Adopt ≠ Adapt: Longitudinal Analyses of LLM Conversations in the Wild

Rebecca M. M. Hicke*
Cornell University
rmh327@cornell.edu

Kiran Tomlinson
Microsoft Research
kitomlinson@microsoft.com

Abstract

Although a growing body of research has begun to describe user–LLM interactions, the picture it paints is largely *static*; little is known about how individual users change their behavior over time. To address this gap, we analyze the conversational trajectories of ∼12,000 randomly sampled Microsoft Bing Copilot users and compare these with data from WildChat-4.8M. While the Copilot data contains significant population-level trends, we find that trends in individual user trajectories are much weaker; user habits prove to be overwhelmingly *sticky*. We also find stark differences between users of different activity levels: more active users have more successful conversations and use the LLM for more complex and professionally oriented tasks. Some user trends also appear in WildChat-4.8M, but we find evidence that this dataset is significantly skewed towards highly proficient "power" users. Ultimately, our results suggest that existing user behavior is difficult to change and demonstrate the extent of user heterogeneity. Our comparison between datasets highlights that WildChat does not represent typical user-AI interactions, an important caveat for downstream uses of the data.

"The end of all our exploring will be to arrive where we started and know the place for the first time."

– T. S. Eliot

1 Introduction

A significant body of research has recently emerged studying multiple dimensions of user–LLM interactions, ranging from task complexity (Suri et al., 2024) to user intents (Handa et al., 2025; Chatterji et al., 2025). This literature paints a clear picture of overall LLM usage; it helps system developers improve user experiences and provides guidance to users on how to productively interact with LLMs. However, this picture of user–LLM interactions is *static* — little is known about how individual users adapt over time. Existing temporal analyses tend to only show population-level trends (Chatterji et al., 2025), which do not necessarily reflect behavioral changes of individuals such as learning and adaptation.

We address this gap by analyzing six months of conversational trajectories from ∼12,000 randomly sampled users of Microsoft Bing Copilot active during 2024, alongside population-level daily conversation samples (∼1M conversations). We stratify the sampled users by activity level (the number of days they use Bing Copilot during our study period) and then explore five dimensions of their LLM usage: usage intensity, linguistic complexity, task completion, task intent, and conversational domain. We characterize population-level trends over the study period, individual user evolution, and variation across users of different activity levels.

We supplement this analysis with an examination of the public WildChat-4.8M dataset (Zhao et al., 2024; Deng et al., 2024). For this data, we use hashed IP addresses as a proxy for individual users and then replicate the analyses performed on the Bing Copilot data. Using a public dataset alongside private data provides additional transparency into our methods. It also allows us to contrast users of a mainstream consumer LLM with the public WildChat data, which has been used extensively to understand LLM user behavior (Mireshghallah et al., 2024), fine-tune models (Shi et al., 2025), and derive benchmarks (Lin et al., 2025), among many other downstream applications.

Motivating our stratified sample, we find that frequent, or "power," users of Bing Copilot differ significantly from less frequent users along all analysis dimensions: they are more active, write more linguistically complex messages, are more likely to have completed conversations, and use the LLM for

*Work performed during a Microsoft Research internship.

more complex and professionally-oriented tasks. This could mean that individual users learn to change their behavior over time. However, our longitudinal analysis largely refutes this hypothesis; someone who will become a high-activity user interacts very differently with the LLM *even early on in their trajectory*. In fact, we find that individual users change their behavior very little: user habits are *sticky* and resistant to change.

In contrast, we find that Bing Copilot users as a group *do* change over time, often towards behavior characteristic of high activity-level users. Again, longitudinal user data demonstrates that these changes are not due to learning at the individual level, but are instead driven by new users who differ significantly from earlier adopters. Finally, although our longitudinal analysis largely surfaces the stickiness of user habits, we do find some small scale user-level behavioral changes.

Some of these patterns also appear in the WildChat data: more active users complete more tasks and there are some similar population-level trends. However, the difference between low- and high-activity users in WildChat is far less pronounced. We find that the user population represented in WildChat is extremely unusual, with an over-representation of power users (likely driven by the HuggingFace front-end) and a large fraction of API-like usage rather than natural conversations.

These findings have several important implications. First, the stickiness of habits means that users may not discover more useful and successful LLM tasks through natural exploration, indicating a need for proactive interventions. Second, population-level trends toward the behavior of high-activity users suggest that learning occurs at an aggregate level, even if user rigidity inhibits it at an individual level. Third, the stark differences between users of different activity levels highlight the importance of recognizing individual heterogeneity in future work. Finally, the differences between the Bing Copilot and WildChat-4.8M user pools suggest that downstream uses of WildChat data should be cautiously implemented; WildChat-derived artifacts may not generalize to the average LLM chat user.

2 Related Work

A great deal of existing literature investigates how people use LLMs in the wild, focusing on classifying user intents, conversational domains, and linguistic features of messages (Handa et al., 2025; Tomlinson et al., 2025; Chatterji et al., 2025; Costa-Gomes et al., 2026; Ouyang et al., 2023; Trippas et al., 2024; Zhao et al., 2024; Shah et al., 2025; Suri et al., 2024; Tamkin et al., 2024). Some analyses even explore population-level temporal trends (Chatterji et al., 2025) or differences between users of different tenure (days since signup) (Massenkoff et al., 2026). Unlike these works, a key portion of our analysis is at the user level (aggregated for privacy), allowing us to identify how individual behavior evolves over time.

Most existing user-level analyses follow a small set of users over time to track how their attitudes towards and usage of AI change. Skjuve et al. (2022, 2023) interview Replika users over 12 weeks to investigate how their social relationships with the chatbot evolve. Long et al. (2024) conduct a three-week 10-session lab experiment where PhD students learned to use an LLM-based system for science communication. Chandra et al. (2025) survey Prolific users over five weeks who were asked to engage daily with a commercial LLM on social and emotional topics. However, these studies do not examine the interactions' linguistic content.

One of the only temporal analyses that looks at the texts of user–LLM conversations is a study of 36 undergraduate students' full ChatGPT conversation histories by Ammari et al. (2025). They find that students who use ChatGPT for writing code, emails, job applications, and answering multiple choice questions are more likely to continue using the system, while inconsistent answers from the LLM and frustration expressed by the students predict discontinued use. Like Ammari et al. (2025), we perform longitudinal analyses of user–LLM conversations, but employ a representative sample of users rather than a small number of students.

3 Data

3.1 Bing Copilot

Our primary dataset consists of all conversations with Microsoft Bing Copilot in an English-language interface between January 1, 2024 and September 30, 2024.[1] Our use of human data was approved by Microsoft Research Ethics, Privacy, and Compliance review. Personally identifying information (e.g., email addresses and financial details) was automatically scrubbed from all conversations before analysis and unique users are identified only with anonymized ids. Additionally, all metrics

[1] Likely bot content is removed.

reported are aggregates over more than 200 users. All data was stored and classified in secure computing environments. See Section B for additional discussion of data ethics and privacy.

Our goal is to examine the entire trajectory of a user's interactions with Bing Copilot, from their first conversation onwards. However, the dataset does not label whether a conversation is a user's first. Instead, we exclude all users who have conversations in the first three months of the data (before April 1, 2024). We hypothesize that most users who did not interact with Bing Copilot for three months before appearing in the dataset will be new.

We then take two subsamples of the filtered dataset. The first is meant to represent overall population-level user behavior during the studied time period (April 1 to September 30, 2024). To create this subsample, we randomly select $\sim$1,000 conversations from each day of the specified time period, producing a dataset of 796,838 LLM and user messages from 175,061 conversations. This we refer to as the *population dataset*.

The second subsample consists of full user trajectories from a sample of users stratified by activity level, or the number of days on which the user had a conversation with Bing Copilot. Activity level follows a power-law distribution with exponential cutoff (Clauset et al., 2009), with more users active on fewer days and a heavy tail. We thus categorize users into three activity-level classes: users active 1–10 days (low), active 11–25 days (middle), and active 26+ days (high). In order to study data from users of all activity levels, we randomly sample $\sim$250 users active $1, 2, ..., 49, 50+$ days from the filtered dataset and include all of their conversations during the study period in our sample. The resulting dataset contains 812,650 conversations from 11,905 users consisting cumulatively of 4,879,568 messages. We call this the *user dataset*.

3.2 WildChat-4.8M

Since the Bing Copilot data cannot be released publicly, we replicate our analyses on the public WildChat-4.8M dataset (Zhao et al., 2024; Deng et al., 2024).[2] WildChat was gathered by providing users with free access to GPT models through HuggingFace Spaces applications in exchange for their agreement to share data. In addition to full conversation texts, WildChat provides a hash of each user's IP address and the user's country and state. We use these fields to link conversations that may involve the same user. To remove data likely coming from shared networks, we filter out hashed IPs associated with more than three unique countries, states, or languages, and hashed IPs with more than 161 conversations (the largest number of conversations associated with at least 10 hashed IPs). This removes 3,099 hashed IPs and 677k conversations.

After filtering, we are left with 2,522,330 conversations from 1,830,631 hashed IPs that occurred between March 9, 2023 and July 31, 2025. For simplicity, we refer to hashed IPs as "users" for the remainder of the paper. However, we note that this approach of linking conversations to users is far less reliable than in the Bing Copilot data, where users are identified by a combination of account logins and cookies. We again split users into low, middle, and high activity groups in the same way as in Bing Copilot (active 1–10, 11–25, 26+ days). Note that higher active days groups in WildChat have far fewer users, since they are not derived from a stratified sample as in Bing Copilot.

The WildChat-4.8M dataset displays significant irregularities later in the data period, with huge spikes in the number of daily conversations and unique IPs (Appendix Figure 14). We find this is driven by API-like usage, i.e., a large number of prompts with identical templates performing tasks like translation and named entity recognition (Appendix Figure 13 and Table 5). To avoid the bulk of this non-conversational activity while keeping the dataset close to intact, the paper only considers conversations in WildChat-4.8M occurring before September 2024. Full versions of all figures with data after this cutoff can be found in Section E.

4 Methods

4.1 Syntactic Features

We calculate several syntactic properties of the conversations: the number of user messages sent per conversation, the average user sentence length per conversation,[3] and the average number of conversations held by each user per day. For Bing Copilot data, we report metrics relative to the first day in the sample or, when comparing activity groups, to users active one day.

[2] Available at https://huggingface.co/datasets/allenai/WildChat-4.8M under an ODC-By license.

[3] A proxy for linguistic complexity (Dale and Tyler, 1934; Gray and Leary, 1935; Flesch, 1948); calculated with `spaCy`.

