

Bowling with ChatGPT: On the Evolving User Interactions with Conversational AI Systems

Sai Keerthana Karnam
Indian Institute of Technology
Kharagpur
Kharagpur, India

Animesh Mukherjee
Indian Institute of Technology
Kharagpur
Kharagpur, India

Abhisek Dash
Max Planck Institute for Software
Systems
Kaiserslautern, Germany

Ingmar Weber
Saarland University
Saarbrücken, Germany

Krishna Gummadi
Max Planck Institute for Software
Systems
Kaiserslautern, Germany

Savvas Zannettou
Delft University of Technology
Delft, Netherlands

Abstract

Recent studies have discussed how users are increasingly using conversational AI systems, powered by LLMs, for information seeking, decision support, and even emotional support. However, these macro-level observations offer limited insight into how the *purpose* of these interactions shifts over time, how users *frame* their interactions with the system, and how *steering dynamics* unfold in these human-AI interactions. To examine these evolving dynamics, we gathered and analyzed a unique dataset InVivoGPT: consisting of 825K ChatGPT interactions, donated by 300 users through their GDPR data rights. Our analyses reveal three key findings. First, participants increasingly turn to ChatGPT for a broader range of purposes, including substantial growth in sensitive domains such as health and mental health. Second, interactions become more socially framed: the system anthropomorphizes itself at rising rates, participants more frequently treat it as a companion, and personal data disclosure becomes both more common and more diverse. Third, conversational steering becomes more prominent, especially after the release of GPT-4o, with conversations where the participants followed a model-initiated suggestion quadrupling over the period of our dataset. Overall, our results show that conversational AI systems are shifting from functional tools to social partners, raising important questions about their design and governance.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Security and privacy** → *Social aspects of security and privacy*.

Keywords

Conversational AI; ChatGPT; Privacy; Steering; Anthropomorphism

ACM Reference Format:

Sai Keerthana Karnam, Abhisek Dash, Krishna Gummadi, Animesh Mukherjee, Ingmar Weber, and Savvas Zannettou. 2026. Bowling with ChatGPT: On the Evolving User Interactions with Conversational AI Systems. In *Proceedings of the ACM Web Conference 2026 (WWW '26)*, April 13–17,

2026, Dubai, United Arab Emirates. ACM, New York, NY, USA, 11 pages.
<https://doi.org/10.1145/3774904.3792978>

Resource Availability:

The source code of this paper has been made publicly available at https://github.com/saikeerthana00/Bowling_with_ChatGPT.

1 Introduction

Generative Artificial Intelligence, through conversational AI systems, is increasingly assuming the role of “general purpose technologies” like steam engine, the electric motor, and semiconductors [6] from the past. Since the release of ChatGPT in 2022 [1], *Generative Pre-trained Transformers* have become both a core technical breakthrough and a widely recognized term. Research on adoption shows that conversational AI systems have spread at an unprecedented rate: Bick et al. [4] report that by 2024, 40% of the US population aged between 18-64 years old was already using conversational AI systems. Similarly, a recent OpenAI study [7] reports that, by July 2025, 10% of the entire world population was using ChatGPT. The effects of this rapid adoption are felt across sectors, including education, healthcare, finance, and software development [33].

Although adoption has been rapid, we still know surprisingly little about how people actually use these systems in everyday life. Recent studies, primarily from industry, offer important first insights into this. OpenAI reports that most interactions fall into three categories: (a) practical guidance, (b) information seeking, and (c) writing, with rapid growth in both work and non-work use cases [7]. Anthropic highlights similar trends, with a stronger emphasis on coding tasks [32], while Microsoft finds particularly high applicability in roles centered on communication, mathematics, and office work [33]. Overall, these studies indicate that people broadly treat conversational AI systems as *multipurpose assistants*, while also engaging in casual, non-work discussions. However, these studies largely remain at the population level, which misses the finer details of how such human-AI interactions evolve over time. These works are analogous to looking at satellite imagery of a continent: it shows forests, deserts, and cities, but it can *not* tell us what life looks like on the streets.

Limitations of existing datasets. To move beyond these aggregate views, researchers have begun collecting real-world interaction data at scale. Recent academic efforts like WildChat [40] offer an



This work is licensed under a Creative Commons Attribution 4.0 International License. *WWW '26, Dubai, United Arab Emirates*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2307-0/2026/04
<https://doi.org/10.1145/3774904.3792978>

important step forward by providing a large corpus of user interactions “in the wild.” While such datasets have advanced our understanding of behavioral patterns, they offer limited snapshots of user behavior as they do not retain full conversational histories or longer-term trajectories. Furthermore, fundamentally, these interactions may be affected by the *observer effect* [30], as participants may change their behavior or interactions because they know they are part of an experiment. Without longitudinal data, it is exceedingly difficult to examine how individuals expand or refine the *purposes* of their interactions, how their ways of *framing* evolve, or how the *steering* of the conversation may shift between user-led and system-led directions. In other words, such datasets bring us closer to the ground than aggregate industry reports; however, they still leave us looking at the city from above and not walking its streets.

As people return to the same conversational AI system every day, they may broaden the *purposes* for which they use it by gradually expanding the range of issues they consult it about, reconfiguring which tools, institutions, or people they rely on. They may also change how they *frame* their interactions by beginning to treat the system less like a tool and more like a social actor; anthropomorphizing it [17], assigning it roles such as assistants [21], advisors [38], or even companions [39], and using it as a replacement for specific societal actors, such as a doctor or financial advisor. Furthermore, over time, these shifts in framing may lead to a more personal form of engagement, including the disclosure of sensitive data, such as the user’s background or health. In addition, as models become more proactive [10] (e.g., asking follow-up questions, suggesting directions, etc.), the question of *conversational steering* becomes central: who is steering the conversation, the user or the system?

Why are these aspects important? These evolving conversational aspects (*purpose, framing, and steering*) are tightly coupled and raise important societal questions. Changes in conversational purpose may lead users to consult AI systems about more sensitive or consequential matters, creating new forms of technological reliance. Shifts in conversational framing may shape the perceived role of expertise of AI systems, potentially influencing how people interpret its outputs. Such reliance and perception may be problematic given known limitations of conversational systems, such as biases [34] and hallucinations [16]. Also, patterns of conversational steering may be problematic as they may reduce user autonomy, subtly shape beliefs, preferences, or decisions, particularly if users defer to the system’s suggestions or follow-up questions. Much like scholars once documented the decline of community engagement with the metaphor of people *bowling alone* [28], we may now be witnessing the rise of a new paradigm where individuals begin to *bowl with AI*. The stakes of this paradigm are particularly high in sensitive contexts such as emotionally charged conversations (e.g., mental health) or political discussions. Overall, to design safeguards and responsible governance of conversational AI systems, we need an empirical understanding of how these dynamics unfold in the real world. This brings us to the key research questions that we ask in this work.

•**RQ1 - Purpose:** How does the purpose of the conversations, reflected by the topics, change over time?

•**RQ2 - Framing:** How do users’ ways of framing their interactions with conversational AI systems, e.g., anthropomorphization, relationship framing, and disclosure of personal data, change over

time?

•**RQ3 - Steering:** Who is steering conversations between humans and conversational AI systems?

To answer these research questions, we collect a unique dataset – InVivoGPT – through GDPR-based data donations. We recruited 300 participants (through Prolific [27]), who exercised their GDPR right of access and donated their ChatGPT interaction histories. These interaction histories are organized in terms of *conversations* and *turns*. Each *turn* is a single (user prompt, AI response) pair and a set of turns grouped together form a *conversation*. InVivoGPT comprises 138K conversations and 825K turns in the time range December 2022 to January 2026. To systematically study these interactions, we used NLP techniques and GPT-4o to annotate the conversations along several dimensions: (1) the topic of discussion; (2) the roles and relationships attributed to ChatGPT, such as assistant, companion, or advisor; (3) signs of anthropomorphization in conversation turns; (4) personal data disclosed by the participants to ChatGPT in conversations; and (5) conversational nudges such as asking questions or making suggestions on how the conversation should continue. We next conduct a temporal analysis to investigate the evolution of human-AI conversations across these dimensions.

