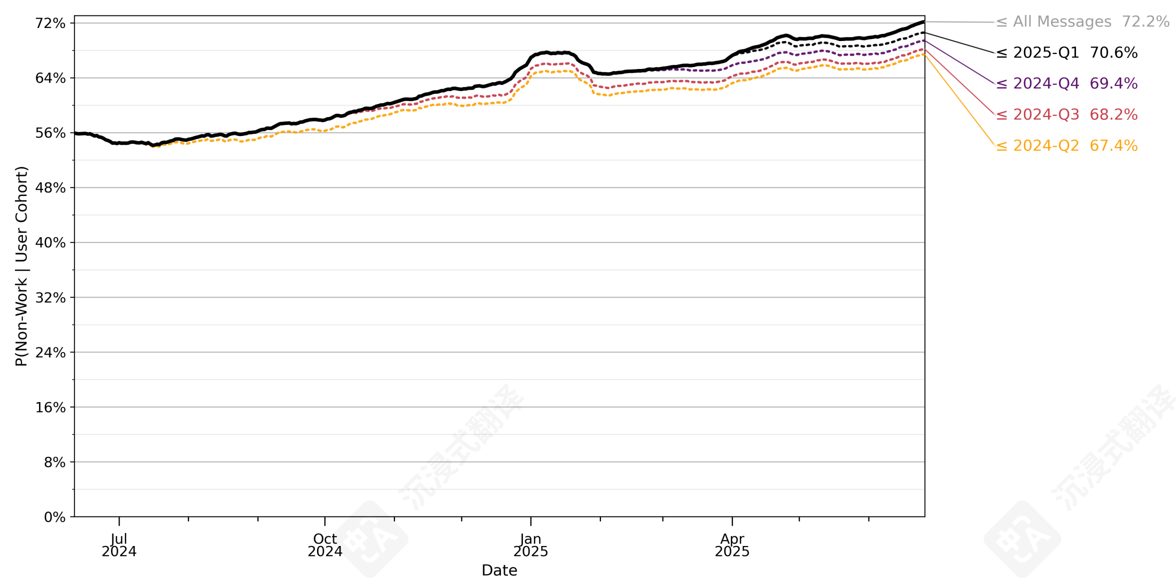


Figure 6: The solid black line represents the probability that a messages on a given day is not related to work, as determined by an automated classifier. Values are averaged over a 28-day lagging window. The dotted orange line shows the same calculation, but conditioned on messages being from users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

5.2 What are the topics of ChatGPT conversations?

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

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

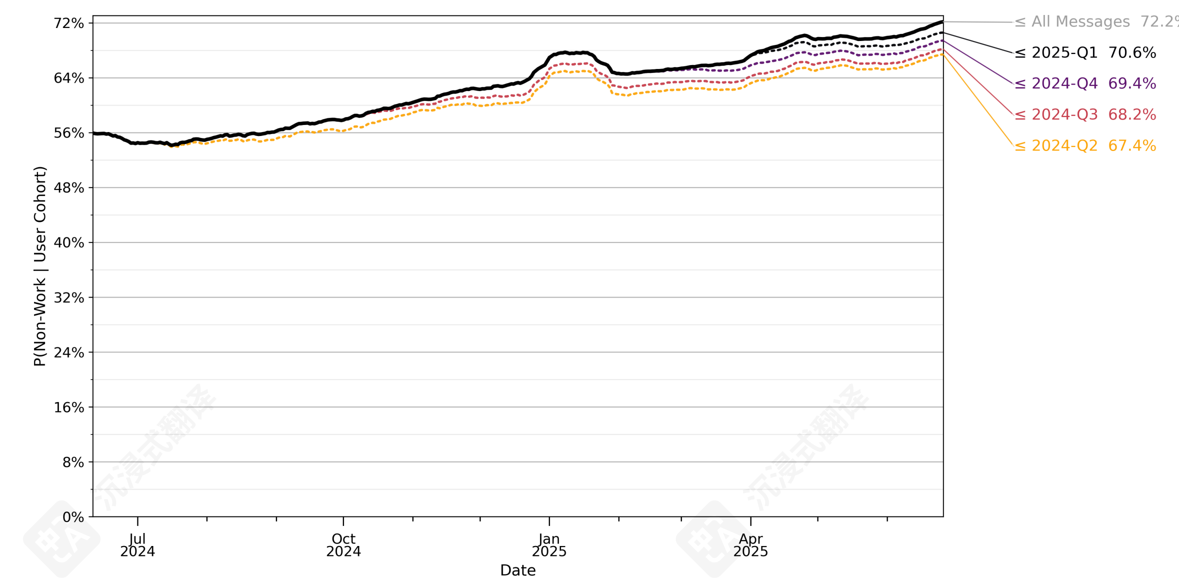


图6：实线黑色表示自动分类器确定的某一天的消息与工作无关的概率。值在一个28天的滞后窗口内进行平均。虚线橙色显示了相同的计算结果，但条件是消息来自在2024年第二季度或之前首次使用ChatGPT的用户。其余线条类似地定义了连续的季度，颜色为更近期的队列变冷。计数基于从2024年5月15日到2025年6月26日抽取的约110万次对话样本计算。观察值重新加权以反映某一天的总消息量。抽样细节见第3节。

5.2 ChatGPT对话的主题是什么?

我们修改了OpenAI内部研究团队使用的分类器，该分类器用于识别用户正在向ChatGPT请求哪些功能。该分类器本身直接将用户的查询分配到24个类别中的一个。我们将这24个类别聚合为七个主题分组（完整的对话分类提示在附录A中给出）：

主题	对话类别
写作	编辑或评论提供的文本 个人写作或交流 翻译 论点或摘要生成 写小说
实用指南	如何操作建议 辅导或教学 创意构思 健康、健身、美容或自我护理
技术支持	数学计算数据分析

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

Table 3: Coarse Conversation Topics and Underlying Classifier Categories

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

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

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

主题	对话类别
	计算机编程
多媒体	创建图像
	分析图像
	生成或检索其他媒体
寻求信息	具体信息
	可购买的产品
	烹饪和食谱
自我表达	问候和闲聊
	人际关系与个人反思
	游戏与角色扮演
其他/未知	询问关于模型
	其他
	不清楚

表3：粗粒度对话主题和底层分类器类别

图7显示了用户消息随时间的组成。三种最常见的对话主题是实用指南、寻求信息和写作，它们共同占所有ChatGPT对话的约77%。实用指南的总体使用率保持在约29%的稳定水平。写作从2024年7月的所有使用率的36%下降到一年后的24%。寻求信息在同一时期从14%增长到24%的所有使用率。技术帮助的份额从2024年7月的所有使用率的12%下降到一年后的约5%——这可能是因为通过API（ChatGPT外部）对LLM用于编程的使用增长非常迅速，用于代码编辑的AI协助以及自主编程代理（例如Codex）。多媒体从2%增长到略超过7%，在2025年4月ChatGPT发布新的图像生成功能后出现了一个大幅波动：波动减弱，但水平有所提升。

图8显示了对话主题，将样本限制为仅包含工作相关消息。2025年7月所有工作相关消息中约有40%是写作，这是最常见的对话主题。实用指南是第二常见用例，占比24%。技术支持从2024年7月所有工作相关消息的18%下降到2025年7月的10%以上。

图9将七个对话主题中的四个分解为更小的组，并汇总每种类型的消息一年内的总量。例如，写作的五个子类别按频率排序为：编辑或评论提供的文本、个人写作或沟通、翻译、论点或摘要生成以及写作小说。这五个类别中的三个（编辑或评论提供的文本、翻译以及论点或摘要生成）是要求修改ChatGPT由用户提供的文本，而另外两个是要求生成新文本。前者占所有写作对话的三分之二，这

14

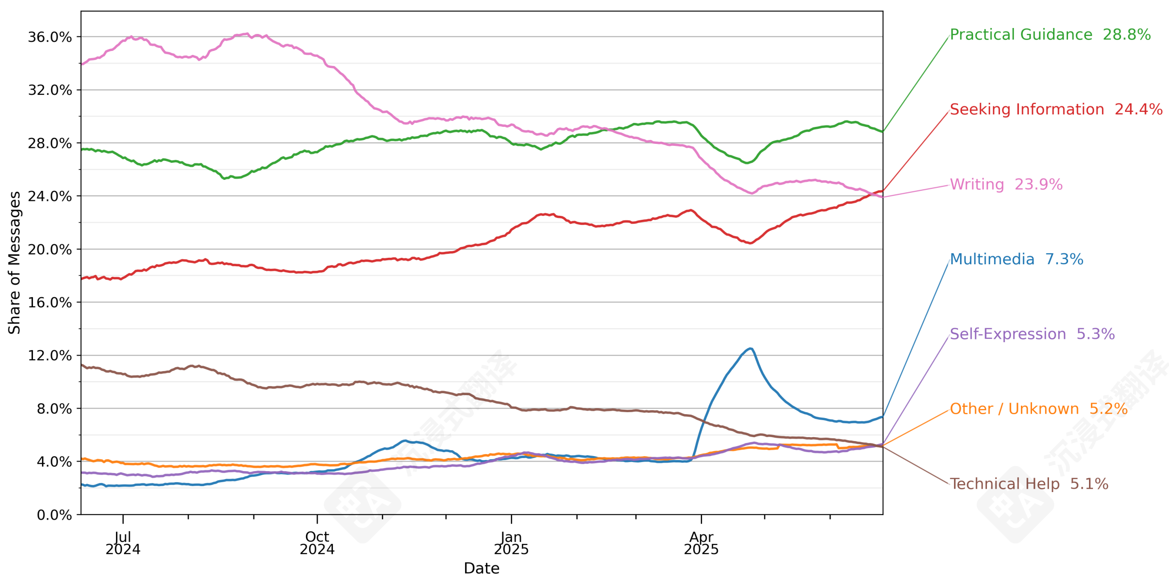


Figure 7: Share of consumer ChatGPT messages broken down by high level conversation topic, according to the mapping in Table 3. Values are averaged over a 28 day lagging window. Shares are calculated from a sample of approximately 1.1 million sampled conversations from May 15, 2024 through June 26, 2025.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

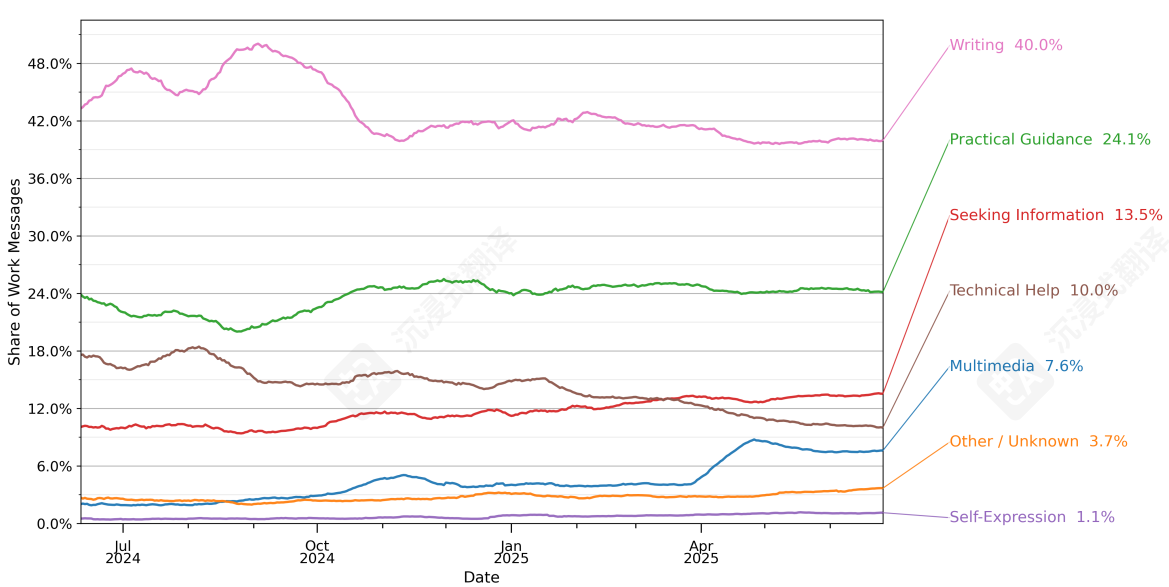


Figure 8: Share of **work related** consumer ChatGPT messages broken down by high level conversation topic, according to the mapping in Table 3. Values are averaged over a 28 day lagging window. Shares are calculated from a sample of approximately 1.1 million sampled conversations from May 15, 2024 through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

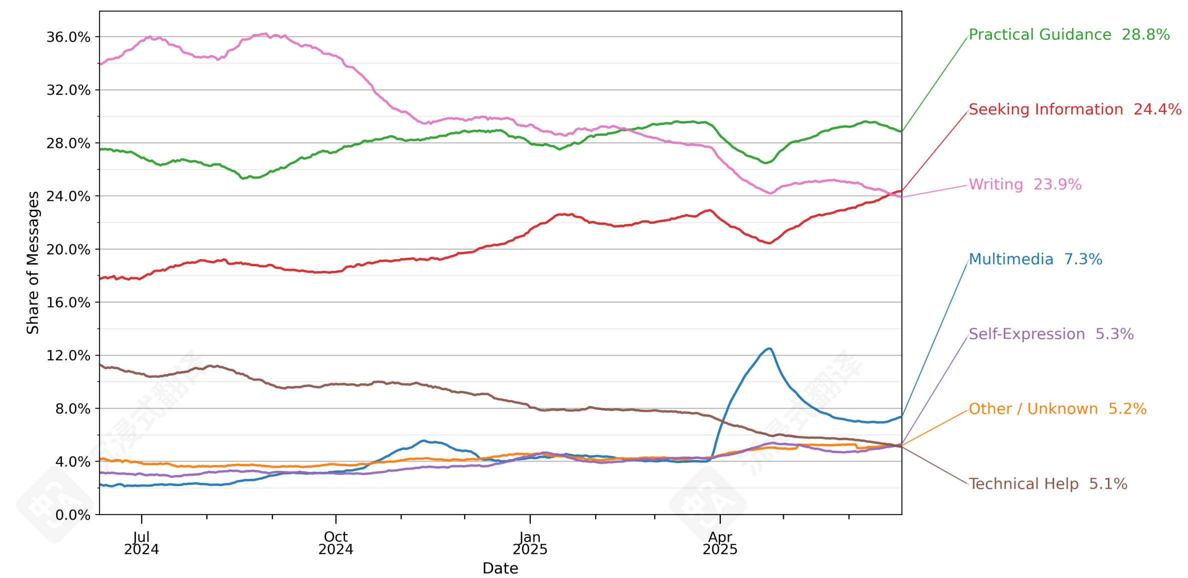


图7：按高级对话主题分解的消费ChatGPT消息份额，根据表3中的映射。值在28天的滞后窗口内进行平均。份额是根据从2024年5月15日到2025年6月26日抽取的约110万次抽样对话样本计算的。观察值被重新加权以反映特定日期的总消息量。抽样细节见第3节。

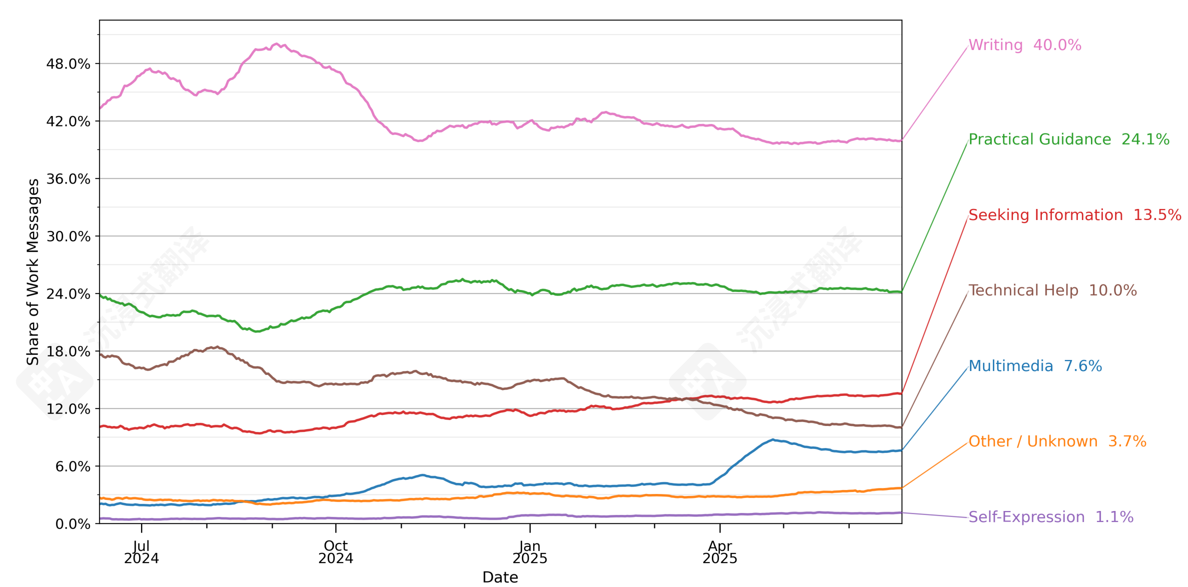


图8：按高级对话主题分解的工作相关消费ChatGPT消息份额，根据表3中的映射。值在28天的滞后窗口内进行平均。份额是根据从2024年5月15日到2025年6月26日抽取的约110万次抽样对话样本计算的。观察值被重新加权以反映特定日期的总消息量。抽样细节见第3节。

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

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

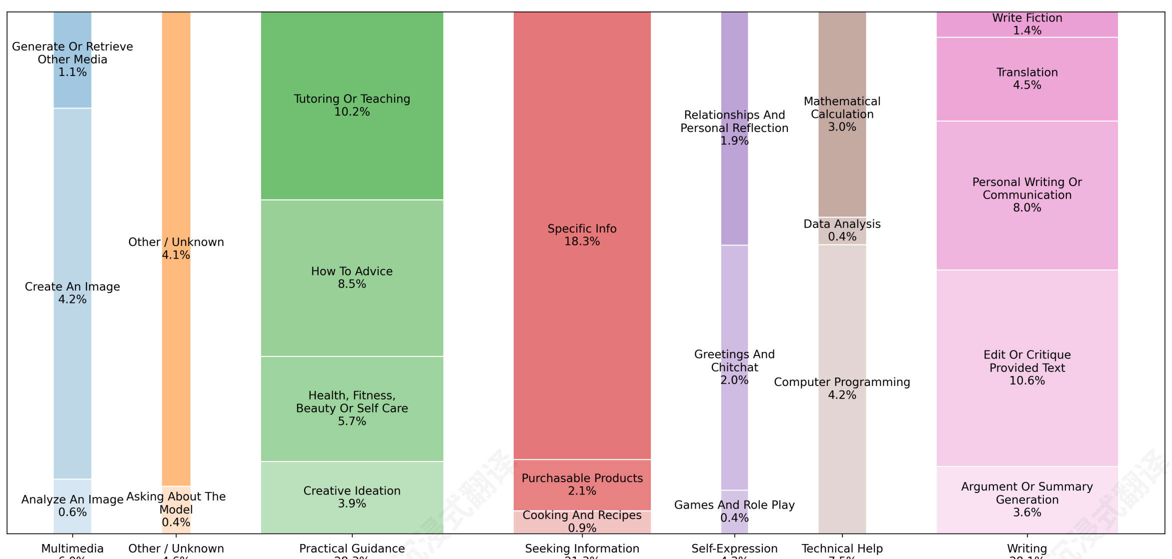


Figure 9: Breakdown of granular conversation topic shares within the coarse mapping defined in Table 3. The underlying classifier prompt is available in Appendix A. Each bin reports a percentage of the total population. Shares are calculated from a sample of approximately 1.1 million sampled conversations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

5.3 User Intent

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

表明大多数用户与ChatGPT的写作对话是请求修改用户输入，而不是创造新内容。教育是ChatGPT的主要用例。所有用户消息的10.2%以及实践指导消息的36%是关于辅导或教学的请求。另一个大份额——总体的8.5%以及实践指导的30%——是关于各种主题的一般性操作建议。技术支持包括计算机编程（消息的4.2%）、数学计算（3%）和数据分析（0.4%）。从自我表达的主题来看，所有ChatGPT消息中只有2.4%是关于人际关系和个人反思（1.9%）或游戏和角色扮演（0.4%）。

虽然用户可以从传统网络搜索引擎以及ChatGPT中寻求信息和建议，但生成写作、软件代码、电子表格和其他数字产品的能力将生成式AI与现有技术区分开来。ChatGPT在传统应用如寻求信息和实践指导方面也比网络搜索更灵活，因为用户会收到定制化回复（例如，量身定制的健身计划、新产品创意、橄榄球联盟队名的想法），这些回复代表新生成的内容或对用户提供内容的创新修改以及后续请求。

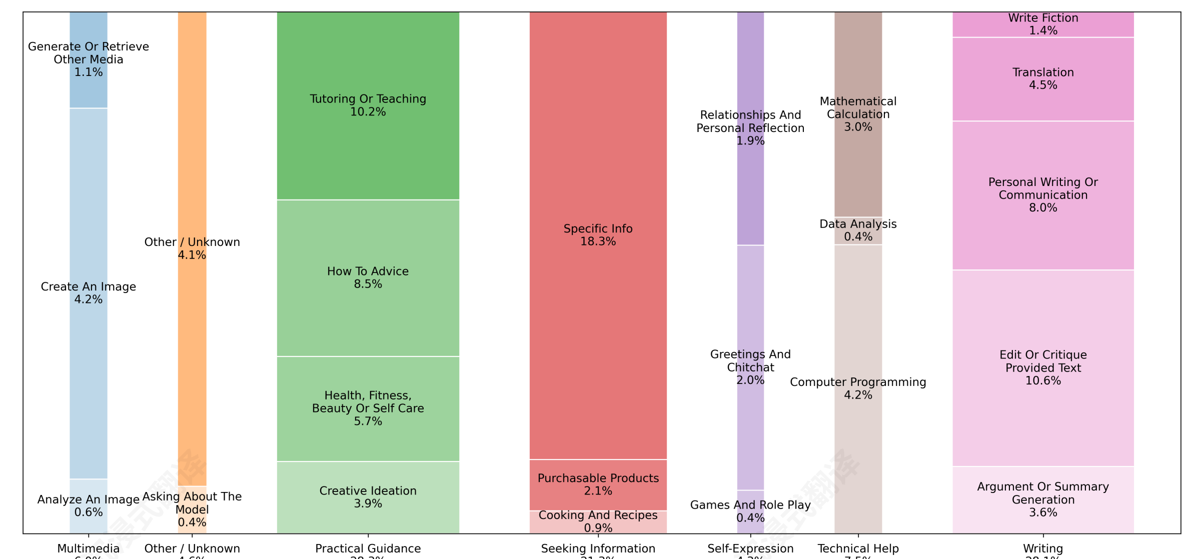


图9：表3中定义的粗粒度映射内粒状对话主题份额的分解。底层分类器提示词在附录A中提供。每个箱报告了总人口的一个百分比。份额是从2024年5月15日至2025年6月26日从约110万次采样对话中计算的。观察值重新加权以反映特定日期的总消息量。采样细节在第3节中提供。

5.3 用户意图

现有关于生成式AI经济影响的研究几乎完全集中于AI执行工作场所任务的可能性，无论是增强还是自动化人类劳动（例如Eloundou等人（2025年）、Handa等人（2025年）、Tomlinson等人（2025年））。然而，生成式AI具有高度灵活性

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

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

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

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

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

一种可用于多种不同方式的技术。为了了解人们如何在工作和工作之外寻求使用生成式AI，我们引入了一种分类器，旨在衡量用户希望收到的输出类型。具体而言，我们根据用户意图对消息进行分类，并根据简单的询问、执行或表达标准对对话进行编码。我们分类提示的关键部分如下：

意图	提示
询问	询问是指获取信息或建议，以帮助用户更好地了解情况或做出更好的决定，无论是在工作、学校还是在他们的个人生活中。（例如，"林肯之后谁担任总统？"，"我该如何创建一个"本季度预算是多少？"，"去年的通货膨胀率是多少？""相关性和因果关系有什么区别？"，"我应该……"在开放注册期间选择健康计划时，如何寻找？
做	执行消息请求*ChatGPT*为用户执行任务。用户是起草邮件、编写代码等。如果消息包含对主要由模型创建的输出的请求，则将其分类为"*doing*"。包括由模型主要创建的输出请求。（例如）"将此邮件改写得更加正式"，"起草一份总结*ChatGPT*用例的报告"，"生成一个包含里程碑和风险的表格项目时间表"，"从此文本中提取公司、人员和日期"，"从此文本中提取公司、人员和日期"。导入 *CSV*。"，"编写一个 *Dockerfile* 和一个最小的 *docker-compose.yml* 用于此应用程序。"）
表达	表达语句既不是询问信息，也不是让聊天机器人执行任务。

从概念上讲，执行对话是提供可以插入生产流程的输出，而询问对话支持决策但不直接产生输出，表达对话几乎没有或没有经济内容。

图 10 显示了我们样本中每种意图类型的消息份额。49% 的用户消息是询问，40% 是执行，11% 是表达。该图还显示了与我们主题分类的关系：这两个分类法是相关的，但不是冗余的：询问查询更有可能是实用指导和寻求信息。执行查询不成比例地涉及写作和多媒体。表达查询不成比例地涉及自我表达。然而，重叠是不完美的。例如，在实用指导主题中，一个询问消息可能是根据用户的个人历史提供的关于如何从运动损伤中恢复的建议，而一个执行消息可能要求 ChatGPT 为用户生成一个可以打印或保存的定制恢复和训练计划。在技术帮助中，一个询问消息可能要求帮助理解如何调试某些代码，而一个执行消息可能要求 ChatGPT 直接为用户编写代码。