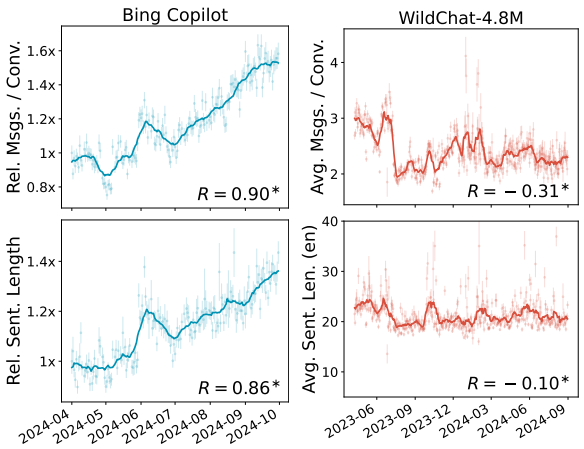


Figure 1: **Population-level user activity and linguistic complexity rise over time in Bing Copilot, but not WildChat.** Solid lines are 14-day averages, while points are daily metrics with standard error. Metrics in Bing Copilot are reported relative to the first day in the sample. Asterisks here and in all other plots indicate Pearson's R is significantly different from 0 ($p < 0.05$).

4.2 Semantic Features

We also evaluate several high-level semantic properties of each conversation using LLM classifiers: user intent, conversational domain, and task completion. User intent is defined as the user's purpose for conversing with Bing Copilot and falls into one of nine categories, which include summarization and information lookup. Conversational domain is defined as the general subject area a conversation falls under and is labeled as one of thirty categories, including law and politics and entertainment. See Section A for prompts, including the full list of intent and domain labels and their definitions.

Task completion describes whether the primary task the user intends to accomplish during the conversation is completed by its end (e.g., if the model is prompted to describe photosynthesis, it does so). In human annotation of WildChat-4.8M conversations, we found that instances labeled as "incomplete" were largely driven by cases where the user had no concrete task: e.g., the user said only "Hi" (see Section C for annotation details). We therefore interpret completion with caution and view it as a proxy for task concreteness. We use GPT-4o-mini with temperature 0 and chain-of-thought prompting (Wei et al., 2022) for all semantic classification tasks.

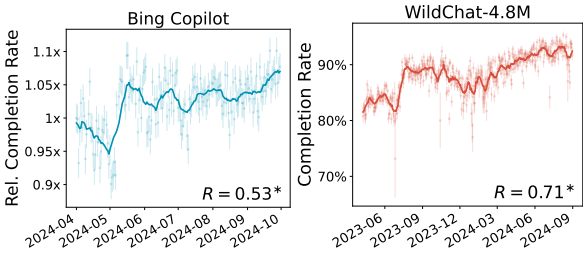


Figure 2: **Population-level task completion increases over time.** Solid lines are 14-day averages, while points are daily metrics with standard error. For Bing Copilot, completion rate is reported relative to the first day in the sample.

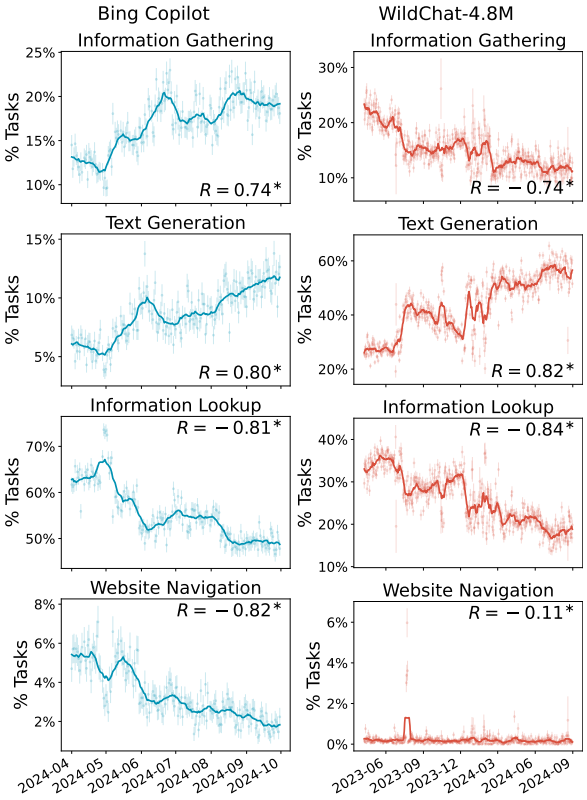


Figure 3: **For Bing Copilot, open-ended tasks rise in population-level popularity over the study period while some simpler tasks fall; trends only occasionally hold in WildChat.** Solid lines are 14-day averages, while points are daily metrics with standard error. Trajectories for the remaining intents can be found in Section D.1, Figure 9.

5 Population-Level Trends

We first examine overall trends in the behavior of both user groups during the relevant study periods. In the Bing Copilot population dataset, the average number of user messages per conversation and user message complexity both increase about 1.5× throughout the study period (Figure 1). The WildChat data does not follow the same trends;

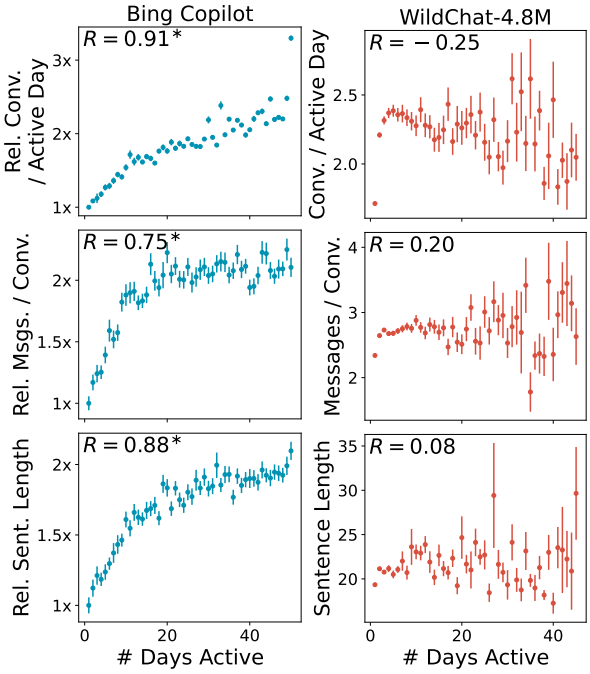


Figure 4: **More active Bing Copilot users by # days are more active by two additional metrics and write more linguistically complex messages; the same trends do not hold for Wildchat users.** Bing Copilot metrics are reported relative to users active one day. Error bars represent standard error. WildChat likely shows weaker relationships and more noise at high activity levels due to a lack of stratified sampling.

average user messages per conversation and average length of user sentences in English both drop slightly. Task completion increases significantly in both user groups over time (Figure 2). However, these changes are small and the absolute completion rate is high throughout the WildChat data.

The frequency with which both user populations pursue different task intents also changes over time. Throughout the study period, Bing Copilot users increasingly attempt more complex tasks like information gathering (on average 12.7% of tasks during the first ten days to 19.2% during the last ten days) and text generation (6.1% to 11.6%) and engage less frequently with simpler tasks like information lookup[4] (62.4% to 49.2%) and website navigation (5.3% to 1.7%) (Figure 3). Some similar trends are found in the WildChat data: text generation gains popularity (22.8% to 55.3%) whereas information lookup loses it (32.2% to 16.6%). However, information gathering actually becomes *less* common in the WildChat data over time (22.7% to 11.3%) and website navigation is almost never attempted.

[4] *Information lookup* describes simple factual queries, while *information gathering* describes more complex research.

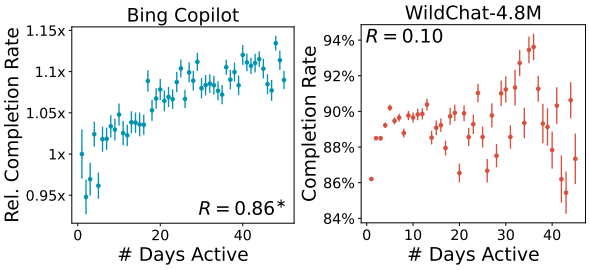


Figure 5: **More active Bing Copilot users complete more tasks; the same does not hold for WildChat users.** Bing Copilot metrics are reported relative to users active one day. Error bars represent standard error.

Overall, the population-level frequency of intents varies considerably between the datasets.

6 Differences by Activity Level

Next, we examine whether differences exist between users of varying activity levels. To do this, we first average feature values for each user on every day they are active. Then, we average over all days a user is active. Finally, we average all users in a single activity level group (defined by the number of days the user is active). Thus, each day a user is active is weighted equally, although the weight of individual conversations may vary based on a user's activity level per day.

We find clear differences in the behavior of Bing Copilot users of different activity levels. However, similar differences largely do not exist between WildChat users. Bing Copilot users who are active on more days also hold more conversations per day, send more messages per conversation, and write more linguistically complex messages (Figure 4). None of these trends are found in the WildChat data, although a jump in all three metrics occurs between users who are active on only one day and those who are active on $\geq$2 days.

More active Bing Copilot users also complete more tasks (Figure 5). This pattern largely holds even when broken down by task intent; high activity Bing Copilot users have significantly more completed tasks than low activity users of all intents types but website navigation (Table 1). While the differences are rarely significant, the same pattern often holds for middle activity users. Interestingly, the tasks for which the differences in completion rates between users of different activity levels are largest are often more complex.

In contrast, task completion only increases with activity level for WildChat users active $\leq$5 days

Intent	Bing Copilot		WildChat-4.8M		
	Middle (pp)	**High** (pp)	**Low** (%)	**Middle** (pp)	**High** (pp)
Summarization	+15.7	**+19.3**	91.8	+1.2	**+3.2**
Translation or Conversion	+7.7	**+15.1**	89.8	**+2.9**	**+3.0**
Open-Ended Discovery	+6.2	**+12.8**	70.0	−2.8	+0.0
Analysis	+2.8	**+7.9**	86.2	−0.3	+0.9
Information Gathering	**+5.9**	**+7.5**	86.1	**+0.4**	+1.7
Text Generation	+2.1	**+4.6**	94.4	**−1.8**	**−2.2**
Information Lookup	**+2.4**	**+3.6**	85.8	**+1.2**	**+1.4**
Image Creation	+0.0	**+0.3**	83.4	**+3.6**	**+0.5**
Website Navigation	−1.2	−4.8	62.4	**+6.6**	+15.2

Table 1: **More active users have higher completion rates for most intents, particularly in Bing Copilot.** The table shows the percentage point (pp) difference in completion rates per intent between middle/high and low activity groups. Baseline low-activity completion rates are also reported for WildChat. Bold numbers indicate a significant two-proportion Z-test vs the low activity group ($p < 0.05$ after Bonferroni correction within each dataset).

