The AI Observatory: A Public Measure of Real-World AI Use

Shayne Longpre[1,*] **Anka Reuel**[2,*] **Dayeon Ki**[16,*]
Cathy Mengying Fang[1,†] Megan Richards[11,†] Zhiping Zhang[3,†] Chuanyang Jin[4,†]
Jenn Mickel[6] Cedric Deslandes Whitney[7] Ahmet Üstün[8]
Niloofar Mireshghallah[9] Victor Ojewale[10] Ariel Lee[12]
Alex (Sandy) Pentland[1,§] Tianshi Li[3,§] Yuntian Deng[13,§]
Mykel Kochenderfer[2,§] Sara Hooker[15,§] Sanmi Koyejo[2,§]
[1]MIT [2]Stanford University [3]Northeastern University [4]Johns Hopkins University
[6]EleutherAI [7]University of California, Berkeley [8]Cohere
[9]Carnegie Mellon University [10]Brown University [11]New York University
[12]Code Metal [13]University of Waterloo [15]Adaption Labs [16]University of Maryland
[*]Co-first author [†]Core contributor [§]Advising contributor
Correspondence: shayne.r.longpre@gmail.com, anka.reuel@stanford.edu

Abstract

Understanding the benefits, risks, and impacts of general-purpose AI assistants requires looking at real conversations—not curated benchmarks or surveys. Yet AI use narratives rely on limited proprietary reports, or on few data sources, fragmented across platforms, models, and time. We introduce the AI Observatory, a public measurement platform that aggregates the most comprehensive set of real AI conversation sources, and develop a common taxonomy of 145 features, spanning function, topic, sensitive use, interaction style, multi-turn dynamics, and conversation structure. Across 23,158 conversations and 85,633 turns, we find that observed AI use is highly heterogeneous: sources differ substantially in tasks, structures, and sensitive-use distributions, with no single source that generalizes. We further show that occupational analyses such as the initial Anthropic Economic Index capture an important but incomplete slice of use: applied to our sources, its occupational filter removes 47.9% of the conversations we analyze as non-occupational, omitting substantial personal, social, cultural, and safety-relevant interaction. AI use is also not static: we find tokens and turns per conversation have been on the rise since 2023, and we show that model variants within the same developer support distinct usage regimes. The Observatory provides a new taxonomy, reusable annotation tools, and a public platform for reproducible measurement of how AI assistants are used and are evolving in the wild: https://ai-observatory.org.

1 Introduction

General-purpose AI assistants have rapidly become embedded in everyday workflows, used by hundreds of millions of people each week across work, education, and personal life [1–3]. Yet public evidence on how people actually use it in the real-world remains extremely limited and fragmented. Researchers rely on single-source conversation datasets [4, 5], surveys [3, 6], social-media interaction studies [7], or proprietary reports [1, 8–10]. Each source provides a partial or redacted picture; together, they do not yet add up to a broad, reusable, and publicly inspectable account of AI use.

Preprint.

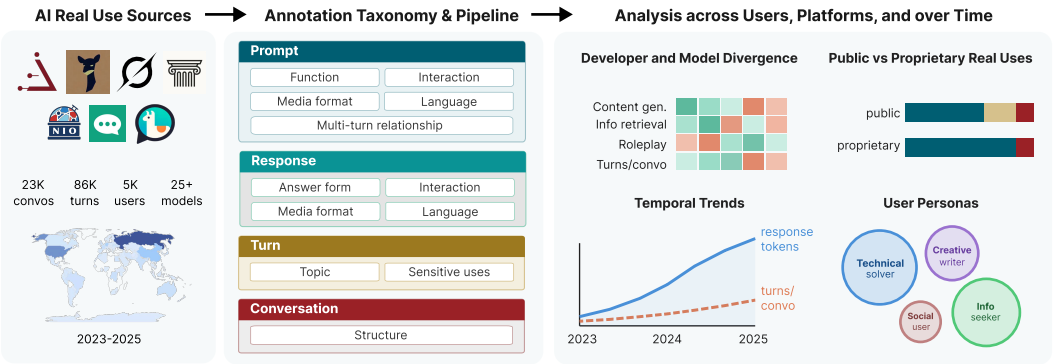


Figure 1: **Overview of the AI Observatory.** We aggregate seven sources of real AI conversations (left; 23,158 conversations, 85,633 turns, 5K users, 25+ models, from 2023–2026) and annotate each conversation with a unified 145-feature taxonomy at four levels—prompt, response, turn, and conversation (middle). The annotated corpus supports cross-source, model-stratified, temporal, and user-level analyses, including public vs. proprietary comparisons, developer/model divergence, temporal trends, and user-persona clustering. Graphs are illustrative (right).

Understanding how people use AI assistants is a prerequisite for almost every downstream question we want to answer about these systems: how to design them effectively, which capabilities to evaluate, which risks to mitigate, which policies to design, and which claims about AI's effects on work, education, and well-being are supported. Recent evidence shows that AI assistants are increasingly used by young people [3], function as companions and emotional confidants [11, 12], and inform real-world decisions such as health and legal questions [13, 14]. Voluntary developer reports have become the most prominent and widely recognized source of real usage information, but they necessarily release analyses that are limited in depth and reproducibility [1, 8, 9]. Independent public monitoring is thus essential: without shared measurement tools, claims about what people use AI for remain difficult to reproduce, compare, or update.

To address this need, we introduce the AI Observatory (Figure 1), a public aggregation of seven sources of AI usage, a conversation taxonomy spanning 145 independent features, and tools for annotating, and comparing real AI conversations across sources, models, and time. Existing single-platform analyses provide a useful starting point but only a partial view: OpenAI's recent large-scale analysis of ChatGPT conversations, for example, reports that nearly 80% of usage falls into three categories (practical guidance, seeking information, and writing), with relational and emotional uses accounting for under 2% [1]. Our cross-source analysis questions this picture: the prevalence of these categories varies substantially across public datasets, and personal, relational, and safety-relevant content is far more prominent in conversations that occupational taxonomies classify as non-task use. The Observatory makes this kind of comparison possible at scale; specifically, we contribute the AI Observatory platform, which includes:

1. **A taxonomy and annotation pipeline for AI use.** We develop a comprehensive annotation framework that captures 145 independent conversation features, spanning media formats, user functions, topics, sensitive uses, interaction behaviors, refusal types, multi-turn dynamics, and conversation structure. We release the taxonomy and annotation pipeline, designed for extending to new annotation tasks or sources, at `https://ai-observatory.org`.
2. **The AI Observatory dataset and dashboard.** We annotate 23,158 conversations (171,266 messages) from seven AI real-use sources under our unified annotation framework. To our knowledge, this is the largest aggregated public collection of real-use AI conversations analyzed under a common taxonomy. We release the resulting conversation feature annotations, tools to reconstruct the corpora, and an interactive analysis dashboard at `https://project-ai-observatory.vercel.app/`.

Using this platform, we surface four major empirical findings:

1. **Real-use sources diverge dramatically in AI conversation features.** We show that real-use datasets differ significantly in task mix, sensitive-use prevalence, and conversation structure.

These differences persist across similar underlying models, suggesting that platform interface and user composition shape how AI is used as much as the model itself, and that evaluations focused on model behavior alone may systematically misestimate the risks that surface in deployed systems.

2. **Proprietary occupational summaries omit nearly half of AI use.** Applying the Economic Index's [8, 9] analysis pipeline to our sources shows that nearly 48% of conversations are filtered, classified as non-occupational. We show this view likely omits significant personal, social, cultural, and safety-relevant content.
3. **AI use shifts substantially across time, model developers, and model variants.** For instance, WildChat conversations grow longer and more code-oriented between 2023 and 2025, and OpenAI model variants support distinct usage regimes (short, template-like exchanges with GPT-3.5; longer, more iterative assistance with GPT-4o; and long, one-shot technical problem solving with reasoning-oriented models like o1). These shifts impact which functions, risks, and interaction styles appear.
4. **Inter-user heterogeneity is significant and specialized.** Use analyses have systematically neglected user-centric frames, that are missed in aggregate trends. We cluster regular WildChat users into personas groups, showing that within-user behavior is largely stable across months and that what distinguishes personas is in their interaction features and sensitive-use flags rather than in topic mix alone. WildChat is the only source in our corpus that supports this analysis, which indicates how limited the public record currently is.

The Observatory is not a representative survey of all AI use; but it is a meaningful step in this direction, building a public measurement layer for comparing observable real-use sources and auditing how conclusions change under source, model, time, and taxonomy choices. Taken together, these results indicate that AI use is shaped substantially by these variables, and that no single public dataset or proprietary summary yet captures the full picture.