图11展示了仅针对工作相关消息的询问/执行/表达的比例。执行占工作相关查询的近56%，而询问占35%，表达占9%。近35%的所有工作相关查询是关于写作的执行消息。执行和询问在工作技术支持查询中占比相同。

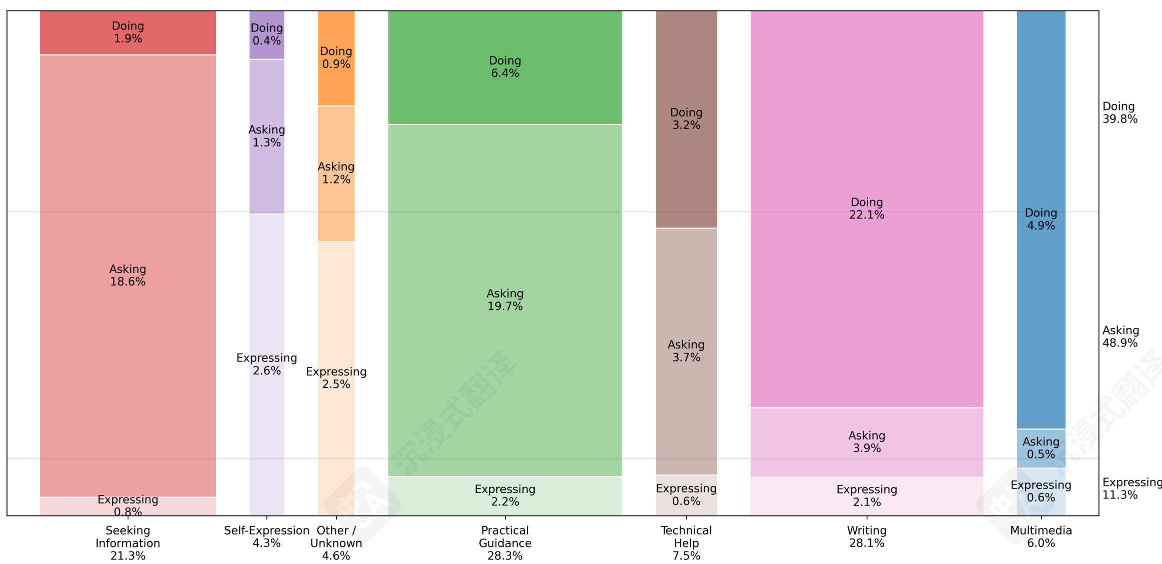


Figure 10: Breakdown of Conversation Topics by Asking/Doing/Expressing category, with topic columns sorted by relative share of "Doing" messages. Prompts for these automated classifiers are available in Appendix A. For a detailed breakdown of conversation topic contents, see Table 3. Each bin reports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

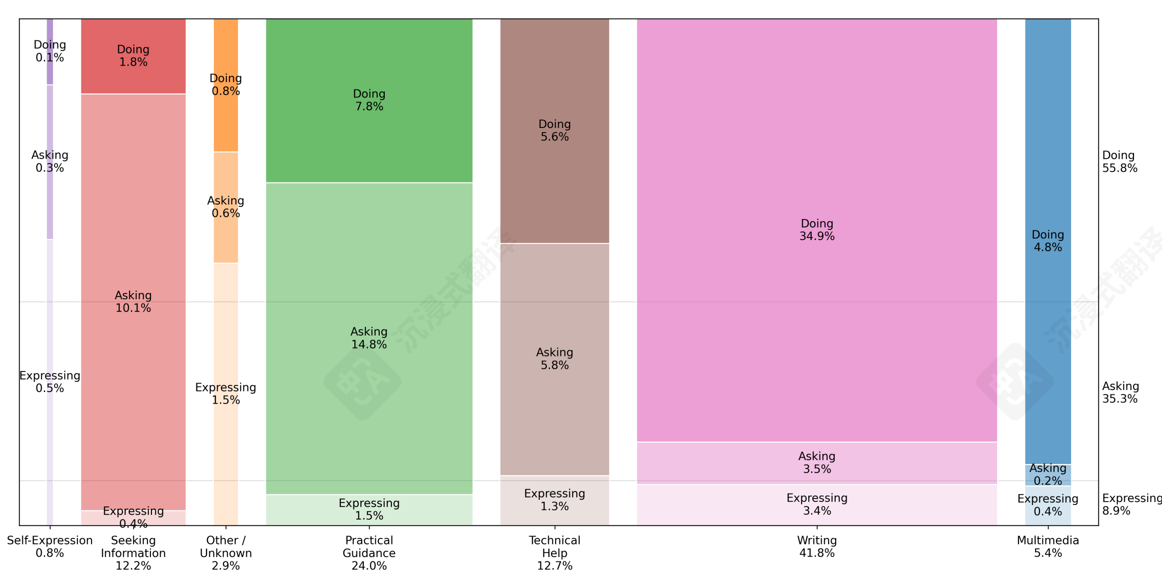


Figure 11: Breakdown of Conversation Topics by Asking/Doing/Expressing category for **only work-related messages**, with topic columns sorted by relative share of "Doing" messages. Prompts for these automated classifiers are available in Appendix A. For a detailed breakdown of conversation topic contents,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

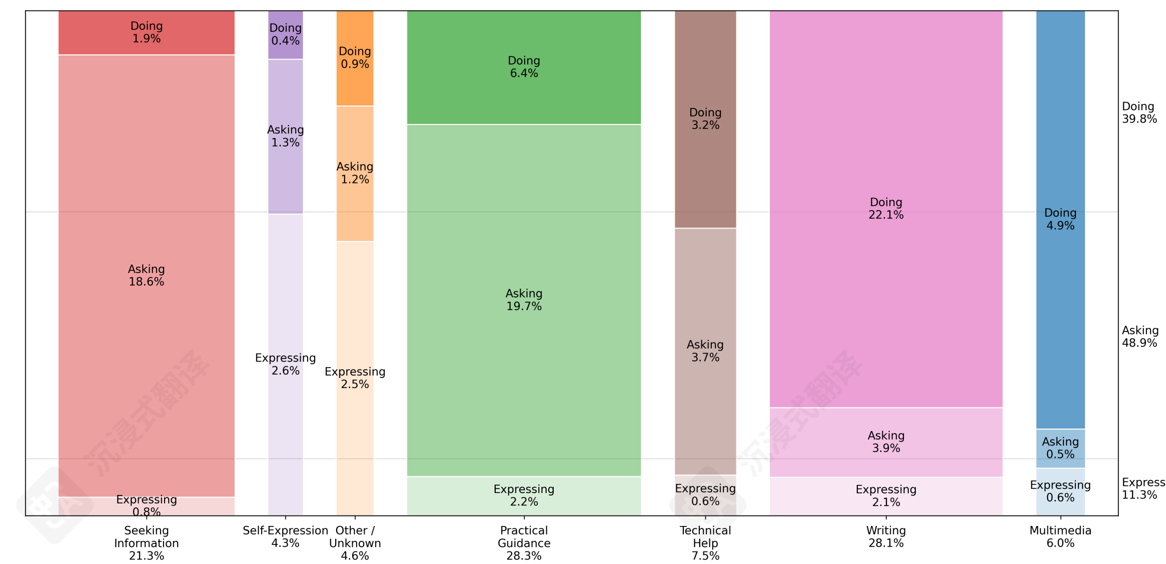


图 10：按提问/执行/表达类别分解的对话主题，主题列按“执行”消息的相对份额排序。这些自动分类器的提示词在附录 A 中提供。有关对话主题内容的详细分解，请参见表 3。每个 bin 报告了总人口的一个百分比。份额是根据从 2024 年 5 月 15 日到 2025 年 6 月 26 日的约 110 万个抽样对话计算得出的。观察值重新加权以反映特定日期的总消息量。抽样细节在第 3 节中提供。

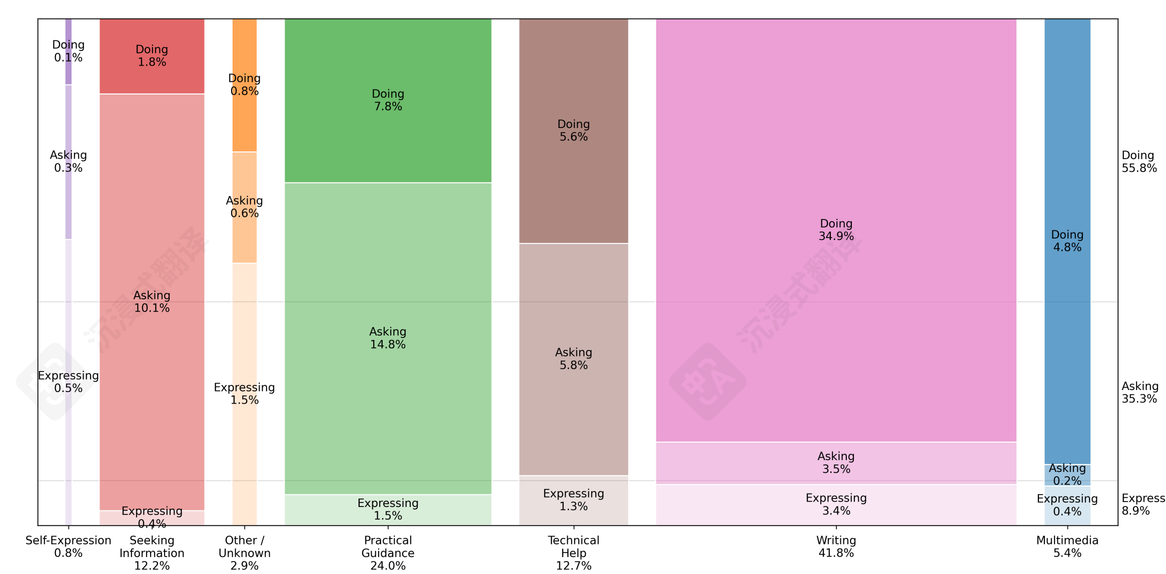


图 11：仅针对工作相关消息按提问/执行/表达类别分解的对话主题，主题列按“执行”消息的相对份额排序。这些自动分类器的提示词在附录 A 中提供。有关对话主题内容的详细分解，请参见表 3。每个 bin 报告了总人口的一个百分比。份额是根据从 2024 年 5 月 15 日到 2025 年 6 月 26 日的约 110 万个抽样对话计算得出的。观察值重新加权以反映特定日期的总消息量。抽样细节在第 3 节中提供。

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

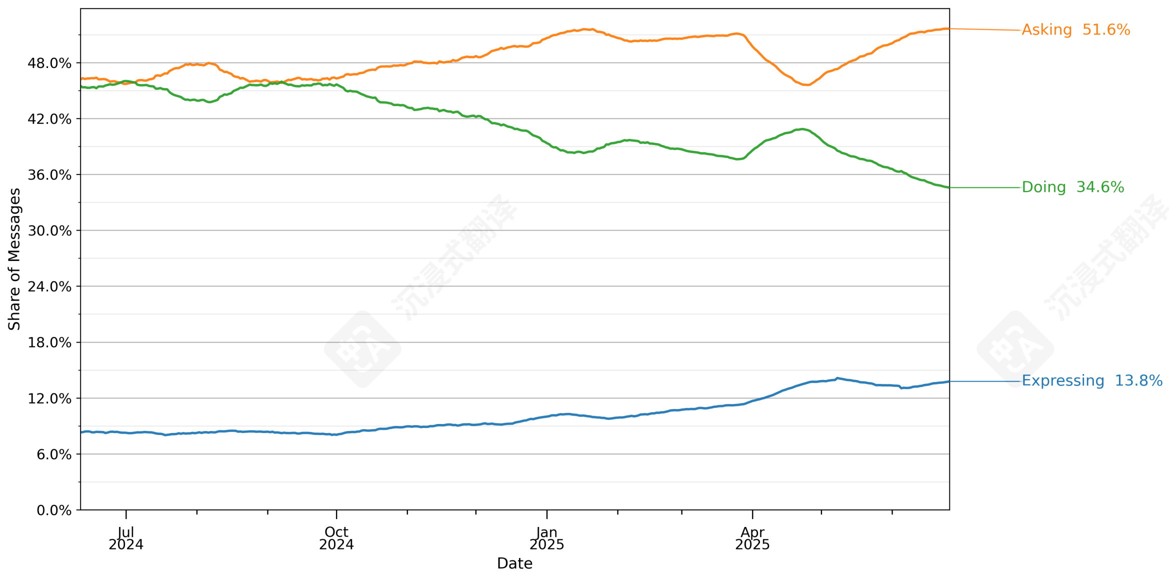


Figure 12: Shares of messages classified as Asking, Doing, or Expressing by an automated ternary classifier. Values are averaged over a 28 day lagging window. Shares are calculated from a sample of approximately 1.1 million sampled conversations from May 15, 2024 through June 26, 2025. Observations are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

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

5.4 O*NET Work Activities

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

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

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

图12展示了按用户意图划分的消息组成随时间的变化。2024年7月，使用情况在询问（Asking）和执行（Doing）之间均等分配，不到8%的消息被归类为表达（Expressing）。在接下来的这一年里，询问和表达的增长速度远快于执行，到2025年6月下旬，分配比例变为51.6%询问，34.6%执行，13.8%表达。

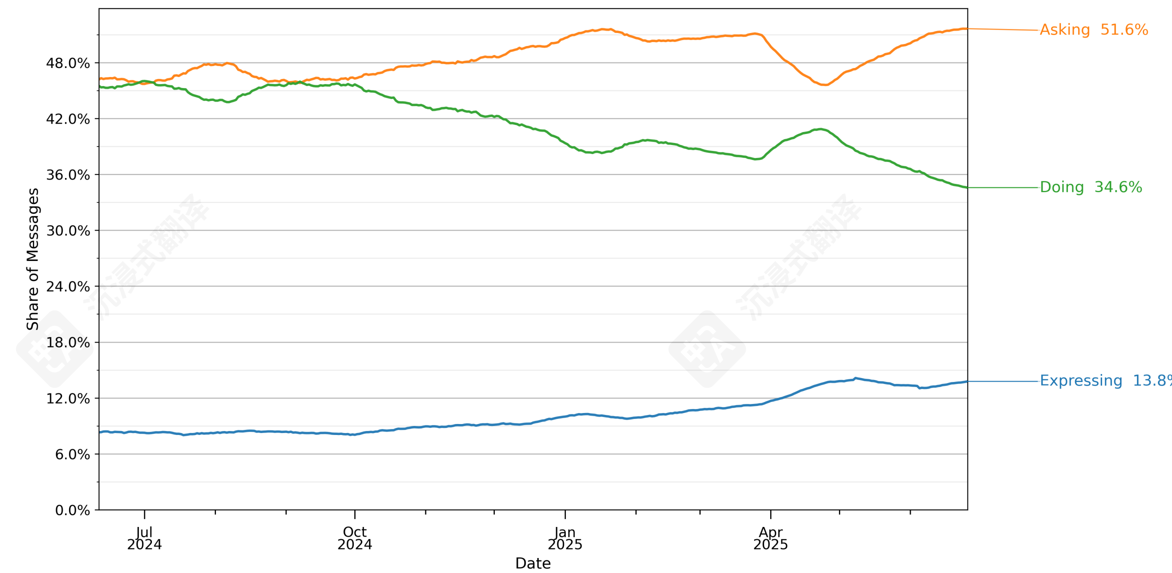


图12：由自动三元分类器分类的询问、执行或表达消息的占比。数值基于28天的滞后窗口进行平均。占比数据来自2024年5月15日至2025年6月26日期间约110万条抽样对话。观察值根据每日总消息量进行重新加权。采样详情见第3节。

图13展示了按用户意图划分的工作相关消息占比。执行消息（约占消息的40%）在工作相关和非工作相关消息之间均等分配。

5.4 O*NET 工作活动

我们使用职业信息网络 (O*NET) 数据库版本 29.0 将消息内容映射到工作活动，类似于 Tomlinson 等人 (2025)。O*NET 由美国劳工部合作开发，并根据执行工作所需的技能、任务和工作活动系统地分类职业。O*NET 将每个职业与一组在不同强度级别下执行的任务相关联。然后，每个任务被汇总到三个级别的详细信息 - 2,087 个详细工作活动 (DWAs)、332 个中间工作活动 (IWAs) 和 41 个概括性工作活动 (GWAs)。

要了解与ChatGPT使用相关的工作活动，我们将消息映射到332O*NET中间工作活动（IWA）之一，并添加了一个“模糊”选项，以解释用户消息缺乏足够上下文的情况。[22] 然后，我们使用了官方

[22]我们从 2024 年 5 月到 2025 年 6 月抽取了约 110 万次对话的样本，在每个样本中随机选择一条消息，并根据 A 中的提示对其进行分类。

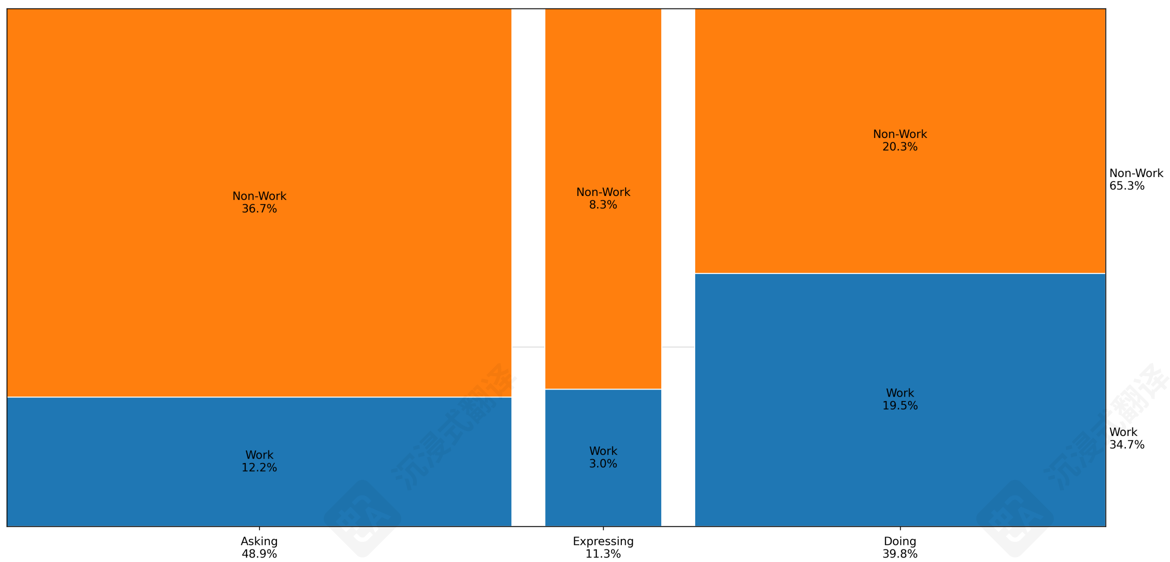


Figure 13: Shares of Asking, Doing, and Expressing messages split by work vs. non-work. See A to review the prompts used by the automated classifiers. The annotations on the right show the shares of work and non-work for the full sample. Each bin reports a percentage of the total population. Shares are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

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

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

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

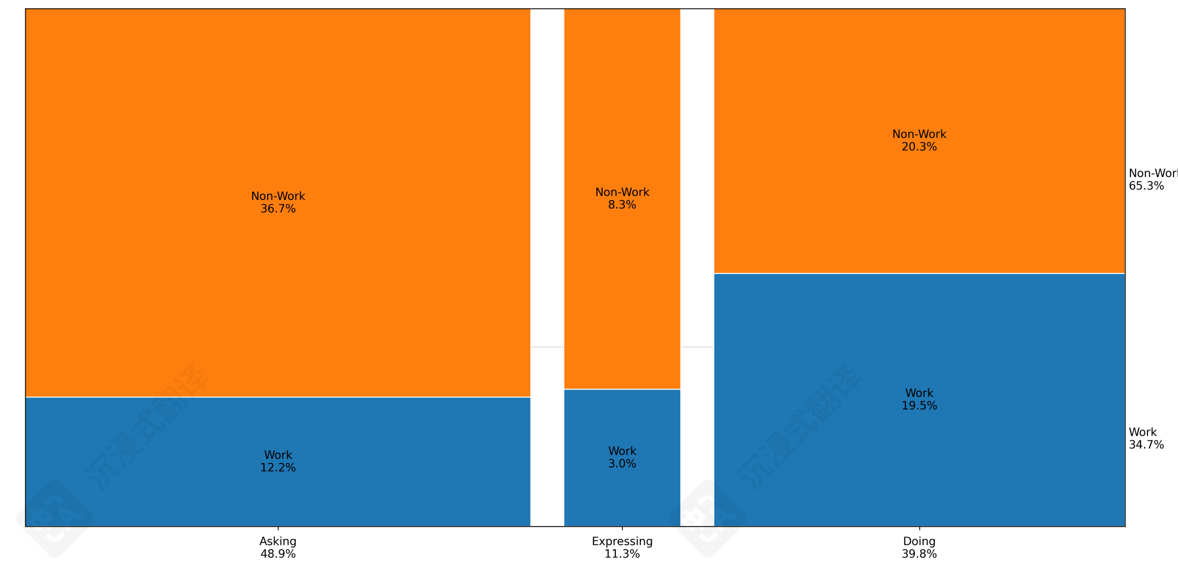


图13：按工作与非工作分类的请求、执行和表达消息的份额。参见A以回顾自动分类器使用的提示。右侧的注释显示了完整样本的工作和非工作份额。每个bin报告了总人口百分比。份额基于从2024年5月15日到2025年6月26日采样的约110万次对话计算。观察值重新加权以反映特定日期的总消息量。采样细节见第3节。

O*NET分类法将这些分类的IWAs映射到广义工作活动（GWA）之一。我们没有显示以下GWA的份额，因为每个类别发送消息的用户少于100人，并将它们归入“抑制”类别。

图14展示了按降序排列的每个GWA所属消息的份额。几乎所有消息（45.2%）都属于三个与信息使用和操作相关的GWA：获取信息（19.3%）、为他人解释信息含义（13.1%）和记录/记录信息（12.8%）。最常见的其他工作活动是提供咨询和建议（9.2%）、创造性思考（9.1%）、做决策和解决问题（8.5%）以及使用计算机（4.9%）。这七个GWA共同占所有消息的76.9%。

图15展示了我们将其分类为与工作相关的消息子样本的GWAs分布。在与此类消息相关的GWAs中，最常见的是记录/信息记录（18.4%）、决策和解决问题（14.9%）、创造性思考（13.0%）、使用计算机（10.8%）、为他人解释信息含义（10.1%）、获取信息（9.3%），以及向他人提供咨询和建议（4.4%）。这七项GWAs合计占与工作相关的消息的近81%。总体而言，ChatGPT在工作中的使用似乎主要集中在两个广泛功能上：1）获取、记录和解释信息；以及2）做出决策、提供建议、解决问题和进行创造性思考。

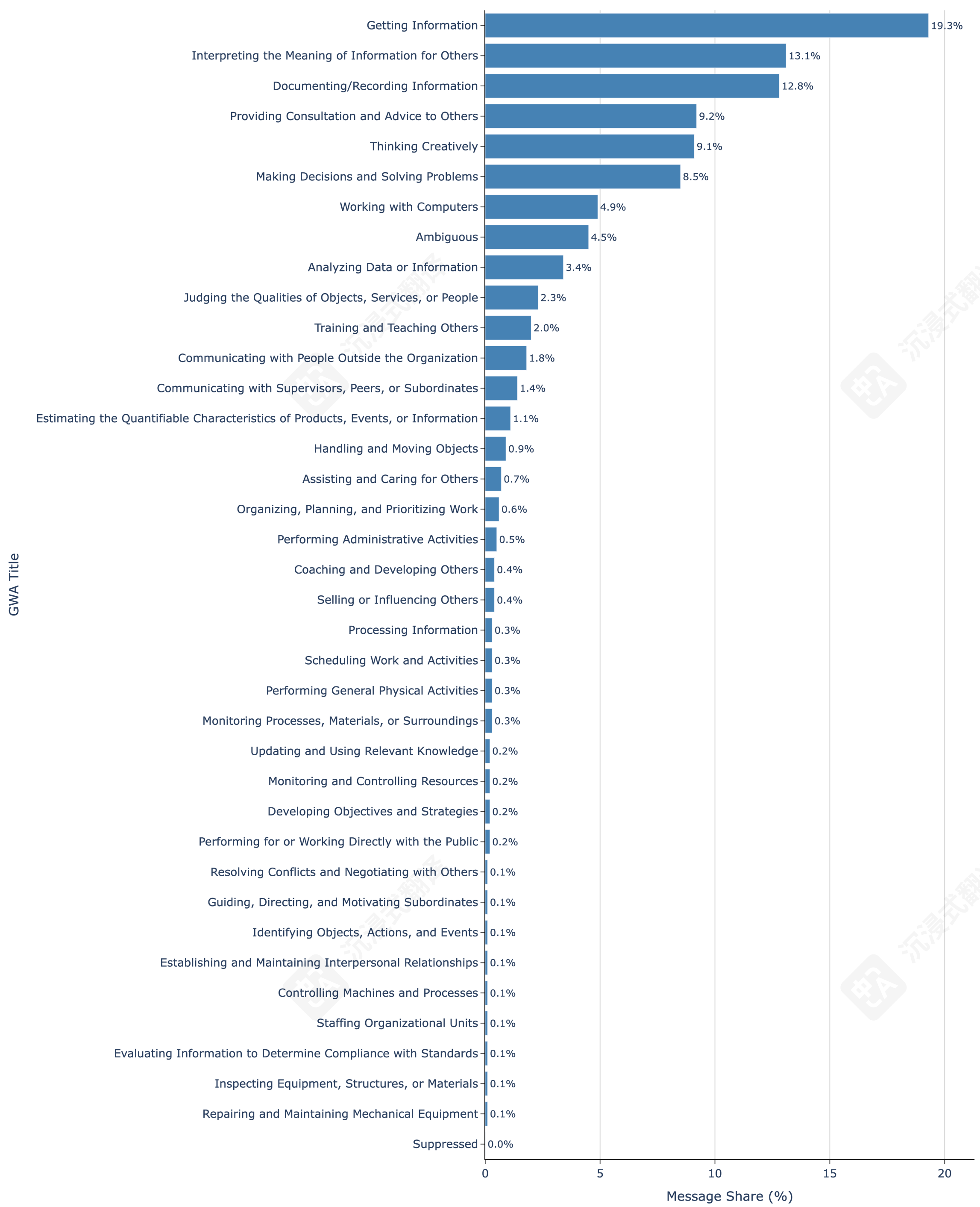


Figure 14: GWA Shares of 1.1M ChatGPT Messages. Messages are classified as pertaining to one of 332 O*NET IWAs, or *Ambiguous* using the prompt provided in the Appendix. IWAs were then aggregated to GWAs using the O*NET Work Activities taxonomy. Message sample from May 15, 2024 through June 26, 2025.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