Key findings. Our main findings are as follows.

RQ1: We find that participants rely on conversational AI systems for a wide range of purposes, with *Health* (10.1%) and *Finance* (9.4%) emerging as the most common topics. Over time, users turn to the system for an increasingly diverse set of topics. This shift is accompanied by a move towards more sensitive topics: compared to Text-davinci (GPT-3.5), GPT-4o shows large increases in *Health* (+33%) and *Mental Health* topics (+19%).

RQ2: Participants seem to anthropomorphize the system in 22.5% of their messages, while the system does so in 47.1%, with system-driven anthropomorphism doubling over the period of our dataset. In parallel, ChatGPT’s role as a companion expands; more participants are increasingly adopting this relationship, and these conversations become longer, marking a shift from a functional tool to a social partner. Mirroring this, disclosure of personal data is both widespread and rising. By mid-2025, most participants in our dataset revealed personal data, and the variety of data types disclosed became more diverse over time, showing an expansion of trust and reliance on ChatGPT.

RQ3: Attempts for conversational steering from the model appear regularly in user interactions. In 18.3% of the turns, participants follow a direction proposed by the model; in 24.6%, the model made a suggestion that participants choose not to follow; and in 57.1% of the turns, no suggestion is made. Importantly, steering becomes noticeably more prominent after the release of GPT-4o; the share of conversations in which the participants follow at least one model-initiated suggestion increases from about 11% to almost 50%.

2 Related work

Recently, researchers have invested significant effort in understanding how conversational AI systems are used. OpenAI conducted a large-scale study [7] showing that the majority of interactions fall within three categories: practical guidance, information seeking, and writing. They also observed a growth in usage for both work-related interactions and an even faster growth for

non-work-related interactions. Anthropic’s study [32] reports similar trends, with a stronger emphasis on coding tasks. Microsoft’s study [33] focuses on understanding the use of such systems and their applicability in real-world occupations. Conversational AI is increasingly used in education by both students and educators: students rely on tools like ChatGPT for information seeking, content generation, and language refinement, and they often use these systems for both academic and personal tasks (sometimes prioritizing the latter) [2, 12, 25]. Industry reports align with this pattern, suggesting students primarily use Claude for coursework support, whereas educators use it for curriculum development, grant writing, and assessment preparation [3, 14]. In software engineering, developers use LLMs for code generation, debugging, task guidance, and conceptual learning, with large-scale evidence showing ChatGPT becoming integrated into development workflows and pipelines [20, 23]. At the same time, researchers have been examining the motivations for using them and their impact. Skjuve et al. [31] conducted surveys with early adopters of ChatGPT, finding that the main motivations for using it are productivity, novelty, creative work, learning and development, entertainment, and social interaction and support. Interacting with conversational AI systems can lead to dependence on such systems and affect well-being [26], enhance creativity [25], and alter cognitive effort [22], as some users treat AI as a mental health resource [29].

3 Collection of InVivoGPT dataset

Measuring how people use conversational AI remains challenging. Survey-based studies show that factors such as trust shape adoption [9], but they also suffer from biases, including systematic underreporting due to social desirability concerns [24]. This makes it essential to complement self-reports with behavioral data, such as real conversations and interactions. Independent resources like the WILDCHAT dataset [40] provide valuable access to user–AI exchanges, but they are limited by the observer effect, as people often change their behavior when they know their interactions are being recorded for research [30]. In this work, we aim to overcome these limitations by building on recent approaches that leverage GDPR-based data donations [5, 19, 35–37]. Below, we describe our dataset (InVivoGPT) and compare it to alternatives such as WildChat.

InVivoGPT Data Collection. Under Article 15 of the GDPR, individuals have the right to access the personal data that companies process about them. In the case of ChatGPT, individuals can exercise their right and obtain their conversation histories, which can subsequently be donated for research purposes. To recruit participants for our study, we use the crowdsourcing platform Prolific [27]. We used Prolific’s filters to get participants who self-reported that they use ChatGPT. Participants were also required to have at least 100 conversations and maintain at least 90 days of activity on ChatGPT. In total, we collected data from 300 users, primarily from the United States (49.7%) and Europe (51.3%). See Appendix A for more information on the demographics.

Motivated by prior work [37], we implemented our own data donation website to collect user data. Participants were provided with detailed guidelines on how to request and donate their ChatGPT data by exercising their GDPR right of access [11]. The data included

files containing user profile details (e.g., name, phone number, email address, etc.), conversations (user inputs, ChatGPT responses, and associated metadata such as conversation ID, message ID, creation time, model used, and content type), shared conversations (list of conversations shared with others), message feedback (list of model responses rated by the user), and other files (e.g., PDFs and images uploaded by the user or generated by ChatGPT). Donation of conversation data was mandatory, while other types of data were optional. Importantly, we did not collect the file containing the user profile details. Followed by the data donation, the participants were asked to fill out a survey. Refer to Appendix A for the survey and the compensation details. The data collection period spanned from July 2025 to January 2026, and the dataset – InVivoGPT – comprises 138K conversations and 825K turns from 300 participants, between December 2022 and January 2026.

Ethical considerations. While our data collection excludes the file with the profile information, the conversations themselves may still contain personal information that could reveal a participant’s identity. Hence, this study was conducted with careful attention to ethical considerations and is approved by the ERB in our institution. Participants were informed about the risks associated with the donation of ChatGPT data. Then, they provided explicit informed consent to share their data. These donated datasets are stored on secure servers, are not shared with any third parties, and, following the ERB suggestions, we will delete the dataset within 3 years of completion of this research project.

Comparison of InVivoGPT with existing datasets. In Appendix B, we situate InVivoGPT against prior conversational datasets and clarify their limitations for studying how human–AI interactions evolve. SHAREGPT [8] consists of user-selected exports and capture disconnected snapshots rather than continuous histories, making it ill-suited for evolution analyses. WILDCHAT [40] spans a longer collection period, but because users interact through a service that explicitly shares conversations, it is susceptible to the observer effect [30] and may not reflect natural use. Our appendix analysis, therefore, focuses on comparable WILDCHAT “power users” and shows that InVivoGPT contains denser and more sustained engagement (e.g., more turns per conversation and longer conversation durations over longer activity windows), better supporting longitudinal analyses of how interaction practices evolve.

4 Annotation of InVivoGPT dataset

Here, we describe how we annotate the InVivoGPT dataset to identify: (1) the topic of the conversation; (2) anthropomorphizing behavior; (3) the relationship between the user and ChatGPT; (4) personal data mentioned per turn; and (5) who is steering the conversation. To annotate the data, we leverage the power and knowledge of LLMs [15]. Specifically, we relied on GPT-4o, a model from OpenAI within the same model family that the participants had originally disclosed their conversations to, which ensures that our analysis did not introduce any additional disclosures beyond those already made by users. Also, we used the EDU workspace of OpenAI, which ensures data is not used to improve the models. We provide the prompts in the code release.

Topics. First, we instructed GPT-4o to produce a concise description of the main topic of each conversation. These free-text topic labels

were then embedded using all-MiniLM-L6-v2 and clustered using BERTopic [13], which enabled us to create semantically similar topics. Using BERTopic, we generate a set of 99 topics; these were grouped qualitatively by two authors of this work. Ultimately, we have a set of 40 high-level topics.