2 The AI Observatory: Data, taxonomy, and annotation

Data. The AI Observatory aggregates real-world AI conversations from conversation releases, user-shared archives, and model-comparison interfaces. It currently includes seven] sources: the full unfiltered (only partially public) WildChat logs [5], ShareGPT [15], a private sample of AI Archive [16], a crawl of Grok invocations from X [17, 18], LMSYS-Chat-1M [19], Chatbot Arena [20], and AI conversations from the National Internet Observatory(NIO) [21]. NIO is a privacy-preserving, browser-instrumented dataset with restricted access; raw conversations cannot be redistributed. We obtained access under NIO's data-use agreement and report only summary statistics derived from our taxonomy. This is the most comprehensive aggregation of public, and private sources of AI real-use logs assembled to-date.

Table 1: **AI Observatory data sources.** Counts are annotated samples used in this draft unless otherwise noted. NIO contributes manually labeled, privacy-preserving annotations only; raw conversations are not redistributed. ✓: Annotated; ✗: Not annotated from the source population.

Dataset	Users (✓)	Dialogs (✓)	Turns (✓)	Dialogs (✗)	Models	Time	Source
WildChat	4,439	14,648	50,152	4,789,542	ChatGPT variants	Apr 2023–Jul 2025	API intermediary
ShareGPT	–	1,000	8,438	89,665	ChatGPT	2023	Shared chat exports
AI Archive	–	2,076	15,955	7,964	Multiple providers	2023–2025	Public conversation archive
Grok	505	1,234	3,750	1,754	Grok	2025 snapshot	Public shared chats
LMSYS-Chat-1M	–	2,000	3,900	998,000	25 models	Apr–Aug 2023	Vicuna demo / Arena
Chatbot Arena	–	2,000	2,526	269,268	Model pairs	2023–2024	Human preference battles
NIO	–	200	912	4,329	ChatGPT, Gemini	Jun 2024–Apr 2025	Browser instrumentation

Conversation features & taxonomy. A core contribution of the Observatory is a shared taxonomy for describing real human–AI interaction beyond task labels alone. Rather than deriving this schema from a single dataset, we seeded it from prior work on real-user prompts and challenging real-world uses [22], conversational search and question-answering formats [23, 24], occupational and research uses of LLMs [25], privacy disclosure and socially situated LLM interaction [12], data-driven topic grouping efforts [26], acceptable-use and safety-policy analysis [27], multi-turn dialogs [28–31], and

Table 2: **Annotation taxonomy summary.** Each conversation is labeled at four levels (prompt, response, turn, conversation). Full label sets and definitions in Subsection D.1.

Family	Level	Labels	Description	Example labels
Structure	Conversation	-	Turns, Tokens, User dialogs	*Turns, prompt tokens*
Language	Prompt, Response	72	Detected language(s)	*English, Chinese*
Media format	Prompt, Response	11	Content types & formats	*Code, math, lists, URLs*
Prompt function	Prompt	39	What the user is trying to do	*Info retrieval, creative writing, code gen.*
Prompt interaction	Prompt	6	Social behaviors in the prompt	*Role assign., courtesy, jailbreak*
Multi-turn relation	Prompt	5	How prompt relates to prior turns	*First request, revision, deepen*
Answer form	Response	6	How the model responds	*Answer, refusal, clarification*
Response interaction	Response	9	Social or stylistic features	*Apology, empathy, hedging*
Topic	Turn	37	Subject matter of the turn	*Health, technology, education*
Sensitive uses	Turn	24	Potentially harmful or restricted content	*Sexual, harassment, privacy*

work on abstention, refusal, and anthropomorphic dialogue behavior [32–34]. We then refined it over five rounds of pilot annotation and adjudication by ten trained author annotators, all active researchers with backgrounds spanning privacy, safety, security, and evaluation science, using stratified examples drawn from multiple sources rather than a single platform.

After each round, observed phenomena that were absent from the taxonomies were added; ambiguous, overlapping, or source-specific labels were merged, dropped, or redefined; labels were retained only if they recurred across sources or were strongly motivated by prior literature or safety policy. The resulting hierarchical schema, summarized in Table 2 and detailed in Subsection D.1, captures 145 conversation features spanning prompt media, user function, multi-turn relations, interaction behaviors, response form, response media, topics, sensitive uses, and conversation structure.

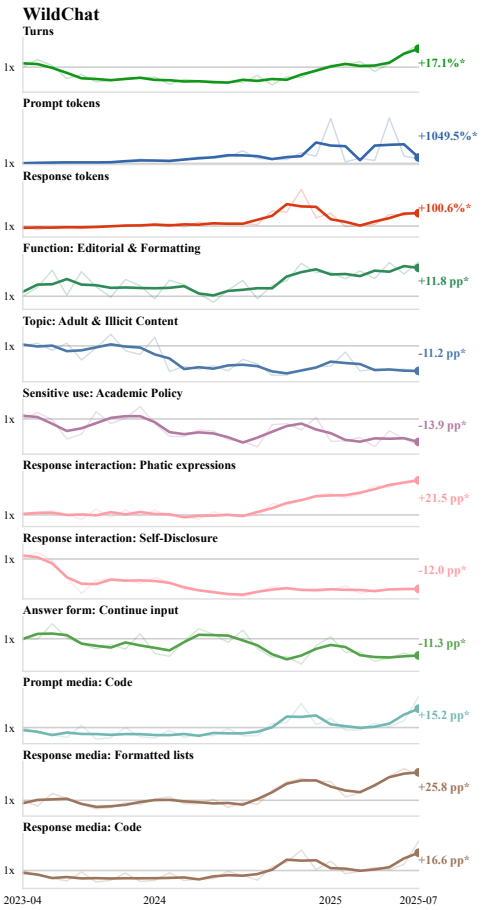


Figure 2: **WildChat shifts over time.** Monthly trajectories show temporal changes; pp denotes prevalence-point change and * marks FDR-significant trends.

Annotation pipeline. Our automatic annotation pipeline uses GPT-4.1 with JSON-formatted conversation histories and task-specific instructions, optimized against a human expert validation set independently double-annotated with disagreements adjudicated by a third annotator (Subsection D.13). Across nine taxonomy families, the F1 scores ranged from 0.756 to 0.942 (median 0.856). Full agreement metrics and taxonomy instructions are in Table 4 and section D. The validation set comprises 120 conversations (594 turns) annotated by 10 human experts, with automatic annotation of the full validation set incurring a total cost of $182.

Statistical testing. We estimate whether feature rates show statistically significant differences between slices of the data, sources, over time, or between models. We estimate these feature differences with regression-based feature screens, using heteroskedasticity-robust standard errors and controlling the false discovery rate within each dataset via Benjamini–Hochberg [35–37]. We use this framework because it yields comparable effect sizes across many sparse categorical and continuous features while accounting for unequal sample sizes and temporal composition. Full implementation details are given in Subsection D.16.

Sampling. To enable temporal analyses (Subsection 3.3), we construct a stratified temporal sample for each dataset by sampling uniformly across months where timestamps are available. We also record monthly population counts so that estimates can be reweighted back to the true source distribution when reporting pooled or population-representative rates. When user IDs are available, we separately construct a user sample by filtering users with $10 \leq K \leq 200$ conversations to avoid unengaged or automatic users. We then sample up to 40 conversations from each of up to 300 users. This user sample supports user/persona-cluster analyses (Subsection 3.4), separate from the conversation-level population summaries. We adopt this stratified design rather than annotating full source populations to make annotation tractable at the scale required for cross-source comparison. Our total annotation cost is $5,680 ($3,904 for WildChat, $329 for ShareGPT, $627 for AI Archive, $466 for Grok, $176 for LMSYS-Chat-1M, and $177 for Chatbot Arena). Additional sampling details are in Subsection C.1, and Table 1 summarizes the annotated samples and source context for each dataset.

3 Real AI use in conversation data

Real use of AI systems is poorly understood. Public knowledge is heavily reliant on only a couple major sources that are proprietary, and that have limited methodological transparency or depth of analysis [2, 8, 9]. We begin by asking how the range of real uses we collect vary. And how features vary across sources (Subsection 3.1), time (Subsection 3.3), model variants (Subsection 3.3), and users (Subsection 3.4). We also expose systematic differences in how the Anthropic Economic Index [9], which applies an occupational filter on top of the Clio tool [8], reports uses, compared to our cluster of real-use sources.