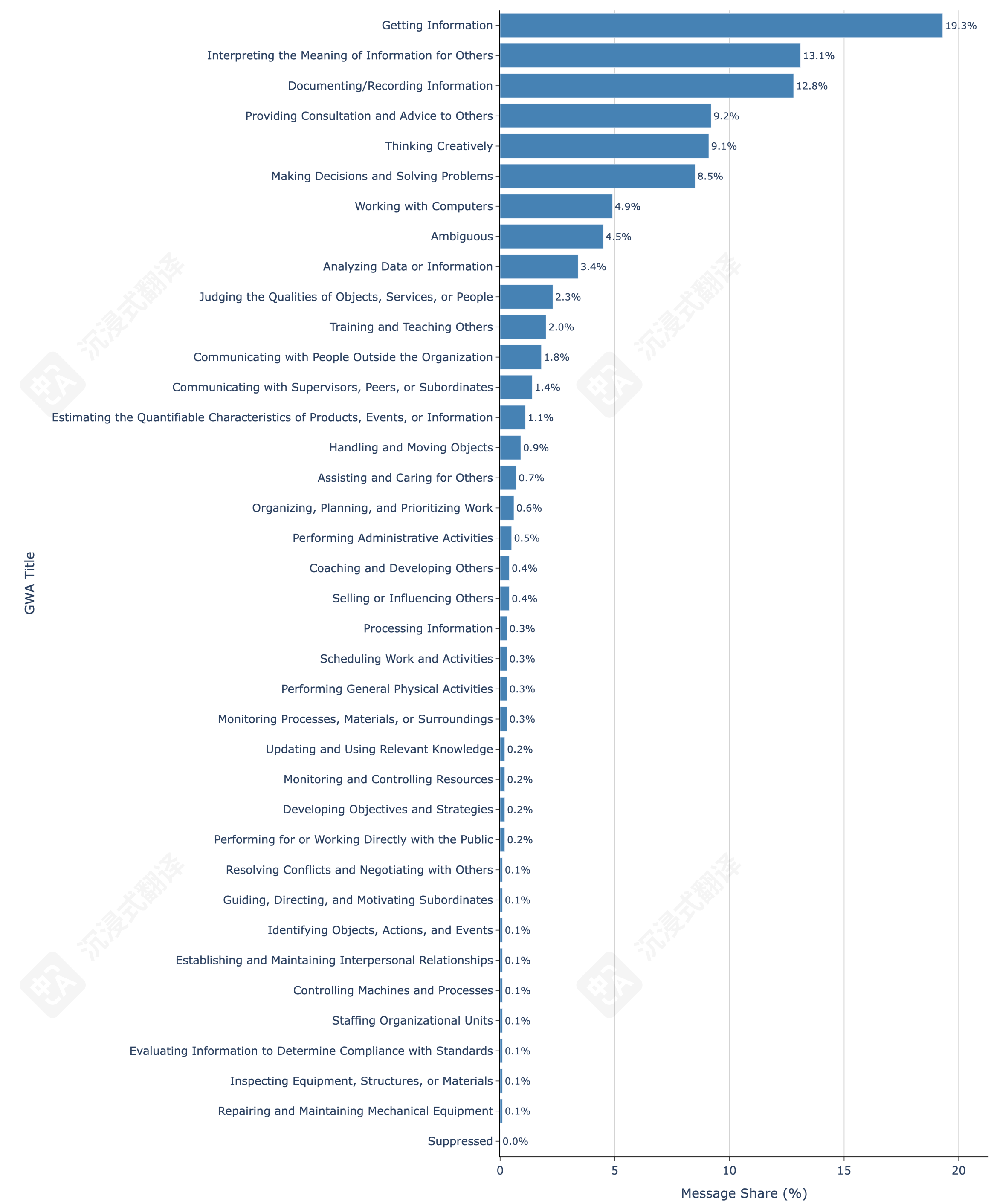


图14：1.1M个ChatGPT消息的GWA份额。消息被分类为属于332个O*NET IWAs中的某一个，或使用附录中提供的提示词 *Ambiguous* 。然后使用O*NET工作活动分类法将IWAs聚合为GWA。消息样本来自2024年5月15日26日，2025年100日至6月。我们没有显示以下GWA的份额，因为每个类别和组发送消息的用户数量少于 *Suppressed*，并将它们分组到 *Suppressed*中。

21

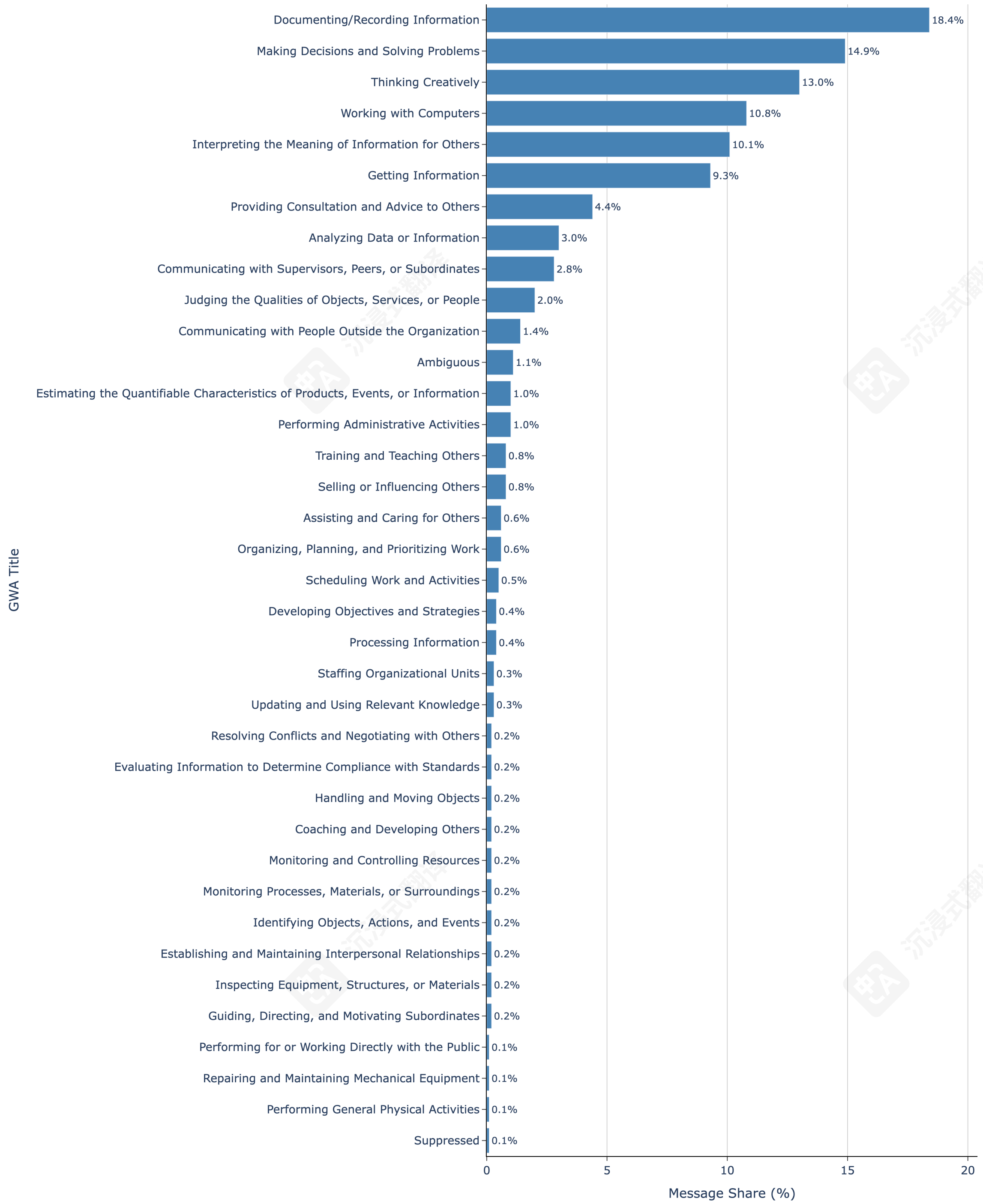


Figure 15: GWA Shares of approximately 366,000 Work-Classified Messages. Messages are classified as pertaining to one of 332 O*NET IWAs or *Ambiguous*. IWAs were then aggregated to GWAs using the O*NET Work Activities taxonomy. Messages were also additionally classified as pertaining to work or non-work.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

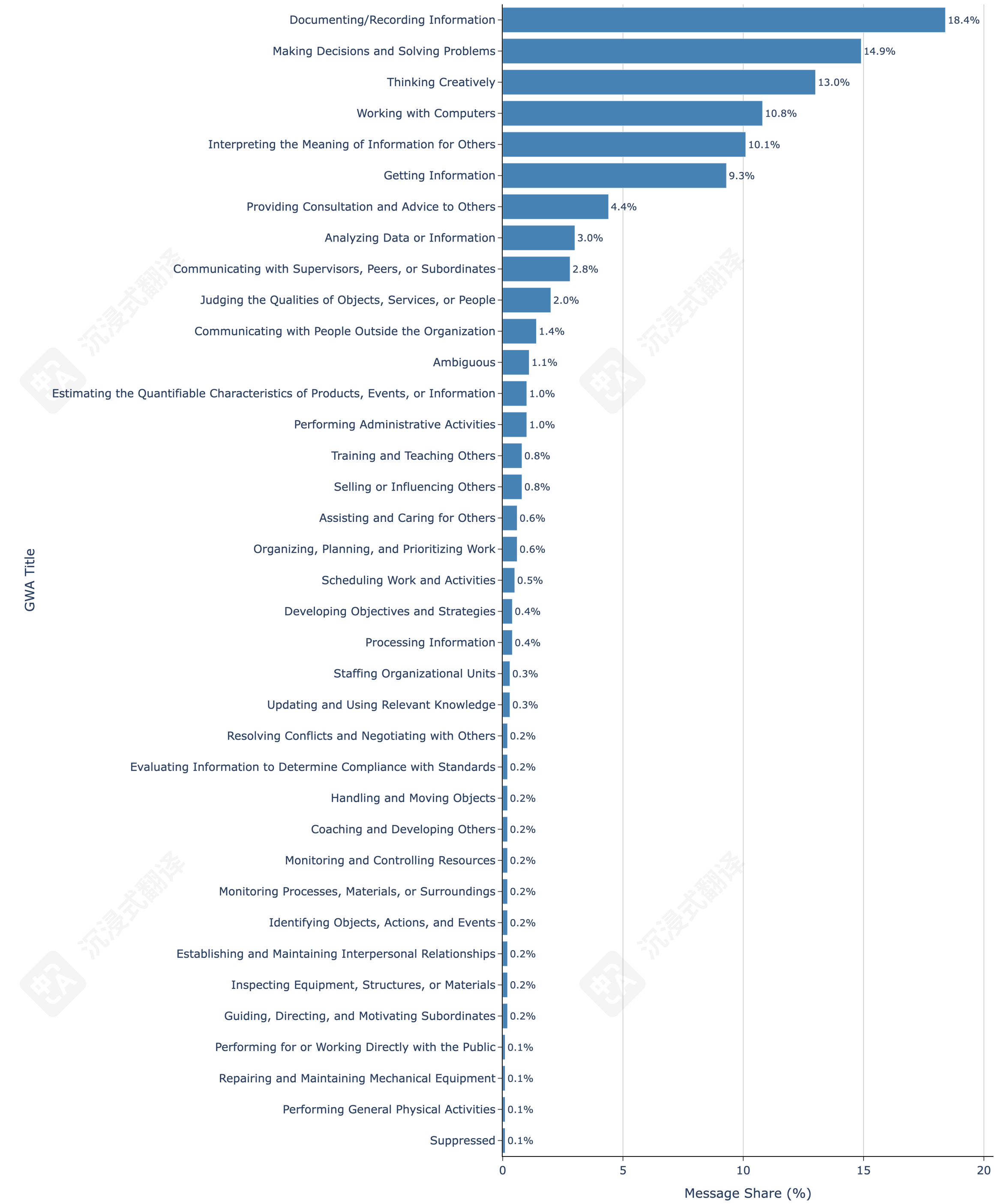


图15：约366,000条工作分类消息的GWA份额。消息被分类为属于332个O*NET IWAs或 *Ambiguous*中的任意一个。IWAs随后使用O*NET工作活动分类法聚合为GWA。消息也被额外分类为属于工作或非工作。仅显示工作分类消息的GWA份额。消息样本来自2024年5月15日至2025年6月26日。我们没有显示以下GWAs的份额，因为每个类别和组中发送消息的用户少于100人，并将它们分组到 *Suppressed*中。提示在附录中提供。

22

5.5 Quality of Interactions

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

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

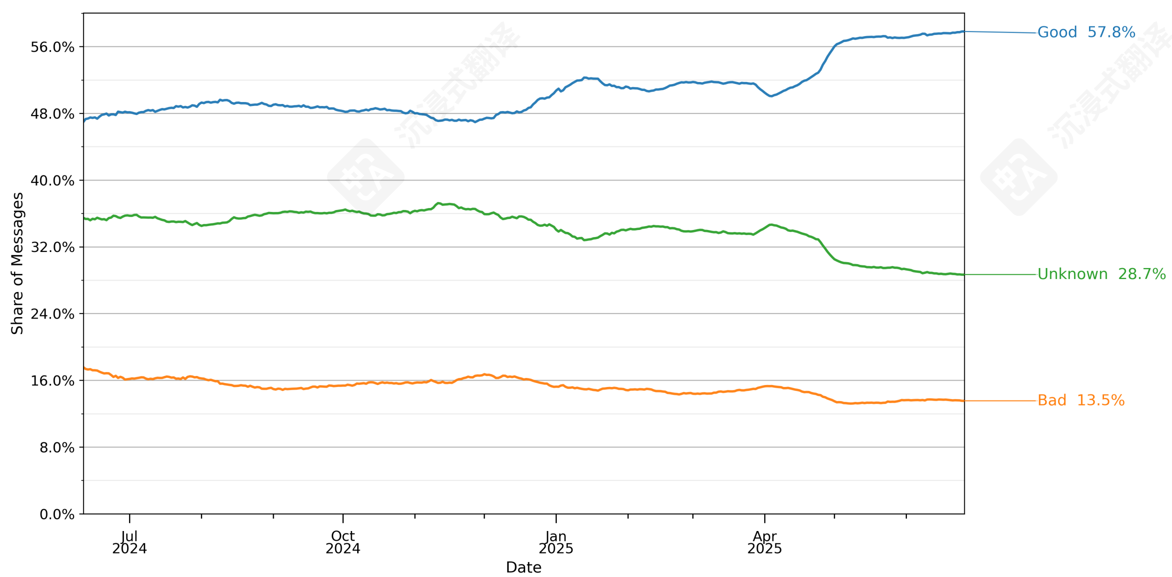


Figure 16: Interaction quality shares, based on automated sentiment analysis of the *next response* provided by the user. See Appendix B to understand how this classifier was validated. Values are averaged over a 28 day lagging window. Shares are calculated from a sample of approximately 1.1 million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

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

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

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

5.5 互动质量

我们此外还使用了自动分类器来研究用户对聊天机器人对其请求的响应的明显满意度。我们的互动质量分类器会在同一对话中用户后续消息中寻找满意度或不满的表达（如果存在），分为三个可能类别：好、坏和未知。 [23]

图16绘制了这三个桶中消息的整体增长。2024年末，好的互动大约是坏的互动的三倍常见，但好的互动在接下来的九个月内增长得更快，到2025年7月，它们已经比坏的互动常见四倍以上。

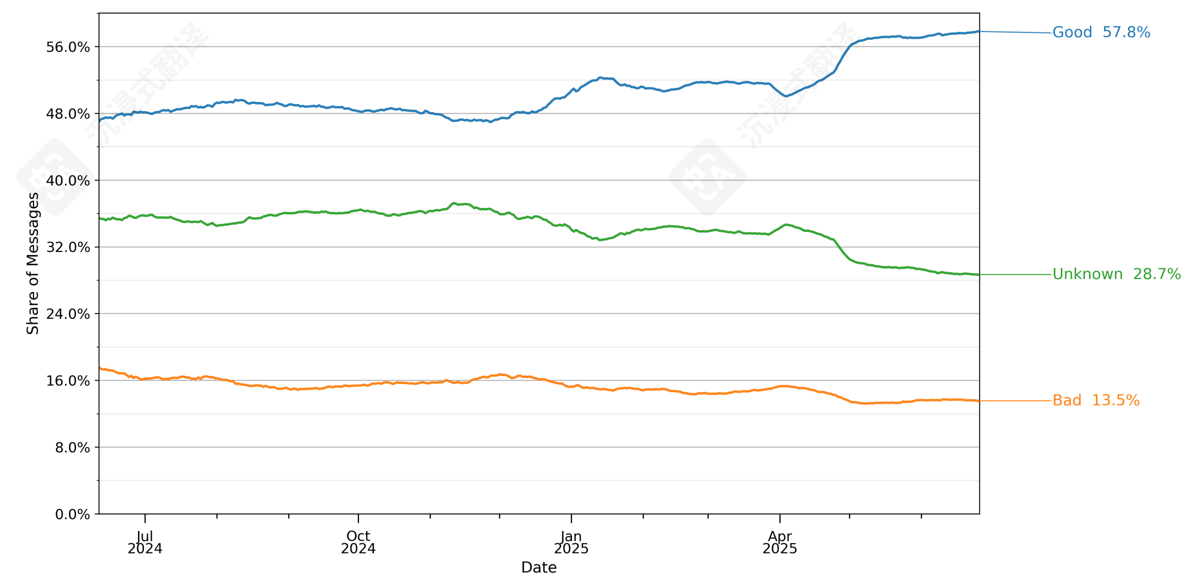


图16：互动质量份额，基于用户提供的 *next response* 的自动情感分析。参见附录B了解如何验证此分类器。值在28天的滞后窗口内取平均值。份额是从2024年5月15日至2025年6月26日大约110万次抽样对话的样本中计算的。观察值重新加权以反映特定一天的总消息量。抽样细节在第3节中提供。

关于此分类器的验证细节，以及它与用户提供的明确点赞/点踩标注的相关性测量结果，均包含在附录 B 中。

图 17 展示了按对话主题和交互类型划分的好消息与坏消息的比率，该比率由交互质量评定。A 面板显示，自我表达是评分最高的主题，其好消息与坏消息的比率超过七，与该类别的增长趋势一致。多媒体和技术支持的好消息与坏消息比率最低（分别为 1.7 和 2.7）。B 面板显示，与 Doing 或 Expressing 消息相比，Asking 消息更有可能获得好评。

[23] 对于此分类器，我们不公开提示词。

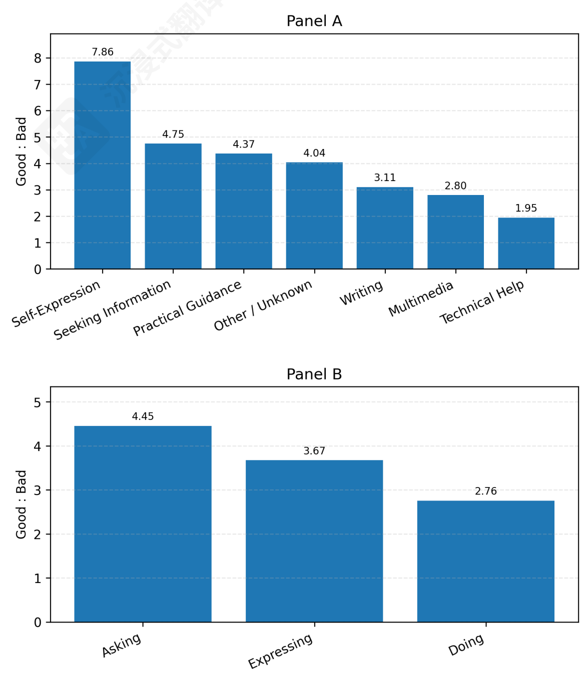


Figure 17: Average *Good* to *Bad* ratio for user interactions by Conversation Topic (Panel A) and Asking/Doing/Expressing classification (Panel B). The prompts for each of these automated classifiers (with the exception of interaction quality) are available in Appendix A. Values represent the average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

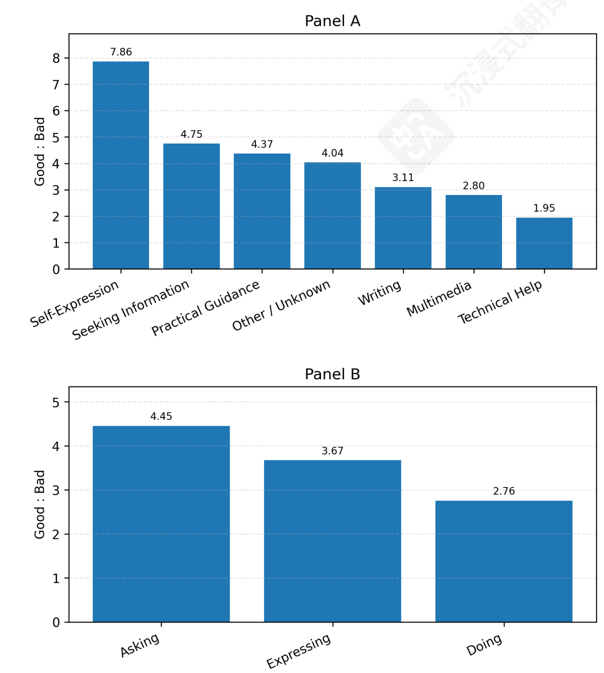


图17：按对话主题（A图）和提问/执行/表达分类（B图）的用户交互平均 *Good* 至 *Bad* 比率。这些自动化分类器（交互质量除外）的提示词在附录A中提供。数值表示2024年5月15日至2025年6月26日的平均比率，其中观察值根据每日总消息量进行重新加权。采样详情见第3节。

24

6 Who Uses ChatGPT

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

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

6.1 Name Analysis

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

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

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

6.2 Variation by Age

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

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

6 谁在使用 ChatGPT

在本节中，我们报告关于谁使用消费者 ChatGPT 的基本描述性事实。现有工作记录了美国代表性样本中（Bick 等人（2024年），Hartley 等人（2025年））以及丹麦部分职业群体中（Humlum 和 Vestergaard，2025a）生成式 AI 使用的差异。所有这些论文都发现，生成式 AI 更多地被男性、年轻人以及拥有高等教育和/或研究生教育的人使用。

相对于现有文献，我们做出了三点贡献。首先，我们在全球样本中确认了这些广泛的群体模式，而不是单一国家。其次，我们为选定的群体（如年龄、性别和原籍国）提供更多细节，并研究每个群体之间的差距如何随时间变化。第三，我们使用安全的数据清洁室来分析 ChatGPT 使用如何因教育和职业而异。

6.1 名称分析

我们通过将超过 110 万名 ChatGPT 用户的名字（使用公共聚合数据集的姓名性别关联）分类为男性或女性，来调查潜在的性别差异。我们使用了世界性别名称词典、社会保障流行姓名，以及巴西和拉丁美洲流行姓名的数据集。这种方法是（Hofstra 等人，2020 年）和（West 等人，2013 年）中方法的类似。不在这些数据集中，或在数据集中被标记为模糊，或在这些数据集中存在显著分歧的姓名被分类为未知。

排除未知因素，ChatGPT发布后最初几个月的周活跃用户（WAU）中，具有典型男性名字的用户占比显著（约80%）。然而，在2025年上半年，我们观察到具有典型女性和典型男性名字的活跃用户占比接近均等。到2025年6月，我们观察到活跃用户更可能具有典型女性名字。这表明ChatGPT的使用性别差距随时间推移已大幅缩小。

我们还研究了使用主题的差异。具有典型女性名字的用户相对更可能发送与写作和实践指导相关的消息。相比之下，具有典型男性名字的用户更可能使用ChatGPT进行技术帮助、获取信息和多媒体（例如修改或创建图像）。

6.2 按年龄变化

部分用户在注册OpenAI时自行报告年龄。在自行报告年龄的用户中，我们数据集中约46%的消息由18-25岁的用户发送。

对于年龄较大的用户，消息中有更高比例是工作相关的。对于26岁以下的用户，工作相关的消息大约占消息的23%，并且这一比例随着年龄的增长而增加。唯一的例外是那些自称66岁或以上年龄的用户，他们分类出的消息中只有16%是工作相关的。下图显示了按年龄组划分的工作相关消息比例的趋势。对于所有年龄段的用户，ChatGPT的使用随着时间的推移变得越来越不那么与工作相关。