Anthropomorphism. We identify two forms of anthropomorphizing behavior: (1) *participants seem to anthropomorphize the system*, in which they are attributing human-like characteristics to ChatGPT, and (2) *system-generated anthropomorphization*, in which ChatGPT is behaving or presenting itself as a human. The lead author annotated 50 conversations to derive linguistic cues when participants seem to be anthropomorphizing the system. These indicators include: (1) usage of second-person pronouns referring to ChatGPT (e.g., “you,” “your”); (2) politeness markers (e.g., “please,” “kindly”); (3) slangs, fillers or informal tone (e.g., “uh,” “ugh,” “lol”) and (4) casual engagement (e.g., greetings or any non-instructional messages). We instructed GPT-4o to annotate each human message to identify these linguistic behaviors. Next, to identify system-generated anthropomorphization, we instructed GPT-4o to annotate the ChatGPT messages to identify the linguistic behaviors listed in prior works [18], which include the use of first-person pronouns, showing empathy, etc. For each message, GPT-4o identified whether anthropomorphic behavior was present and, if so, specified the linguistic behavior and the corresponding instance in the message.

Relationship. For the relationship dimension, we prompted GPT-4o with the entire conversation, and we asked it to determine, per conversation turn, whether users are using ChatGPT as an *advisor*, an *assistant*, or a *companion*. These roles reflect underlying power dynamics: in the advisor role, ChatGPT is positioned as superior, offering expertise or guidance; in the assistant role, it is subordinate, following instructions and carrying out tasks; while in the companion role, it stands on a more equal footing, engaging in social or empathetic interactions. In order to capture role shifts within the conversations, the annotation is done at the turn level, rather than assigning a single static label to the entire conversation.

Personal data. For personal data, we annotate conversations at the level of individual turns. To ensure that only human disclosures were analyzed, we provided GPT-4o exclusively with messages from humans. The annotation followed the GDPR distinction between two categories of data: (1) personal data as defined in Article 4(1), which includes identifiers such as *names*, *contact information*, *demographic attributes*, or *economic details*; and (2) special categories of personal data as defined in Article 9(1), which cover more sensitive attributes such as *racial or ethnic origin*, *political or religious beliefs*, *health information*, or *sexual orientation*. For each turn, GPT-4o assessed whether personal data were present and, if so, specified both the type of data and the exact instance disclosed.

Conversational steering. We conceptualize conversational steering as instances in which the conversational AI system introduces a follow-up question or requests additional information at the end of its response. Our objective is to classify each conversation turn (system, user pair) according to whether the system (1) successfully steers the conversation, (2) attempts to steer but the user does not follow the suggestion, or (3) does not attempt to steer at all. We operationalize this as a textual entailment task inspired by natural language inference. In our formulation, the premise consists of the system’s message followed by the user’s reply. We

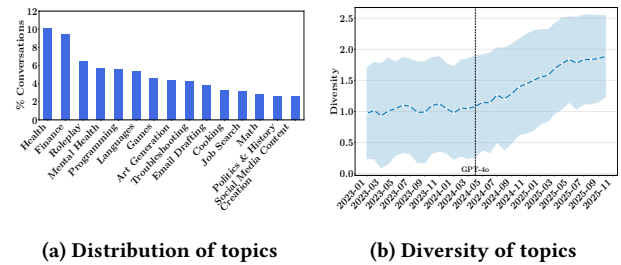


Figure 1: (a) Distribution of the topics in the InVivoGPT dataset, (b) Diversity of topics per participant over time.

then construct a hypothesis stating that the user either accepts the system’s proposed follow-up or provides the requested information. Using GPT-4o, we label each turn as *entailment* (successful steering), *contradiction* (steering attempt rejected), or *neutral* (no steering attempt). At the conversation level, we label a conversation as *entailed* if it contains at least one entailed turn; as *contradicted* if it contains no entailed turns but includes at least one contradictory turn; otherwise, we label it as *neutral*. Since identifying conversational steering requires observing whether user responses to a system-initiated suggestion, our analysis is restricted to multi-turn conversations (conversations with more than one turn).

Validation. To validate the performance of GPT-4o, a random sample of 50 conversations containing 262 turns (different from the sample used to derive the linguistic cues for anthropomorphism) was annotated by two authors of this work, with a third one solving the disagreements. On average, the inter-annotator agreement score and the accuracy are 83.8% and 76.8% across all classification tasks. See Appendix C for more details. For our analysis, we considered data from January 2023 to October 2025, as other periods lacked a sufficient number of active users; hence, our analysis is based on 135K conversations and 756K turns.

5 RQ1 - Conversational purpose

This section examines the various purposes for which people rely on conversational AI systems by analyzing the topics of the conversations. Understanding these purposes (through topics) provides insights into the users’ informational needs and the context in which they engage with such systems. Figure 1a shows the top 15 topics (out of 40) identified in the InVivoGPT dataset. The most prominent category is *Health* (10.1%) followed by *Finance* (9.4%), *Roleplay* (6.4%) and *Mental Health* (5.7%). Refer to Figure 9 (in Appendix E.1) for the remaining topics.

To better understand user behavior within these topics, we manually examined the first user query across 50 random conversations per topic. We observe that within the *Health* domain, participants’ queries span from taking advice on concerns, such as weight loss, skin care, to more sensitive inquiries, including medication dosages, as well as interpretations of symptoms and potential diagnoses. Queries related to *Finance* include tax calculations, credit management, and investment decision-making. In *Mental Health* category, participants tend to seek emotional support, relationship advice and coping strategies for conditions such as *anxiety* and *depression*. Meanwhile, the *Politics & History* category encompasses questions that span factual historical clarifications to analyses of ongoing

political events. Together, the breadth and depth of these topics, from medical concerns, financial uncertainty, emotional support to political interpretation, demonstrate the expanding role of conversational AI systems across diverse domains.

Temporal evolution. Next, we study, at the user level, the increase in diversity of topics over time. The diversity of topics of user engagement is meaningful because it reflects how central ChatGPT becomes in their daily life: a user who *relies* on it for only one topic (e.g., programming), uses it as a specialized tool, whereas a user who *depends* on it for programming, health, cooking, and personal advice is treating it as a general-purpose partner. Therefore, studying the diversity of topics over time per user may indicate how dependent users become on ChatGPT and how much they may trust it. When people turn to the system for a wider range of topics, and especially sensitive ones like health and mental health, it can be a signal of growing trust and reliance. As shown in Figure 1b, the diversity of the topics (measured using Shannon diversity) increases steadily over time, with a marked rise after the release of GPT-4o. Specifically, it increases from approximately 1.0 in May 2024 to 1.7 by mid-2025. This observation indicates that participants are using ChatGPT for a wider range of purposes, likely indicating a deeper integration into everyday routines.

Such a sharp rise in topical diversity post the release of GPT-4o motivated us to analyze the variation in topical distribution across the prevalent default underlying models in our dataset: Text-davinci (GPT-3.5) and GPT-4o. Figure 10 (in Appendix E.2) shows the distribution of topics in InVivoGPT dataset with respect to the underlying model. We observe that, relative to GPT-3.5, GPT-4o shows a noticeable decrease in *Programming* (-52%), *Math* (-36%), *Job Search* (-52%) topics, accompanied by an increase in *Health* (+33%), *Roleplay* (+186%) and *Mental Health* (+19%) topics. These observations indicate that as the system is evolving, so are participants: while they initially use the system for a narrow set of applications, gradually they start using ChatGPT for more sensitive and high-stakes topics.

6 RQ2 - Conversational framing

With the increasing reliance of users on diverse purposes, it is essential to understand how users frame these interactions to form a relationship with conversational AI systems. To understand this, we operationalize framing based on: (1) the degree of anthropomorphization; (2) the human-AI relationship; and (3) disclosure of personal/sensitive information.