3.1 How real-use sources vary

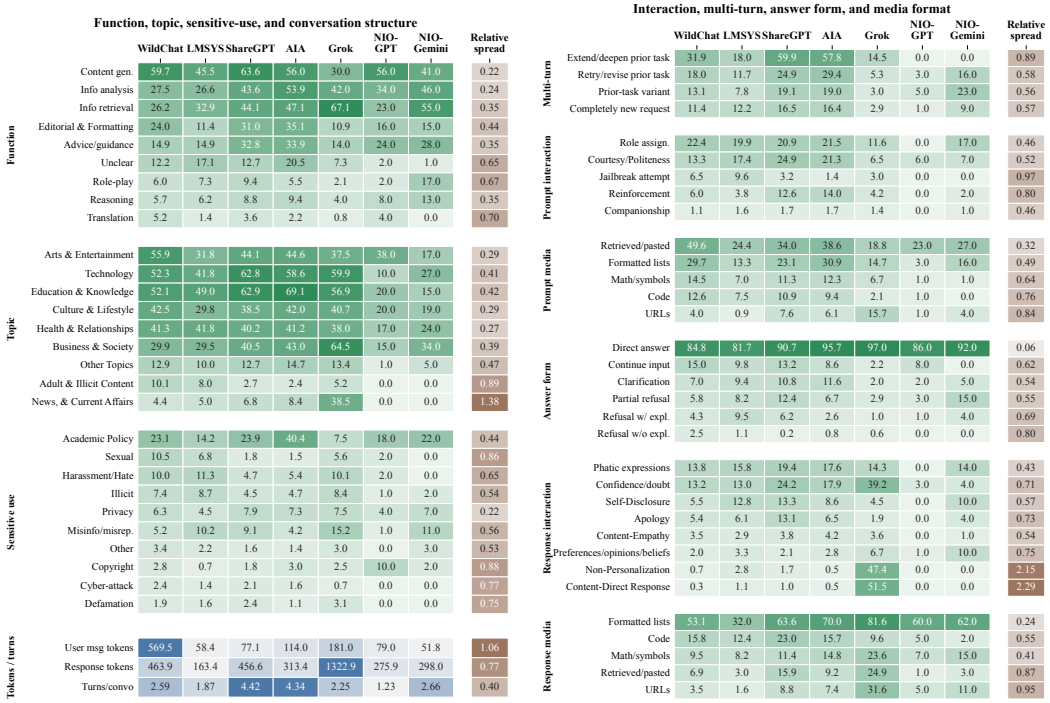
Conversation datasets are often analyzed and interpreted beyond their original collection context—assumed to reflect broad real-world use. We test whether this generalization is warranted by comparing our seven sources that differ in collection mechanism, hosting platform, time period, model availability, privacy constraints, and user self-selection along our annotation dimensions. All seven are opt-in or user-shared, so they are neither representative of AI use overall nor exchangeable with each other, and this design cannot attribute an observed difference to any one of these properties. We therefore interpret the comparisons below as evidence of measurement sensitivity, showing that source choice changes the observed distribution, rather than as estimates of how much any single source property matters. Subsection 3.3 provides evidence that this divergence is not purely a sampling artifact.

Real-use sources diverge on function and topic distribution. How AI systems are used, and the topics of conversation, diverge significantly across sources. Figure 3(a) shows: Grok and NIO-Gemini leans heavily toward information retrieval (67.1% and 55.0% versus 26.2% in WildChat), AI Archive toward information analysis (53.9%), ShareGPT toward content generation (63.6%) and editorial/formatting (31.0%). Interestingly, Grok and NIO are the two notable outliers, in different ways. Grok heavily prioritized for News & Current Affairs (38.5%) and Business & Society (64.5%), likely due to the public and social nature of its creation, whereas NIO is more distributed and sparser across uses and topics (as well as from shorter mean conversations). These differences may reflect a combination of temporal, model choice, or other source selection factors, which we cannot separate here, so we do not attribute them to platform. The important observation is their magnitude, i.e., which source a study draws on substantially changes the observed picture of use.

Sensitive-use prevalence varies, so any single source misestimates specific risks. Academic-policy issues (such as probable cheating on homework) range from 23.1% (WildChat) to 40.4% (AI Archive), and misinformation concentrates in Grok (15.3% versus 4.2–10.2% elsewhere), see Figure 3. Lower-prevalence categories like sexual content, harassment/hate, illicit content, privacy, copyright, cyber-attacks, and defamation show high relative spread (0.51–0.99), indicating that risk concentrations differ sharply by source.[1]

Prompt length, response length, and turn count differ significantly across sources. Conversation structure varies as much as topical content, with WildChat showing much longer prompts (569.5

[1] Since these sources consist of voluntarily contributed or opt-in conversations, we suspect prevalence rates of sensitive-use categories are even higher in the overall AI user population. See Section E.



(a) Function, topic, sensitive use, & structure. (b) Interaction, multi-turn, answer form, & media.

Figure 3: **Real-usage datasets often show significant distributional differences.** Cells report the conversation-level, multi-label incidence rate: the percentage of conversations in each source where that label appears at least once. Token and turn rows report raw means. The Relative Spread column is the standard deviation across sources divided by the row mean, floored at 1.0 for percentage rows.

tokens versus 51.8–181.0 elsewhere), Grok showing much longer responses (1322.9 tokens), and ShareGPT and AI Archive carrying more turns per conversation (4.42 and 4.34) than the others (Figure 3a). We return to structural variation in Subsection 3.3, where temporal and model-stratified analyses suggest that much of this spread is driven by collection period and model mix rather than source-intrinsic properties.

3.2 Comparison to occupation-only frames

Anthropic's Economic Index is the most well known, cited, and detailed release of AI uses [9]. The March 27, 2025 Index first filters for Claude.ai conversations that are relevant to occupations, and then uses Clio to map these conversations to occupational task clusters [8, 9], providing one of the most prominent public accounts of what AI is used for. We reimplement that occupational gate from Anthropic's released system prompts and taxonomy and apply it to our sources to observe what an occupational frame captures, and what it filters out before any task distribution is computed. Results in Figure 4 compare the occupational task mix in our pooled real-use data against Anthropic's initially published Economic Index distribution and characterize the conversations classified as non-occupational. Note that later versions of the Index do not apply the occupational filter, and find similar use distributions, which underscores that our differences might be due to a mix of product, temporal, and source shifts.

The occupational filter excludes roughly half of the AI use we observe before task mix is measured. We run the gate on the six sources whose raw conversations we can process; NIO is excluded because its data-use agreement permits only pre-agreed aggregate annotations to leave the enclave, so the pipeline cannot be applied there. Across these sources, 47.9% of conversations are classified as non-occupational (Figure 4, left). At minimum, 34.2% non-occupational for AI Archive, and at maximum, 61.9% in LMSYS. The rate varies with model, platform, and collection period. We do not claim it as an estimate for Claude.ai traffic or for representative AI use overall. What the

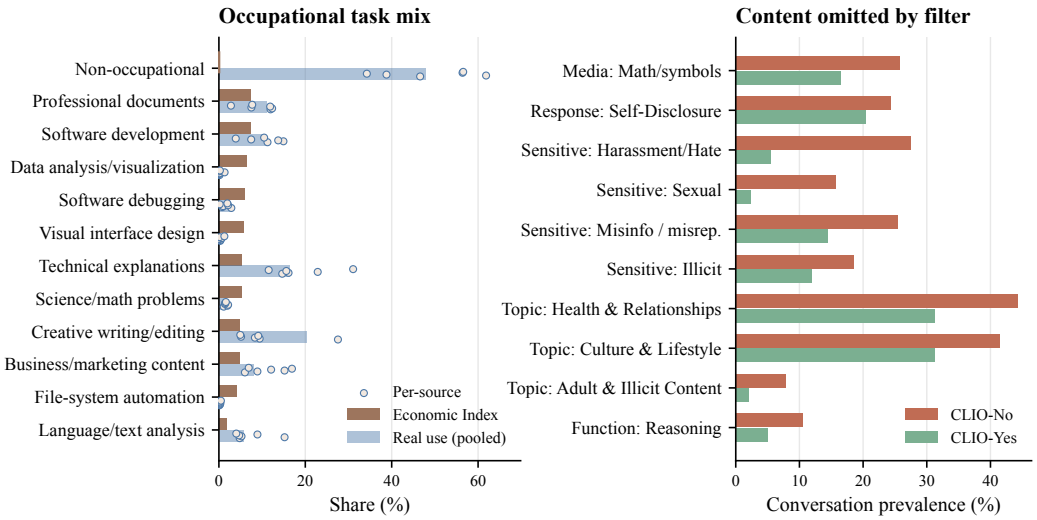


Figure 4: **An occupational lens both reshapes and neglects certain facets of AI use.** *Left:* bars compare the Economic Index's published occupation-use distribution, using O*NET [38], with our pooled real-use sources. The non-occupational row uses all conversations as the denominator and is zero for the Economic Index by construction. Bar dots show source-level real-use values. *Right:* rows show the FDR-significant features that are most overrepresented among conversations that the Economic Index's gate removes.