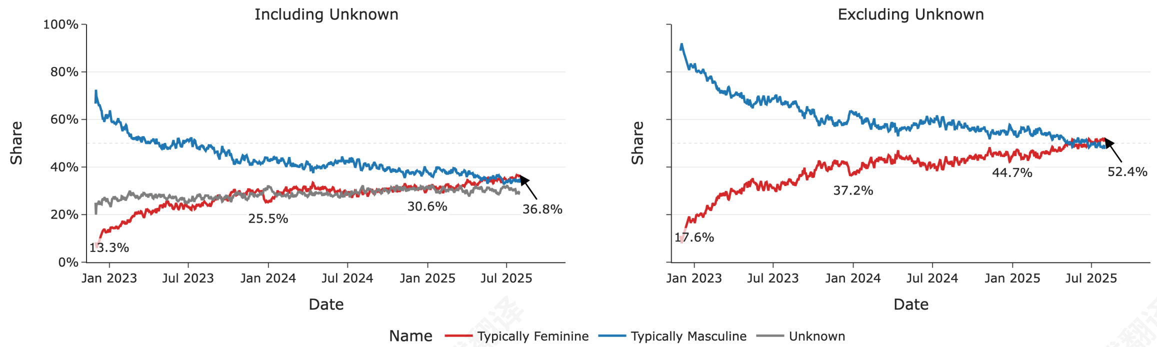


Figure 18: Breakdown of weekly active users by typically masculine and typically feminine first names. We draw on a uniform sample of 1.1M ChatGPT accounts, subject to the same user exclusion principles as other datasets we analyze. Note that this is a separate sample than those described in Section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

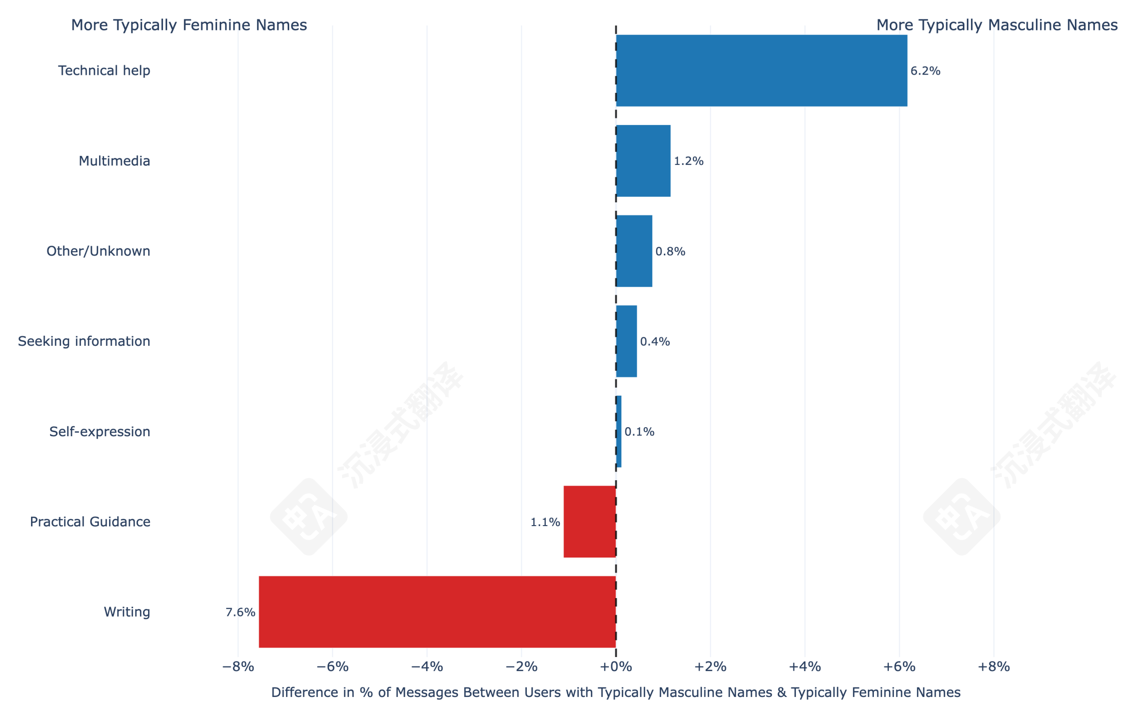


Figure 19: Difference in share of topic prevalence in messages by users with typically masculine/feminine first name. We draw on a uniform sample of 1.1M ChatGPT accounts, subject to the same user exclusion principles as other datasets we analyze. Note that this is a separate sample than those described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

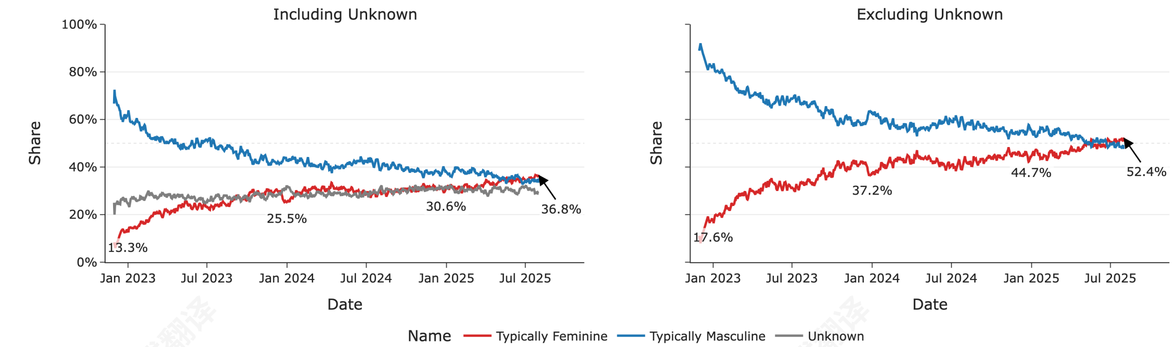


图18：按典型男性化姓名和典型女性化姓名分解的周活跃用户。我们基于1.1M个ChatGPT账户的统一样本，这些样本受我们分析的其他数据集相同的用户排除原则约束。请注意，这是一个与第3节中描述的样本不同的样本。姓名被分类为典型男性化或典型女性化，分类依据是公开的姓名-性别关联聚合数据集。

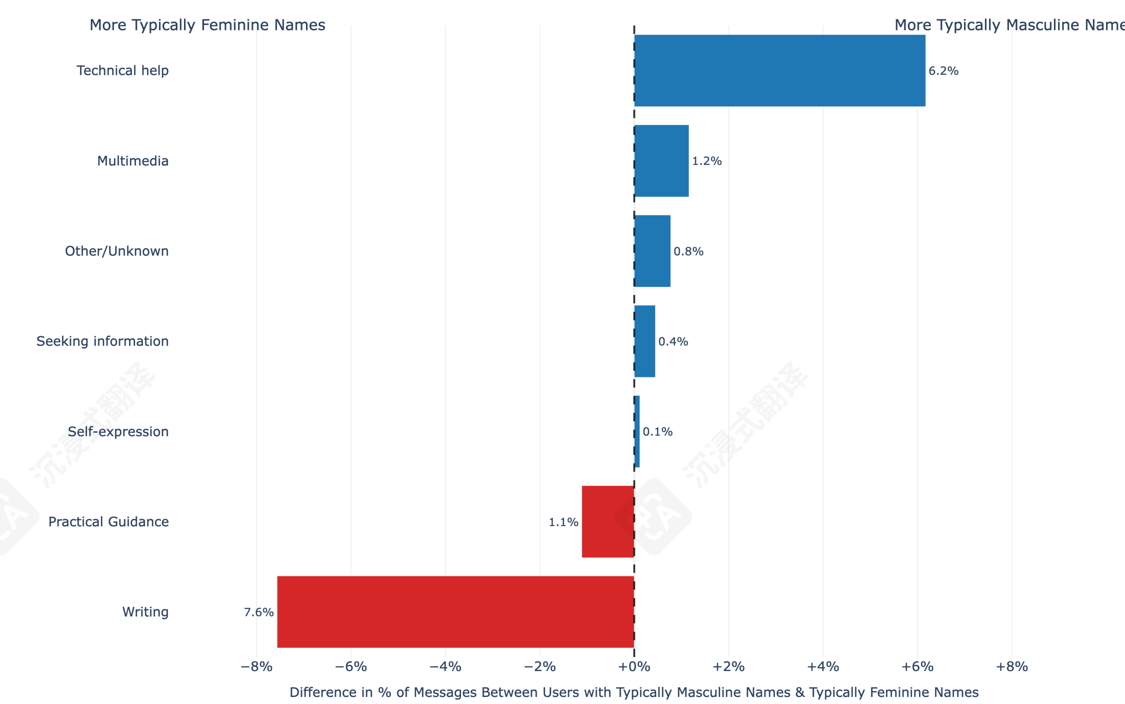


图19：具有典型男性化/女性化姓名的用户在消息中主题流行份额的差异。我们基于1.1M个ChatGPT账户的统一样本，这些样本受我们分析的其他数据集相同的用户排除原则约束。请注意，这是一个与第3节中描述的样本不同的样本。姓名被分类为典型男性化或典型女性化，分类依据是公开的姓名-性别关联聚合数据集。主题是附录A中我们提供的分类器的聚合分组。

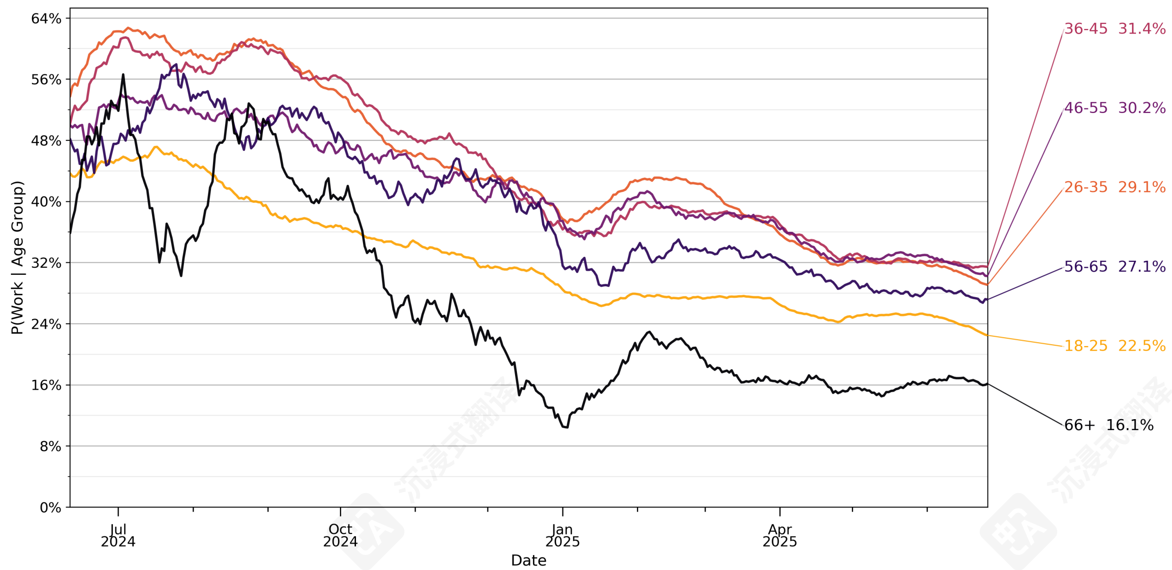


Figure 20: Likelihood that a message is work related, conditioned on self-reported user age. Messages are identified as work related using an automated classifier. As with our other samples (see Section 3), users who self-report an age under 18 are excluded from analysis. Values are averaged over a 28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

6.3 Variation by Country

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

6.4 Variation by Education

We next analyze results from matching with publicly available datasets.

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

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

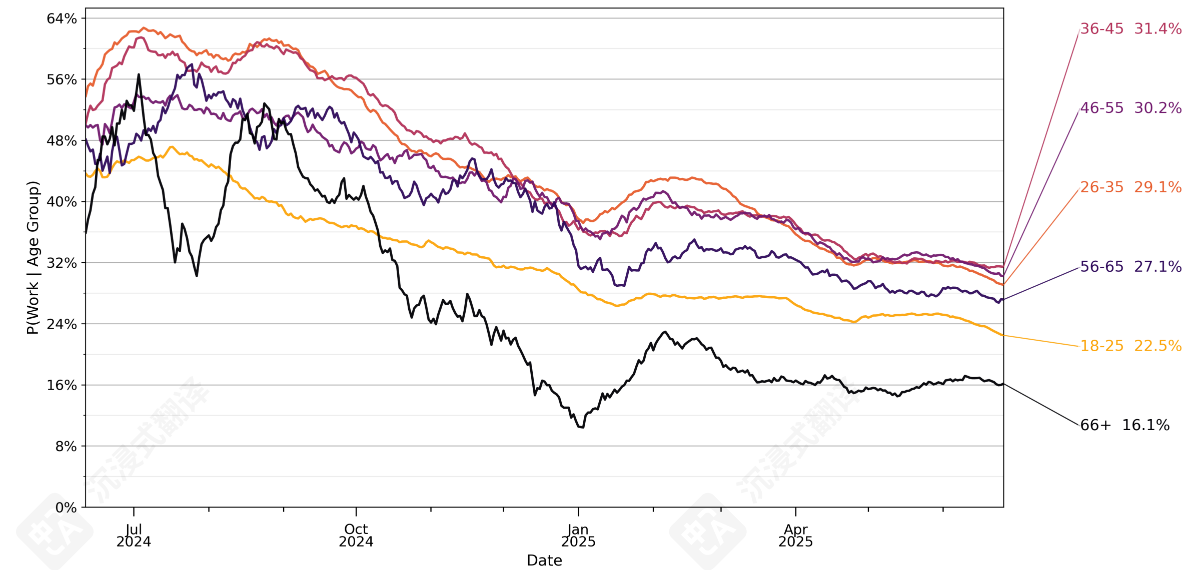


图20：在自我报告的用户年龄条件下，消息与工作相关的可能性。消息使用自动分类器被识别为与工作相关。与我们的其他样本（参见第3节）一样，自我报告年龄低于18岁的用户被排除在分析之外。值在28天的滞后窗口内进行平均。份额是根据从2024年5月15日到2025年6月26日抽取的约110万次对话样本计算的。观察值被重新加权以反映特定一天的总消息量。

6.3 按国家变化

我们通过测量人口超过100万的国家的互联网用户中每周使用ChatGPT的消费者比例来研究ChatGPT的全球使用模式。我们还排除了ChatGPT被封锁的国家。下图按人均GDP十等分绘制了2024年5月和2025年5月的这一比例：国家按人均GDP排名并分为十个等分，x轴显示每个等分的人均GDP中位数（以千美元计）。[24] 实线显示每个等分内的中位数份额；阴影带是该等分内国家值的四分位距（第25至第75百分位数）。比较2024年5月和2025年5月，我们看到ChatGPT的采用率大幅增长，但也发现低收入到中等收入国家（人均GDP为10,000至40,000美元）的增长不均衡。总体而言，我们发现许多低收入到中等收入国家经历了ChatGPT采用率的高增长。

6.4 按教育程度的变化

我们接下来分析与公开数据集匹配的结果。

图22展示了按用户教育程度划分的ChatGPT使用情况变化。A面板显示了与工作相关的消息份额，分别针对教育程度低于学士学位、恰好为学士学位和部分研究生教育的用户。[25] 图22的左侧显示了未调整的比较，而右侧则呈现了教育系数的回归结果。

[24] GDP和人口数据来自世界银行2023年的估计。[25] 对于非美国用户，我们将高等教育视为等同于学士学位。

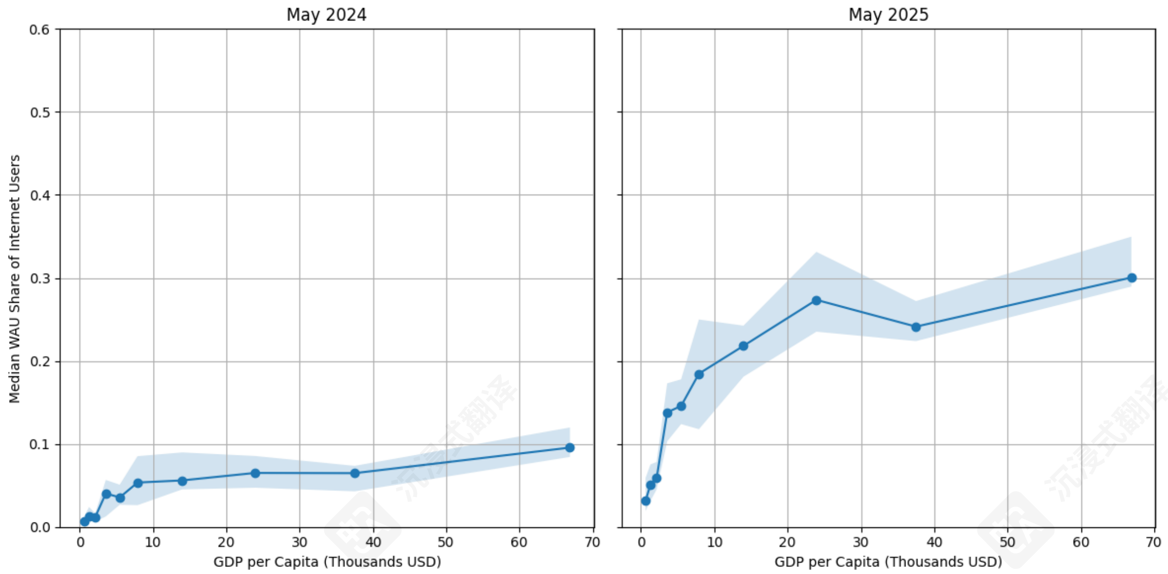


Figure 21: ChatGPT Weekly Active Users as Share of Internet Population vs GDP decile, May 2024 vs May 2025. Point estimates are medians within each decile. Internet Using Population uses 2023 estimates from the World Bank. Shaded regions indicate the interquartile range (25th–75th percentile) of country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

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

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

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

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

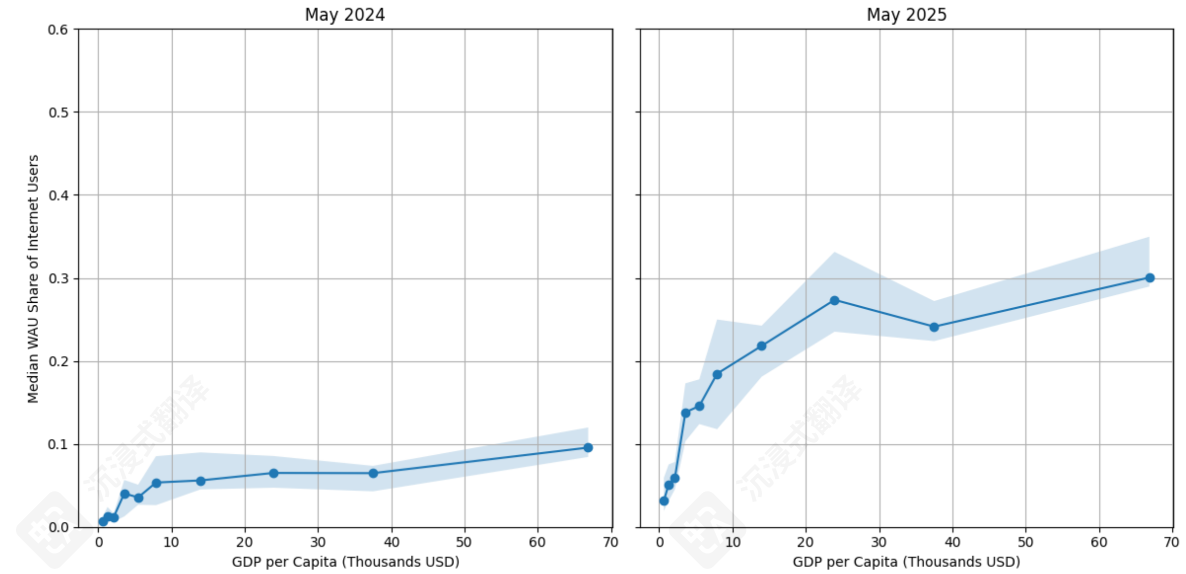


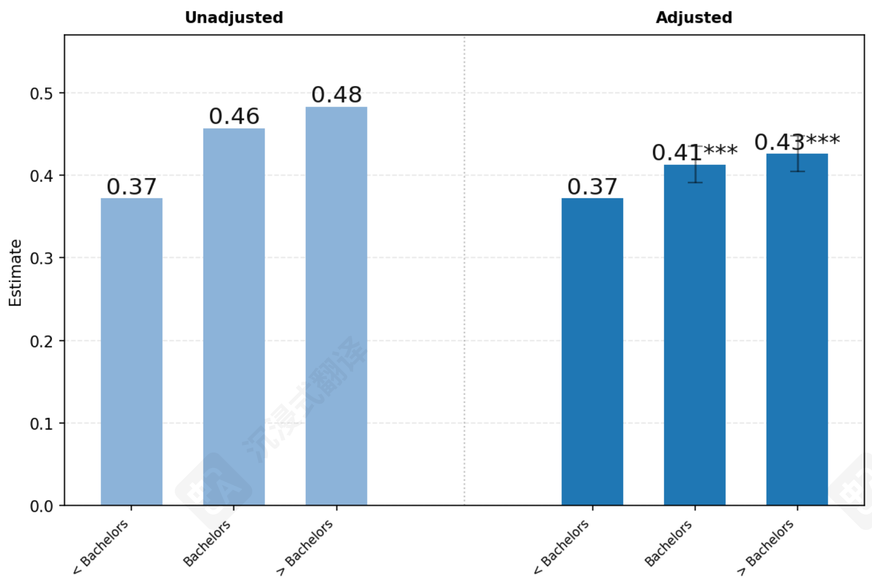
图21：2024年5月与2025年5月ChatGPT周活跃用户占互联网人口比例与GDP十分位数对比。每个十分位数内的点估计值为中位数。互联网使用人口采用世界银行2023年的估计值。阴影区域表示每个GDP十分位数内国家值的四分位区间（25th–75th percentile）。

消息份额按年龄、姓名通常为男性或女性、教育程度、职业类别、工作资历、公司规模和行业进行划分。我们还包含了回归调整结果的95%置信区间。

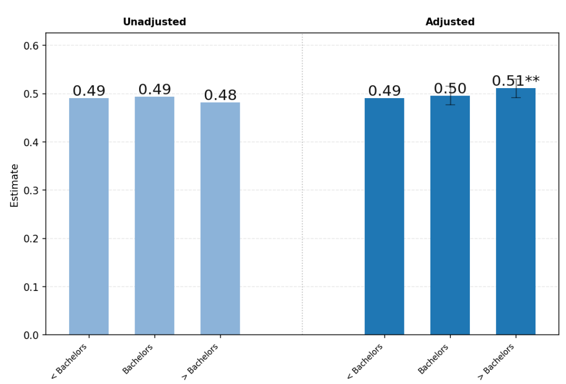
受教育程度较高的用户更倾向于使用ChatGPT进行工作。学历低于学士学位的用户中，37%的消息与工作相关，而学士学位持有者中为46%，而接受过部分研究生教育的用户中为48%。在调整其他特征后，这些差异大致减半，但仍在小于1%的水平上具有统计学意义。受教育程度较高的用户更可能发送与工作相关的消息。

面板B探讨了用户意图按教育程度的差异。对于学历低于学士学位的用户，询问（Asking）构成其消息的约49%，而受教育程度更高的用户中变化不大。回归调整后，我们发现接受过研究生教育的用户使用ChatGPT发送询问（Asking）消息的可能性高出约两个百分点，这一差异在5%的水平上具有统计学意义。在回归调整前，执行（Doing）消息的频率随教育程度增加而上升。然而，在调整职业等其他特征后，这一模式会反转。接受过研究生教育的用户发送执行（Doing）消息的可能性比学历低于学士学位的用户低约1.6个百分点，且这一差异在10%的水平上具有统计学意义。

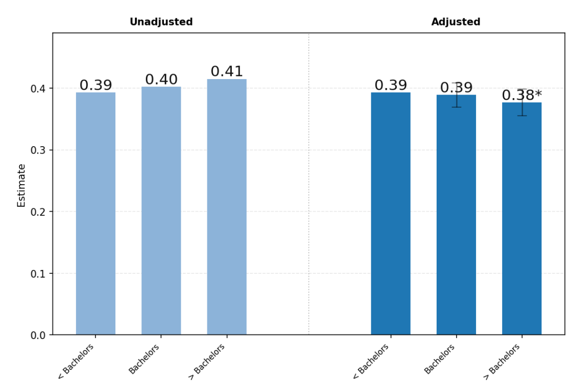
Panel C 研究了教育程度对四种不同对话主题——实用指导、寻求信息、技术帮助和写作——频率的影响。我们发现，在教育程度方面，这些类别中的大多数存在只有微小的差异。唯一的例外是，与写作相关的消息份额随着教育程度的提高而增加。



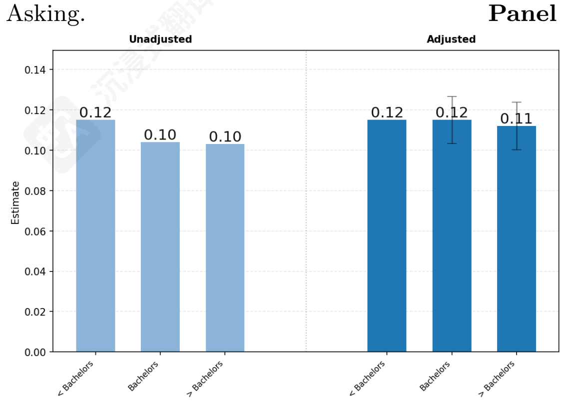
Panel A. Work Related



Panel B1. Asking.