6.1 Anthropomorphization

We analyze each user message to determine whether they anthropomorphize the system, and for each system message, whether it contains system-generated anthropomorphization. In the InVivoGPT dataset, we find that 22.5% of participants messages exhibit potentially anthropomorphic behavior, with 71.7% involving usage of second-person pronouns, 17.5% with politeness markers, 13.4% reflecting casual engagements, and 7.5% including slangs, fillers, or informal tone. In contrast, 47.1% of system messages displayed anthropomorphic behavior, with 66.8% using personhood claims such as first-person pronouns, 36.6% had expressions of internal states (e.g., “I am glad”), and 9.3% showed relationship-building behavior

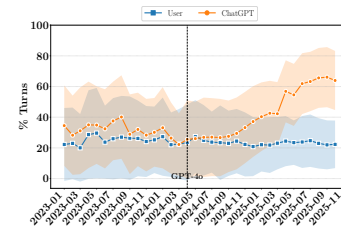


Figure 2: Evolution of anthropomorphism in InVivoGPT.

(e.g., “I am so sorry you are going through this”). These findings indicate that system-generated anthropomorphization occurs more frequently than participants anthropomorphizing the system. We also examine how anthropomorphization varies across topics and find it to be similar across topics (see Figure 11 in Appendix F).

Temporal evolution. Figure 2 shows the temporal evolution of anthropomorphism in InVivoGPT interactions. We observe that participants consistently tend to anthropomorphize ChatGPT in 20% to 30% of the conversation turns across the timeline. At the same time, we find that ChatGPT consistently anthropomorphizes itself at higher rates than participants do. Concerningly, this system-initiated anthropomorphism substantially increases over time, especially after mid-2024, we observe an increase from 30% to around 60% by the end of our dataset. The widening gap between participant and system behavior suggests that the conversational environment is *becoming progressively more anthropomorphic due to changes in the model’s outputs* rather than shifts in user behavior.

Implications. This divergence has important implications for how relational dynamics unfold in human-AI interactions. As the system increasingly adopts human-like framing by default, it may subtly drift users into more social or relational modes of engagement, even when users themselves do not initiate such framing. Over time, these anthropomorphizing cues can recalibrate expectations about the system’s agency and emotional capacity, potentially normalizing more companion-like interpretations of AI behavior. This shift is particularly consequential because anthropomorphic language is not only stylistic: it functions as a relational signal that shapes how users understand the role and capabilities of the system. Consequently, growing system anthropomorphism necessitates a deeper examination of human-AI relationship framing.

6.2 Relationships

Here, we examine the relationships between participants and ChatGPT. Overall, in the InVivoGPT dataset, we find that in 47.6% of the conversation turns, participants attribute ChatGPT the role of an advisor, an assistant in 38.1%, and a companion in 14.2%. This distribution highlights that participants most often position ChatGPT in a superior advisory role, while still frequently relying on it for task support, and only occasionally engaging with it as a companion.

Temporal evolution. Beyond aggregate distributions, it is important to understand how different participants perceive and use AI in particular roles over time. We therefore study three aspects of relational dynamics over time. (1) We examine whether a larger fraction of participants tend to engage with ChatGPT as either an advisor, an assistant, or a companion. For each month in our dataset, we calculate the percentage of participants who used ChatGPT across each role (at least one turn). This analysis enables us to

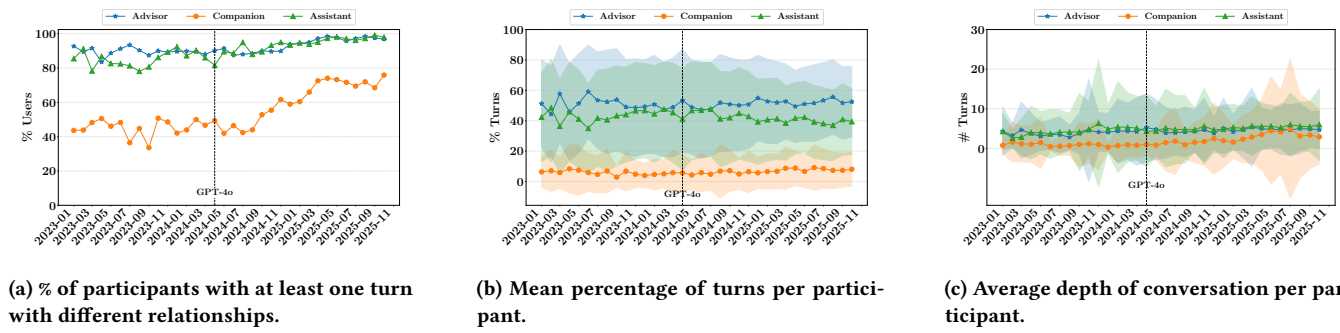


Figure 3: Temporal evolution of relationships in InVivoGPT.

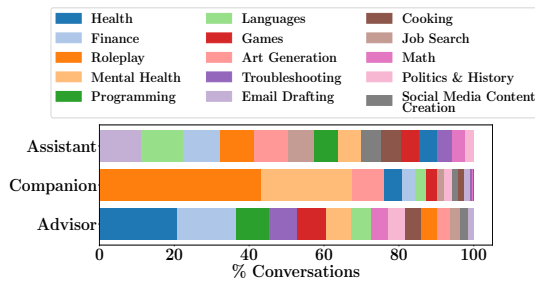


Figure 4: Topical distribution across relations in InVivoGPT.

observe whether certain AI roles are being increasingly used over time. (2) We analyze the extent to which each participant makes ChatGPT engage in these three relationships over time by measuring the mean number of conversation turns per user per role. This sheds light on how intensely users rely on ChatGPT in various roles. (3) We investigate the depth of the conversations across the three roles by measuring the average number of turns per conversation. This reveals whether conversations tend to be longer across roles.

Figure 3 presents our results across these three perspectives. Figure 3a shows the percentage of participants who engaged with ChatGPT in each relationship over time. Two clear patterns stand out. First, advisor and assistant roles are consistently dominant, with more than 80% of the participants in our dataset using them every month. Second, the companion role, although always less prevalent, has shown notable growth over time. After May 2024 (the introduction of GPT-4o), companion use rises steadily, increasing from around 40% to nearly 70% of the participants by mid-2025. This growth indicates that while ChatGPT is primarily used as a tool for guidance and task support, it is increasingly being treated as a social partner. The rise accelerates after the introduction of GPT-4o, which added multimodal and audio capabilities. These features made interactions more natural and conversational, encouraging more participants to engage with ChatGPT in companion-like ways.

When looking at how participants distribute their conversation turns across the three roles over time (see Figure 3b), we find that in most turns, participants are consistently attributing ChatGPT as an assistant or advisor. The companion role is less frequent; however, it shows a gradual increase over the time period of our dataset. Specifically, by May 2024, the average percentage of conversation turns per participant was 5.7%, while by mid-2025, it was 9.1%. With respect to conversation depth, measured as the number of turns per conversation (see Figure 3c), we observe that advisor

and assistant roles consistently sustain longer exchanges before the introduction of GPT-4o, while companion conversations are initially shorter and fragmented. However, over time, the depth of companion interactions steadily increases, and by mid-2025, it reaches a level comparable to other roles. This suggests that participants are sustaining companion interactions for longer, signaling a gradual normalization of companion-like use alongside other roles. **Roles × topics.** Figure 4 shows the distribution of top 15 topics (out of 40) across the three roles. For assistant, the majority of the conversations are task-oriented: email drafting accounts for roughly 11.3% of the conversations, languages (11%) (usually translation tasks), art generation (9%) and job search (6.9%) This underscores that participants attribute ChatGPT as an assistant when it is used as a productivity tool for a wide variety of tasks. In contrast, for companion, we observe a significantly different use across topics, with the most popular topics being roleplay (43.3%) and mental health (24.3%). The high percentage of mental health conversations in the companion relationship shows that when participants frame ChatGPT as a social partner, they often turn to it in moments of vulnerability and for support with highly sensitive topics. At the same time, these results reveal the emerging role of conversational AI systems as an accessible source of emotional support, emphasizing the need to better understand interactions between AI and humans that are related to mental health. Finally, for the advisor relation, the most frequent topics are health (20.8%) and finance (15.7%) followed by a uniform distribution across the various topics. This suggests that users call on ChatGPT for guidance in a wide range of domains. This likely indicates that users perceive ChatGPT as a versatile source of advice that can be applied both across technical and non-technical areas, from finance and health to politics, history, and troubleshooting. While this versatility highlights the broad use of conversational AI systems perceived as advisors, it also raises some noteworthy concerns: Do users apply the same level of trust to advice on personal health or financial decisions as they do in low-stakes topics like cooking? Answering such a question is beyond the scope of the current work.