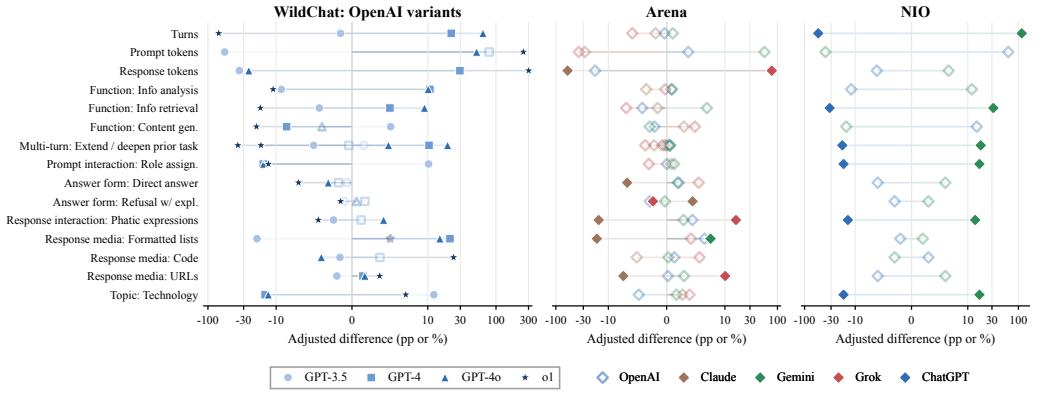


Figure 5: **Observed AI uses differ significantly by model variant, even within identical source and time periods.** We selected conversation features with the most large or FDR-significant model heterogeneity. Points show adjusted differences from other eligible models in the same source-month when timestamps are available, or within the same annotation sample for NIO; pp denotes prevalence-point differences and % denotes mean token/turn differences. Color indicates provider, marker shape distinguishes model families, and filled markers are FDR-significant.

within-source spread does establish is that the gate is not a neutral preprocessing step; the differences suggest occupational filters can shift the narrative around how AI is being used quite significantly.

Conversations outside the occupational frame are concentrated in personal, cultural, and safety-relevant use. Holding source and annotations fixed and splitting on the gate alone, the removed conversations are much more likely to involve health and relationships (44.2% versus 31.2%), adult or illicit topics (7.9% versus 2.1%), harassment/hate (27.5% versus 5.6%), and sexual content (15.7% versus 2.4%) (Figure 4, right). These contrasts are properties of the gate applied to our sources. They indicate that an occupational frame such as the Economic Index's may give a sharp view of labor-market-facing use, while systematically setting aside a set of consequential AI interactions

occurring outside that frame, suggesting that public infrastructure such as the AI Observatory provides a broader perspective into evolving human-AI interactions.

3.3 Observed AI use across time and model variants

Cross-source variation can reflect platform differences, model differences, and/or temporal differences. We disentangle these by holding the source fixed and asking whether observed use changes within a single dataset over time, or across model variants. WildChat is the only source with a multi-year timestamped horizon, so we report the temporal results below based on this source only; WildChat, Chatbot Arena, and NIO record model identity per conversation and support model-stratified comparisons. Within-source variation is substantial along both dimensions, indicating that collection period and model mix significantly affect what real-use measurement captures, with the sharing population held constant.

Temporal shifts: WildChat conversations have grown substantially more elaborate over time. Between April 2023 and July 2025, mean prompt tokens rose by 1,049.5%, mean response tokens by 100.6%, and turns per conversation by 17.1% (Figure 2). Later WildChat conversations are longer, more developed, and contain enough context for richer task labels and response formatting; a static WildChat sample therefore mixes multiple temporal regimes. The interaction style also changes: phatic expressions rise by 21.5 points, while self-disclosure falls by 12.0 points. Formatting and code increase as well, with formatted lists up 25.8 points, and response code up 16.6 points. These interaction changes are consistent with evolving expectations and interfaces between humans and chatbots, though more sources with a temporal dimension are necessary to confirm this pattern more broadly.

Model variations: OpenAI variants within WildChat support distinct usage regimes rather than interchangeable ChatGPT behavior. In short, the model causes users to treat assistants differently: some as a quick-answer tool, an iterative collaborator, or a high-effort technical solver. Our findings provide the most detailed clear evidence of users' inter-model specialization, even within a developer.

Differences between model developers emerge mostly in interaction style and continuation behavior. In Arena, Gemini and Grok produce much longer responses than the pool (+81.0% and +78.0%), while Claude is much shorter (-59.8%); Grok is also more phatic and URL-heavy, whereas Claude shows fewer formatted lists, and more explanation-bearing refusals. In NIO, these differences extend beyond response style into user behavior: Gemini conversations are more than twice as long as ChatGPT conversations, more retrieval-oriented, and more likely to deepen, or retry a prior task, with several contrasts remaining FDR-significant. This matters because it shows that model developers shape not just outputs, but distinct interaction regimes that influence how users continue and structure conversations.

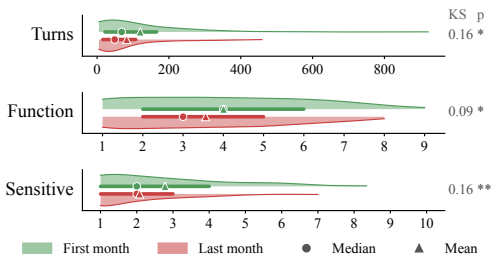


Figure 6: **Within-user shifts from first to last active month.** Ridge half-violin plots compare each user's first (green, top) and last (red, bottom) active month, across prompt tokens, annotated turns, and features. KS statistics and Wilcoxon paired-test significance are shown at right (** $p < 0.01$, * $p < 0.05$).

3.4 Returning users cluster into distinct personas and narrow their use over time

Source-, time-, and model-level variations tell us how aggregate use shifts, but they tell us little of how individual users may vary. This is a critical gap, thus far neglected in AI use analyses, since actual user tendencies and evolution can be masked by aggregate trends. We address this by focusing on returning WildChat users, the only one of our seven sources with both stable user identifiers and sufficient temporal history, both of which are necessary for tracking a user's behavior over time. We characterize cross-user variation by clustering users into behavioral personas (implementation details at Subsection D.15) and within-user variation by comparing each user's first and last active month.

Returning users narrow their use over time. Among returning WildChat users, three dimensions shift significantly from the first to the last active month (Figure 6): turns per window compress (KS 0.16, $p < 0.05$), distinct function labels drop from 4 to 3 (KS 0.09, $p < 0.05$), and distinct sensitive-use flags drop from 2 to 1 (KS 0.16, $p < 0.01$); prompt tokens, media format, interaction features, multi-turn relationship, and topic spread show no significant change. Returning users therefore do not write longer prompts or engage more heavily over time, so the aggregate WildChat expansion in Subsection 3.3 is driven by user-composition shifts rather than within-user intensification. This suggests individual users may specialize their interests over time.

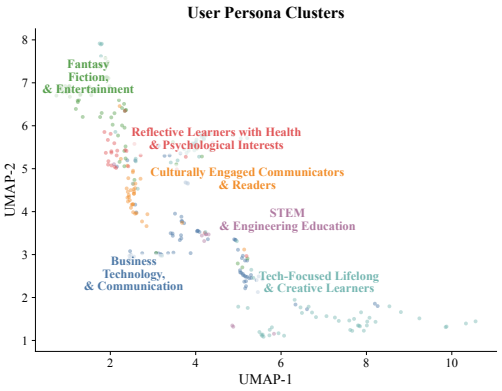


Figure 7: **UMAP projection of user clusters.**

Cross-user variation resolves into six recurring personas with distinct behavioral signatures. Clustering returning users by topic prevalence identifies 24 personas grouped into six parent groups (Figure 7; methodology in Subsection D.15), each with a distinctive profile relative to the dataset baseline (Figure 8). For instance: *Fantasy, Fiction, and Entertainment Enthusiasts* concentrate role-play (+10.4 pp), with frequent sexual content (+29.5 pp), and jailbreak attempts (+20.0 pp). *Tech-Focused Lifelong and Creative Learners* are marked by code in responses (+27.4 pp) and prompts (+14.5 pp) alongside more privacy-related turns (+14.8 pp). Understanding unique user groups like this helps paint a more thorough picture of user interest specialization.