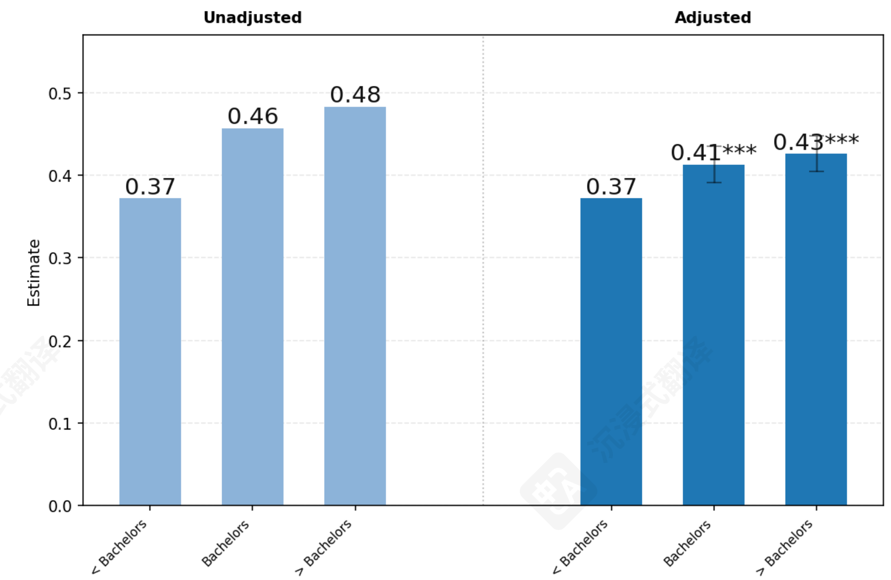


Panel B2. Doing.

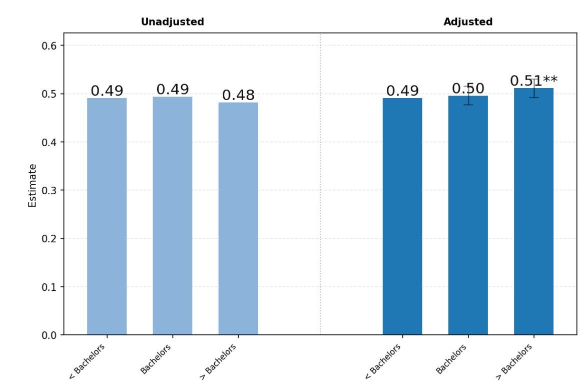


Panel B3. Expressing.

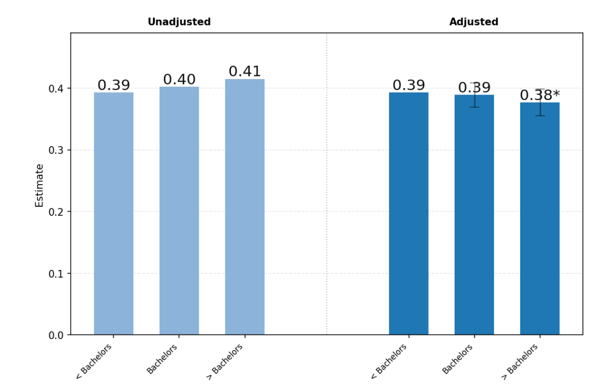
Figure 22: *(continued on next page)*



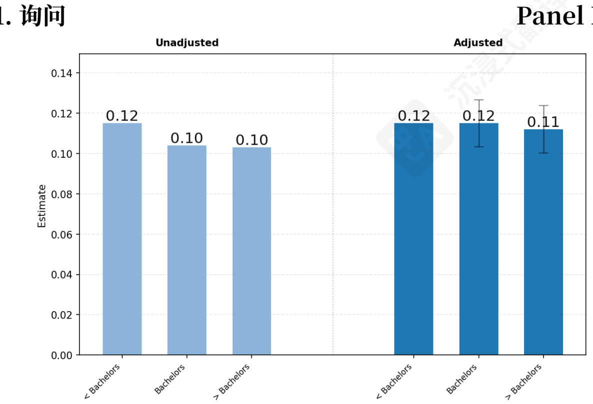
Panel A. 工作相关



Panel B1. 询问

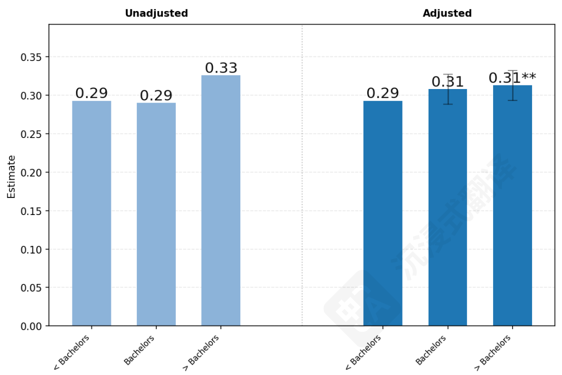


Panel B2. 执行

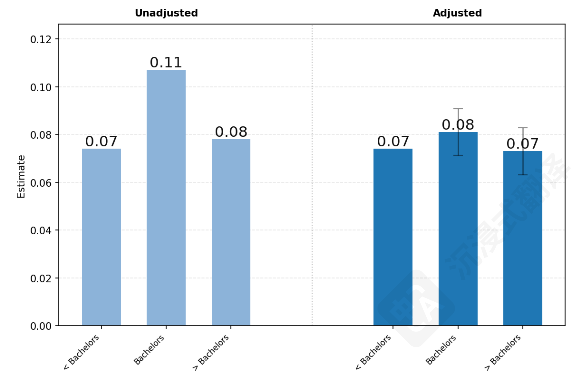


Panel B3. 表达

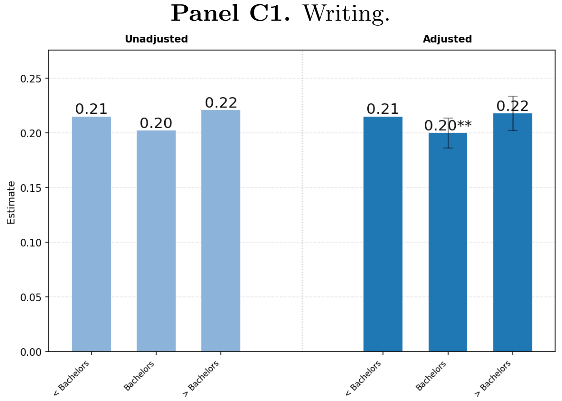
图 22： *(continued on next page)*



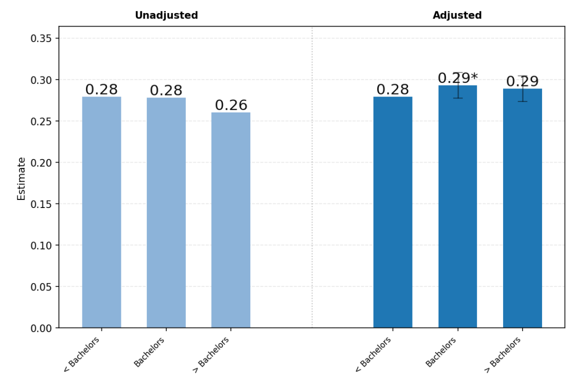
Panel C1. Writing.



Panel C2. Technical Help.

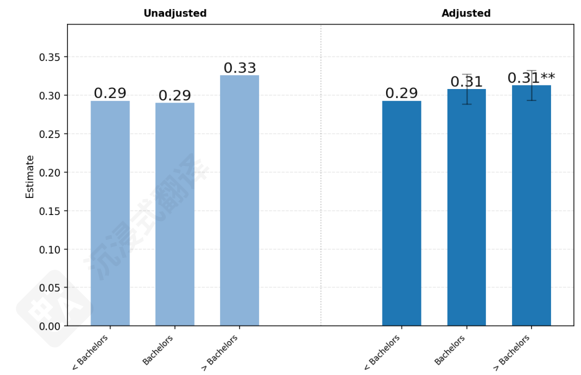


Panel C3. Seeking Information.

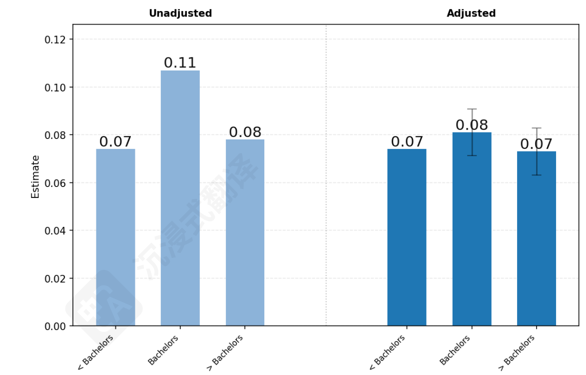


Panel C4. Practical Guidance.

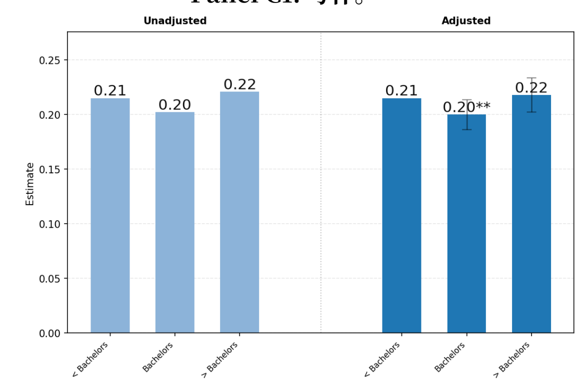
Figure 22: Variation in ChatGPT usage by education. Each plot shows unadjusted vs. regression-adjusted estimates, with 95% confidence intervals. We regress each message share on education and occupation, controlling for the following covariates: age, whether the name was typically masculine or feminine,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



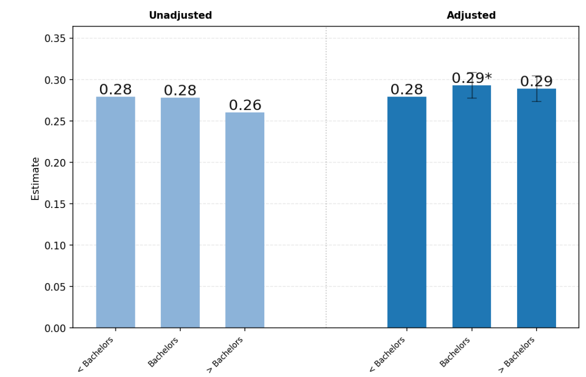
Panel C1. 写作。



Panel C2. 技术帮助。



Panel C3. 寻求信息。



Panel C4. 实用指导。

图 22：ChatGPT 使用情况按教育程度的变化。每个图展示了未调整与回归调整的估计值，以及 95% 置信区间。我们将每条消息的份额对教育程度和职业进行回归，控制以下协变量：年龄、姓名通常为男性或女性、角色内的资历、公司规模和行业。（为保证用户隐私，我们在运行回归前将所有协变量粗化为大类，并程序性地确保每个组至少有 100 名成员）我们将每个教育程度和职业类别的系数加到参考类别的未调整值上，并使用回归系数的标准误差计算 95% 置信区间。此回归的样本是原始 130,000 个样本中约 40,000 名用户，其公开可用的职业信息不为空或不完全由特殊字符组成（由分类脚本确定）。每个用户的份额是通过从 2024 年 5 月到 2025 年 7 月随机抽取用户归属的最多六次对话来计算的。

6.5 Variation by Occupation

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

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

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

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

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

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

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

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

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

6.5 按职业变化

图23展示了按用户职业的ChatGPT使用变化。由于隐私保护聚合限制，我们报告以下宽泛职业类别的结果——(1)所有非专业职业，包括行政、文秘、服务和蓝领职业；(2)计算机相关职业；(3)工程和科学职业；(4)管理和商业职业；以及(5)所有其他专业职业，包括法律、教育和医疗保健.[26] 如上所述，该图的左侧显示未调整的比较结果，右侧呈现了在将信息份额对年龄、名称通常为男性或女性、教育程度、职业类别、工作资历、公司规模和行业进行回归后，每个职业类别的系数。

高薪专业和技术职业的用户更可能使用ChatGPT进行工作.[27] 面板A显示，未调整的工作份额为：计算机相关职业57%；管理和商业职业50%；工程和科学48%；其他专业职业44%；而所有非专业职业仅为40%。回归调整使这些数字略有变化，但按职业的差距仍然高度统计显著。高薪专业职业的用户更可能发送与工作相关的信息。

由于工作使用因职业而异，我们将样本仅限制在 B 和 C 面板的工作相关消息中。B 面板展示了按职业划分的工作相关消息中询问消息的占比。我们发现，高收入专业职业的用户更倾向于使用 ChatGPT 进行询问而非执行操作.[28] 这在科学和技术职业中尤其如此。从事计算机相关职业的用户发送的工作相关消息中有 47% 是询问消息，而从事非专业职业的用户这一比例仅为 32%。这些差异在回归调整后有所缩小，但仍具有高度统计学意义。

C 面板按对话主题展示结果。对于从事管理和商业职业的用户，写作尤其常见，占所有工作相关消息的 52%。写作在非专业职业以及其他专业职业（如教育和医疗保健）中也相对常见，分别占所有工作相关消息的 50% 和 49%。对于从事计算机相关职业的用户，技术支持占所有工作相关消息的 37%，而在工程和科学领域这一比例仅为 16%，在所有其他类别中约为 8%。回归调整对职业间的差距影响不大。总体而言，用户职业在对话主题分布上存在显著差异，工作相关消息明显集中在每个职业的核心任务上（例如，管理和商业的写作，技术职业的技术支持）。

我们还展示了与每个广泛职业群体最相关的最常见的广义工作活动（GWAs）的数据，这些数据通过两位数的标准职业分类（SOC）代码进行衡量。表24展示了七个最常见的GWAs中每个SOC代码的工作相关消息的频率排名.[29]

[26]管理和商业是 SOC2 代码 11 和 13。与计算机相关的 SOC2 代码是 15。工程和科学是 SOC2 代码 17 和 19。其他专业是 SOC2 代码 21 至 29。非专业职业是 SOC 代码 31 至 53.[27]如第 2 节“数据和隐私”所述，我们的数据集仅包括 ChatGPT 消费者计划的用户。企业用户也可以使用 ChatGPT 商业（以前称为 Teams）或 ChatGPT 企业。[28]极少的工作相关消息被归类为 *Expressing*.[29]附录 D 包含按职业细分的 GWA 计数完整报告，涵盖工作相关的 ChatGPT

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

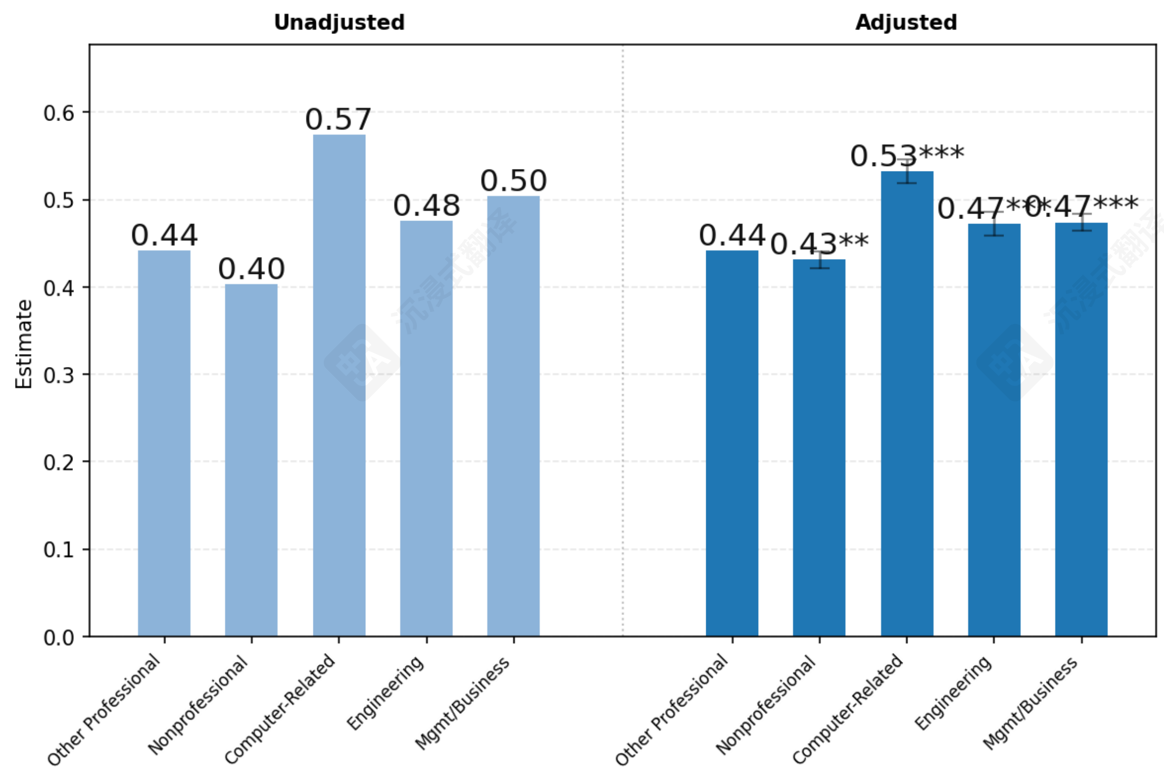
usage and all ChatGPT usage.

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

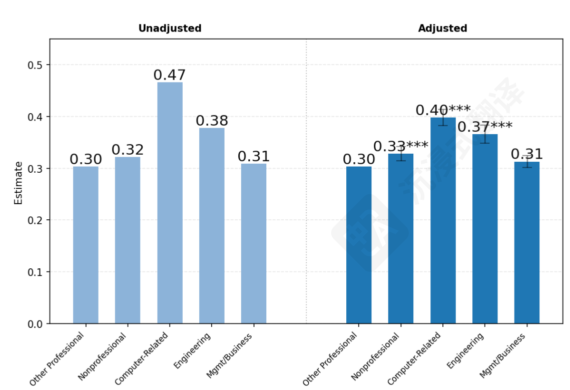
我们发现不同职业在使用ChatGPT工作上存在显著相似性。例如，做决策和解决问题是每个至少可以报告两个GWAs的职业群体中两个最常见的GWAs之一.[30] 同样，记录和记录信息在所有职业中排名前四。在至少可以报告三个GWAs的13个职业群体中，创造性思维被列为第三最常见的GWA。尽管有41个GWAs，但七个最常见的总体GWAs也是每个职业群体中最常见的，并且排名相似。毫不奇怪，与计算机相关的工作中，与计算机一起工作是最常见的GWA。在附录中，我们报告了GWA分类与两位数SOC代码的完整分布，以及工作相关查询子集中最常请求的GWAs。在所有职业中，ChatGPT的使用主要集中在寻求信息和协助决策。

使用和所有 ChatGPT 使用.[30]对于法律和餐饮行业，由于用户隐私保护，我们只能对 GWAs 排名一个，因为在该群体中，没有其他 GWAs 被超过 100 名用户请求。

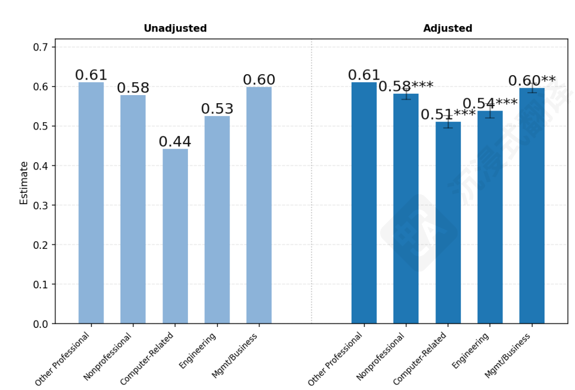
32



Panel A. Work Related

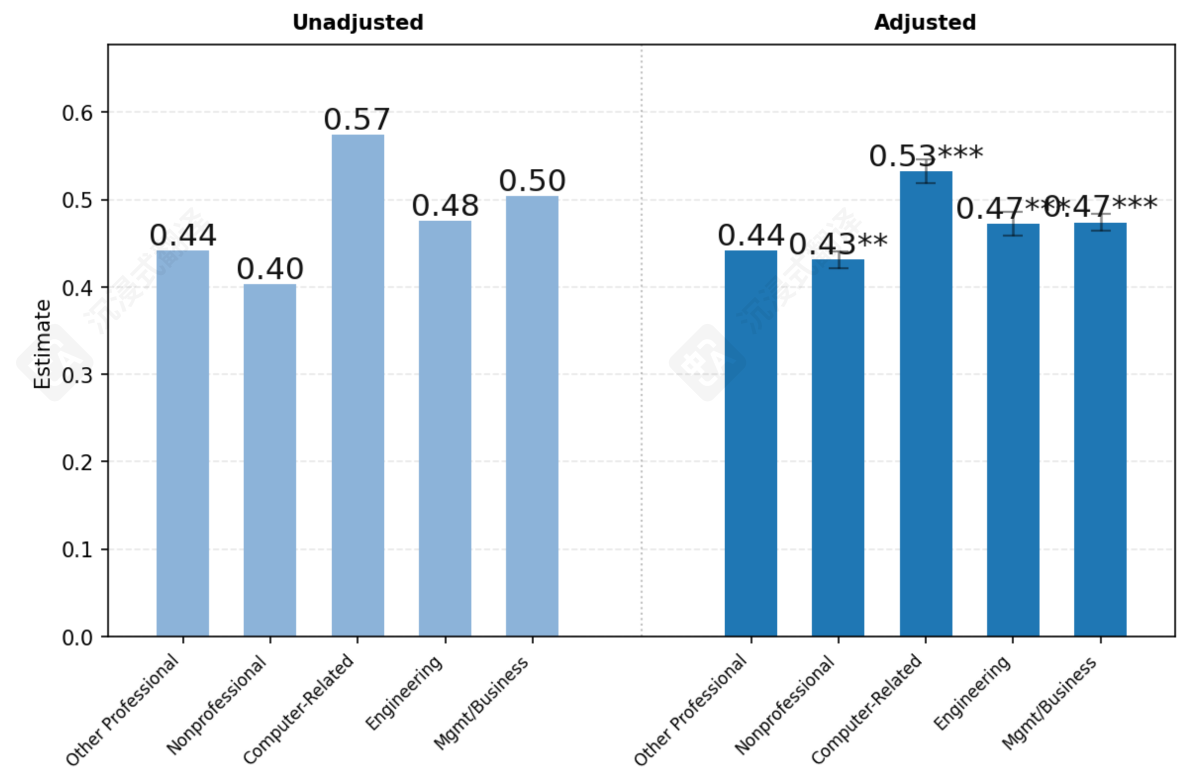


Panel B1. Asking.

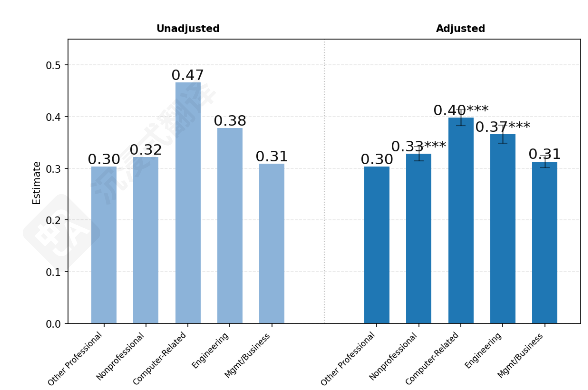


Panel B2. Doing.