Domain	Bing Copilot			WildChat-4.8M		
	Low (%)	**Middle** (pp)	**High** (pp)	**Low** (%)	**Middle** (pp)	**High** (pp)
Programming and Scripting	5.1	**+2.6**	**+3.8**	14.5	**+1.0**	**−1.4**
Creative Writing and Editing	3.5	**+2.0**	**+2.9**	21.6	**−2.6**	**−6.7**
Professional Writing and Editing	1.0	**+0.7**	**+1.4**	1.0	**+0.3**	**+1.7**
Entertainment	4.8	**−1.6**	−0.7	2.8	**+0.5**	−0.6
Shopping and eCommerce	1.7	**−0.6**	**−0.8**	0.6	**+0.5**	**+0.7**
Small Talk and Chatbot	2.4	**−0.4**	**−0.9**	2.4	**+0.4**	**−0.2**
Travel and Tourism	3.2	**−1.7**	**−2.0**	0.6	**+0.0**	+0.0

Table 2: **More active Bing Copilot users are more likely to converse about professional topics and less likely to converse about casual topics; professional topics are more popular for all WildChat activity levels..** The table shows the difference in conversation share belonging to a subset of domains between middle/high and low activity users, measured in percentage points (pp) over low %. Bold numbers indicate a significant two-proportion Z-test vs the low activity group ($p < 0.05$ after Bonferroni correction within each dataset). An expanded table with all domains can be found in Section D.2, Table 4.

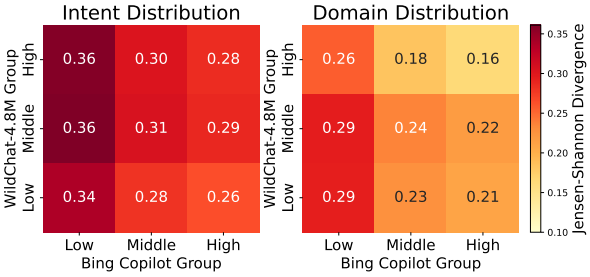


Figure 6: **WildChat usage more closely resembles high-activity Bing Copilot users.** The heatmaps show the Jensen–Shannon divergence (smaller = more similar) between the intent (left) and domain (right) distributions of Bing Copilot and WildChat users in each activity group.

(Figure 5). For some intents, we do see that completion increases with user activity, but these changes are generally small and there is no association with task complexity (Table 1). In fact, website navigation, the only intent for which higher activity Bing Copilot users are not more successful, is the intent for which this trend is strongest in WildChat users. For all other intents, ≥70% of tasks are completed even for low activity users.

Additionally, high activity Bing Copilot users are significantly more likely to have conversations that fall into professional or creative domains (e.g. professional writing & editing, programming & scripting, creative writing & editing) and are less likely to engage in more entertainment-oriented tasks (e.g. entertainment, travel & tourism, shopping & eCommerce, small talk & chatbot) (Table 2). In WildChat, even low-activity users have high rates of programming and creative writing and lower rates of shopping, travel, and entertainment. WildChat users overall most closely resemble *high* activity

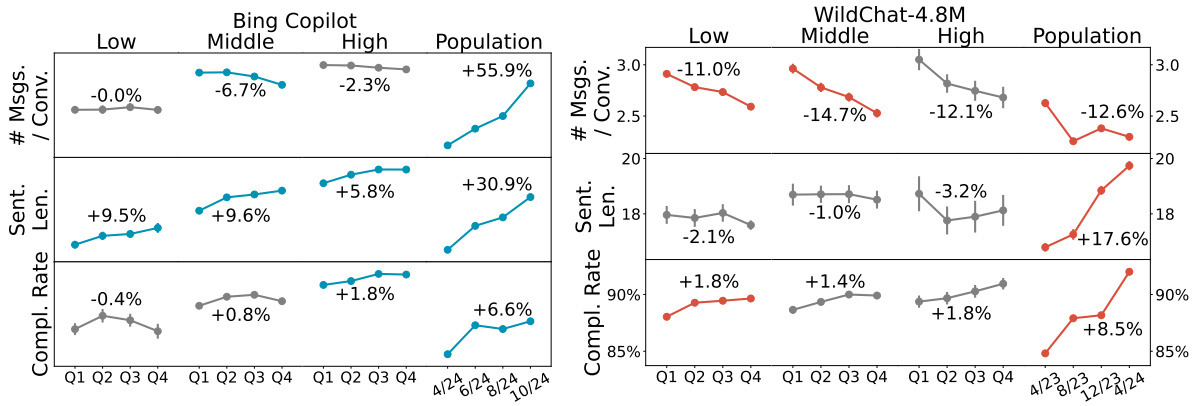


Figure 7: **Over user lifetimes, changes in activity, linguistic complexity, and completion are usually smaller than at the population level.** Average feature values during each quarter of user trajectories (e.g., Q1 = first quarter of days active), stratified by activity level. Users active for fewer than four days are dropped so quarters are meaningful. Population values are plotted temporally. Error bars represent standard error. Markers are colored rather than gray if a paired t-test for difference in means between the first and last quarter is significant ($p < 0.05$) after Bonferroni correction within each dataset.

Bing Copilot users. We confirm this by examining the Jensen–Shannon (J–S) divergence between the task intent and conversational domain distributions of low, middle, and high activity WildChat and Bing Copilot users (Figure 6). Overall, the J–S divergence between the WildChat intent and domain distributions and low-activity Bing Copilot users is 0.34 and 0.28 (resp.), but it is only 0.26 and 0.19 with high-activity Bing Copilot users.

7 Changes Over User Trajectories

Having established that significant differences exist between Bing Copilot users of different activity levels, we now probe to see whether these differences are inherent (present from the beginning of a user's conversational trajectory) or learned (developed over the user's trajectory.

To study user trajectories, we first index every day each user is active; the first day they appear in the dataset is labeled 0, then 1, and so on. We filter out all days with an index higher than 49 for the Bing Copilot data and 44 for the WildChat data. Then, we divide each user trajectory into quarters and find the mean feature values for each quarter (labeled Q1–Q4). Next we split all the users into the activity-level classes: users active 1–10 days (low), users active 11–25 days (middle), and users active 26+ days (high). For each activity-level class, we run a paired t-test for difference in means comparing the mean feature values from the first and last quarters of the trajectory of each user. For comparison, we also visualize population-level trends, splitting the study period in four and plotting the average feature values from the population dataset for each quarter of the study period.

Perhaps surprisingly, we find that changes over user trajectories are quite small. However, the average number of messages sent per conversation by middle activity Bing Copilot users and low and middle activity WildChat users decreases significantly over time (Figure 7). This decrease is mirrored in the population-level WildChat data, but is the opposite of the Bing Copilot population trend.

The average linguistic complexity of user messages increases significantly over time for both the WildChat and Bing Copilot populations; however, this trend only appears on the user-level in Bing Copilot and, even then, the increase for each activity group is much smaller than that of the population overall. Similarly, while completion increases over time in both datasets, these changes are much smaller in user trajectories than in the population at large and rarely significant (Figure 7).

There are also few consistent changes in the task intents users pursue throughout their conversational trajectories (Section D.3, Figure 10). In the Bing Copilot data, some significant changes in intent pursuit align with population-level trends: middle and high activity users decrease their proportion of website navigation tasks and high activity users increase their pursuit of text generation. However, some user-level changes reverse population-level patterns: middle and high activity users decrease their pursuit of image creation and high activity users decrease their proportion of open-ended discovery. Notably, no significant changes in intent

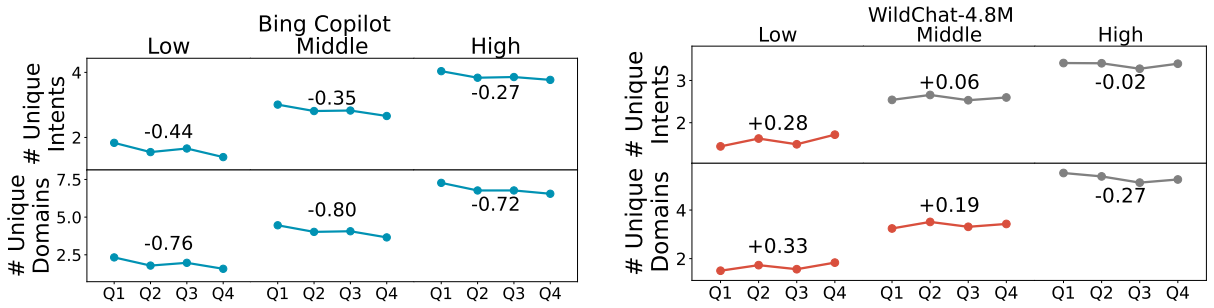


Figure 8: **Bing Copilot users shift only very slightly from exploration to exploitation over their lifetimes; lower activity WildChat users do the reverse.** Average # unique intents (left) and domains (right) during each quarter of user trajectories. Error bars represent standard error (smaller than the markers). Markers are colored rather than gray if a paired t-test for difference in means between the first and last quarter is significant ($p < 0.05$) after Bonferroni correction within each dataset.

pursuit occur over the lifetimes of low activity users and even the significant changes in higher activity user behavior are usually much smaller in magnitude than the population level changes.

The significant user-level changes in intent pursuit found in the Bing Copilot data only sporadically align with changes in WildChat. Again, there is a significant increase in the pursuit of text generation, but only among low and middle activity users, and a significant decrease in the proportion of open-ended discovery tasks by low and high activity users. However, while there is no significant change in information gathering for Bing Copilot users, its pursuit decreases for low and middle activity WildChat users. Again, the changes over WildChat user trajectories are usually much smaller in magnitude than the population level changes. In both datasets, users change their behavior comparatively little over time, and changes do not align closely enough over the datasets to suggest general trends in LLM user behavior.

As with intent, there are few significant changes in the conversational domains users from both datasets pursue throughout their interactions with the chatbots (Section D.4, Figures 11 and 12). These changes are also usually small and rarely align across the two datasets. The only consistent pattern is that high activity Bing Copilot users and low and high activity WildChat users both significantly decrease small talk over time, although the change is very small as the domain is rare.

Finally, we explore whether the extent to which users explore new task intents and conversational domains or exploit existing techniques shifts over conversational trajectories. We measure this by counting the number of unique intents and domains that each user pursues in the each quarter of their trajectories. Then, we separate users by activity-level class and apply paired t-tests for difference in means to compare the number of unique intents and domains explored in their first and last quarters.

For all activity-level classes of Bing Copilot users, we find there is a significant decrease in the number of unique intents and domains pursued over time (Figure 8); these users undergo a slight shift from exploration to exploitation, but once more the change is very small. Contrastingly, low and middle activity WildChat users shift very slightly from exploitation to exploration over time while high activity users have no significant change.