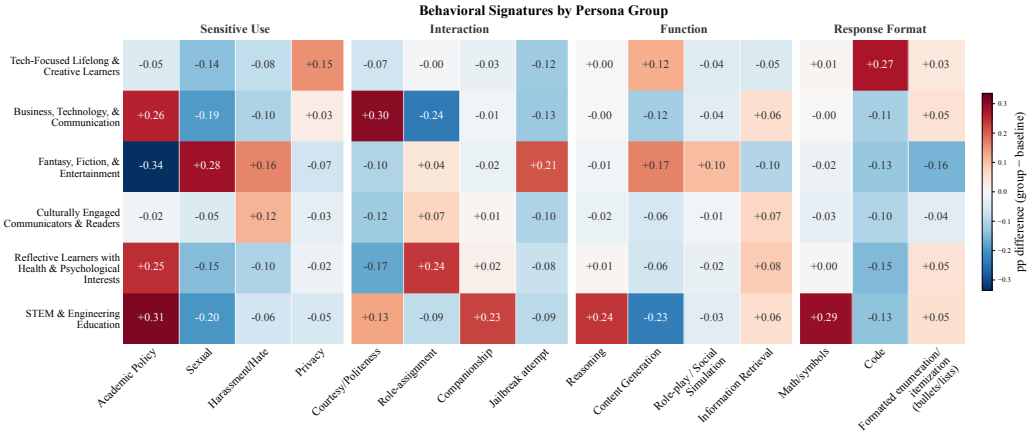


Figure 8: **Behavioral signatures by user persona group.** Each row of the heatmap is a parent persona group (ordered by size); columns are the most discriminative labels across four annotation families (sensitive-use flags, prompt interaction features, function purpose, and response media format). Cell values show weighted-mean difference (group share minus dataset baseline).

4 Discussion

Observed AI use is sensitive enough to source choice that it changes the story. A central empirical result of this paper is that observed AI use is neither narrow nor distributionally interchangeable across public sources. Across the Observatory's seven-source corpus, we find a shared assistant-use core, but also substantial variation in functions, topics, sensitive uses, interaction styles, and conversation structure. Some of this variation is intuitive: different platforms attract different users and incentives for sharing. Because all of our sources are opt-in, we cannot identify which property drives any particular difference, but the magnitude matters regardless. Source choice changes not only which tasks appear most common, but also how much personal, social, cultural, and safety-relevant interaction becomes visible. Many current narratives about "what AI is used for" are therefore more source-dependent than they appear.

Subtle measurement choices greatly impact interpretations of real use—proprietary reports offer incomplete perspectives. Our comparison to the Economic Index's occupational gate makes this especially clear. Occupational analyses recover an important slice of assistant use, but the gate is itself a measurement choice with a large effect: nearly half of the conversations in our pooled corpus are classified as non-occupational, and these omitted conversations are not random residue. They are enriched for personal, relational, cultural, and several safety-relevant forms of interaction. This matters because a framework can look comprehensive while still systematically filtering out uses that are socially important, policy-relevant, or disproportionately tied to harm. Real-use measurement should therefore be treated as a problem of representational scope, not merely one of scale.

Observed use is not static: it shifts with time, models, and surrounding interaction regimes. In WildChat, conversations become longer, more iterative, and more structurally elaborate over time; model-stratified analyses likewise show that even closely related model variants support meaningfully different usage regimes. These are not cosmetic differences. They alter the prevalence of retrieval, formatting, direct answering, refusal, code-related use, and social interaction features. Any static account of AI use will therefore age quickly, and model comparisons that ignore interface, developer, or deployment context risk mistaking properties of the surrounding interaction regime for properties of the model itself.

Future evaluation, safety analysis, and policy work should begin from observed usage regimes rather than abstract averages. Benchmark designers should be explicit about which parts of real use their evaluations are intended to approximate, instead of assuming that one dataset or one task family stands in for deployment broadly. Safety researchers should expect risk exposure to vary across sources, model families, and time, and should be cautious about treating prevalence estimates from any single corpus as universal. Policy analysts should likewise distinguish between proprietary summaries, public source slices, and broader claims about real-world use. The point of the AI Observatory is not to offer a final map of AI use, but to make this kind of measurement cumulative, comparable, and updateable as systems, interfaces, and user behavior continue to change.

Acknowledgements

We would like to acknowlege Deb Raji and the National Internet Observatory team for their advice, feedback, and support throughout this project. We would also like to thank Kun Cao and Ali Ajam for contributing conversation logs from aiarchives.org.

AR and SK are partially supported by NSF 2046795, 2205329, and 2504264, NIH, ARPA-H, the MacArthur Foundation, Good Ventures, Schmidt Sciences, the Hasso Plattner Förderstiftung, and Stanford HAI. AR is further supported by the Stanford Interdisciplinary Graduate Fellowship. YD acknowledges support from an NSERC Discovery Grant (RGPIN-2024-05178). CJ is supported by the Amazon AI PhD Fellowship. VO is supported in part by the MacArthur Foundation and the Heising-Simons Foundation.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2234660 (MR). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) with a grant funded by the Ministry of Science and ICT (MSIT) of the Republic of Korea in connection with the Global AI Frontier Lab International Collaborative Research.

References

[1] Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use ChatGPT. Technical Report 34255, National Bureau of Economic Research, sep 2025. URL https://www.nber.org/papers/w34255.

[2] OpenAI. Unlocking economic opportunity: A first look at ChatGPT-powered productivity. https://cdn.openai.com/global-affairs/be0fe9e0-eb97-43d1-9614-99f2bd948bcc/OpenAI_Productivity-Note_Jul-2025.pdf, jul 2025. Productivity note.

[3] Colleen McClain, Monica Anderson, Olivia Sidoti, and William Bishop. How teens use and view AI. https://www.pewresearch.org/internet/2026/02/24/how-teens-use-and-view-ai/, feb 2026. Pew Research Center report.

[4] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-chat-1m: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BOfDKxfwt0.

[5] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

[6] Shreyasi Roy Chowdhury and Kiran Garimella. How people use ChatGPT: Conversation-level evidence from india, nigeria, brazil and pakistan. https://gvrkiran.github.io/content/How_people_use_ChatGPT.pdf, 2026. Manuscript.

[7] Katelyn Xiaoying Mei, Robert Wolfe, Nicholas Weber, and Martin Saveski. Grok in the wild: Characterizing the roles and uses of large language models on social media, 2026. URL https://arxiv.org/abs/2602.11286.

[8] 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, Jared Kaplan, and Deep Ganguli. Clio: Privacy-preserving insights into real-world ai use, 2024. URL https://arxiv.org/abs/2412.13678.

[9] 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. Which economic tasks are performed with ai? evidence from millions of claude conversations, 2025. URL https://arxiv.org/abs/2503.04761.

[10] Kiran Tomlinson, Sonia Jaffe, Will Wang, Scott Counts, and Siddharth Suri. Working with ai: Measuring the applicability of generative ai to occupations, 2025. URL https://arxiv.org/abs/2507.07935.

[11] Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. *arXiv preprint arXiv:2407.11438*, 2024.

[12] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery, 2024. doi: 10.1145/3613904.3642385.

[13] Bright Huo, Amy Boyle, Nana Marfo, Wimonchat Tangamornsuksan, Jeremy P. Steen, Tyler McKechnie, Yung Lee, Julio Mayol, Stavros A. Antoniou, Arun James Thirunavukarasu, Stephanie Sanger, Karim Ramji, and Gordon Guyatt. Large language models for chatbot health advice studies: A systematic review. *JAMA Network Open*, 8(2):e2457879–e2457879, 02 2025. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2024.57879. URL https://doi.org/10.1001/jamanetworkopen.2024.57879.

[14] Inyoung Cheong, King Xia, KJ Kevin Feng, Quan Ze Chen, and Amy X Zhang. (a) i am not a lawyer, but...: engaging legal experts towards responsible llm policies for legal advice. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pages 2454–2469, 2024.

[15] RyokoAI. ShareGPT52K. `https://huggingface.co/datasets/RyokoAI/ShareGPT52K`, 2023. Dataset.

[16] Kun Cao and Ali Ajam. A.i. archives: Save, share, and cite generative a.i., 2026. URL `https://aiarchives.org/`. Accessed: 2026-05-05.

[17] Rebecca Bellan. Thousands of Grok chats are now searchable on Google, August 2025. URL `https://techcrunch.com/2025/08/20/thousands-of-grok-chats-are-now-searchable-on-google/`. TechCrunch. Accessed: 2026-05-06.

[18] Katelyn Xiaoying Mei, Robert Wolfe, Nicholas Weber, and Martin Saveski. Grok in the wild: Characterizing the roles and uses of large language models on social media, 2026. URL `https://arxiv.org/abs/2602.11286`.

[19] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. Lmsys-chat-1m: A large-scale real-world llm conversation dataset, 2024. URL `https://arxiv.org/abs/2309.11998`.

[20] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL `https://arxiv.org/abs/2403.04132`.

[21] National Internet Observatory. National internet observatory, 2025. URL `https://nationalinternetobservatory.org/`. Accessed: 2025-01-27.

[22] Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild, 2024. URL `https://arxiv.org/abs/2406.04770`.

[23] Chirag Shah and Emily M. Bender. Situating search. In *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '22, pages 221–232, 2022. doi: 10.1145/3498366.3505816.

[24] Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. Question answering is a format; when is it useful? *CoRR*, abs/1909.11291, 2019. doi: 10.48550/arXiv.1909.11291.