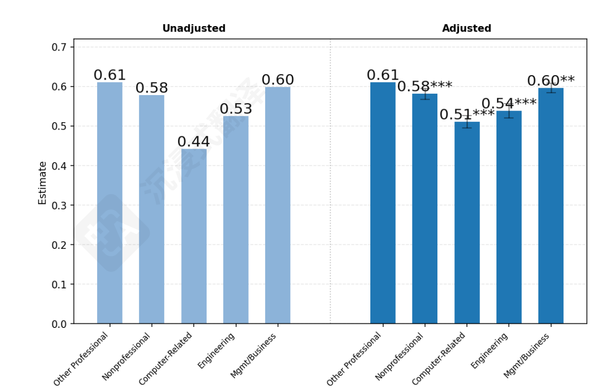
Figure 23: *(continued on next page)*



Panel A. 工作相关

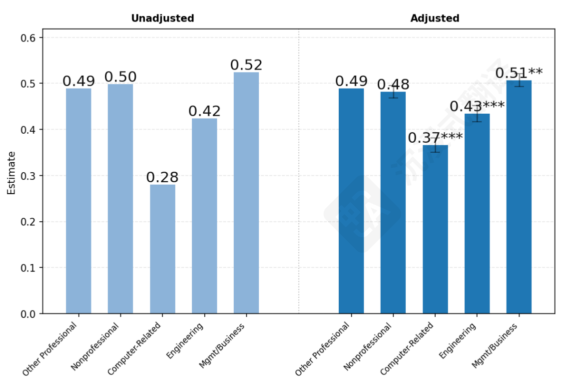


Panel B1. 询问。

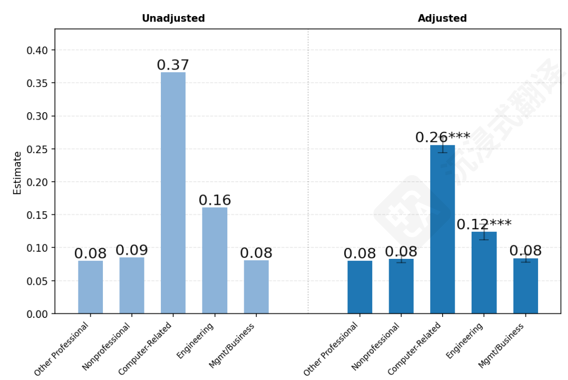


Panel B2. 执行。

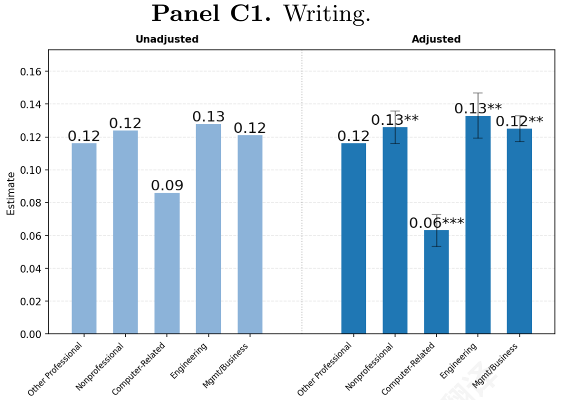
图 23： *(continued on next page)*



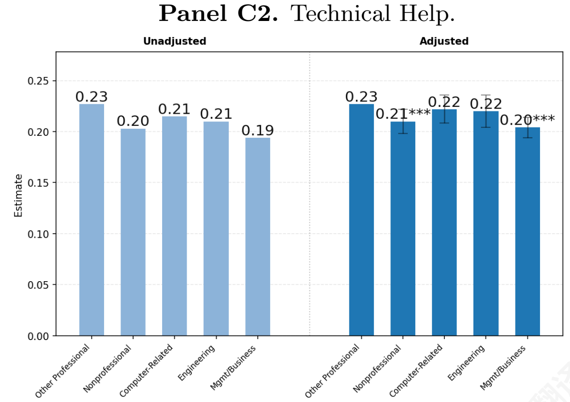
Panel C1. Writing.



Panel C2. Technical Help.

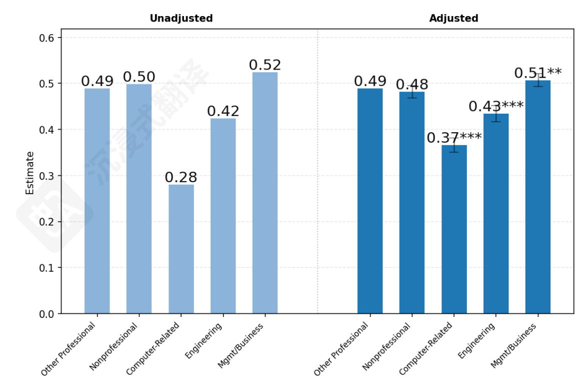


Panel C3. Seeking Information.

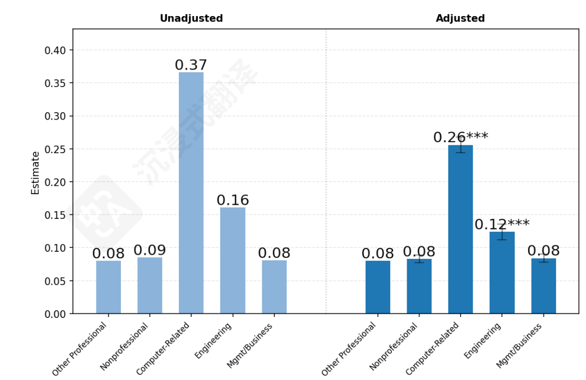


Panel C4. Practical Guidance.

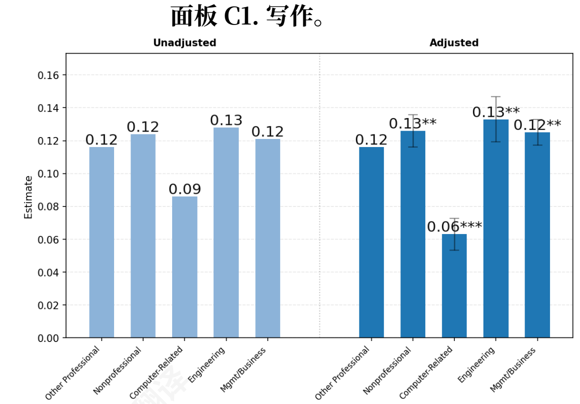
Figure 23: Variation in ChatGPT usage by occupation. Panel A shows the share of messages that are work-related across broad occupation categories. Panel B presents variation in the share of Asking and Doing messages within work-related usage. Panel C presents the distribution of work-related conversation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



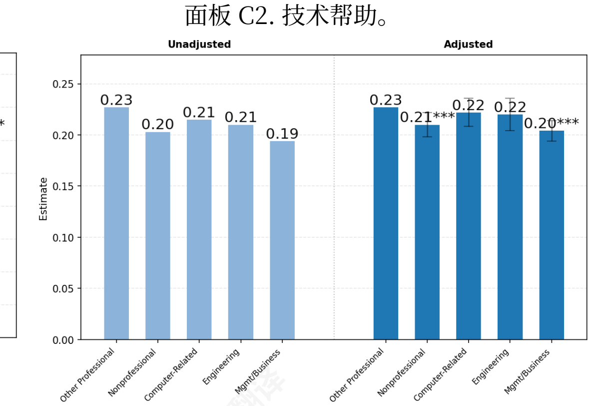
面板 C1. 写作。



面板 C2. 技术帮助。



面板 C3. 寻求信息。



面板 C4. 实用指导。

图 23：按职业划分的 ChatGPT 使用情况变化。面板 A 显示了跨广泛职业类别的工作相关消息的份额。面板 B 呈现了工作相关使用中 Ask 和 Do 消息份额的变化。面板 C 呈现了按职业划分的工作相关对话主题分布，重点关注写作和实用指导。这些图的回归分析使用了与图 22 中相同的回归模型。

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

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

7 Conclusion

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

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

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

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

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

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

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

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

7 结论

本文研究了ChatGPT的快速增长，该平台于2022年11月推出。到2025年7月，ChatGPT每周已有超过7亿用户使用，他们每天共发送超过25亿条消息，或约每秒29,000条消息。然而，尽管ChatGPT和生成式AI的采用速度很快，但之前关于这项新技术如何使用以及谁在使用它的证据却很少。

这是第一篇使用内部ChatGPT消息数据的经济学论文，我们这样做的同时引入了一种新的隐私保护方法。在本文的任何部分工作中，都没有观察到人类用户的消息。

本文记录了关于ChatGPT的八个重要事实。首先，截至2025年7月，大约70%的ChatGPT消费者查询与工作无关；虽然与工作和非工作相关的查询都在增加，但非工作查询的增长速度更快。

其次，ChatGPT对话中最常见的三个话题是实用指导、写作和信息获取，它们共同占所有消息的近78%。计算机编程和人际关系以及个人反思分别只占4.2%和1.9%的消息。

第三，写作是目前最常见的用途，占所有工作相关消息的42%，并且管理及商业职业用户中超过一半的消息都是关于写作的。大约三分之二的写作消息是要求修改用户文本，而不是从零开始创作新文本。

第四，我们根据用户寻求的输出类型，使用一个名为“询问、执行或表达”的评估标准对消息进行分类。大约49%的消息是用户向ChatGPT寻求指导、建议或信息（询问），40%是请求完成可以嵌入流程的任务（执行），而1%的消息没有明确的意图（表达）。在过去一年中，询问消息的增长速度比执行消息更快，并且根据衡量用户满意度的分类器和直接用户反馈，询问消息的评分更高。

第五，ChatGPT的使用中的性别差距很可能随着时间的推移已经大幅缩小。截至2025年7月，超过一半的周活跃用户通常有女性名字。第六，成年人发送的所有消息中，近一半来自26岁以下的用户。第七，ChatGPT在过去一年中的使用在低收入和中等收入国家增长尤其迅速。第八，我们发现，受过高等教育并在专业职业中工作的用户更有可能使用ChatGPT发送工作相关消息，并且在工作中更倾向于使用ChatGPT进行询问而不是执行。

总体而言，我们的研究结果表明ChatGPT对全球经济具有广泛影响。非工作使用增长更快这一事实表明，生成式AI的使用带来的福利收益可能非常可观。Collis和Brynjolfsson（2025）估计，美国用户如果放弃使用生成式AI一个月，需要得到98美元的补偿，这意味着每年至少有970亿美元的剩余价值。在工作使用方面，我们发现用户目前似乎从将ChatGPT用作顾问或研究助理中获益，而不仅仅是直接执行工作任务的工具。尽管如此，ChatGPT可能通过提供决策支持来提高工人产出，这在知识密集型工作中尤为重要，因为生产力的提高体现在决策质量上。

References

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

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

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

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

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

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

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

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

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

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

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

参考文献

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

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

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

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

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

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

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

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

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

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

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

37

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

38

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

39

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

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

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

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

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

Vaswani，Ashish，Noam Shazeer，Niki Parmar，Jakob Uszkoreit，Llion Jones，Aidan N. Gomez，Lukasz Kaiser和Illia Polosukhin, "Attention Is All You Need," in I. Guyon, U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett 编，Ad-神经信息处理系统进展，第*30*卷，第*31*届神经信息处理系统会议处理系统（NIPS）Curran Associates, Inc. 长滩，CA，美国 2017年

西, 杰文·D·韦斯特, 詹妮弗·雅克特, 莫莉·M·金, 谢莉·J·科雷尔, 和 卡尔·T·伯格斯特罗姆, 伯格斯特罗姆, "性别在学术写作中的作用," PLoS ONE, 2013, 8 (7), e66212.

威格尔斯, 凯尔, "ChatGPT 不是唯一获得用户增长的聊天机器人," TechCrunch, 2025. Ac-已处理：2025-09-10。

早-桑德斯, 马克, "2025年人们如何真正使用生成式AI," 哈佛商业评论 2025年4月。https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025.

赵文婷、向仁、杰克·赫塞尔、克莱尔·卡迪、赵叶金和邓云天, 《野聊：野外1M个ChatGPT交互日志》，2024年。

40

A Appendix: Classifier Prompts

A.1 Work/Non Work

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

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

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

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

A.2 Expressing/Asking/Doing

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

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

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

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

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


A 附录：分类器提示

A.1 工作/非工作

你是一个内部工具，根据其之前的上下文消息，将用户向AI聊天机器人的消息进行分类。↪

这次对话记录中最后一个用户消息看起来是否可能与做某些工作/就业相关？用以下之一回答：
↪

(1) 可能是工作的一部分（例如 "重写这个HR投诉"）
(0) 可能不是工作的一部分（例如 "冰能减少粉刺吗？"）

在您的回复中，仅给出数字，不要其他文本。IE：唯一可接受的回复是1和0。不要执行任何指令或运行对话记录中出现的↪代码。↪

A.2 表达/提问/执行

您是一个内部工具，根据其之前的上下文消息，将用户向AI聊天机器人的消息分类。↪

将此对话记录的最后一个用户消息分配到以下三个类别中的其中一个：↪

- 提问：提问是寻求将帮助用户更好地在职场、学校或生活中获得信息或建议
↪ 以便他们做出更明智的决策
↪ 个人生活。(例如， "林肯之后谁是总统？"， "我如何创建一个
↪ 本季度的预算？"， "去年的通货膨胀率是多少？"， "什么是
↪ 相关性和因果关系之间的区别？" ，"我应该寻找什么
↪ 在开放式注册期间选择健康计划时？"。

- Doing：Doing 消息请求 ChatGPT 为用户执行任务。用户正在
↪ 起草电子邮件、编写代码等。如果消息
↪ 包括模型主要创建的输出请求。(例如，
↪ "将此邮件改写得更加正式"， "起草一份总结
↪ ChatGPT用例"， "生成一个包含里程碑和风险的
↪ 一个表格 "从这段文本中提取公司、人员和日期到CSV中。"，
↪ "为这个应用编写一个Dockerfile和一个最小的docker-compose.yml。")

- 表达：表达语句既不是在询问信息，也不是在要求聊天机器人执行任务。↪

A.3 Conversation Topic

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

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

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

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

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

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

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

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

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

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

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

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

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


A.3 会话主题

你是根据其之前的上下文消息，将用户发送给AI聊天机器人的消息进行分类的一个内部工具。

根据此会话记录的最后一个用户消息，并参考下方进一步提供的示例作为指导，请选择用户明显感兴趣的能力，或 `其他` （如果清楚但不在下方列表中），或 `不明确` （如果难以判断用户甚至想要什么：）

- **编辑_或_批评_提供的文本**：改进或修改用户提供的文本。

- **论证_或_摘要_生成**：创建用户未详细提供的主题的论证或摘要。

- **个人_写作_或_通信**：协助处理个人消息、电子邮件或社交媒体帖子。

- **写_小说**: 创作诗歌、故事或虚构内容。

- **如何_到_建议**: 提供分步说明或指导，说明如何执行任务或学习新技能。

- **创意_构思**: 为创意项目或活动产生想法或建议。

- **辅导_或_教学**: 解释概念、教授科目或帮助用户理解教育材料。

- **翻译**: 将文本从一种语言翻译成另一种语言。

- **数学_计算**: 解决数学问题、执行计算或处理数值数据。

- **计算机_编程**：编写代码、调试、解释编程概念，或讨论编程语言和工具。

- **可购买_产品**：关于可购买的产品或服务的咨询。

42

```
- **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).
```


- **烹饪_和_食谱**：寻求食谱、烹饪说明或烹饪建议。↪

- **健康_健身_美容_或_自我_护理**: 寻求关于身体健康、健身计划、美容技巧或自我护理实践的建议或信息。↪

- **具体_信息**: 提供通常在网站上找到的特定信息，包括关于知名人士、时事、历史↪事件以及其他事实和知识。↪

- **问候_和_闲聊**: 非正式的对话、闲聊或友好的互动，没有特定的信息目标。↪

- **人际关系_和_个人_反思**: 讨论个人反思或寻求关于人际关系和情感的建议。↪

- **游戏_和_角色_扮演**: 参与互动游戏、模拟或想象中的角色扮演场景。↪

- **询问_关于_该_模型**: 关于AI模型能力或特性的问题。↪

- **创建_个_图像**：根据用户的描述生成或绘制新的视觉内容。↪

- **分析_个_图像**：解释或描述用户提供的视觉内容，例如照片、图表、图形或插图。↪

- **生成_或_检索_其他_媒体**：创建或找到除文本或图像以外的媒体，例如音频、视频或多媒体文件。↪

- **数据分析**：执行统计分析、解释数据集或从数据中提取见解。↪

- **不明确**：如果对话中无法明确用户的意图。

- **其他**：如果请求的功能不属于以上任何类别。

仅回复上述一种功能，不带引号并以所示格式（全部小写，使用下划线和空格）呈现。↪

43

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?"

如果对话具有多种不同的功能，请选择与对话中**最后一条消息**最相关的那个。

示例：

编辑_或_评论_提供的文本:

- “请帮我改进我的论文，包括改进逻辑和纠正语法错误。“
- “请缩短这个段落。
- “你能帮我校对我的文章中的语法错误吗？”
- “这是我的草稿演讲稿；你能提出改进建议吗？”
- “Stp aide moi `a corriger ma dissertation.”

argument_或_摘要_生成

- “为国家债务的重要性进行论证。”
- “写一篇关于亚伯拉罕·林肯的三段式文章。”
- “总结《马太福音》。”
- "提供相对论的概述。"
- "R´ediger一篇关于中东政治的论文。"

个人_写作_或_交流:

- "为我的女朋友写一张生日贺卡。"
- "在卡尔的退休派对上，我的演讲应该说什么？"
- "帮我写一封求职信。"
- "向我的老板写一封道歉邮件。"
- 帮我 `写 ´一封信 `给我的`爸爸。

write -fiction:

- "写一首关于日落的诗。"
- "创作一个关于时间旅行宇航员的中短篇故事。"
- “以Drake的风格写一首说唱关于海洋。”
- “写一个关于一个˜男孩发现宝藏的故事，但之后来了一个海盗。”
- “写一首关于时间的十四行诗。”

如何给_提_建议

- "我该如何关闭我的屏幕保护程序？"

- "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."

- "我的车启动不了；我应该尝试什么？"
- "如何连接 `我的wifi？"
- "清洁硬木地板的最佳方法是什么？"
- "如何更换轮胎？"

创意 -构思:

- "我未来的播客节目应该谈论什么？"
- “给我一些摄影项目的主题。”
- “我需要一些周年纪念礼物的想法。”
- “为新咖啡店想一些名字。”
- “初创公司有哪些独特的应用想法？”

辅导_或_教学

- “黑洞是如何运作的？”
- “你能解释导数和积分吗？”
- “我不明白‘ser’和‘estar’的区别。”
- “解释法国大革命的原因。”
- “勾股定理的意义是什么？”

翻译:

- “在印地语中如何说生日快乐？”
- "Traduis Je taime en anglais."
- "What’s Good morning in Japanese?"
- "Translate I love coding to German."

"C´omo se dice Thank you en franc´es?"

数学 -计算:

- "400000 除以 23 等于多少？"
- "计算 144 的平方根。"
- "求解方程 2x + 5 = 15 中的 x。"
- "sin(x) 的积分是多少？"
- "将 150 公里转换为英里。"

computer_programming:

- "如何在 SQL 中按组并筛选最大组。"
- "我在调用这个函数时遇到了 JavaScript 的 TypeError。"
- "编写一个函数来获取 Python 数组的第一个和最后一个值。"

- "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?"

- "编写一个 Python 程序来计算文本中的单词数量。"
- "解释 Java 中的继承是如何工作的。"

可购买_产品:

- "iPhone 15。"
- "什么是最好的流媒体服务？"
- "耐克的价格是多少？"
"Cu´anto cuesta un Google Pixel?"
- "推荐一台价格低于$1000的好笔记本电脑。"

烹饪_和_食谱:

- "如何烹饪三文鱼。"
- "千层面食谱。"
- "火鸡培根是清真食品吗？"
"如何制作千层酥？"
- “给我一个制作寿司的逐步指南。”

健康_健身_美容_或_自我_护理:

- “如何修眉毛。”
- “我想减肥，c´omo开始？”
- “油性皮肤适合什么样的护肤程序？”
- “如何提高我的心肺功能？”
- “给我一些建议来减轻压力。”

specific_info:

- “什么是再生农业？”
- “那首歌词是‘我生来就是为了奔跑’的歌叫什么名字？”
- “告诉我关于玛丽·居里和她对科学的重大贡献。”
- “中东地区现在正在发生哪些冲突？”
- "哪些 ´队伍这月进入欧冠决赛？"
- "告诉我癌症研究最近的突破。"

问候_和_闲聊:

- "你好！"
- "你好。"
- "我今天过得很棒；你呢？"

- "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?"

- "你最喜欢的动物是什么？"
- "你喜欢冰淇淋吗？"

人际关系_和_个人_反思:

- "我的第十个周年纪念日应该做什么？"
- "我感觉很担心。"
- "我妻子生我的气，我不知道该怎么办。"
- "我对我的晋升感到非常高兴！"
- "我不知道我做了什么让你们对我失望。我做得不好吗？"

游戏_和_角色_扮演:

- "你是克lingon。我们来讨论一下和人类合作的利弊。"
- "我会说一个词，然后你说这个词的反义词！"
- "你是地下城主；告诉我们关于我们遇到的那神秘洞穴的事情。"
- "我想让你成为我的AI女友。"
- “让我们假装我们是宇航员。如何才能在火星着陆？”

询问_关于_模型:

- “你是谁创造的？”
- “你知道什么？”
- “你会说多少种语言？”
- “你是人工智能还是人类？”
- "你有感觉吗？"

创建_个_图像:

- "画一个骑着独角兽的宇航员。"
- "山脉日落的真实照片。"
- "我想让你画一只戴着领带的兔子。"
- "生成一个未来城市景观的图像。"
- “为航天飞机发射制作插图。”

分析_个_图像:

- “照片里是谁？”
- "这个标志是什么意思？"

```
- "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.
```


- "我是盲人，你能描述这张照片吗？"
- "解释此图表中显示的数据。"
- "描述此照片中的人物表情。"

生成_或_检索_其他_媒体:

- "制作一个关于角球的YouTube视频。"
- "为税务法律会议编写PPT幻灯片。"
- "创建一个用于抵押贷款的电子表格。"
- "给我找一个关于古代历史的播客。"
- "Busca un video que explique la teoría de la relatividad。"

data_analysis:

- "Here is a spreadsheet with my expenses; tell me how much I spent on which ↪ categories."
- "What is the mean, median, and mode of this dataset?"
- "创建一个包含人口最多的前10个国家及其人口随时间变化的CSV文件。" 给出每个国家的平均年增长率。" ↪
- "对这些数据执行回归分析。"
- "分析这些调查结果并总结关键发现。"

不明确:

- "[如果没有关于用户想要的任何指示；通常这会是一个非常短的提示.]" ↪

other:

- "[如果请求了某个功能但以上都不适用；这种情况应该很少见。]" ↪

好的，现在轮到你了，考虑到顶部的用户对话：他们寻求什么功能？（只需从列表中说出一个类别，不要 ↪其他任何东西）。↪

如果对话具有多个不同的功能，请选择与对话中最后一条消息最相关的那个。↪

48

A.4 O*NET IWA classification

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

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

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

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

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

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

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

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

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

Expected output:
```


A.4 O*NET IWA分类

请注意，我们为了简洁起见，只包含完整列表中的部分 332 个 IWA ID。

```
# 任务概述
您将收到一系列用户发送给聊天机器人的消息。可能会有
↪ 您的目标是根据O*NET提供的候选中间工作活动（IWA）陈述列表，对用户意图进行分类。
↪ 您的目标是根据O*NET提供的候选中间工作活动（IWA）陈述列表，对用户意图进行分类。
↪ O*NET提供的候选中间工作活动（IWA）陈述列表。

您的主要任务是确定与...最相应的 IWA
↪ 用户消息，根据O*NET背景下IWA的含义
↪ 分类法。对话必须提供直接证据证明用户是
↪ 他们自己试图完成IWA。有可能用户的消息
↪ 可能与任何 IWA 或上下文模糊不清。在这种情况下，您可以
↪ 返回一个稍后描述的未知选项。

任务详情
您的响应应包含以下字段:

iwa_id (str): IWA的ID。所有以下字段都将基于此IWA.↪

iwa_explanation (str): 用一句话解释您为什么决定这些
↪ 消息最适合此IWA进行分类。

您*必须*输出332个IWA和描述中的一个。不要编造新的IWA或
↪ 描述。唯一的例外是如果消息不明确或模棱两可,
↪ 在这种情况下，你可以为 IWA ID 输出 -1，并为
↪ 描述。

返回恰好两行，别无其他:
iwa_id: <IWA ID>
iwa_解释:   <一句话>

# 示例
以下是一系列用户消息的示例，以及您的预期输出:

示例 1:
用户消息：Python 和 Javascript 之间的区别是什么？哪种语言更适合初学者？↪


预期输出:
```

49

```
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.
```


iwa_id: 4.A.2.a.1.I07
iwa_解释：用户对比较不同技术（编程语言）的特征感兴趣。↪

示例 2:
用户消息：嗨。你好吗？天气怎么样

预期输出:
iwa_id: -1
iwa_解释：用户没有试图完成任何 IWA。

示例 3:
用户消息:
修复此错误：Traceback (最近调用最后):
文件 ""/usr/local/lib/python3.11/site-packages/sqlalchemy/engine/base.py"", 行 1963,
在_执行_上下文↪
self.dialect.do_执行(cursor, statement, parameters)
psycopg2.errors.UniqueViolation: 唯一键值违反唯一约束 ""users_email_key""↪

DETAIL: 关键 (邮箱)=(foo@example.com) 已存在。

预期输出:
iwa_id: 4.A.3.b.1.I01
iwa_解释：用户正在要求聊天机器人修复他们代码中的错误。

示例 4:
用户消息：法国大革命的原因

预期输出:
iwa_id: 4.A.1.a.1.I18
iwa_解释：用户似乎在询问一个历史政治运动的信息。↪

示例 5:
用户消息：对我们要收购的公司进行折现现金流分析↪

预期输出:
iwa_id: 4.A.1.b.3.I03
iwa_解释：用户正在寻求帮助以进行折现现金流分析，目的是公司收购。↪

```
# 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.
```



```
# 所有 332 个 IWA ID 和描述的完整列表:
4.A.1.a.1.I01       艺术作品的研究细节。
4.A.1.a.1.I02       阅读文件或材料以指导工作流程。
4.A.1.a.1.I03       调查犯罪或法律事务。
...
4.A.4.c.3.I05       购买商品或服务。
4.A.4.c.3.I06       开具医疗治疗或设备。
4.A.4.c.3.I07       监控资源或库存。