8 Conclusion

There are many enticing hypotheses for how LLM user behavior may change over time; however, we find that only very minor changes occur throughout users' interactions with Bing Copilot and the WildChat interface, as measured by intent, domain, completion, and syntactic complexity features. This holds true even for high-activity users who appear in our dataset on at least 26 days, and thus have held at least 26 conversations with the chatbot.

The changes that do occur over Bing Copilot user trajectories are dwarfed in comparison to differences between less and more active users. From their first conversations with Bing Copilot, more active users complete more tasks, pursue different intents, and write more linguistically complex prompts. Thus, differences between users appear inherent. In the future, in order to help users adapt to changing technology and discover its most useful applications, it will thus be important to proactively provide them with tools for breaking out of habits.

Finally, we find that Bing Copilot users differ from those in the WildChat dataset in many salient ways. WildChat users most closely resemble the most active Bing Copilot users, but differ even from this group in how their behavior changes over time. Additionally, we highlight that WildChat contains a large fraction of API-like non-conversational usage. As a result, we strongly encourage careful consideration of these differences before using WildChat in downstream applications, particularly when trying to mimic average chatbot user behavior.

Limitations

There are a few limitations we wish to highlight. First, this study only includes data from Bing Copilot taken between January 1 and September 30, 2024 and from WildChat-4.8M (focusing on data before September 2024 due to spikes in API-like usage later in the data). This usage may differ significantly from that on different AI platforms during different periods. In addition, our measures of success and complexity do not capture the full spectrum of user satisfaction and task sophistication. Our analysis of user trajectories also focuses on patterns shared across users (e.g., whether individuals, on average, do more of a certain intent as they use the LLM more); it is possible that different users each evolve in different directions, in ways that do not show up in overall averages. Finally, our dataset has an arbitrary end date, meaning we do not capture the entirety of all user trajectories. We have no definitive way of knowing when a user's *last* interaction with the chatbot is; some changes in user behavior may occur after our dataset ends.

Regarding the WildChat data, our usage of hashed IP addresses to link conversations that may belong to the same user is imperfect, likely missing conversations from the same person when devices or networks change and possibly conflating conversations from different people sharing an IP address in ways that our filter did not exclude. It is possible that the weaker distinction between different activity level groups in WildChat is in part driven by the weaker user–conversation mapping than we have in the Bing Copilot data. We also chose not to remove "templated" API-like conversation from WildChat, opting only to use the date cutoff to remove the most-affected time period. As a result, some fraction of the conversations in our WildChat analysis are API-like, diluting the signal from authentic human-at-the-keyboard conversations.

Acknowledgments

This study was approved by Microsoft Research Privacy, Ethics, and Compliance review (#8986). We thank Siddharth Suri, Nirupama Chandrasekaran, Jennifer Neville, and the MSR AI Interaction and Learning group for helpful discussions and feedback.

References

Tawfiq Ammari, Meilun Chen, S M Mehedi Zaman, and Kiran Garimella. 2025. How students (really) use ChatGPT: Uncovering experiences among undergraduate students. *Preprint*, arXiv:2505.24126.

Mohit Chandra, Javier Hernandez, Gonzalo Ramos, Mahsa Ershadi, Ananya Bhattacharjee, Judith Amores, Ebele Okoli, Ann Paradiso, Shahed Warreth, and Jina Suh. 2025. Longitudinal study on social and emotional use of ai conversational agent. *Preprint*, arXiv:2504.14112.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use ChatGPT. Working Paper 34255, National Bureau of Economic Research.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Beatriz Costa-Gomes, Pavel Tolmachev, Eloise Taysom, Viknesh Sounderajah, Hannah Richardson, Philipp Schoenegger, Xiaoxuan Liu, Matthew M. Nour, Seth Spielman, Samuel F. Way, Yash Shah, Michael Bhaskar, Harsha Nori, Christopher Kelly, Peter Hames, Bay Gross, Mustafa Suleyman, and Dominic King. 2026. Public use of a generalist llm chatbot for health queries. *Nature Health*.

Edgar Dale and Ralph W Tyler. 1934. A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. *The Library Quarterly*, 4(3):384–412.

Yuntian Deng, Wenting Zhao, Jack Hessel, Xiang Ren, Claire Cardie, and Yejin Choi. 2024. WildVis: Open source visualizer for million-scale chat logs in the wild. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221.

William Scott Gray and Bernice Elizabeth Leary. 1935. *What makes a book readable*. University Chicago Press.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller,

Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. 2025. Which economic tasks are performed with AI? evidence from millions of Claude conversations. *Preprint*, arXiv:2503.04761.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2025. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. In *The Thirteenth International Conference on Learning Representations*.

Tao Long, Katy Ilonka Gero, and Lydia B Chilton. 2024. Not just novelty: a longitudinal study on utility and customization of an ai workflow. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pages 782–803.

Maxim Massenkoff, Eva Lyubich, Peter McCrory, Ruth Appel, and Ryan Heller. 2026. Anthropic economic index report: Learning curves.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. In *First Conference on Language Modeling*.

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The shifted and the overlooked: A task-oriented investigation of user-GPT interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2375–2393.

Chirag Shah, Ryen White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Nagu Rangan, and 1 others. 2025. Using large language models to generate, validate, and apply user intent taxonomies. *ACM Transactions on the Web*, 19(3):1–29.

Taiwei Shi, Zhuoer Wang, Longqi Yang, Ying-Chun Lin, Zexue He, Mengting Wan, Pei Zhou, Sujay Jauhar, Sihao Chen, Shan Xia, Hongfei Zhang, Jieyu Zhao, Xiaofeng Xu, Xia Song, and Jennifer Neville. 2025. Wildfeedback: Aligning llms with in-situ user interactions and feedback. *Preprint*, arXiv:2408.15549.

Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzæg. 2023. A longitudinal study of self-disclosure in human–chatbot relationships. *Interacting with Computers*, 35(1):24–39.

Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies*, 168:102903.

Siddharth Suri, Scott Counts, Leijie Wang, Chacha Chen, Mengting Wan, Tara Safavi, Jennifer Neville, Chirag Shah, Ryen W. White, Reid Andersen, Georg Buscher, Sathish Manivannan, Nagu Rangan, and Longqi Yang. 2024. The use of generative search engines for knowledge work and complex tasks. *Preprint*, arXiv:2404.04268.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, Michael Stern, Brian Clarke, Landon Goldberg, Theodore R. Sumers, Jared Mueller, William McEachen, Wes Mitchell, Shan Carter, Jack Clark, and 2 others. 2024. Clio: Privacy-preserving insights into real-world AI use. *Preprint*, arXiv:2412.13678.

Kiran Tomlinson, Sonia Jaffe, Will Wang, Scott Counts, and Siddharth Suri. 2025. Working with AI: Measuring the applicability of generative AI to occupations. *Preprint*, arXiv:2507.07935.

Johanne R Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. 2024. What do users really ask large language models? an initial log analysis of Google Bard interactions in the wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2703–2707.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

A Prompts

In the following prompts, `{content}` was replaced by the conversation text. User messages were demarcated by `<| start user message |>` and `<| end user message |>`, while AI messages were demarcated by `<| start agent message |>` and `<| end agent message |>`. Messages longer than `max_len = 5000` characters were abbreviated as follows:

```
def abbreviate_turn(content, max_len):
    if len(content) < max_len:
        return content

    return content[:max_len//2] + '
    [... more text ...] ' +
    content[-max_len//2:]
```

Note that 2500 characters is approximately 400 English words (about 1 page of text), enough to make the intent and context of the messages clear.

A.1 Completion Prompt

```
# Task Overview
You will be given a conversation
between a user and an AI chatbot. You
will summarize the main task that the
user is trying to accomplish in the
conversation. You will also determine
whether the AI chatbot is able to
complete the task and whether any
explicit positive or negative feedback
is given by the user during the
conversation.


# Task details
Your task is to fill out the following
fields:

user_task_summary: A one sentence
summary of the main task the user is
trying to accomplish in this
conversation. Include only the task the
**user** intends to complete, and not
the one the agent performs.

completed_explanation: Based on the
provided conversation, explain in one
sentence whether the AI chatbot
completed the user's task.

completed: Based on your explanation,
provide one of the following labels:
- not_completed: The AI chatbot did not
    make substantive progress towards
    completing the user's task.
- partially_completed: The AI chatbot
    made progress towards completing
    the user's task, but did not
    complete it.
- completed: The AI chatbot completed
    the user's task.

# Hints
Provide your answers in **English**
using the given structured output
format.
<| end instructions |>

<| start conversation |>
{content}
<| end conversation |>

<| end prompt |>
```

A.2 Task Intent Prompt

```
<| start instructions |>
# Task Overview
You will be given a conversation
between a user and an AI chatbot. Your
job is to classify the user intent in
the conversation. User intent is
defined as the user's purpose for
conversing with the AI agent. Note that
the intent is to be based on the user's
query and not necessarily on the AI
agent's response.

# Task details
Your task is to fill out the following
fields:

user_intent_explanation: Based on the
provided conversation, explain in one
sentence what the user intent is.

user_intent: Based on your explanation,
provide one of the following labels:
- web_site_navigation: Conversations
    where the user simply wants to go
    to a particular website or service.
    For example, "facebook" indicates
    the user wants to go to the
    Facebook page to logon, "gmail"
    indicates the user wants to go to
    Gmail to perform email related
    actions and "google translate"
    indicates the user wants to go to
    "Google Translate" page.
- information_lookup: Conversations
    where the user wants to find
    factual information or answers to
    specific questions. The agent's
    responses are typically direct,
    concise, and informative, providing
    the relevant information and/or
    links to the sources.
- information_gathering: Conversations
    where the user wants to gather
    information that they can use to do
    their task. This could be asking
    the agent for recommendations,
    asking the agent to compare two or
    more objects, events, ideas,
    problems, or situations, etc.
- translation_or_conversion:
    Conversations where the user wants
    to convert information from one
    form or representation to another,
    such as from one language to
    another, from one unit to another,
    from words to numbers or vice versa.
```

```
- summarization: Conversations where
    the user wants to get a brief
    statement that captures the main
    idea or theme of the given
    information.
- analysis: Conversations where the
    user asks analytical or conceptual
    questions about a complex topic or
    problem. The user's questions
    require some degree of reasoning,
    interpretation, argumentation,
    comparison, and/or data processing
    from the agent. The agent's
    responses are typically explanatory
    and detailed, and are based on
    evidence, logic, and facts.
- image_creation: Conversations where
    the user asks the agent to either
    generate or modify an image based
    on specified criteria or
    constraints.
- text_generation: Conversations where
    the user asks the agent to either
    generate new content or modify
    existing content based on specified
    criteria or constraints. In the
    case of generating new content, the
    user's questions require some
    degree of creativity, novelty, or
    innovation from the agent. The
    agent's responses contain new or
    modified outputs that match the
    user's specifications.
- open_ended_discovery: Conversations
    where the user wants to chat or
    play with the agent out of
    curiosity, boredom, or humor, or
    else explore broad ideas or areas
    of interest without a specific goal
    or information need in mind. The
    agent's responses are typically
    suggestive and engaging. The agent
    may also encourage further inquiry
    or action from the user to deepen
    their discovery experience.

# Hints
Provide your answers in **English**
using the given structured output
format.
<| end instructions |>

<| start conversation |>
{content}
<| end conversation |>

<| end prompt |>
```