[25] Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. Llms as research tools: A large scale survey of researchers' usage and perceptions. *CoRR*, abs/2411.05025, 2024. doi: 10.48550/arXiv.2411.05025.

[26] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, Deb Roy, and Sara Hooker. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *CoRR*, abs/2310.16787, 2023. doi: 10.48550/arXiv.2310.16787.

[27] Kevin Klyman. Acceptable use policies for foundation models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 752–767, 2024. doi: 10.1609/aies.v7i1.31677.

[28] Verena Rieser and Johanna Moore. Implications for generating clarification requests in task-oriented dialogues. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 239–246, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219870. URL https://aclanthology.org/P05-1030/.

[29] Gary Ren, Xiaochuan Ni, Manish Malik, and Qifa Ke. Conversational query understanding using sequence to sequence modeling. In *Proceedings of The Web Conference 2018*, pages 1715–1724, 2018. doi: 10.1145/3178876.3186081.

[30] Corby Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. Leading conversational search by suggesting useful questions. In *Proceedings of The Web Conference 2020*, pages 1160–1170, 2020. doi: 10.1145/3366423.3380140.

[31] Hyunwoo Kim, Yoonseo Choi, Taehyun Yang, Honggu Lee, Chaneon Park, Yongju Lee, Jin Young Kim, and Juho Kim. Using llms to investigate correlations of conversational follow-up queries with user satisfaction. In *CEUR Workshop Proceedings*, volume 3752, pages 66–91, 2024.

[32] Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models, 2025. URL https://arxiv.org/abs/2407.18418.

[33] Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. Mirages. on anthropomorphism in dialogue systems. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.290. URL https://aclanthology.org/2023.emnlp-main.290/.

[34] Lujain Ibrahim, Canfer Akbulut, Rasmi Elasmar, 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. *CoRR*, abs/2502.07077, 2025. doi: 10.48550/arXiv.2502.07077.

[35] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980. doi: 10.2307/1912934.

[36] James G. MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325, 1985. doi: 10.1016/0304-4076(85)90158-7.

[37] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[38] National Center for O*NET Development. O*NET OnLine. https://www.onetonline.org/, 2025. U.S. Department of Labor, Employment and Training Administration (USDOL/ETA). Licensed under CC BY 4.0.

[39] Craig Silverstein, Monika Henzinger, Hannes Marais, and Michael Moricz. Analysis of a very large web search engine query log. *ACM SIGIR Forum*, 33(1):6–12, 1999.

[40] Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, and Tefko Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001.

[41] Yuntian Deng, Wenting Zhao, Khyathi Chandu, Jack Hessel, Claire Cardie, and Yejin Choi. WildVis: Open source visualizer for million-scale chat logs in the wild. https://huggingface.co/spaces/allenai/WildVis, 2024. Visualization system for WildChat and related chat logs.

[42] Chuanyang Jin, Jing Xu, Bo Liu, Leitian Tao, Olga Golovneva, Tianmin Shu, Wenting Zhao, Xian Li, and Jason Weston. The era of real-world human interaction: Rl from user conversations. *arXiv preprint arXiv:2509.25137*, 2025.

[43] Chuanyang Jin, Binze Li, Haopeng Xie, Cathy Mengying Fang, Tianjian Li, Shayne Longpre, Hongxiang Gu, Maximillian Chen, and Tianmin Shu. Thoughttrace: Understanding user thoughts in real-world llm interactions. *arXiv preprint arXiv:2605.20087*, 2026.

[44] Marzyeh Ghassemi, Abeba Birhane, Mushtaq Bilal, Siddharth Kankaria, Claire Malone, Ethan Mollick, and Francisco Tustumi. ChatGPT one year on: who is using it, how and why? *Nature*, 624:39–41, 2023.

[45] Kristina McElheran, J. Frank Li, Erik Brynjolfsson, Zachary Kroff, Emin Dinlersoz, Lucia Foster, and Nikolas Zolas. AI adoption in america: Who, what, and where. *Journal of Economics & Management Strategy*, 2024. doi: 10.1111/jems.12576.

[46] Francesco Filippucci, Peter Gal, Cecilia Jona-Lasinio, Alvaro Leandro, and Giuseppe Nicoletti. The impact of artificial intelligence on productivity, distribution and growth: Key mechanisms, initial evidence and policy challenges. Technical report, Organisation for Economic Co-Operation and Development, April 2024. URL https://www.oecd.org/en/publications/the-impact-of-artificial-intelligence-on-productivity-distribution-and-growth_8d900037-en.html.

[47] 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, 2024. URL https://arxiv.org/abs/2404.14901.

[48] Qiwei Pang, Mengze Zhang, Kum Fai Yuen, and Mingjie Fang. When the winds of change blow: an empirical investigation of ChatGPT's usage behaviour. *Technology Analysis & Strategic Management*, 37:3113–3127, 2025.

[49] Natalie Grace Brigham, Chongjiu Gao, Tadayoshi Kohno, Franziska Roesner, and Niloofar Mireshghallah. Developing story: Case studies of generative AI's use in journalism, 2024. URL https://arxiv.org/abs/2406.13706.

[50] Victor Ojewale, Ro Encarnación, Suresh Venkatasubramanian, and Danaé Metaxa. Monitrllm: A community-centered evaluation infrastructure for large language models, 2026. URL https://arxiv.org/abs/2608.02409.

[51] Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, and Verena Rieser. Mirages: On anthropomorphism in dialogue systems, 2023. URL https://arxiv.org/abs/2305.09800.

[52] Lujain Ibrahim, Canfer Akbulut, Rasmi Elasmar, 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. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZAx4c4ZH5Y.

[53] Shayne Longpre, Stella Biderman, Alon Albalak, Hailey Schoelkopf, Daniel McDuff, Sayash Kapoor, Kevin Klyman, Kyle Lo, Gabriel Ilharco, Nay San, Maribeth Rauh, Aviya Skowron, Bertie Vidgen, Laura Weidinger, Arvind Narayanan, Victor Sanh, David Adelani, Percy Liang, Rishi Bommasani, Peter Henderson, Sasha Luccioni, Yacine Jernite, and Luca Soldaini. The responsible foundation model development cheatsheet: A review of tools & resources, 2025. URL https://arxiv.org/abs/2406.16746.

[54] Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, and Alexander D'Amour. Bias in language models: Beyond trick tests and toward ruted evaluation, 2025. URL https://arxiv.org/abs/2402.12649.

[55] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From live data to high-quality benchmarks: The arena-hard pipeline. https://lmsys.org/blog/2024-04-19-arena-hard/, 2024. LMSYS blog.

[56] Nahema Marchal, Rachel Xu, Rasmi Elasmar, Iason Gabriel, Beth Goldberg, and William Isaac. Generative ai misuse: A taxonomy of tactics and insights from real-world data. *arXiv preprint arXiv:2406.13843*, 2024.

[57] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. URL https://arxiv.org/abs/2402.04249.

[58] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. Jailbreakbench: an open robustness benchmark for jailbreaking large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

[59] Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.

[60] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1671–1685, 2024.

[61] Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=n5R6TvBVcX.

[62] Merlin Stein and Connor Dunlop. Safe beyond sale: post-deployment monitoring of AI. https://www.adalovelaceinstitute.org/blog/post-deployment-monitoring-of-ai/, jun 2024. Ada Lovelace Institute policy analysis.

[63] Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. Visibility into AI agents, 2024. URL https://arxiv.org/abs/2401.13138.

[64] Aylin Caliskan and Kristian Lum. Effective AI regulation requires understanding general-purpose AI. https://www.brookings.edu/articles/effective-ai-regulation-requires-understanding-general-purpose-ai/, 2024. Brookings commentary.

[65] Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, Kevin Klyman, Christopher Klamm, Hailey Schoelkopf, Nikhil Singh, Manuel Cherep, Ahmad Mustafa Anis, An Dinh, Caroline Chitongo, Da Yin, Damien Sileo, Deividas Mataciunas, Diganta Misra, Emad Alghamdi, Enrico Shippole, Jianguo Zhang, Joanna Materzynska, Kun Qian, Kush Tiwary, Lester Miranda, Manan Dey, Minnie Liang, Mohammed Hamdy, Niklas Muennighoff, Seonghyeon Ye, Seungone Kim, Shrestha Mohanty, Vipul Gupta, Vivek Sharma, Vu Minh Chien, Xuhui Zhou, Yizhi Li, Caiming Xiong, Luis Villa, Stella Biderman, Hanlin Li, Daphne Ippolito, Sara Hooker, Jad Kabbara, Sandy Pentland, and Data Provenance Initiative. Consent in crisis: the rapid decline of the ai data commons. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

[66] A.I. Archives. A.I. Archives | Save, Share, and Cite Generative A.I., 2025. URL https://aiarchives.org/.