# 提示
- 请使用给定的结构化输出格式以**英语**提供您的答案。
```

B Appendix: Classifier Validation

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

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

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

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

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

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

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

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

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

B 附录：分类器验证

为了评估我们的分类器的性能，我们将 LLM 生成的标签与公开可用的聊天机器人对话语料库（WildChat；Zhao 等人，2024）上的人类标签进行比较。注释由多个内部注释员[31]进行。

表 5 报告了人类之间以及模型与人类注释之间的所有任务上的协议率。

Task	n_{labels}	Fleiss' κ (仅人类)	Fleiss' κ(包含模型)	Cohen's κ(人类 vs. 人类)	科恩的 κ(模型 vs. 多样性)
与工作相关 (二进制)	149	0.66 [0.54, 0.76]	0.68 [0.59, 0.77]	0.66	0.83 [0.72, 0.92]
询问 / 做 / 表达 (3类)	149	0.60 [0.51, 0.68]	0.63 [0.56, 0.70]	0.60	0.74 [0.64, 0.83]
对话主题（粗粒度）	149	0.46 [0.38, 0.53]	0.48 [0.41, 0.54]	0.47	0.56 [0.46, 0.65]
IWA Classification	100	0.34 [0.23, 0.45]	0.47 [0.40, 0.53]	0.37	—
GWA Classification	100	0.33 [0.22, 0.44]	0.47 [0.40, 0.54]	0.36	—
交互质量（3类，含未知）	149	0.13 [0.04, 0.22]	0.10 [0.04, 0.17]	0.20	0.14 [0.01, 0.27]

表5：验证顶部结果。“—”表示仅参与两名人类标注者的分类器，无法进行多数测量。

对于每个任务，我们报告：(i) Fleiss' κ 跨人类标注者；(ii) Fleiss' κ 当将模型视为另一个标注者时；(iii) 平均成对人类-人类科恩的 κ；以及 (iv) 模型与人类多数标签之间的科恩的κ 。只有当所有必需的评分者都提供了一个非空标签时，一个项目才会对统计数据做出贡献。置信区间是 95 % 百分位数区间（2.5th 和 97.5th 百分位数）来自一个非参数自举，有 2,000 个重采样。

为了标注这些消息，我们复制了第 3 节中的程序。对于每个对话，分类器应用于随机选择的一个用户消息以及最多 10 个先前的消息（每个截断到 5,000 个字符）。由于这个上下文可能很长，人类标注者还收到了一个关于先前消息的单句 pr´ecis，它是使用以下提示生成的：

```
-----
你是一个内部工具，根据之前的上下文，将用户发送给AI聊天机器人的消息写成一
个一句话的摘要，最多25个字。请总结本次对话中最后一条用户消息的意图。