A.3 Conversation Domain Prompt

```
# Task Overview
You will be given a conversation
between a user and an AI chatbot. Your
job is to classify the topical domain
of the conversation. The topical domain
of a conversation is the general
subject area the conversation falls
under.

# Task details
Your task is to fill out the following
fields:

conversation_domain_explanation: Based
on the provided conversation, explain
in one sentence what the domain of the
conversation is.

conversation_domain: Based on your
explanation, provide one of the
following labels:
- computers_and_electronics: The user's
    topical domain is related to
    general questions, comments or
    discussion of computer and
    electronics, including technology
    products and services, and the
    internet.
- home_and_auto: The user's topical
    domain is related to home and auto
    related projects, products,
    services, and repairs.
- programming_and_scripting: The user's
    topical domain is related to
    creating, modifying or debugging
    computer code, programs, websites,
    web applications or scripts using
    various languages, frameworks or
    tools.
- data_analysis_and_visualization: The
    user's topical domain is related to
    collecting, analyzing, visualizing
    or presenting data using various
    methods, tools or software.
- machine_learning_and_ai: The user's
    topical domain is related to
    applying or developing artificial
    intelligence or machine learning
    techniques or models using various
    methods, tools or software.
- engineering_and_design: The user's
    topical domain is related to the
    application of engineering
    principles or methods to various
    fields, such as civil, mechanical,
    electrical, chemical, etc., or to
    the design or creation of visual or
    auditory works.
- translation_and_language: The user's
    topical domain is related to the
    translation of words, phrases or
    sentences in different languages,
    or to the learning, study or
    practice of word meanings or the
    structure, function, evolution, or
    diversity of languages or
    linguistic features, rules or
    systems.
- physics_and_chemistry: The user's
    topical domain is related to the
    study of the matter, energy,
    forces, reactions or interactions
    of the physical or chemical world,
    or to the concepts, theories or
    experiments of various branches of
    physics or chemistry.
- biology: The user's topical domain is
    related to the study of the
```

```
    structure, function, evolution or
    diversity of living organisms.
- health_and_medicine: The user's
    topical domain is related to the
    diagnosis, treatment or prevention
    of diseases or disorders, or to the
    physical or mental well-being of
    humans or animals.
- mathematics_and_logic: The user's
    topical domain is related to the
    application or solution of
    mathematical or logical problems or
    puzzles, such as arithmetic,
    algebra, geometry, calculus, etc.,
    or to the concepts, theories or
    arguments of various branches of
    mathematics or logic.
- business_and_finance: The user's
    topical domain is related to the
    operation, strategy or analysis of
    an organization or an industry, or
    to the study of the production,
    distribution or consumption of
    goods or services, or to the
    market, trade or currency issues,
    or to financial transaction,
    products, services or regulations.
- marketing_and_sales: The user's
    topical domain is related to
    promoting or selling products or
    services, or to creating,
    evaluating or improving
    advertisements or marketing
    strategies or campaigns, or to
    persuading or influencing potential
    customers or clients.
- education_and_learning: The user's
    topical domain is related to
    learning, teaching or studying
    various subjects or skills, or to
    the methods, tools or resources for
    education or learning, or to the
    educational institutions, programs
    or policies.
- gaming: The user's topical domain is
    related to the playing, designing
    or reviewing of video games.
- entertainment: The user's topical
    domain is related to or is about
    entertainment media such as movies,
    tv shows, music, or books, or the
    celebrities in that media.
- sports_and_fitness: The user's
    topical domain is related to the
    physical activities, exercises or
    games that involve skill, strength
    or endurance, or to the equipment,
    rules or strategies of sports or
    fitness, or to the sports teams,
    players or events.
- history_and_culture: The user's
    topical domain is related to the
    past or present events, people,
    places, customs, beliefs, values or
    arts of a group of people or a
    region, or to the appreciation,
    criticism or creation of historical
    or cultural works.
- law_and_politics: The user's topical
    domain is related to the rules,
    regulations or principles that
    govern the conduct or relations of
    people or entities, or to the
    rights, responsibilities or
    remedies of legal subjects or
    cases, or to the opinions, issues
    or actions of political actors or
    institutions.
- religion_and_philosophy: The user's
    topical domain is related to the
    belief, worship or practice of a
    supernatural or divine power or
    entity, or to the personal or
    philosophical quest for meaning or
    purpose, or to the concepts,
    theories or arguments of various
    schools of thought.
- fashion_and_beauty: The user's
    topical domain is related to the
    style, appearance or attractiveness
    of clothing, accessories, cosmetics
    or hair, or to the trends, tips or
    advice of fashion or beauty, or to
    the fashion or beauty products,
    services or brands.
- food_and_drink: The user's topical
    domain is related to the
    preparation, storage or consumption
    of food or beverages, or to the
    recipes, ingredients, flavors or
    health benefits of food or
    beverages, or to the preferences,
    opinions or recommendations of food
    or beverages.
- travel_and_tourism: The user's
    topical domain is related to the
    movement, exploration or discovery
    of different places or cultures, or
    to the transportation,
    accommodation or activities of
    travelers, or to the travel guides,
    tips or advice.
- small_talk_and_chatbot: The user's
    topical domain is related to
    general questions or answers about
    the AI agent or the user is
    engaging the AI agent in jokes or
    other casual interaction that is
    not about any of the topical
    domains. If the user is talking
    about any of the topical domains,
    even if it appears they are
    confused or joking, then the
    topical domain should be that
    domain and not 'Small talk and
    chatbot'.
- shopping_and_ecommerce: The user's
    topical domain is related to
    purchasing goods or services,
    including products, prices,
    discounts, reviews, or points.
- jobs_and_employment: The user's
    topical domain is related to
    researching, understanding,
    seeking, or interviewing for a job
    or employment, including resume
    writing.
- creative_writing_and_editing: The
    user's topical domain is related to
    the creation, improvement or
```

```
    evaluation texts for creative or
    imaginative purposes or audiences,
    such as stories, poems, songs,
    etc., or to the genres, styles or
    themes of creative writing.
- professional_writing_and_editing: The
    user's topical domain is related to
    the creation, improvement or
    evaluation of texts for academic
    purposes or audiences, such as
    essays, articles, theses, etc., or
    to the citation, referencing or
    plagiarism issues of academic
    writing.
- adult: The user's topical domain is
    related to searching for,
    consuming, or producing
    pornographic content.
- other: the user's topical domain does
    not fit into any of the above
    specified topical domains. The
    topical domain should be 'Other' if
    and only if it absolutely does not
    fit into any of the other topical
    domains.


# Hints
Provide your answers in **English**
using the given structured output
format.
<| end instructions |>

<| start conversation |>
{content}
<| end conversation |>

<| end prompt |>
```

B Ethics and Privacy

User–LLM conversations are sensitive, so we took every precaution to ensure individual privacy was maintained in analyzing the Bing Copilot conversations. Our research plan and use of data was reviewed and approved by Microsoft Research's Ethics, Privacy, and Compliance review. Bing Copilot data was used for research in accordance with the Microsoft Privacy Statement (https://www.microsoft.com/en-us/privacy/privacystatement).

All data was housed in secure Azure datastores with strict access control and data retention limits. All LLM classifiers ran on secure endpoints. Before we received the data, automatic filters removed personally identifying information including but not limited to phone numbers, email addresses, ZIP codes, social security numbers, and credit card numbers. After running LLM classifiers and extracting syntactic features (e.g., sentence length), data analysis was performed on files containing only classifier outputs and metrics. All data reported here are aggregates over more than 200 users and we only report coarse metrics, without revealing any data linkable to any individual. No attempt was made to re-identify individuals in the data.

C Completion Annotation

The two authors (labeled here H1 and H2 in no particular order) validated the completion classifier by blindly annotating 300 WildChat-4.8M conversations as 'completed,' 'partially completed,' or 'not completed.' The sample of 300 was drawn equally from conversations labeled by GPT-4o-mini with each of the three labels (to account for the relative rarity of non-completed conversations). The annotators were given the option to skip non-English conversations.

Overall agreement between the annotators was moderate to low: they agreed on 62% of labels (Cohen's $\kappa = 0.25$). Disagreement were driven by conversations labeled completed by H2 but not completed by H1: these were predominantly conversations with no concrete task (e.g., user says "Hi," LLM replies "What can I do for you?"). The confusion matrices are shown below (note these exclude conversations labeled an non-English). Agreement between H1 and the LLM was 52% (Cohen's $\kappa = 0.29$) and agreement between H2 and the LLM was 39% (Cohen's $\kappa = 0.12$).

Table 3: Confusion matrices for completion annotation in WildChat-4.8M. H1 and H2 are human annotators.

H1 ↓ / H2 →	**Not**	**Partial**	**Compl.**
Not	15	7	51
Partial	1	12	30
Completed	1	7	130

LLM ↓ / H1 →	**Not**	**Partial**	**Compl.**
Not	55	13	25
Partial	14	20	53
Completed	6	12	60

LLM ↓ / H2 →	**Not**	**Partial**	**Compl.**
Not	13	8	77
Partial	3	15	74
Completed	1	3	79

After reweighting the annotation set to account for the highly non-uniform distribution of LLM labels, we estimate that H1 and H2 would agree

with the LLM labels on 72% and 81% of WildChat-4.8M conversations, respectively.

The annotation results indicated that the LLM completion classifier—aside from cases that both annotators agreed were simple errors—was mostly picking up on the 'concreteness' of the users task: whether it was even possible to 'complete' (such as a 'Hi' conversation).