[67] AI Archives. A.i. archives: Share claude, chatgpt, gemini, meta. https://chromewebstore.google.com/detail/ai-archives-share-claude/jagobfpimhagccjbkchfdilhejgfggna, 2025. Browser extension for user-submitted conversation sharing.

[68] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[69] LMSYS. lmsys/lmsys-chat-1m · Datasets at Hugging Face, June 2025. URL https://huggingface.co/datasets/lmsys/lmsys-chat-1m.

Appendix

Table of Contents

A Related Work

Studies on real uses of AI. Early web search log studies showed the value of large-scale behavioral traces for understanding how people seek information, but their evidence was limited to short, mostly single-turn search queries rather than open-ended assistant conversations [39, 40]. Recent public LLM conversation datasets such as LMSYS-Chat-1M and WildChat made real assistant interactions available for research at scale, enabling analyses and derived tools such as WildVis and WildBench [4, 5, 22, 41–43]. Public social-media deployments create another observational setting: Mei et al. study Grok calls on X and show that LLMs in public threads often act not only as information providers, but also as truth arbiters, advocates, adversaries, and dispute mediators [7]. Privacy-preserving analyses of proprietary logs, including Anthropic's Clio and Economic Index, OpenAI's NBER working paper on ChatGPT use, and a Bing Copilot occupational-use study, reveal broad use for coding, writing, education, information work, and non-work daily activity [1, 2, 8–10]. Conversation-level and survey-based studies add complementary detail, including donated ChatGPT exports from India, Nigeria, Brazil, and Pakistan and Pew's nationally representative survey of U.S. teen AI use [3, 6]. Survey, commentary, and econometric work similarly documents rapid but uneven AI adoption across firms, workers, and regions, while emphasizing uncertainty about productivity effects and risks of misuse or overstatement [44–46]. Compared with these studies, our work aggregates multiple real-usage sources rather than relying on one platform, and applies a common taxonomy to compare temporal, geographic, topical, sensitive-use, response, and conversation-structure patterns across datasets.

User–AI interaction behaviors and conversation dynamics. A complementary line of work studies how people use LLMs within particular practices and social contexts, such as software engineering, journalism, and changing everyday usage behavior [47–50]. These studies show that users often rely on assistants for debugging, drafting, summarization, and publication workflows, sometimes exposing confidential or copyrighted material to the model [47, 49]. Human-subject and log-based privacy studies find that users disclose sensitive personal information because of utility, convenience, flawed mental models, and human-like conversational affordances [11, 12]. Work on anthropomorphism argues that dialogue systems can invite human-like interpretations, and recent multi-turn evaluations show that relationship-building, empathy, validation, and other anthropomorphic behaviors often emerge only after several turns [51, 52]. Our analyses extend this interactional perspective by measuring multi-turn relationships, assistant response behaviors, and user clusters across large real-usage corpora rather than a single domain or controlled study.

Benchmark context and downstream Use. Real-use measurement is also relevant to evaluation design, because benchmarks built from selected prompts or synthetic tasks can only be interpreted against some account of deployment context [53, 54]. Benchmarks built from user prompts, including WildBench and Arena-Hard, move closer to real use but still convert selected interactions into evaluation items rather than characterizing the broader distribution of user needs [22, 55]. We treat benchmark auditing as a downstream application of the Observatory rather than the focus of this paper.

Sensitive uses, misuse, and post-deployment monitoring. Safety research has produced taxonomies and benchmarks for misuse, jailbreaks, over-refusal, and harmful requests, often using curated policies or adversarial prompts as the starting point [56–59]. In-the-wild studies add an empirical layer by characterizing jailbreak communities, mining adversarial examples from WildChat and LMSYS, and measuring personal disclosures in real human–LLM conversations [11, 60, 61]. Policy and governance work argues that post-deployment monitoring, visibility into AI agents, and empirical understanding of general-purpose AI use are necessary for effective oversight once systems are deployed at scale [62–64]. Other work highlights risks in the data commons itself, including consent, sexual content, and copyright concerns in user-shared AI conversation data [65]. Our study contributes to this monitoring agenda by estimating sensitive-use prevalence across multiple real-usage datasets with a unified taxonomy and by linking these categories to broader usage, user, temporal, model-specific, and interactional patterns.

B Additional Results & Findings

B.1 Developer & model differences

Developer differences are visible but weaker and more context-dependent. In Arena, the developer differences emerged in assistant responses. Gemini and Grok produce much longer responses (+81.0% and +78.0%, respectively), while Claude produces shorter responses (-59.8%). Grok is also more phatic (+16.2 points) and URL-heavy (+10.1 points), suggesting a more socially styled and web-referential response mode in this setting. Claude moves in the opposite direction on several style rows compared to the mean: fewer formatted lists (-16.5 points), fewer phatic expressions (-15.3 points), fewer direct answers (-6.8 points), and more refusals with explanation (+4.5 points). These developer contrasts are informative and point to differences between model developers but should be read more cautiously than the variant contrasts because Arena usage reflects a comparison interface.

The NIO split suggests that developer differences can also change user continuation behavior. In the NIO sample, Gemini conversations have more than twice as many turns as ChatGPT conversations (+116.3%), are more likely to be information retrieval (+32.0 points), and more often extend or deepen a prior task (+18%). NIO-Gemini also shows more role assignment, phatic expressions, and self-disclosure, while NIO-ChatGPT is more concentrated in arts and entertainment. The differences between turns, information retrieval, and deepen a prior task are FDR-significant indicating that user continuation behavior is impacted and developer difference may contribute to this.

C Real-use dataset details

C.1 Sampling design

Temporal sample. For each real-use dataset, we construct a general STRATIFIED TEMPORAL SAMPLE. When timestamps are available, we sample conversations uniformly within month and record the number of conversations in each month, enabling analyses to reweight from a balanced temporal sample back toward the source population. When timestamp metadata is unavailable, the corresponding sample is a random conversation sample.

User sample. Where user IDs are present, we also define a USER SAMPLE for user-level analyses. We filter to users with between $10 \leq K \leq 200$ conversations, targeting real, continued users rather than experimental accounts or automated processes. We then sample up to $U = 300$ users and up to $U_C = 40$ conversations per user. This sample is reserved for the user/persona-cluster analyses in Subsection 3.4.

C.2 WildChat

WildChat [5] is a dataset of in-the-wild user conversations with ChatGPT. The original release had about one million conversations, and the latest release has about 4.8 million conversations collected between April 9, 2023 to Jul 31, 2025. Users consented to the collection, use, and sharing of their conversation data in exchange for free access to ChatGPT model variants.

C.3 ShareGPT

ShareGPT is a public collection of user-shared ChatGPT conversations. We use a sampled subset of 1,000 conversations and 8,438 turns to align the source with the Observatory annotation pipeline. Since the public exports do not provide stable user identifiers, ShareGPT contributes to conversation-level and source-level analyses but not user-cluster analyses. Of the 1,000 sampled conversations, 968 are successfully annotated. The remaining 32 are intentionally excluded due to malformed or unparseable turn structures, including entries with no corresponding user turns.

C.4 National Internet Observatory

The National Internet Observatory (NIO) contributes a privacy-preserving, manually labeled sample of browser-instrumented AI assistant use [21]. Because the underlying conversations are highly private, we do not ingest or redistribute raw conversation objects for this source. Instead, the analysis uses the

released annotation caches for 100 ChatGPT conversations and 100 Gemini conversations. We pool these annotations as NIO for source-balanced distributional analyses and keep the ChatGPT/Gemini split for source-display and model-differentiation analyses.

C.5 AI Archive

AI Archive is an opt-in, public repository that curates user-shared conversations with popular LLM systems (e.g., Claude, ChatGPT, Gemini and Meta) for the purposes of transparency, research, and education. Users voluntarily contribute chats via a Chrome extension that lets them select and upload specific conversations for public archiving with shareable links and citation metadata.[2] In our work, we sample publicly posted items from the AI Archive website; because submissions are ongoing and user-driven, the corpus is rolling rather than bounded by a fixed date window. Our sampled snapshot contains 2,076 conversation records with timestamps ranging from 2023-02-21 to 2025-08-09.

C.6 LMSYS 1M

LMSYS-Chat-1M [4] is a one-million conversation dataset of real-world chats collected by the LMSYS team from the Vicuna demo and Chatbot Arena platform [68]. The dataset card reports that conversations span *April–August 2023*, cover *25 models*, and include metadata such as timestamps, language, turn counts, and anonymized identifiers [69]. Collection is opt-in via the Arena interface where users compare models and consent to share anonymized transcripts for research. We use the Hugging Face release [69] and preserve the dataset's model identifiers (e.g., "vicuna-13b," "gpt-3.5-turbo," etc.) as provided, which enables cross-model stratification in our analyses.