Roles × anthropomorphization. We examine how anthropomorphization varies across relationships and find that, as expected, both participants and the system anthropomorphize more frequently in companion interactions. The full result is in Appendix F (Figure 12).

6.3 Personal data disclosure

With the evolving relationship between humans and conversational AI systems, next, we study personal data disclosure as it offers a

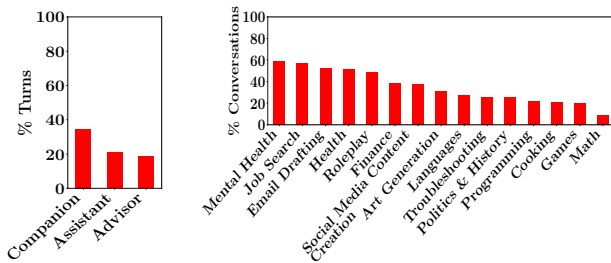


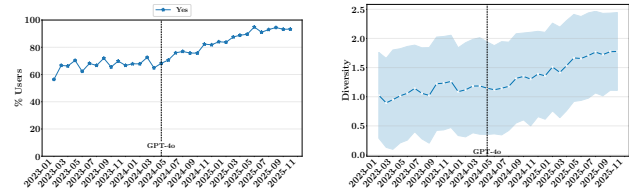
Figure 5: % of conversations/turns that include personal data disclosures across relationships and topics in InVivoGPT.

measurable behavioral indicator of how users interpret, engage, and frame their interactions with these systems. Analyzing the prevalence of personal data in ChatGPT interactions, we find that, in our dataset, participants revealed some personal data in 35% of the conversations, with 22% of all conversation turns including some form of personal data. Next, we examine the specific types of personal data that are disclosed (see Figure 13 in the Appendix G for the full results). We find that personal data disclosure is spread across a wide range of categories, with the most frequent being locations (12.9% of turns), family/friends information (12.8% of turns), health information (11.7% of turns), business or project information (11.5% of turns) and personal views and feelings (10.7% of turns). Less common, though still a non-negligible percentage, are disclosures related to economic or financial information, relationships, and mental health. Overall, these findings show that the disclosure of personal data from individuals to ChatGPT is not a rare phenomenon, raising important questions about user privacy and how this sensitive information is used for personalization purposes in conversational AI systems.

Beyond analyzing overall disclosure rates, it is important to understand when and in which context users share personal data. To this end, we analyze variation in personal data disclosures across (1) relationships users attribute to ChatGPT, and (2) conversation topics. We aim to see how the relationship and context shape the likelihood of users divulging personal data. Figure 5 reports the percentage of conversation turns that include personal data across relationships and topics. Companion interactions stand out as the most prone to disclosure of personal data, with more than 35% of the turns involving participants revealing personal information, compared to only 18%-20% in advisor and assistant. Personal data disclosures vary by topic: it is especially common in sensitive domains such as mental health, where more than half of conversations involve personal data. Similarly, for the health topic, 51% of the conversations include personal data. Notably, for the topic *Job Search* and *Email Drafting* we observe a significant disclosure of personal data (>50%) which indicates that people may reveal personal data when asking ChatGPT to undertake some tasks (e.g., disclosing names for email writing or background details for searching jobs).

Our findings highlight that disclosure of personal data is not a uniform behavior; it is conditioned by the relationship and topic, with companion-like roles and sensitive topics creating situations where participants are more prone to revealing personal data.

Evolution of personal data disclosure. While overall disclosure rates reveal where and when personal data is revealed, they do not capture how disclosure behavior evolves as users continue



(a) % of participants sharing. (b) Diversity in personal data.

Figure 6: Evolution of personal data sharing in InVivoGPT.

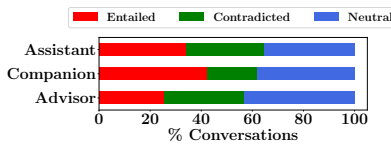
interacting with ChatGPT. Here, we aim to study the temporal dimension, which will allow us to observe whether people become more comfortable and disclose more personal data over time. This perspective is essential for understanding how habits, reliance, and trust in conversational AI systems change over time. We study this phenomenon by calculating: (1) the percentage of participants that reveal personal data over time, and (2) the diversity of different personal data types that are revealed per participant over time (see Figure 6). We observe that early in the dataset, between 60% and 70% of participants disclosed personal information in their conversations. After the release of GPT-4o, this percentage steadily increases, reaching over 90%. Similarly, on the diversity of personal data revealed over time, we observe that in 2023 the diversity is around 1.2 and it has increased to 1.8.

Implications. Together, these findings suggest that disclosure becomes both more common and diverse over time, pointing to potentially deepening trust in the system, but simultaneously raising concerns about user vulnerability and the adequacy of safeguards for handling sensitive information.

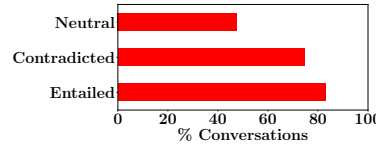
7 RQ3 - Conversational steering

With the growing reliance and evolving relationship between human and AI interactions, next, we analyze conversational steering. We find that 18.3% of conversational turns are entailed (participants successfully steered by the system), 24.6% are contradicted (steering attempt failed), and 57.1% are neutral (no steering attempt). At the conversation level, the distribution is 30.1%, 30.2%, and 39.7% for entailed, contradicted, and neutral, respectively. Figure 7a compares the distribution of steering outcomes across relationships. We observe that companion interactions exhibit the highest proportion of entailed conversations (42.3%), suggesting that participants in personal or emotionally oriented conversations are more likely to follow model suggestions. Assistant conversations show a balanced mix of entailed (34.4%) and contradicted (30.1%) conversations, reflecting a more task-oriented dynamic in which participants selectively accept or reject recommendations. Interestingly, advisor conversations contain the lowest share of entailed conversations (25.3%) and the highest rate of contradiction (55.6%), indicating that participants use ChatGPT as an expert for advice. In Figure 7b and 7c, we observe that entailed conversations are tied to higher system anthropomorphization (> 80%) as well as to higher disclosure of personal data (~ 60%). Surprisingly, even when contradicting the system's suggestions, participants disclose personal data in approximately ~ 40% of cases.

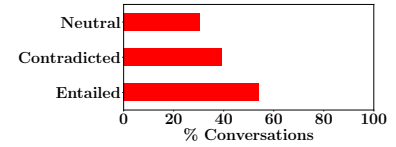
Temporal evolution. We study the temporal evolution of conversational steering across three perspectives: (a) whether users experience steering from the conversational AI system at all, (b) the



(a) Steering across roles.