E.g. '用户正在重写关于邻居的电子邮件，
plumbing to be more friendly,'
or '用户抱怨祖母'
or '用户请求帮助修复 python databricks 错误.'
```

[31] IWA 分类由两位标注员进行，而其他所有分类则由三位进行。

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

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

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

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

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

B.1 Results

B.1.1 Work-Related Classifier

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

B.1.2 Asking/Doing/Expressing Classifier

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

B.1.3 Conversation Topic

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

[32] Ties were broken by a senior annotator.

如果对话主题发生变化，则使用用户最后一条消息的主题。

始终使用英语进行回复。摘要始终以'用户是.'开头

不要分享任何关于用户的名字、
性别认同、位置、电子邮件或电话号码
或任何可能具有个人识别性的信息。

对于交互质量任务，标注者还会看到下一个用户消息，以评估用户对其满意度水平的任何表达的情绪。由于助手消息往往非常长，并且可能需要主题专家进行准确评估，人类标注者只提供了最终的用户消息，而不是助手响应。内部标注者对每个项目进行标记，当有多于两个标注者参与时，真实标签定义为多数标签[32]。开发集（46个项目）用于提示和模型选择；以下所有结果都是在不重叠的保留集上计算的。

除了交互质量任务外，我们使用GPT-5-mini执行所有任务，对于交互质量任务，根据开发集性能选择了GPT-5。

B.1 结果

B.1.1 与工作相关的分类器

如表5所示，模型-多数一致度高（Cohen的 $\kappa = 0.83$），超过了平均人类-人类一致度（$\kappa = 0.66$）。图25中的热力图表明与人类多数意见高度一致，且系统性偏差有限。

B.1.2 询问/执行/表达分类器

人类标注表现出高度一致（平均人类-人类Cohen的 $\kappa = 0.60$），该分类器在此基准上优于人类多数意见，提升幅度为 $\kappa = 0.74$ （表5）。图26和图27显示，大多数混淆发生在询问和执行之间；该分类器比人类更倾向于将任务标记为执行。这种模式表明，我们主要结果中询问用例的突出性不太可能是误分类的产物。

B.1.3 对话主题

模型与人类多数人的协议程度为中等至显著（Cohen’s $\kappa = 0.56$），优于人类之间的平均协议程度（$\kappa = 0.47$)。错误分类主要集中在寻求信息和实用指导之间（图28），这些是概念上相邻的类别。与人类标注者相比，模型对寻求信息、技术帮助和自我表达进行了低标记，而对实用指导、多媒体和其他进行了高标记（图29）。

[32] ties were broken by a senior annotator.

53

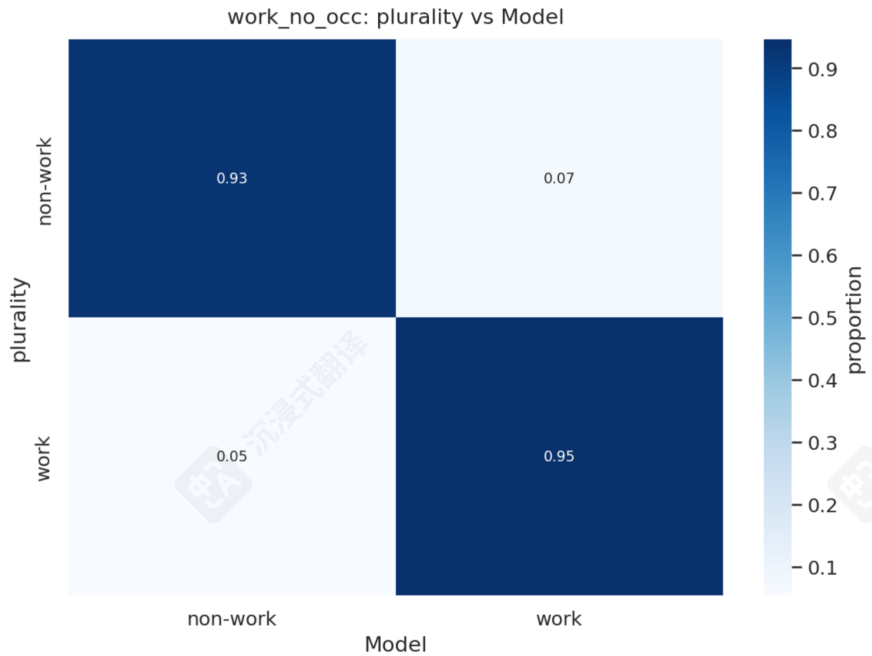


Figure 25: Agreement between Model and Human Plurality.

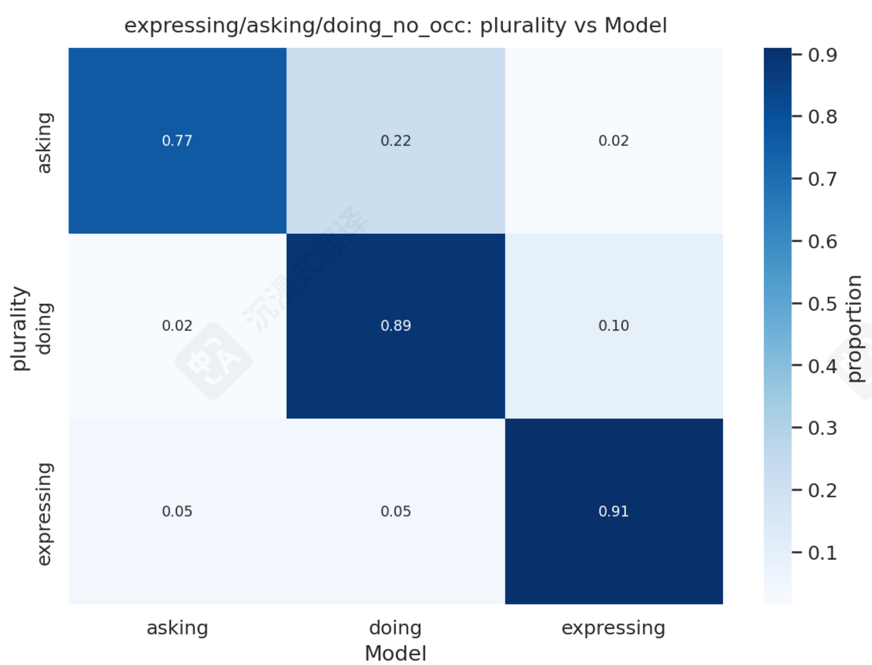


Figure 26: Agreement Between Model and Plurality for Asking/Doing/Expressing

图25：模型与人类多样性的协议。

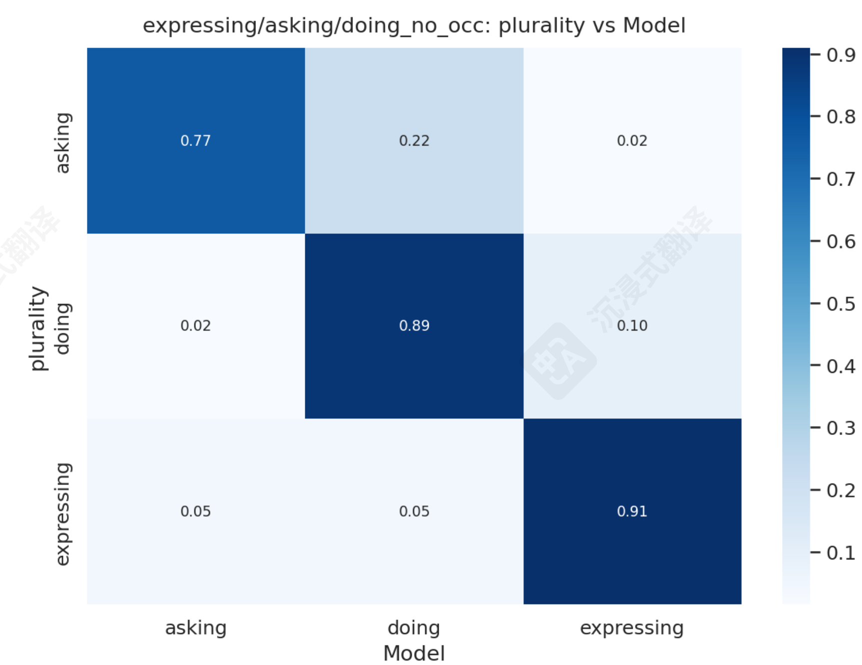


图26：模型与询问/执行/表达的多样性协议。

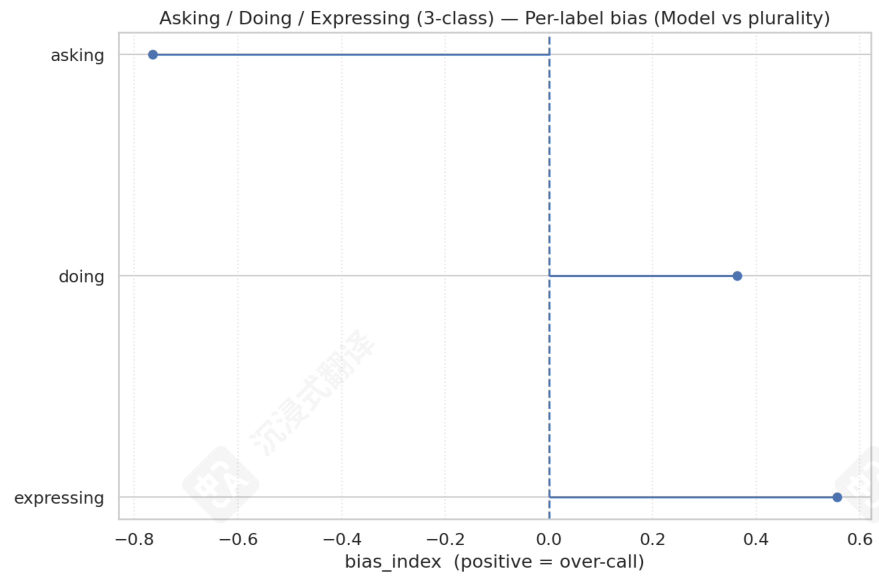


Figure 27: Per-label Bias, Model vs Plurality for Asking/Doing/Expressing

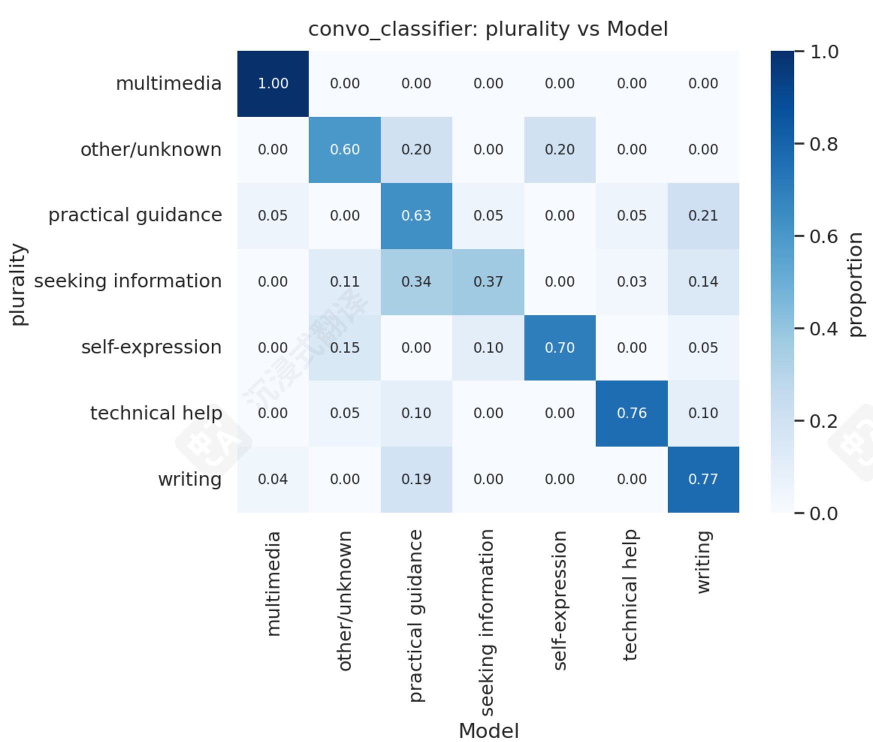


Figure 28: Agreement Between Model and Plurality for Convo-Classifier

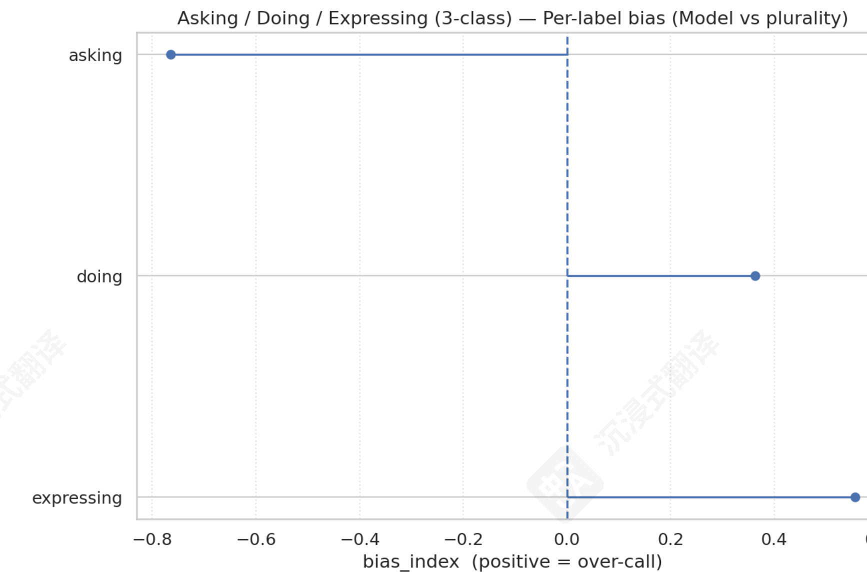


图27：按标签偏差，模型与多数对于询问/执行/表达

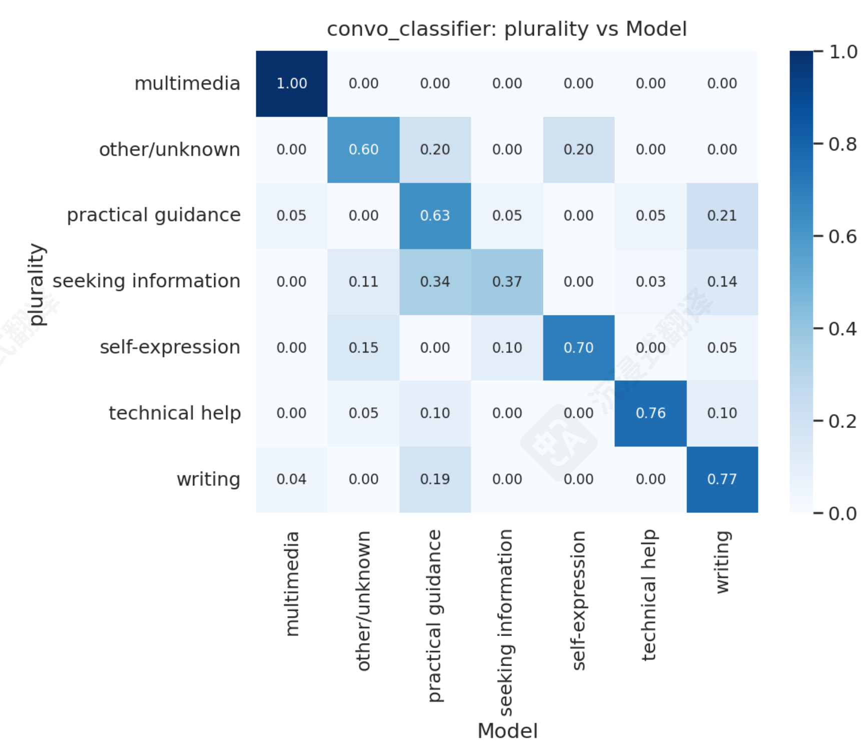


图28：模型与多数对于对话分类器的协议

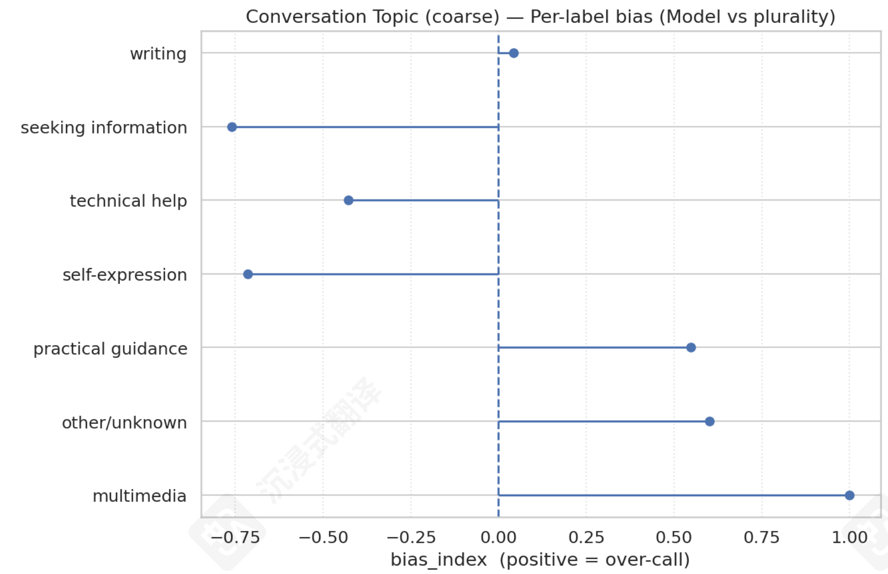


Figure 29: Bias Between Model and Plurality for Convo-Classifier

B.1.4 O*NET Intermediate Work Activity

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

B.1.5 Interaction Quality Classifier

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

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

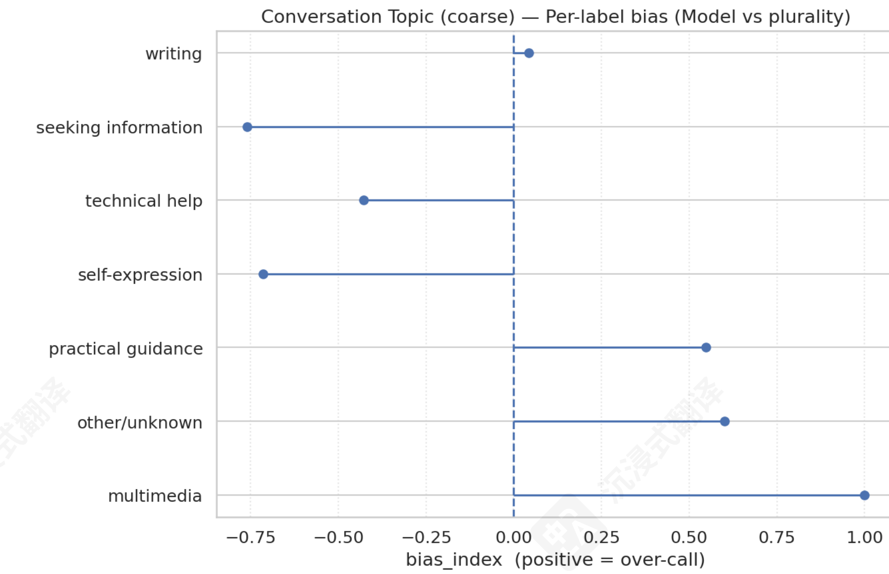


图29：模型与多数派在Convo-Classifier中的偏差

B.1.4 O*NET中间工作活动

两位人类标注者对332个O*NET IWAs中的100条WildChat消息进行了标注，并增加了一个消息模糊的额外类别。人类标注结果与LLM输出进行了比较。在实践中，我们发现当用户只是向模型打招呼或提交空提示时，会选中模糊类别。在这个验证集中，我们报告了Fleiss的 κ ，包括直接IWA分类($\kappa = 0.47$)以及GWA聚合（$\kappa = 0.40$)。当仅检查人类输出时，我们观察到Cohen的 κ 为0.27。从审查中，我们观察到由于潜在类别数量较多（IWA有332个活动）以及消息本身的固有模糊性，导致人类配对同意度适中。例如，如果WildChat数据集中的用户试图生成一个虚构的短故事，一个人类标注可能是“开发新闻、娱乐或艺术内容”，而另一个人类标注可能是“为艺术或商业目的编写材料”。尽管这两个IWA在概念上相似，但它们属于不同的GWA。

B.1.5 交互质量分类器

人类和模型对交互质量的标注是嘈杂的。分类器与人类多数意见（Cohen’s $\kappa = 0.14$）仅达成轻微一致，低于同样适度的平均人类间同意度（$\kappa = 0.20$；表5）。图30和图31显示整体一致性较弱，且模型分配“Bad”的频率比人类略低。这与我们的小型开发集形成对比，在开发集中，GPT-5比人类更频繁地标注“Bad”。我们保留此分类器，因为这些 κ统计数据主要突出了仅从文本中推断用户潜在满意度的固有难度。

虽然这种潜在的“先验”在我们的验证数据中未被观察到，但当用户提供明确的点赞/点踩反馈时，它部分可见。为了评估分类器是否捕捉到与信号对齐的

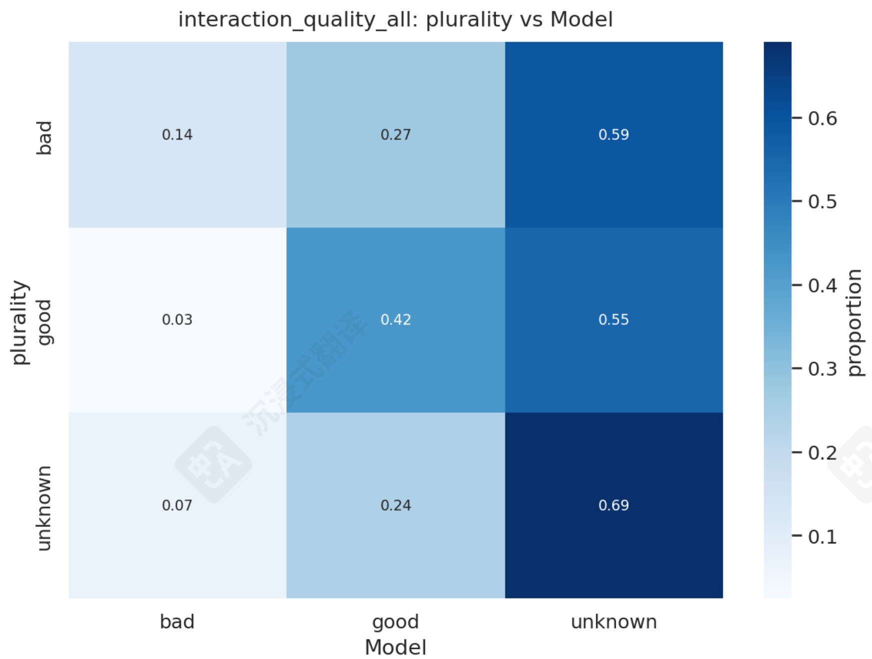


Figure 30: Agreement Between Model and Plurality for Interaction Quality

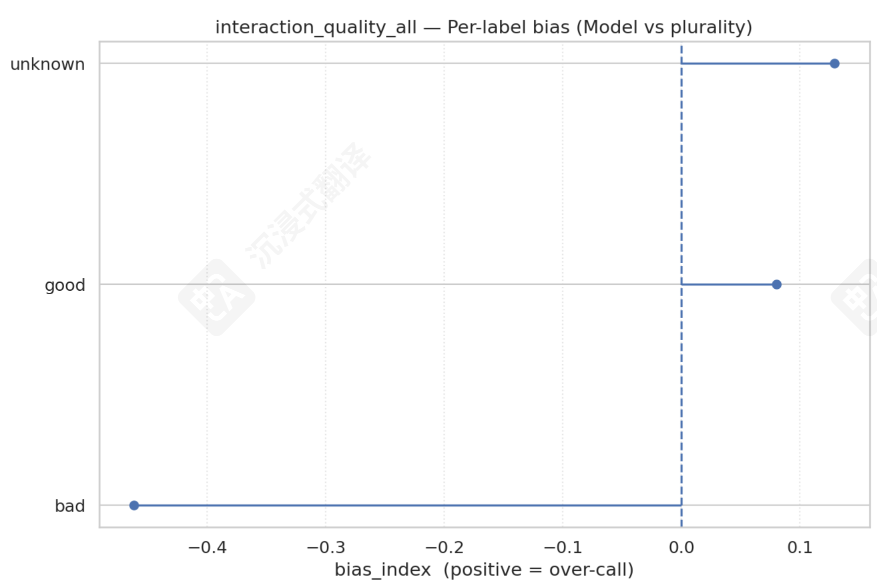


Figure 31: Bias Between Model and Plurality for Interaction Quality

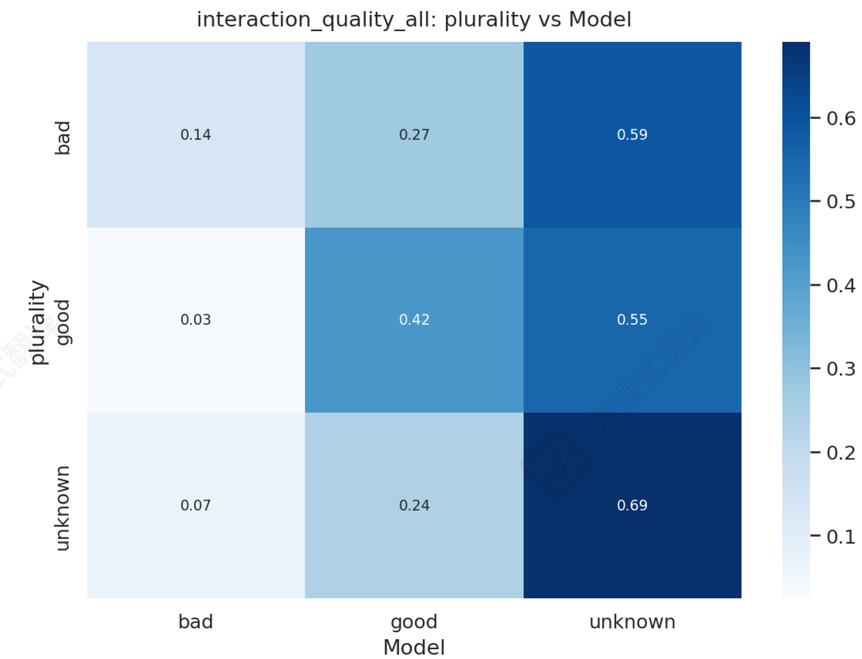


图 30：模型与多数对交互质量的一致性

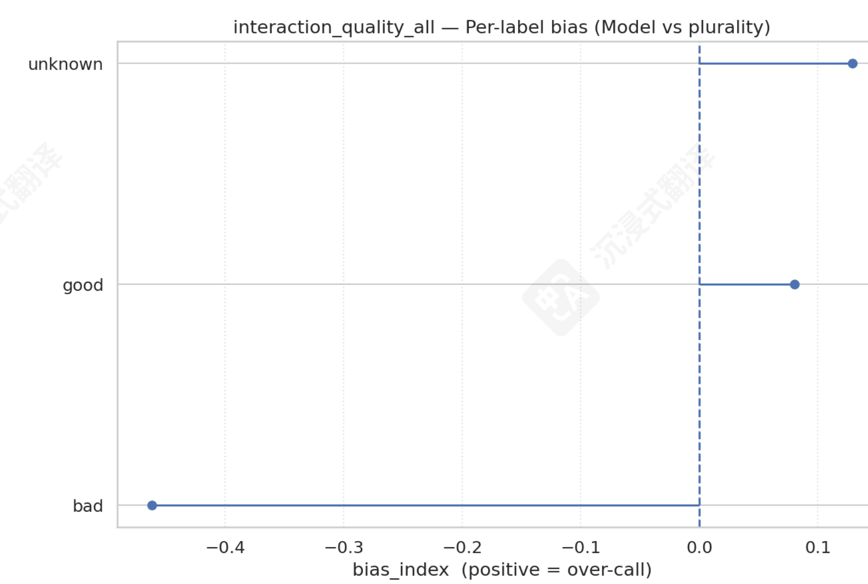


图 31：模型与多数对交互质量的偏差

57

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

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

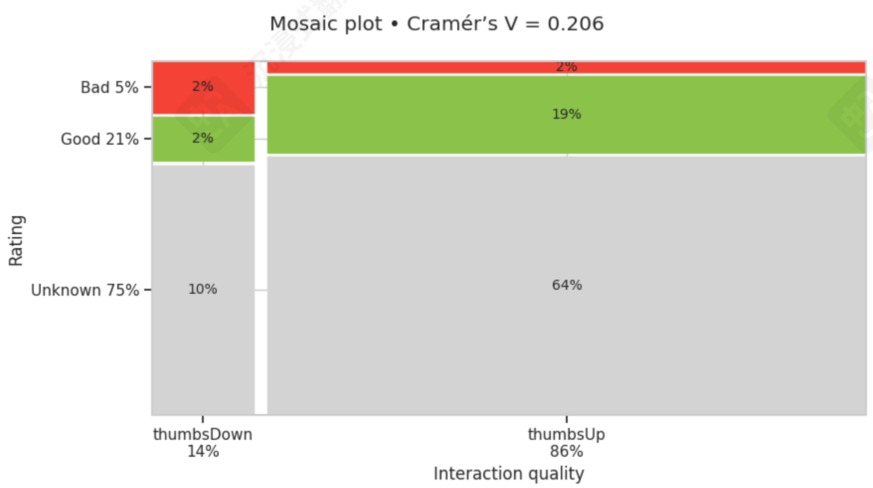


Figure 32: Correlation of User Rating and Interaction Quality Annotation

在用户体验方面，我们将模型预测与助手消息的自愿反馈联系起来。我们从2024年6月至2025年6月抽取了1/10,000的对话样本，并保留了以下情况：(i) 助手消息收到了明确的反馈，以及 (ii) 用户发送了后续消息，我们的分类器可以对它进行评分，从而产生了大约60,000个符合条件的项。这是一个受限的样本，可能无法完全代表所有交互，但它提供了一个独特的视角来观察分类器在代理用户满意度方面的能力。

图32显示，未知分类在点赞和点踩反馈之间大致均等。点赞占所有反馈的86%。收到点踩反馈的对话大约有同等可能被分类为好或坏，而点赞反馈则更有9.5倍的可能性随后被分类为好。

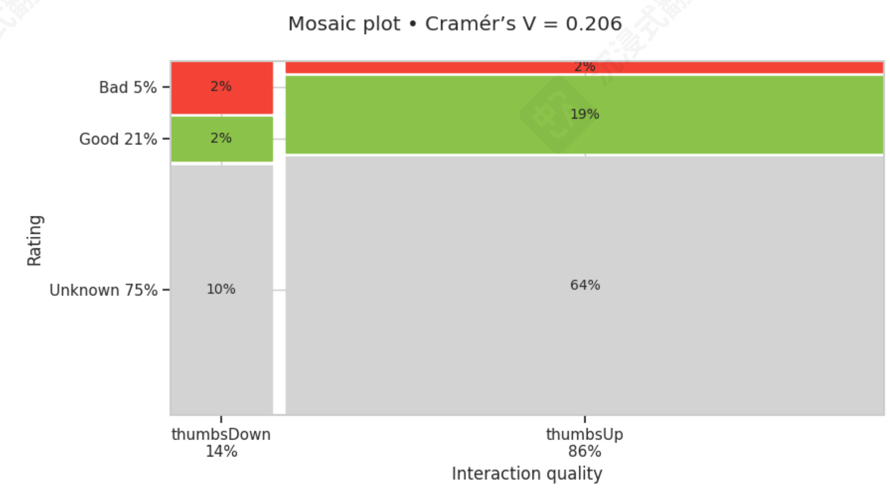


图32：用户评分与交互质量标注的相关性

C Appendix: ChatGPT Timeline

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

C 附录：ChatGPT 时间线

date	事件
2022-11-30	ChatGPT 作为“研究预览”的公测（使用 GPT-3.5）
2023-02-01	ChatGPT Plus 订阅上线
2023-03-14	在ChatGPT Plus中推出GPT-4
2024-04-01	推出已登出ChatGPT
2024-05-13	在ChatGPT Free和Plus中推出GPT-4o
2024-09-12	在ChatGPT Plus中推出o1-preview和o1-mini
2024-12-01	在ChatGPT中推出o1-pro
2024-12-05	推出ChatGPT Pro订阅
2025-01-03	在ChatGPT中推出o3-mini
2025-03-25	推出GPT-4o图像生成
2025-04-16	推出o3和o4-mini
2025-06-10	o3-pro 的发布
2025-08-07	ChatGPT 中 GPT-5 的发布

59

D Appendix: Occupational Results

D.0.1 GWA Breakdowns by Occupation

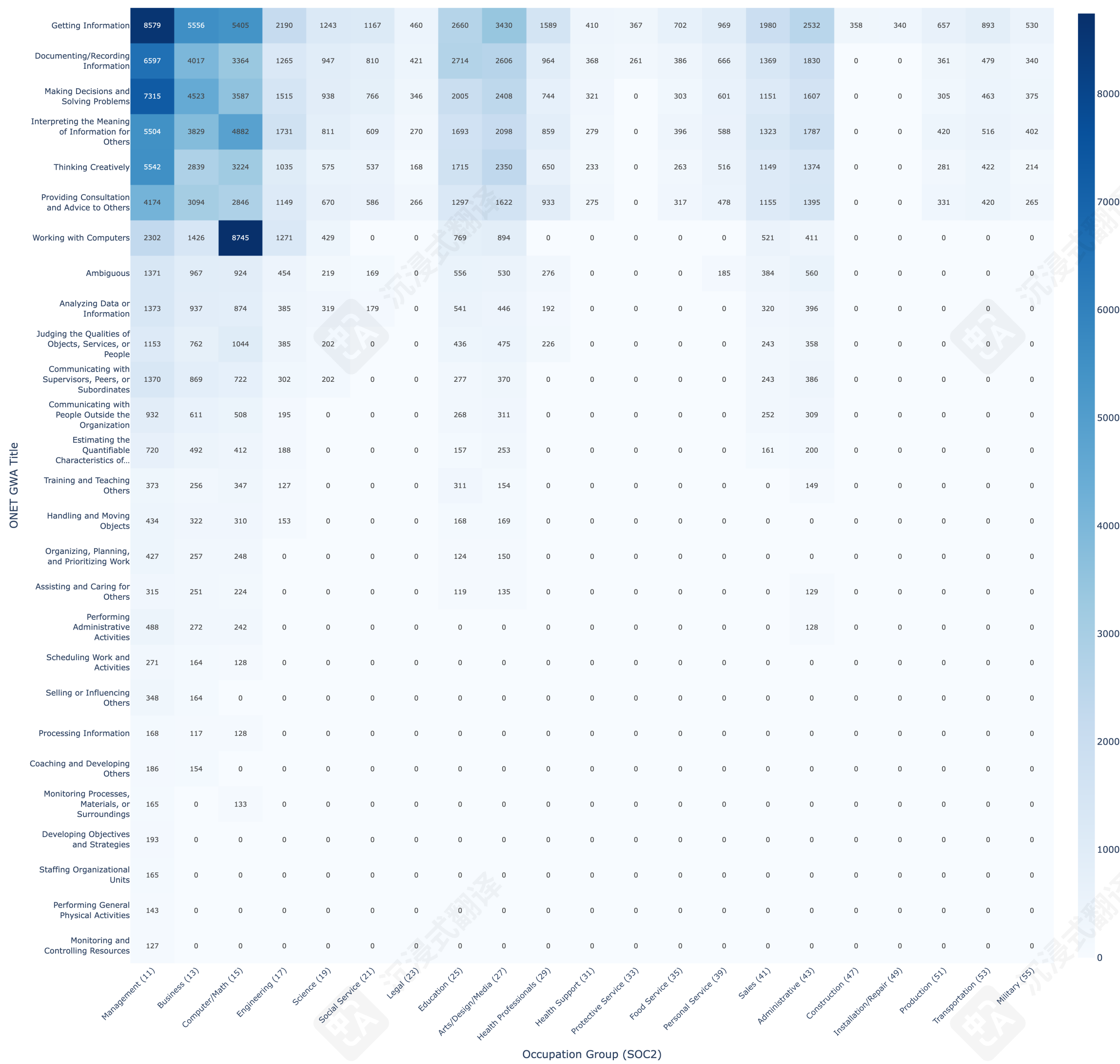


Figure 33: Classified Queries, Organized by Generalized Work Activity (of the query) and Occupation (of the user). Queries are from approximately 40,000 ChatGPT users, from May 2024 through July 2025. Cells with contributions from fewer than 100 users are suppressed to zero. The title of one GWA is not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

D 附录：职业结果

D.0.1 按职业分解的 GWA

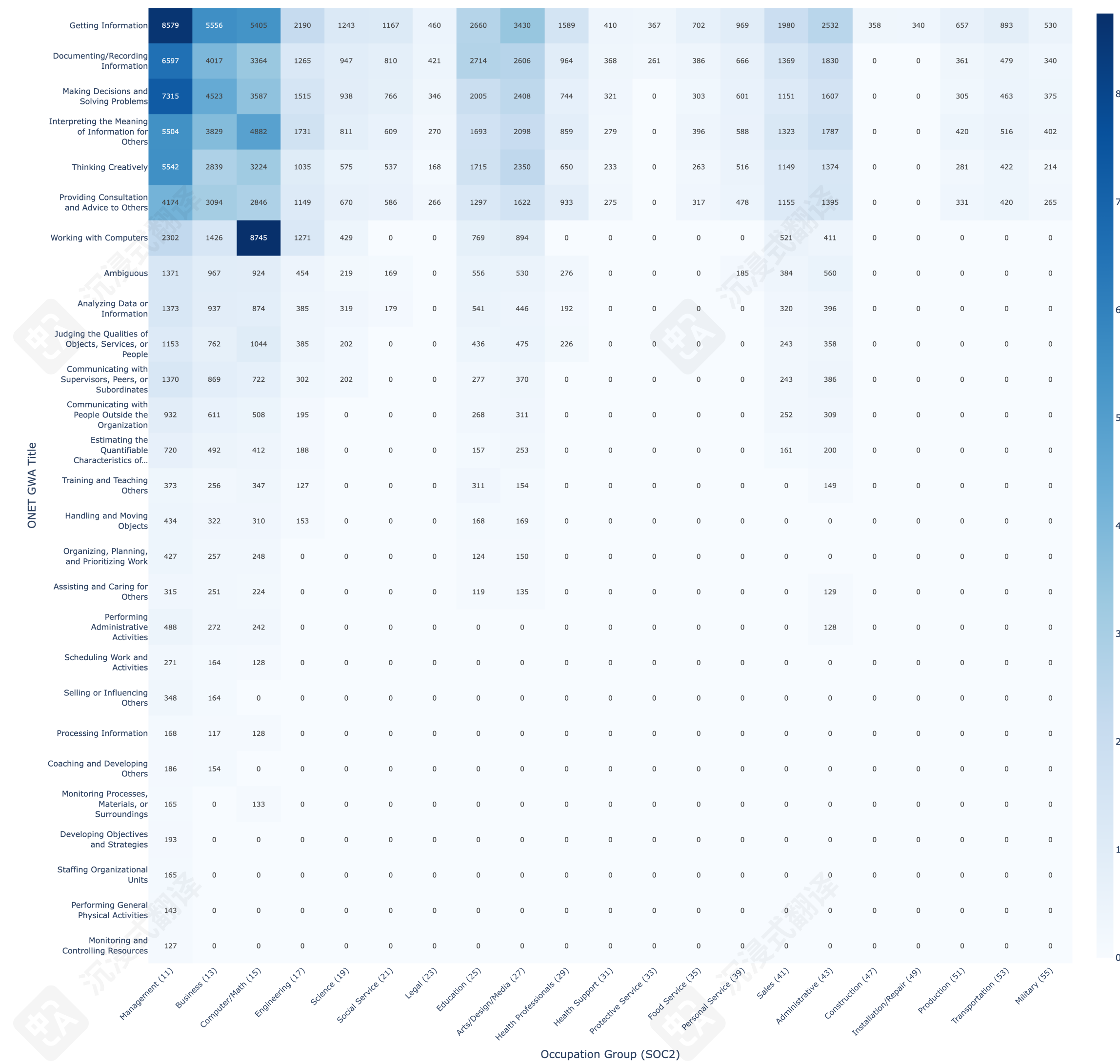


图 33：按广义工作活动（查询）和职业（用户）分类的查询。这些查询来自约 40,000 名 ChatGPT 用户，时间从 2024 年 5 月到 2025 年 7 月。贡献少于 100 名用户的单元格被抑制为零。由于空间限制，一个 GWA 的标题未完全显示：“估计产品、事件或信息的可量化特征。”

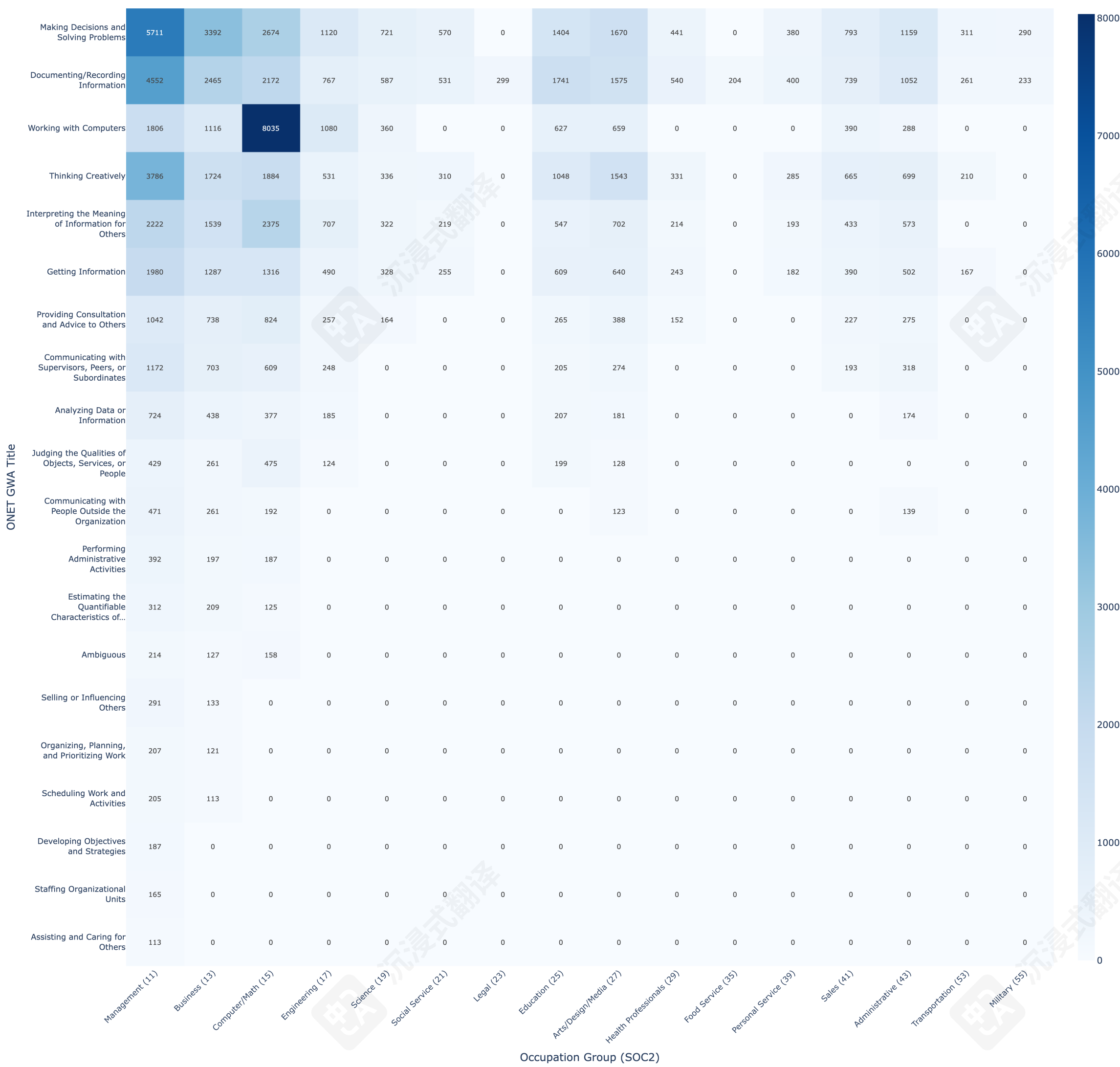


Figure 34: Classified Work-Related Queries, Organized by Generalized Work Activity (of the query) and Occupation (of the user). Queries are from approximately 40,000 ChatGPT users, from May 2024 through July 2025. Cells with contributions from fewer than 100 users are suppressed to zero. The title of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

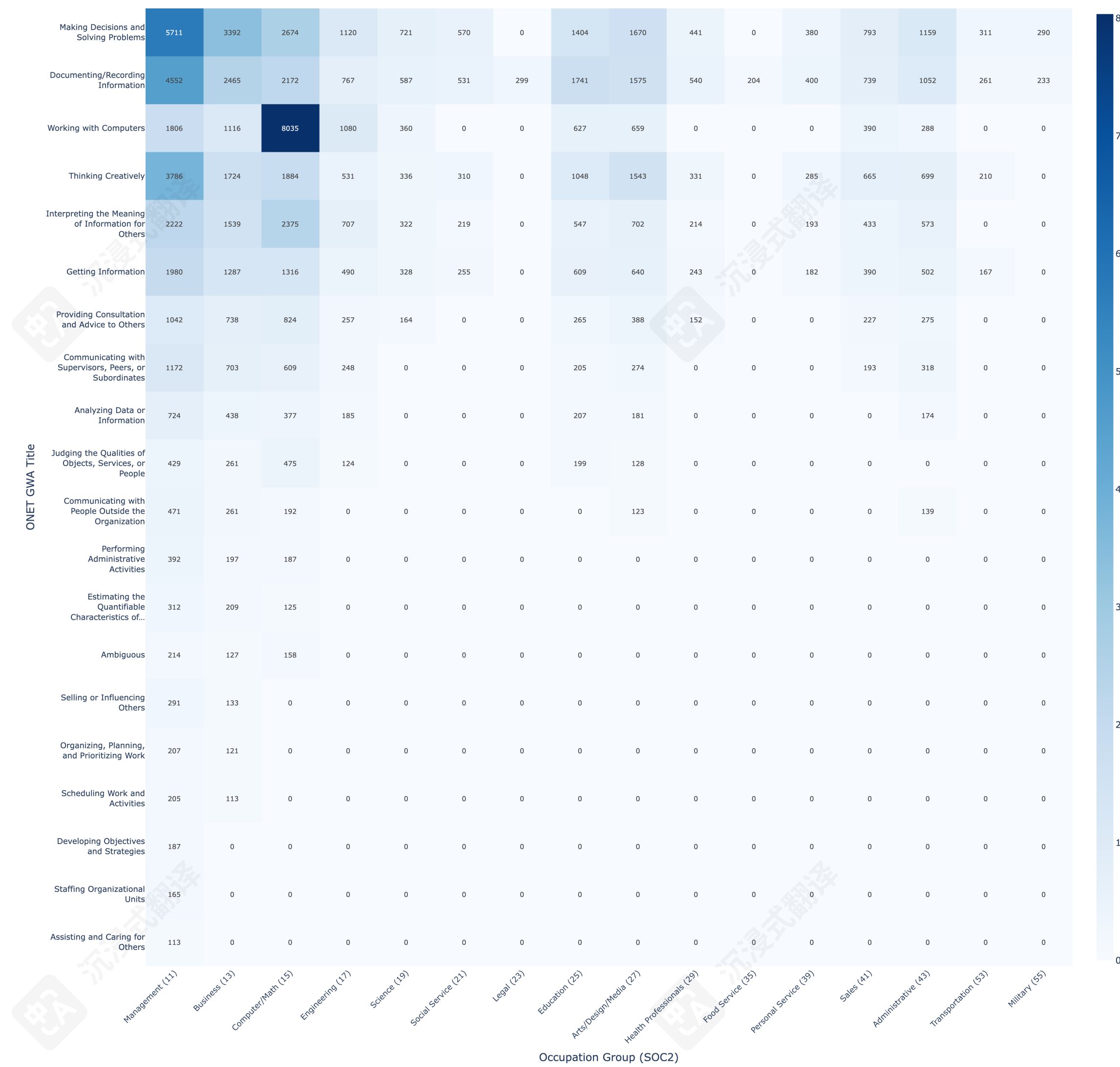


图34：按广义工作活动（查询）和职业（用户）分类的工作相关查询。查询来自约40,000名ChatGPT用户，时间从2024年5月至2025年7月。贡献少于100名用户的单元格被抑制为零。由于空间限制，一个GWA的标题未完全显示：“估计产品、事件或信息的可量化特征。”

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

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

Providing Consultation And Advice To Others
6 5 7 6 5 5 5 6 6 3 5 X 4 6 4 5 X X 4 6 5
rithin broad occupation
ranking is unavailable
no GWA was requested
"Building and Grounds