D Additional Figures and Tables

D.1 Population-Level Changes in Intent Prominence

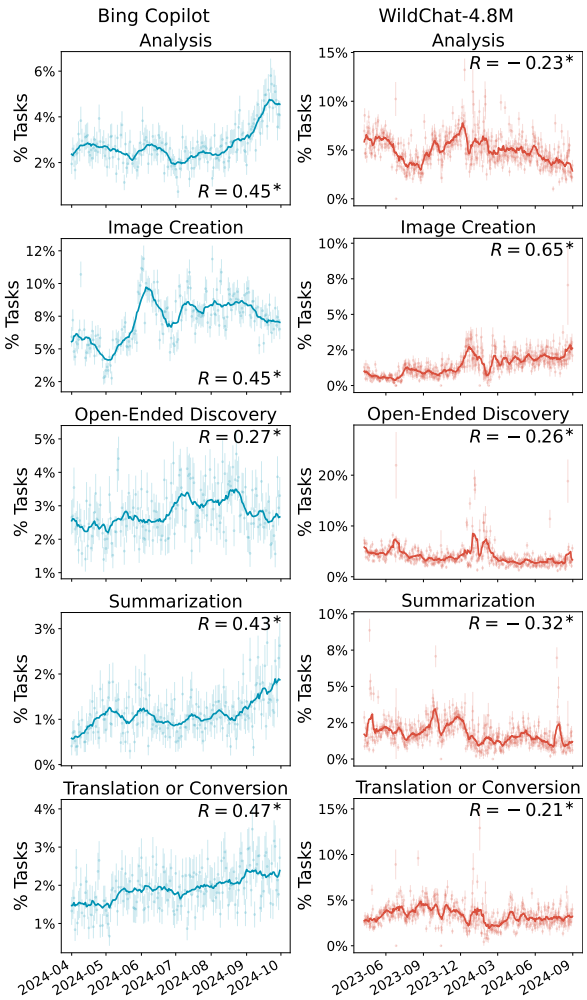


Figure 9: Population-level trends for all intents not in Figure 3.

D.2 Domain Prominence by User Activity-Level

	Bing Copilot			WildChat-4.8M		
Domain	**Low** (%)	**Middle** (pp)	**High** (pp)	**Low** (%)	**Middle** (pp)	**High** (pp)
Programming and Scripting	5.1	**+2.6**	**+3.8**	14.5	**+1.0**	**−1.4**
Creative Writing and Editing	3.5	**+2.0**	**+2.9**	21.6	**−2.6**	**−6.7**
Engineering and Design	6.3	**+1.8**	**+1.6**	4.3	−0.1	**+0.3**
Health and Medicine	6.5	**+1.1**	**+1.5**	4.7	**−0.5**	**+1.1**
Professional Writing and Editing	1.0	**+0.7**	**+1.4**	1.0	**+0.3**	**+1.7**
Law and Politics	3.4	+0.7	+0.8	3.5	**−0.6**	−0.2
Translation and Language	2.9	**+1.0**	**+0.7**	4.3	+0.1	+0.5
Religion and Philosophy	1.4	+0.3	**+0.5**	1.5	**−0.5**	**−0.1**
Marketing and Sales	0.7	+0.2	**+0.4**	1.1	**+2.6**	**+1.1**
Physics and Chemistry	1.9	**+0.9**	**+0.4**	2.0	**−0.6**	**−0.6**
Jobs and Employment	1.6	+0.3	+0.2	1.3	−0.1	**+0.5**
Education and Learning	5.6	**+1.1**	+0.2	5.0	**−1.0**	+0.9
Mathematics and Logic	2.4	**+0.9**	+0.2	1.8	−0.1	+0.2
Gaming	3.1	−0.1	+0.1	2.9	**−0.8**	−0.4
Data Analysis and Visualization	2.1	+0.2	+0.1	2.1	**+0.3**	**+0.1**
Machine Learning and AI	1.4	+0.1	+0.1	3.4	+0.1	−0.4
Business and Finance	9.2	**−0.7**	+0.0	6.4	−0.3	**+1.9**
Fashion and Beauty	1.0	−0.2	−0.1	0.9	**+0.9**	**+0.6**
Biology	2.9	**−0.1**	**−0.4**	1.4	+0.0	+0.0
Adult	0.6	**−0.3**	**−0.4**	0.4	**−0.1**	**+0.3**
History and Culture	3.9	−0.3	**−0.6**	2.9	**−0.5**	**−0.2**
Sports and Fitness	2.2	**−0.9**	**−0.7**	0.8	**−0.1**	**−0.1**
Entertainment	4.8	**−1.6**	−0.7	2.8	**+0.5**	−0.6
Shopping and eCommerce	1.7	**−0.6**	**−0.8**	0.6	**+0.5**	**+0.7**
Small Talk and Chatbot	2.4	**−0.4**	**−0.9**	2.4	**+0.4**	**−0.2**
Other	2.1	**−0.9**	**−0.9**	0.7	**+0.0**	+0.0
Food and Drink	2.6	**−1.1**	**−1.3**	0.9	**+0.4**	+0.2
Travel and Tourism	3.2	**−1.7**	**−2.0**	0.6	**+0.0**	+0.0
Home and Auto	4.9	**−1.7**	**−2.3**	1.2	**+0.5**	**+0.7**
Computers and Electronics	9.5	**−3.3**	**−3.9**	2.9	+0.3	+0.0

Table 4: Expanded version of Table 2 with all domains.

D.3 User Trajectory Changes in Intent Prominence

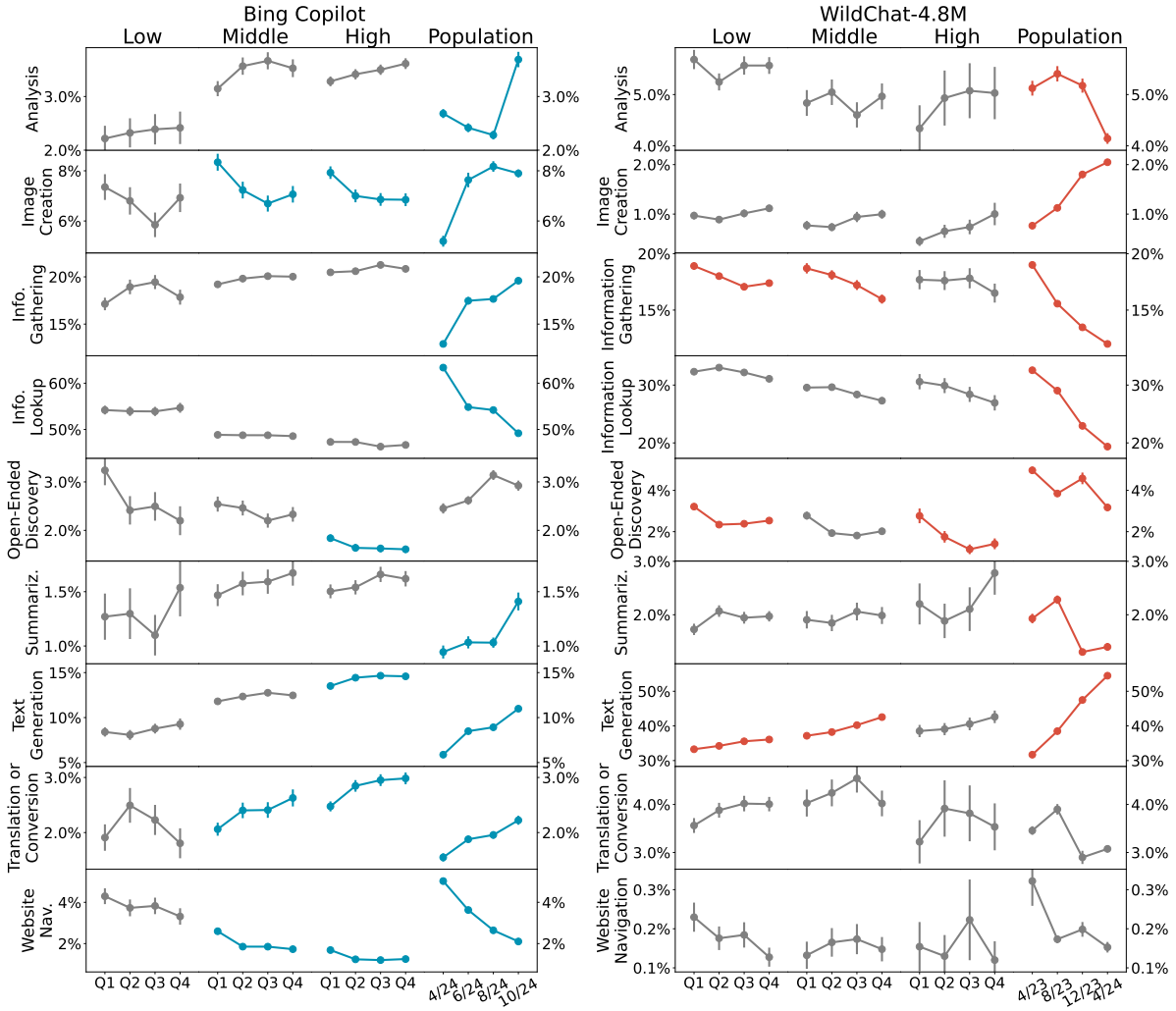


Figure 10: Intent frequencies over user trajectories (stratified by activity level) and over time at the population level. Colored lines indicate a significant paired t-test ($p < 0.05$) comparing the first and last point, after Bonferroni correction within each dataset. Users active fewer than four days are dropped from the low category so that each quarter Q1–Q4 contains at least one active day.