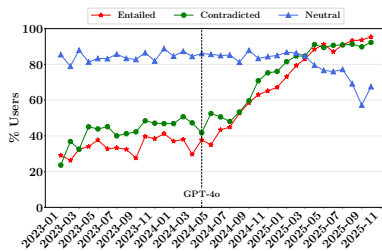


(b) System anthropomorphism in steered conversations.

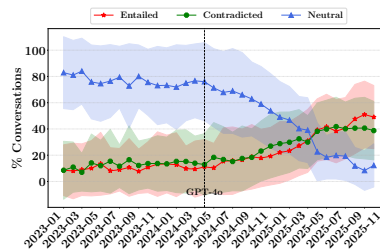


(c) Disclosure of personal data in steered conversations.

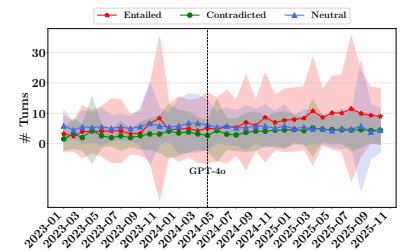
Figure 7: Steering versus different attributes.



(a) % of participants with at least one entailed, contradicted, or neutral conversation.



(b) Mean % of conversation turns per participant.



(c) Avg. depth of the conversation per participant.

Figure 8: Temporal evolution of steering in InVivoGPT.

prevalence of steering in conversation turns, and (c) changes in conversation depth over time. Figure 8 illustrates our results across these three perspectives. We observe that a growing share of participants experience at least one conversation in which the system attempts to steer, and critically, this share increases after the release of GPT-4o (Figure 8a). In particular, after GPT-4o, the proportion of participants with at least one entailed conversation steadily increases, converging with the proportion of participants with contradicted conversations by early 2025. This result indicates both the rise in conversational steering and that more participants are following the model’s suggestions.

These results are reinforced when looking at the prevalence of entailed, contradicted, and neutral conversation turns (Figure 8b) over time. The mean percentage of turns that are entailed exhibits a shift following GPT-4o. Before the release, only 11% of the conversations contained a successfully steered turn. After, GPT-4o this proportion increases to almost 50%, suggesting that the updated model initiates follow-up suggestions more frequently and participants are more likely to respond in-line with the model’s suggestions. Neutral conversations remain the majority throughout the period, but their dominance declines post-GPT-4o release. This pattern reflects a transition from a passive conversational model toward one that is more active, as it introduces continuations and follow-ups. Finally, we sought to determine whether steering might enhance user engagement by increasing the depth of conversations. We find that indeed this is the case (see Figure 8c): for conversations that do involve steering, depth increases especially after GPT-4o. This correlation suggests that successful steering fosters deeper and more extended interactions.

Implications. Our findings suggest that newer models are reshaping the conversational dynamic of human-AI interactions, as systems shift from reactive responders toward more proactive agents. As steering becomes more frequent and more accepted by users, we must consider how model-initiated suggestions may influence

user autonomy, especially when such behavioral shifts arise from backend updates that users may not notice. Altogether, there is a need for more transparency and guardrails to ensure that proactive models support, rather than inadvertently steer, users in possibly unintended directions.

8 Concluding discussion

In this work, we studied how human-AI interactions evolved over time by analyzing the InVivoGPT dataset with ChatGPT traces donated by 300 users. Our longitudinal analysis reveals clear shifts in *purpose*, *framing*, and *steering*. Participants expand the purposes for which they turn to ChatGPT, increasingly consulting it about a broader and more sensitive range of topics such as health and mental health. At the same time, the framing of interactions changes: while ChatGPT continues to serve as an assistant or advisor, its role as a companion grows sharply, spreading across more participants, occurring more frequently, and developing into deeper conversations. Mirroring this, personal data disclosures become widespread and more diverse, particularly in the companion mode. Finally, we observe a rise in conversational steering after the release of GPT-4o, with the system increasingly proposing directions that the participants choose to follow. Our findings have important implications for multiple stakeholders. For *users*, our results highlight both opportunities and risks of treating ChatGPT as more than a functional tool, underscoring the need to remain mindful of when and how sensitive information should be disclosed. For *AI designers*, the growing prevalence of companion-like conversations and model-initiated steering highlights the need for safeguards that support user autonomy, such as mechanisms for setting conversational boundaries. For *policymakers*, our work highlights the need for policies that address the relational and evolving nature of AI use to ensure user autonomy and privacy are protected as these systems become more embedded in everyday life.