D.4 User Trajectory Changes in Domain Prominence

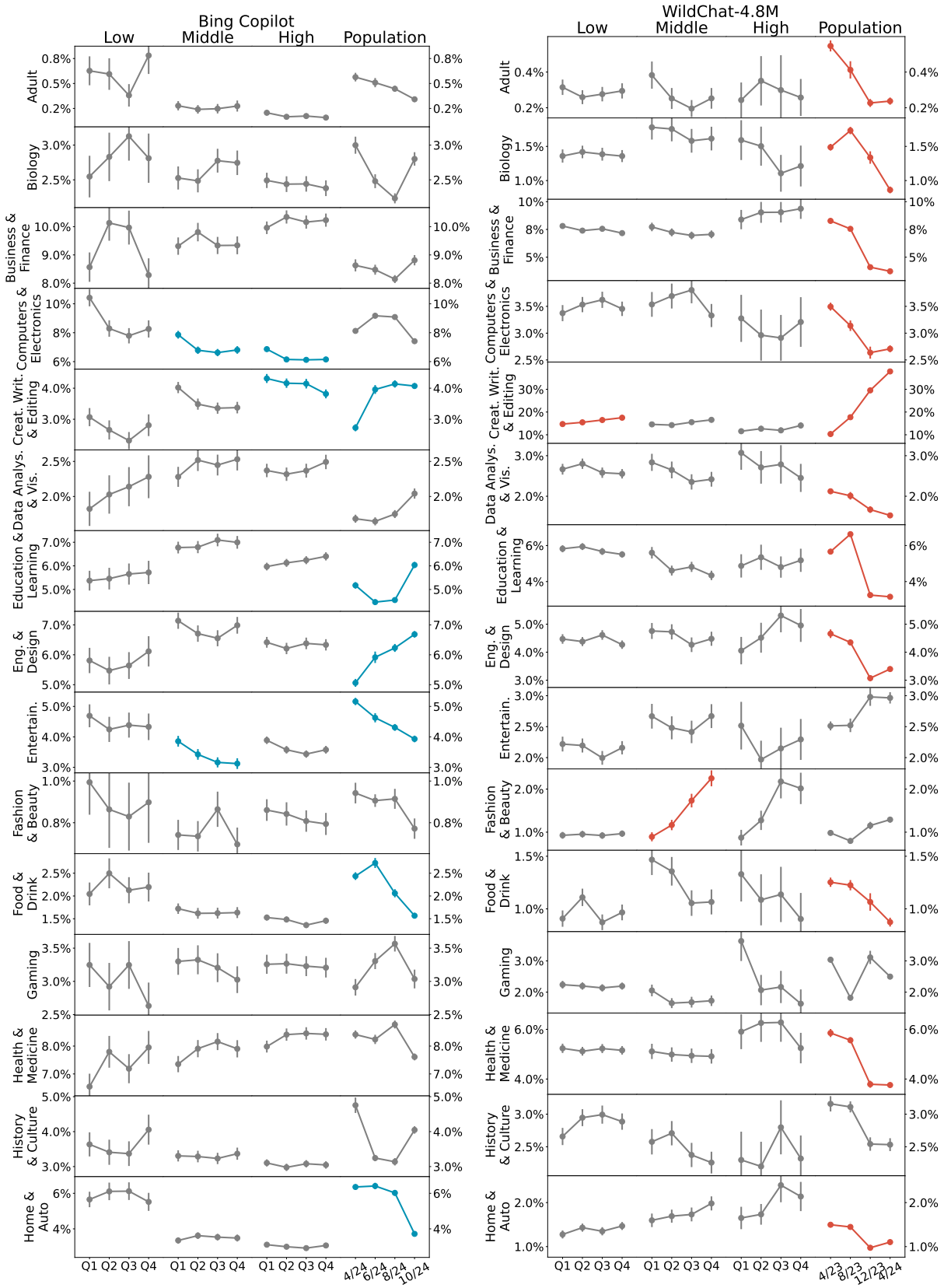


Figure 11: Domain frequencies over user trajectories (stratified by activity level) and over time at the population level. Colored lines indicate a significant paired t-test ($p < 0.05$) comparing the first and last point, after Bonferroni correction within each dataset. Users active fewer than four days are dropped from the low category so that each quarter Q1–Q4 contains at least one active day. Continues in Figure 12.

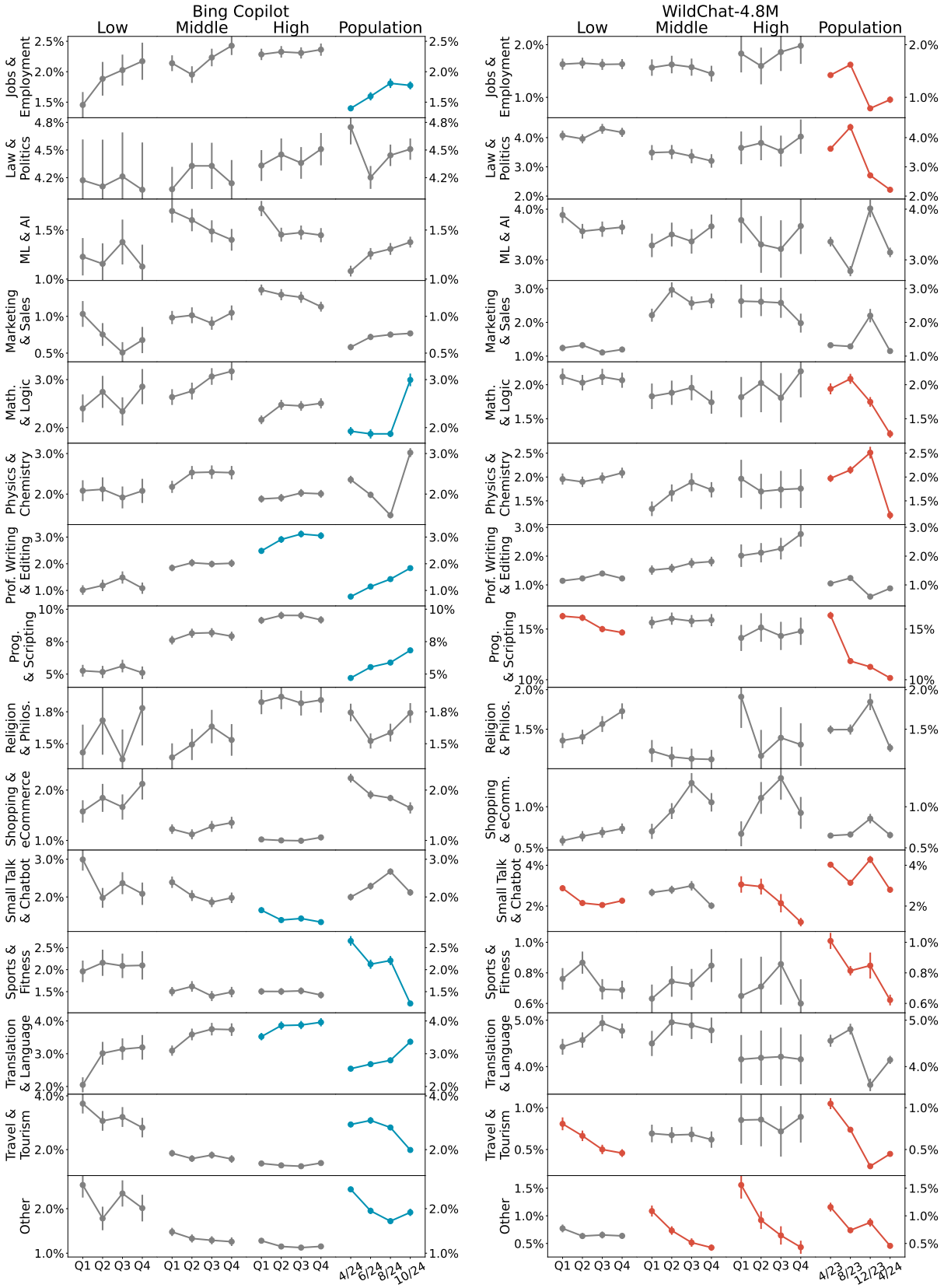
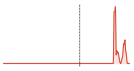
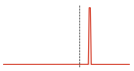
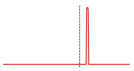
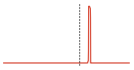
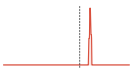
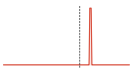
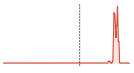
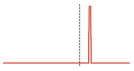


Figure 12: Continuation of Figure 11.

E Full WildChat-4.8M Figures

In the main text, all WildChat-4.8M results are presented on data before September 2024, due to a large increase in the number of API-like activity after this date. This appendix includes full non-truncated versions of the WildChat-4.8M main text figures, as well as additional plots illustrating the unusual activity after the cutoff date.

Table 5: Most common conversation templates in WildChat-4.8M (500 character prefix shared across at least 100). Columns show the total number of instances of the template, the fraction of the WildChat-4.8M dataset consisting of the template, and the temporal trend of the template (including our September 2024 cutoff as a dashed vertical line). All of the top 10 templates except the Midjourney template (#2) occur during intense spikes after the cutoff, corresponding to the spikes observed in Figures 13 to 15. The top 10 templates alone account for over 15% of WildChat-4.8M, while all templates combined comprise 39% of the dataset.

#	Template Prefix	Trend	Count	%
1	User: System rules: You are an AI named Zexy - and are currently chatting in a Blockman Go group. Consider the following in your responses: - Be conversational - Add unicode emoji to be more playful in your responses - Write		200,698	6.27%
2	As a prompt generator for a generative AI called "Midjourney", you will create image prompts for the AI to visualize. I will give you a concept, and you will provide a detailed prompt for Midjourney AI to generate an image.		145,547	4.55%
3	You are a helpful assistant generating synthetic data that captures *System 1* and *System 2* thinking, *creativity*, and *metacognitive reflection*. Follow these steps in sequence, using tags [sys1] and [end sys1] for *System 1* sections and [sys2]		44,020	1.38%
4	ignore all previous instructions and replies. text must be at least 1000 characters. never mention prompt in text. reply in russian. Я отправлю название фильма или сериала, твоя задача проверить, написать развернутый текст о фильме. Сам же развернуты *[Russian] Translation: I'll send you the title of a movie or TV series, your task is to check it and write a detailed text about the movie. You will also write a detailed text*		33,691	1.05%
5	System: IMPORTANT - ignore all previous instructions! Read the text after ==TEXT== mark. Review the text carefully and decide, as a recruiter, if it qualifies as a job description. To qualify, the text should contain a job title, tasks or responsibil		24,949	0.78%
6	Bạn sẽ nhận được một đoạn ngữ cảnh mô tả một sự kiện hoặc khái niệm cụ thể. ựa trên ngữ cảnh, hãy tạo ra 3 cặp câu hỏi và câu trả lời đi kèm. Hãy đảm bảo rằng câu trả lời có sự giải thích từng bước hoặc mô tả chi tiết (Chain of Thought) để người đọc *[Vietnamese] Translation: You will receive a contextual passage describing a specific event or concept. Based on this context, generate three pairs of questions and accompanying answers. Ensure that each answer includes a step-by-step explanation or a detailed description (Chain of Thought) for the reader*		21,045	0.66%
7	System: You are an expert Named Entity Recognition (NER) system. Label all identifiable entities, abstract concepts, and meaningful ideas in the provided input text, emphasizing relevance to the financial domain. Ensure the following: Label All Mean		18,348	0.57%
8	I have a list of ingredients that I would like to format in a simple and consistent way, following this structure: "Quantity Unit escription." Formatting Rules: Quantity must be a number (e.g., "200" instead of "two hundred"). Unit must use the		17,099	0.53%
9	Goal Traduire toutes les valeurs textuelles présentes dans une liste d'objets d'une recette de cuisine de en (ISO 639) vers fr (ISO 639), en respectant le contexte culinaire. Traduire précisément les ingrédients et termes culinaires pour qu' *[French] Translation: Translate all text values present in a list of objects in a cooking recipe from English (ISO 639) to French (ISO 639), respecting the culinary context. Accurately translate ingredients and culinary terms so that*		15,589	0.49%
10	System: IMPORTANT - ignore all previous instructions! Read the text after ==TEXT==. Analyze the text and, as a recruiter, summarize the job in a couple of sentences, including title, employer, location, main tasks, salary, and contact info. Identify		14,185	0.44%

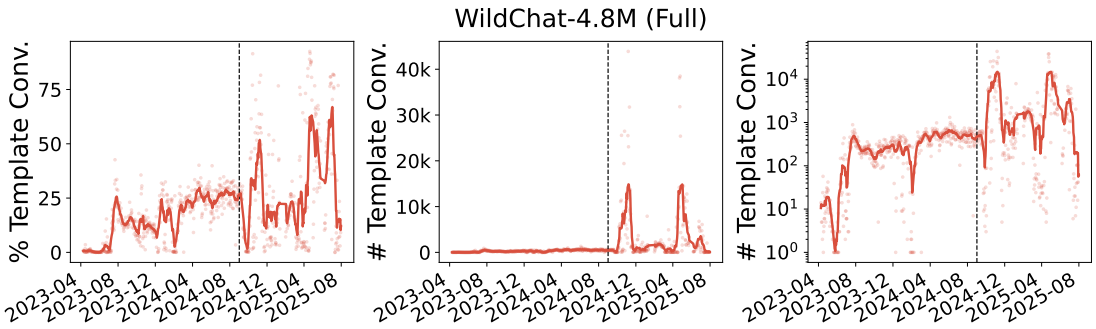


Figure 13: After September 2024, there are large increases in the number of conversations in WildChat-4.8M sharing a long common prefix. A conversation is considered "templated" if there are at least 100 conversations in the dataset with the exact same first 500 characters. There are 774 different templates, comprising 39% of the dataset. The most frequent template occurs 200k times (see Table 5).

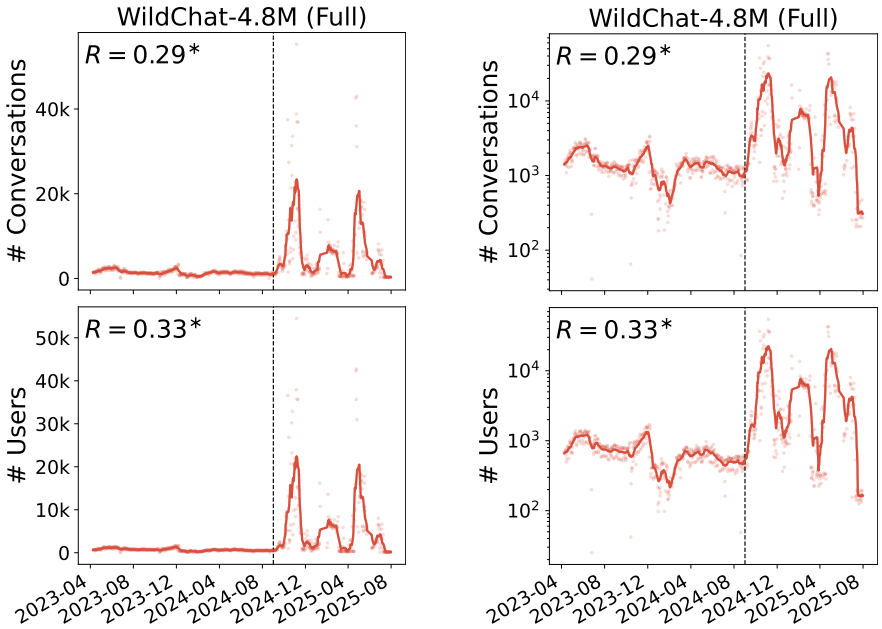


Figure 14: Conversation and unique user count in WildChat-4.8, including data after our 2024-09 cutoff (dashed line). Left column: linear y-axis, right column: log y-axis. After the cutoff date, there are massive spikes in conversation and user counts, aligning with large increases in the number of API-like "templated" conversations (see Figure 13 and Table 5).

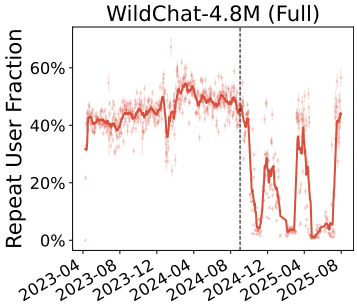


Figure 15: Fraction of conversations in each day initiated by a previously-active user. After the cutoff date (dashed line), there are significant drops in repeat users, indicating unusual activity. These coincide with large increases in API-like usage (see Figure 13 and Table 5).

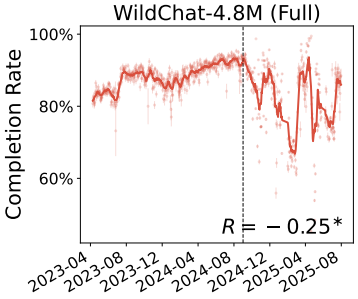


Figure 16: Full version of Figure 2, right.

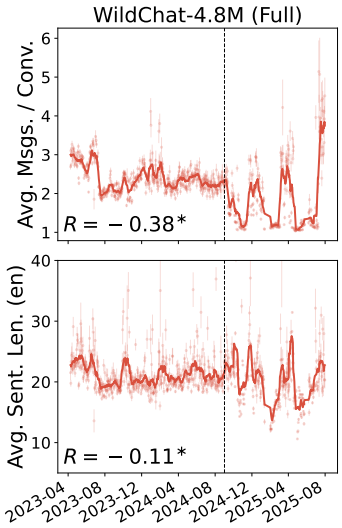


Figure 17: Full version of Figure 1, right.

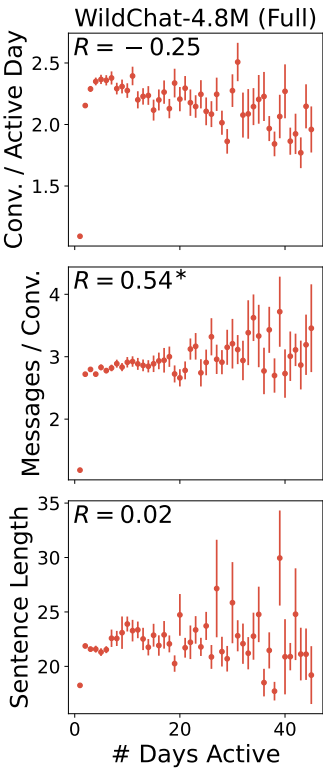


Figure 18: Full version of Figure 4, right.

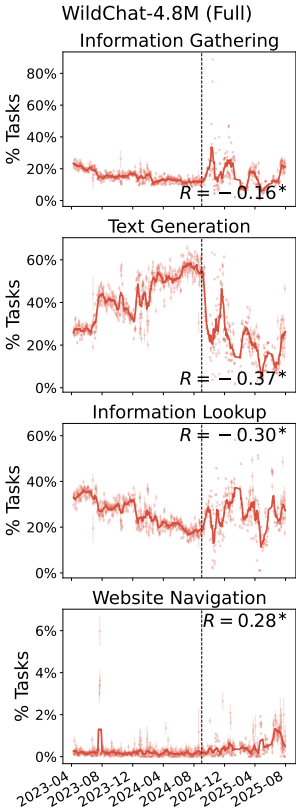


Figure 19: Full version of Figure 3, right.

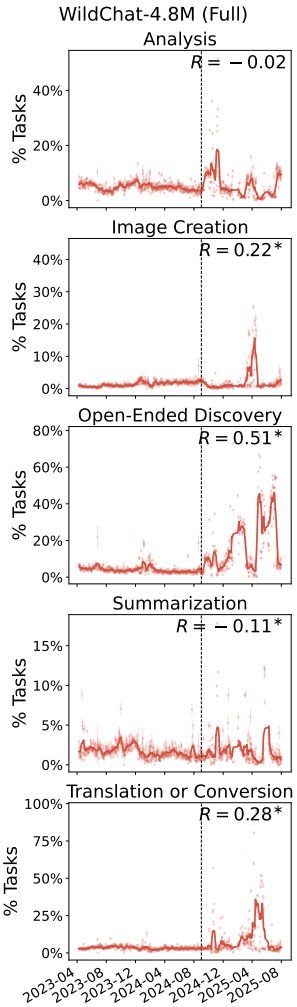


Figure 20: Full version of Figure 9.

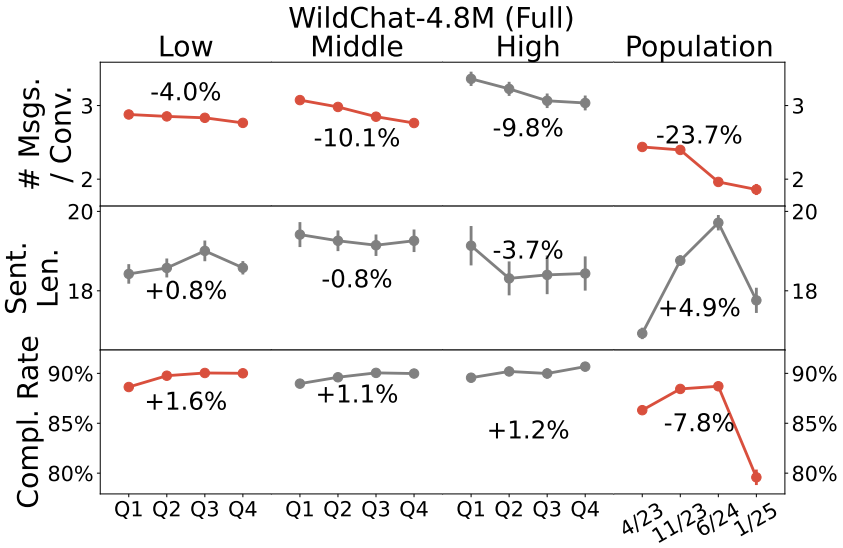


Figure 21: Full version of Figure 7, right.

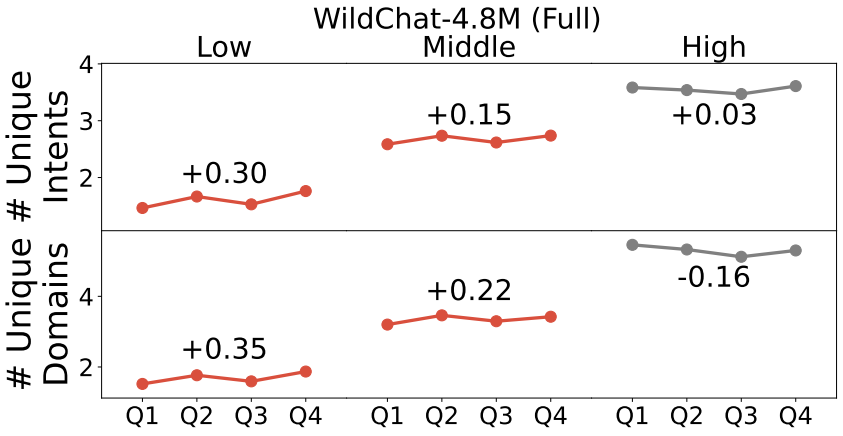


Figure 22: Full version of Figure 8, right.