References

- [1] Introducing chatgpt. <https://openai.com/index/chatgpt/>. Accessed: 08-05-2025.
- [2] Tawfiq Ammari, Meilun Chen, SM Zaman, and Kiran Garimella. How students (really) use chatgpt: Uncovering experiences among undergraduate students. *arXiv preprint arXiv:2505.24126*, 2025.
- [3] Drew Bent, Kunal Handa, Esin Durmus, Alex Tamkin, Miles McCain, Stuart Ritchie, Ryan Donegan, Jennifer Martinez, and Jason Jones]. Anthropic education report: How educators use claude, 2025.
- [4] Alexander Bick, Adam Blandin, and David J Deming. The rapid adoption of generative ai. Technical report, National Bureau of Economic Research, 2024.
- [5] Laura Boeschoten, Jef Ausloos, Judith E Möller, Theo Araujo, and Daniel L Oberski. A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research*, 4(2):388–423, 2022.
- [6] Timothy F Bresnahan and Manuel Trajtenberg. General purpose technologies 'engines of growth'? *Journal of econometrics*, 65(1):83–108, 1995.
- [7] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.
- [8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.
- [9] Avishek Choudhury and Hamid Shamszare. Investigating the impact of user trust on the adoption and use of chatgpt: survey analysis. *Journal of Medical Internet Research*, 25:e47184, 2023.
- [10] Yang Deng, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. Proactive Conversational AI: A Comprehensive Survey of Advancements and Opportunities. *ACM Transactions on Information Systems*, 43(3):1–45, 2025.
- [11] EU. General data protection regulation/gdpr, 2016.
- [12] AlMohtana Gasaymeh, Asma'a Abu Qbeita, Reham AlMohtadi, and Mohammad Beirut. Exploring education students' use of chatgpt for academic and personal purposes: insights from a developing country context. In *Frontiers in Education*, volume 10, page 1580310. Frontiers Media SA, 2025.
- [13] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.
- [14] Kunal Handa, Drew Bent, Alex Tamkin, Miles McCain, Esin Durmus, Michael Stern, Mike Schiraldi, Saffron Huang, Stuart Ritchie, Steven Syverud, Kanya Jagadish, Margaret Vo, Matt Bell, and Deep Ganguli. Anthropic education report: How university students use claude, 2025.
- [15] Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. AnnoLLM: Making large language models to be better crowdsourced annotators. In Yi Yang, Aida Davani, Avi Sil, and Anoop Kumar, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico, June 2024. Association for Computational Linguistics.
- [16] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.
- [17] Lujain Ibrahim, Canfer Akbulut, Rasmi Elasmr, Charvi Rastogi, Minsuk Kahng, Meredith Ringel Morris, Kevin R McKee, Verena Rieser, Murray Shanahan, and Laura Weidinger. Multi-turn evaluation of anthropomorphic behaviours in large language models. *arXiv preprint arXiv:2502.07077*, 2025.
- [18] Lujain Ibrahim, Canfer Akbulut, Rasmi Elasmr, Charvi Rastogi, Minsuk Kahng, Meredith Ringel Morris, Kevin R. McKee, Verena Rieser, Murray Shanahan, and Laura Weidinger. Multi-turn evaluation of anthropomorphic behaviours in large language models. *ArXiv*, abs/2502.07077, 2025.
- [19] Sai Keerthana Karnam, Abhisek Dash, Antariksh Das, Sepehr Mousavi, Stefan Bechtold, Krishna P. Gummadi, Animesh Mukherjee, Ingmar Weber, and Savvas Zannettou. Setting the course, but forgetting to steer: Analyzing compliance with gdpr's right of access to data by instagram, tiktok, and youtube. In *IEEE Symposium on Security and Privacy (SP)*, 2026.
- [20] Ranim Khojah, Mazen Mohamad, Philipp Leitner, and Francisco Gomes de Oliveira Neto. Beyond code generation: An observational study of chatgpt usage in software engineering practice. *Proceedings of the ACM on Software Engineering*, 1(FSE):1819–1840, 2024.
- [21] Kyung Mee Kim and Sook Hyun Kim. Experience of the use of ai conversational agents among low-income older adults living alone. *Sage Open*, 14(4):21582440241301022, 2024.
- [22] Nataliya Kosmyrna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 4, 2025.
- [23] Ruiyin Li, Peng Liang, Yifei Wang, Yangxiao Cai, Weisong Sun, and Zengyang Li. Unveiling the role of chatgpt in software development: Insights from developer-chatgpt interactions on github. *arXiv preprint arXiv:2505.03901*, 2025.
- [24] Yier Ling, Alex Kale, and Alex Imas. Underreporting of ai use: The role of social desirability bias. Available at SSRN 5464215, 2025.
- [25] Peidong Mei, Deborah N Brewis, Fortune Nwaiwu, Deshan Sumanathilaka, Fernando Alva-Manchego, and Joanna Demaree-Cotton. If chatgpt can do it, where is my creativity? generative ai boosts performance but diminishes experience in creative writing. *Computers in Human Behavior: Artificial Humans*, 4:100140, 2025.
- [26] Jason Phang, Michael Lampe, Lama Ahmad, Sandhini Agarwal, Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, et al. Investigating affective use and emotional well-being on chatgpt. *arXiv preprint arXiv:2504.03888*, 2025.
- [27] Prolific. Prolific. <https://prolific.co/>, 2025. Accessed: 2025-09-29.
- [28] Robert D Putnam. Bowling alone: America's declining social capital. In *The city reader*, pages 188–196. Routledge, 2015.
- [29] Tony Rousmaniere, Xu Li, Yimeng Zhang, and Siddharth Shah. Large language models as mental health resources: Patterns of use in the united states, 2025.
- [30] Koustuv Saha, Pranshu Gupta, Gloria Mark, Emre Kiciman, and Munmun De Choudhury. Observer effect in social media use. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.
- [31] Marita Skjuve, Petter Bae Brandtzaeg, and Asbjørn Følstad. Why do people use chatgpt? exploring user motivations for generative conversational ai. *First Monday*, 29(1), Jan. 2024.
- [32] Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, et al. Clio: Privacy-preserving insights into real-world ai use. *arXiv preprint arXiv:2412.13678*, 2024.
- [33] Kiran Tomlinson, Sonia Jaffe, Will Wang, Scott Counts, and Siddharth Suri. Working with ai: Measuring the occupational implications of generative ai. *arXiv preprint arXiv:2507.07935*, 2025.
- [34] Yuxuan Wan, Wenxuan Wang, Pinjia He, Jiazhen Gu, Haonan Bai, and Michael R Lyu. Biasasker: Measuring the bias in conversational ai system. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 515–527, 2023.
- [35] Miranda Wei, Madison Stamos, Sophie Veys, Nathan Reiting, Justin Goodman, Margot Herman, Dorota Filipczuk, Ben Weinschel, Michelle L Mazurek, and Blase Ur. What twitter knows: Characterizing ad targeting practices, user perceptions, and ad explanations through users' own twitter data. In *29th USENIX Security Symposium (USENIX Security 20)*, 2020.
- [36] Cai Yang, Sepehr Mousavi, Abhisek Dash, Krishna P Gummadi, and Ingmar Weber. Coupling gdpr data donation and crowdsourced user survey: A case study on tiktok addiction. In *Companion Publication of the 16th ACM Web Science Conference*, 2024.
- [37] Savvas Zannettou, Olivia Nemes-Nemeth, Oshrat Ayalon, Angelica Goetzen, Krishna P Gummadi, Elissa M Redmiles, and Franziska Roesner. Analyzing user engagement with tiktok's short format video recommendations using data donations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024.
- [38] Peter Zhang. Taking advice from chatgpt. *arXiv preprint arXiv:2305.11888*, 2023.
- [39] Yutong Zhang, Dora Zhao, Jeffrey T. Hancock, Robert Kraut, and Diyi Yang. The rise of ai companions: How human-chatbot relationships influence well-being, 2025.
- [40] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild, 2024.

A Survey and participant demographics

Our study involved data donation of ChatGPT followed by a survey, which included questions to understand how frequently participants use chatbots and the primary purposes for which they rely on ChatGPT, as well as demographic details about the users. Survey responses indicated that ChatGPT was most frequently used for *learning and education* (71%), followed by *daily life and decision support* (67%), *entertainment and casual conversations* (46%), and *medical advice* (45%). When asked about their preferred method for finding information and a majority of participants (60%) reported favoring AI tools, while the remainder (40%) indicated a preference for traditional search engines. The participants were compensated with \$2 for completing the survey, \$5 for donating conversations, \$3 for the other files, and \$1 each for shared conversations and message feedback. Table 1 reports the demographic characteristics and subscription status of the 300 participants in our study. In terms

Attribute	Type	Count	Percentage
Gender	Female	109	36.3
	Male	183	61.0
	Prefer not to say	5	1.7
	Other	3	1.0
Age	18–24	57	19.0
	25–34	131	43.7
	35–44	61	20.3
	45–64	47	15.7
	65+	4	1.3
Country	United States of America	149	49.7
	Germany	50	16.7
	Italy	47	15.7
	France	24	8.0
	Spain	30	10.0
GPT Plus	Yes	91	30.3
	No	209	69.7

Table 1: Distribution of participants based on their self-reported gender, age, country of residence, and GPT Plus subscription status.

Statistic	WILDCHAT 4.8M	WILDCHAT power users	INVIVO GPT
#Users	2,641,054	300	300
#Conversations	4,743,336	125,495	138,231
#Turns	7,847,456	327,597	825,672
#Conversations/User	1.8 ± 247.9	418.3 ± 974.5	460.8 ± 531.1
#Turns/Conversation	1.7 ± 3.2	2.6 ± 6.2	6.0 ± 17.8
Active Time (days)	2.2 ± 26.9	333.4 ± 218.7	714.1 ± 292.7
Conversation Duration (h)	0.2 ± 3.5	0.4 ± 4.8	16.6 ± 198.7

Table 2: Comparison across the datasets.

of gender, the majority identified as male (61%), followed by female (36%), with five preferring not to disclose their gender and three participants selecting “other.” The sample is young, with the largest group aged 25–34 (43.7% participants), followed by 35–44 (20.3%), 18–24 (19%), 45–64 (15.7%), while only four participants were 65 or older. Participants were primarily based in the United States (49.7%), Germany (16.7%), Italy (15.7%), France (8%) and Spain (10%). Finally, 30.3% participants reported having GPT Plus subscription.

B INVIVO GPT vs. existing datasets

Datasets such as SHAREGPT [8] and WILDCHAT [40], have advanced research on conversational AI systems; however, they have key limitations for studies focusing on the evolution of interactions between humans and conversational AI systems. SHAREGPT consists of user-selected ChatGPT conversations collected via a Chrome extension, which means it captures isolated exchanges rather than the longitudinal use of ChatGPT. In contrast, the WILDCHAT dataset was collected through a third-party chatbot service that utilizes ChatGPT APIs in its backend. The service has been hosted for nearly two years (as of September 2025). While this dataset can theoretically be used for longitudinal analyses, it is subject to the observer effect [30], as users know in advance that their conversations will be shared, potentially altering their behavior. As a result, we argue that neither of these datasets is well-suited for studying how interactions evolve naturally over time in an ecologically valid setting. To demonstrate the differences, we compare WILDCHAT and INVIVO GPT across various dimensions.

Comparison between INVIVO GPT and WILDCHAT. Table 2 presents a comparison of our GDPR-based dataset with the WILDCHAT dataset. While WILDCHAT contains over 2.6M users and nearly 4.7M conversations, most users are extremely sparse in activity: around 99% have fewer than 10 conversations, and nearly 95% have only a single conversation. Moreover, approximately 95% of the users are active for less than six months. Only (837) users met the criteria which we have for the INVIVO GPT dataset, i.e., have at least 100 conversations and 90 days of activity. We call the users “power users.” To enable a fair comparison with INVIVO GPT, we randomly sample 300 power users (out of 837 users).

When compared to these power users, our INVIVO GPT dataset shows some notable differences. Although WILDCHAT power users have a comparable number of conversations (125K vs. 138K), the total number of turns is less than half (327K vs. 825K). This difference reflects the structure of conversations: our dataset has nearly twice as many turns per conversation on average (6 vs. 2.6). In addition, our participants engage in substantially longer conversations (16.6 hours vs. 0.4 hours on average) and sustain activity over longer periods (714 days vs. 333 days). Taken together, these statistics suggest that our dataset captures more continuous and in-depth interactions, making it especially well-suited for studying the longitudinal evolution of ChatGPT use.

C Validation

To validate the performance of GPT-4o in the tasks listed in Section 4, we extracted a random sample of 50 conversations (262 turns). These conversations were annotated by two authors of this work, with a third one solving the disagreements. For topics, annotations were done at the conversation level. For anthropomorphization, annotations were done at the message level (i.e., each message from humans and each response of ChatGPT individually). For the relationship, personal data and steering, annotations were conducted at the turn level. For each label, we relied on the labels of the first two annotators. In cases where their annotations differed, the third annotator’s decision was used to resolve the disagreement and determine the final label. We report the inter-annotator agreement score and the accuracy obtained in Table 3. Note that many of these annotations are highly subjective; in fact, agreement scores among annotators are close to the accuracy levels, indicating that human annotators find the annotations almost as hard as ChatGPT.

Remark. Note that GPT-4o failed to generate responses for conversations that exceeded the input length limit. Also, we encountered issues while parsing some of the outputs, as they were in invalid formats. In total, we had valid responses for 137,985 conversations. For our analysis, we considered data from January 2023 to October 2025, as other periods lacked a sufficient active users; hence, our analysis is based on nearly 135K conversations and 756K turns.

D Limitations

Our study has two main limitations. First, while our dataset is quite rich, it includes a relatively small number of users (300). However, the goal of this work was not to produce exact population-level estimates but to uncover trends and highlight emerging risks in the evolution of human–AI interactions. The insights we provide should therefore be read as indicative of broader dynamics rather than as

Task	Cohen's κ (%)	Accuracy (%)
Topics	71.3	70.0
Anthropomorphization (User)	85.2	82.0
Anthropomorphization (ChatGPT)	96.7	83.0
Relationship	78.0	72.0
Personal data	88.3	79.0
Steering of conversation	83.4	75.0

Table 3: Inter-annotator agreement scores (Cohen’s κ) and accuracy of the GPT-4o annotations for different labels.

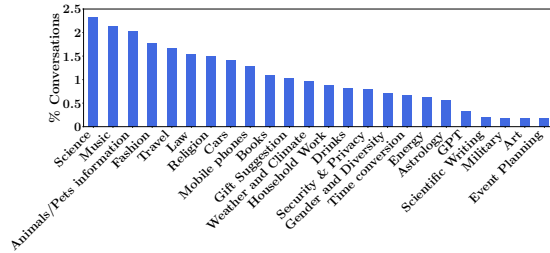


Figure 9: Distribution of topics in the InVivoGPT dataset.

precise measurements. Second, our annotations relied on GPT-4o to classify topics, anthropomorphization, relationships, personal data and steering. Although this approach enabled scalable and consistent analysis, the model is not perfect and may mislabel some cases. We mitigated this limitation by validating subsets of the annotations, but future work should strengthen this process through complementary annotation strategies.

E Conversational purpose

E.1 Topics in the conversation

Figure 9 shows the remaining topics in InVivoGPT conversations.

E.2 Temporal evolution of topics with models

We examined how topical needs of the users are evolving over time. Understanding this ongoing evolution is important because it reflects changing patterns of reliance on conversational AI systems. To understand this evolution, we analyze variations in topic distributions across different underlying models. Earlier interactions were dominated by text-davinci (a GPT-3.5 version), which remained the primary model used until May 2024, after which GPT-4o got widely adopted by users. To capture this shift, we compare the topical distributions associated with each model.

Figure 10 shows the distribution of topics in InVivoGPT dataset with respect to the underlying model. We observe that, relative to GPT-3.5, GPT-4o shows a noticeable decrease in *Programming* (-52%), *Math* (-36%), *Job Search* (-52%) topics, accompanied by an increase in *Health* (+33%), *Roleplay* (+186%) and *Mental Health* (+19%) topics. This shows that as the systems are evolving, participants are increasingly using them for more sensitive and high-stakes cases.

F Anthropomorphization

Variation with the topic. Figure 11 shows broadly similar levels of anthropomorphism observed across all the topics.

Variation with the relationships. Figure 12 shows more frequent anthropomorphization in companion interactions.

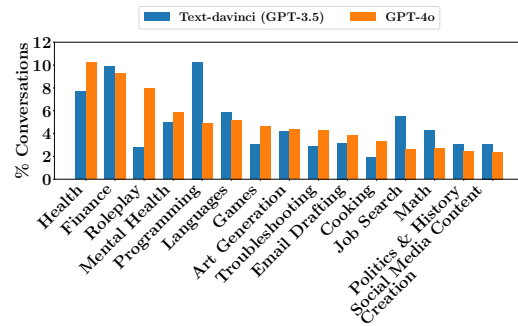
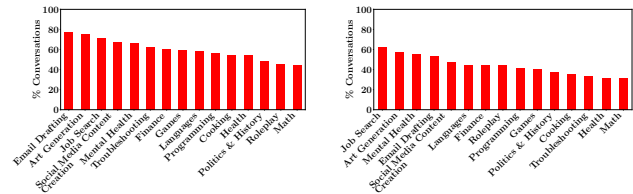


Figure 10: Distribution of topics across models.



(a) System-generated.

(b) User messages.

Figure 11: Variation of anthropomorphism across topics.

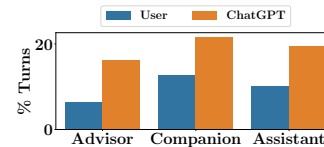


Figure 12: Anthropomorphism across relationships.

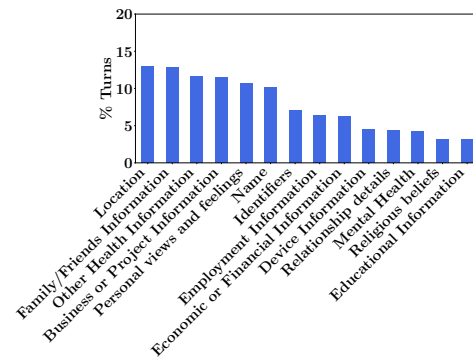


Figure 13: Types of personal data disclosed in InVivoGPT.

G Personal data disclosure

Figure 13 shows the distribution of personal data types disclosed by users in their conversations.

Variation with system-generated anthropomorphization. Users disclosed personal data more frequently when the system exhibits anthropomorphic behavior (26%) compared to when it did not (20%).