C.7 Grok

The Grok source consists of public shared Grok conversations collected from user-visible share links. Our annotated sample contains 1,234 conversations, 3,750 turns, and 505 user identifiers. We treat the dataset as a source-specific view of Grok usage rather than a representative sample of all Grok traffic because sharing and collection are user-selected. This source is complementary to recent work on Grok calls on X, which characterizes public social-media interactions with Grok rather than user-shared assistant conversations [7].

C.8 Chatbot Arena

Chatbot Arena records user interactions from side-by-side model comparison battles [68]. Our annotated sample contains 2,000 conversations and 2,526 turns. Arena is useful for model-stratified comparisons because model identities are recorded, but its interface is evaluative rather than ordinary single-assistant use, so we interpret Arena results as developer-level triangulation rather than a clean estimate of everyday usage.

C.9 Licensing

All datasets, tools, and code packages used in the paper are credited with their original citations, and licenses or terms-of-use governing each asset are documented in Table 3. We respect each source's terms and do not redistribute raw conversations: our release consists of derived annotations and a download-and-formatting script that obtains each source through its original access mechanism and joins it with our annotations, so users can reconstruct the annotated corpus locally. CC-BY attribution is provided for O*NET, the Anthropic Economic Index, and Chatbot Arena prompts. Annotations of Chatbot Arena model outputs are released under CC-BY-NC 4.0 in compliance with the source license; all other annotations are released under CC-BY 4.0, and the pipeline code under the MIT license.

D Annotation Taxonomy & Design

For all our conversation data, we annotate each message (up to 20 conversation turns) in each conversation. Each conversation is then annotated by a human or automatic annotation system, with

[2] See [66] and the Chrome Web Store listing [67] for the extension's description and permissions.

Table 3: **Existing assets used in this paper, with versions, licenses or terms of use, and how each is used.** We respect each asset's terms: gated or restricted-access sources are not redistributed, and derived annotations inherit upstream licenses where applicable (see Section C).

Asset	Version	License / Terms	Notes on Use
WildChat (unfiltered)	`allenai/WildChat-4.8M-Full` (Link)	ODC-BY 1.0 (gated)	Used under gated access; we do not redistribute raw conversations.
ShareGPT	`ShareGPT_Vicuna_unfiltered` (Link)	Apache-2.0	HTML-cleaned split used; Annotations only redistributed.
LMSYS-Chat-1M	`lmsys/lmsys-chat-1m` (Link)	LMSYS-Chat-1M Dataset License Agreement (custom; gated)	We do not redistribute raw conversations; we honor the right-to-request-deletion clause.
Chatbot Arena Conversations	`lmsys/chatbot_arena` (Link)	Prompts: CC-BY 4.0; Outputs: CC-BY-NC 4.0 (gated)	NC restriction propagates to response-side annotations; we release these under CC-BY-NC 4.0.
AI Archive	aiarchives.org (Link); Terms (Link)	Site Terms of Service; user "Share Publicly" consent	Used only publicly-shared conversations; raw conversations not redistributed.
Grok shared chats	Authors' crawl of public X posts, two collection sweeps: September 2025 and April 2026	xAI Terms of Service govern user content; conversations are user-posted-public	We do not redistribute raw conversations; aggregate annotations only.
National Internet Observatory (NIO)	nationalinternetobservatory.org (Link)	NIO Terms of Service; project-specific Data Use Agreement; IRB-equivalent review	Data-enclave access only; only pre-approved aggregate annotations leave the secure enclave.
O*NET	onetonline.org, DB v30.2 (Link); License (Link)	CC BY 4.0	Used for Economic Index reproduction; attribution provided.
Anthropic Economic Index	`Anthropic/EconomicIndex`, release_2025_03_27 (Link)	Data: CC-BY 4.0; Code: MIT	Used as comparison baseline; attribution provided.
GPT-4.1 (`gpt-4.1-2025-04-14`)	OpenAI API	OpenAI Terms of Use	Used for automatic annotation; outputs subject to OpenAI ToS.
`langdetect`	fedelopez77/langdetect (Link)	MIT	Used for prompt and response language detection.

visibility into the conversation history. The system prompt is shown at **??**. Several conversation properties are annotated for, at the user prompt-level, response-level, both prompt and response independently, and turn-level (prompt and response together). Here are the taxonomies that in turn list the properties they label for:

1. **Prompt & Response: Characters & Tokens**: This taxonomy captures basic structural metadata about conversations, including token counts, turn lengths, and the number of conversations per user, which are automatically extracted from the data. These metrics provide fundamental insights into conversation complexity and user engagement patterns. (Subsection D.3)

2. **Prompt & Response: Language**: This taxonomy identifies the language(s) used in prompts and responses, allowing us to understand the multilingual nature of conversations and analyze language-specific usage patterns. (Subsection D.4)

3. **Prompt & Response: Media Format**: This taxonomy categorizes the types of media content present in messages, including natural language, code, mathematical expressions, images, audio, and other formats, enabling analysis of multimodal interaction patterns. (Subsection D.5)
4. **Prompt: Function/Purpose**: This taxonomy characterizes the user's intent and the primary task type of each prompt, ranging from information retrieval and content generation to reasoning and role-play, derived from analysis of major LLM providers' use case documentation. (Subsection D.6)
5. **Prompt: Interaction Features**: This taxonomy labels social and interactive aspects of user prompts, such as role assignments, politeness, praise, and companionship, to understand how users relate to AI systems as social entities. (Subsection D.7)
6. **Prompt: Multi-turn Relationship**: This taxonomy describes how each user prompt relates to previous turns in the conversation, distinguishing between first requests, new topics, revisions, variations, and extensions of prior tasks. (Subsection D.8)
7. **Response: Answer Type**: This taxonomy categorizes how the system responds to user prompts, including direct answers, refusals, partial responses, continuations, and clarification requests. (Subsection D.9).
8. **Response: Interaction Features**: This taxonomy labels social and stylistic features in model responses, such as self-disclosure, expressions of preferences or empathy, apologies, and hedging language, to understand how systems present themselves. (Subsection D.10)
9. **Conversation Turn: Topic**: This taxonomy identifies the subject matter and domain of each conversation turn, covering a broad range of topics from mathematics and technology to entertainment, social issues, and sensitive content areas. (Subsection D.11)
10. **Conversation Turn: Sensitive Uses**: This taxonomy flags potentially harmful, inappropriate, or high-risk content and use cases, including violent content, illegal activities, private information, discrimination, and high-stakes decision-making, based on acceptable use policies from major AI providers. (Subsection D.12)

D.1 Taxonomy Development

Design goals and development process. Our taxonomy was designed to satisfy three goals simultaneously: breadth over the space of real-use interactions, comparability across heterogeneous sources, and reproducibility for future extension. We therefore did not induce labels from a single corpus or from one developer's framing of AI use. Instead, we grounded the taxonomy in publicly available use-case descriptions, benchmark and HCI literature, and developer policy documents, then refined the schema through five rounds of pilot annotation on stratified conversation batches drawn from multiple datasets, including WildChat, ShareGPT, AI Archive, LMSYS, and Grok. The development panel consisted of ten AI researchers with backgrounds spanning privacy, safety, security, and evaluation science. After each round, disagreements were reviewed collectively, label definitions were tightened, and labels that were redundant, too source-specific, or too ambiguous to support reliable annotation were merged or removed.

Taxonomy Development: Function/Purpose. To capture prevalent and intended uses of general-purpose assistants, we examined the top-ranked closed and open models on the HELM leaderboard in December 2024 and identified the major developers represented there.[3] We reviewed publicly advertised use cases, prompt libraries, and model cards from these developers to identify recurring classes of assistant use that might be underrepresented in any one public corpus. We supplemented this with prior work on real-user prompts and conversational tasks, including WildBench's use of real-user challenging prompts, work on conversational search and question answering, and recent studies of how researchers and other users employ LLMs in practice [22–25]. These sources seeded the function/purpose taxonomy, which we then refined during pilot annotation to distinguish recurring classes such as information retrieval, content generation, editing and formatting, information analysis, role-play, advice, and reasoning. To avoid overfitting the schema to early high-volume datasets, we retained function labels only when they either recurred across multiple sources or filled a theoretically important gap not captured by existing categories.

[3] `https://crfm.stanford.edu/helm/`

Taxonomy Development: Topic. Topic labels were designed to capture what conversations are substantively about, independently of what function they serve. We began from the Data Provenance Initiative's topic grouping efforts [26] and expanded them using prior HCI work on privacy-sensitive, socially situated, and identity-relevant LLM interactions [12]. During pilot rounds, we broadened the taxonomy to better capture domains that appeared repeatedly across sources but were not well handled by initial groupings, including lifestyle, relationships, law, politics, and fandom or fictional content. We retained topic labels only when they were either repeatedly observed across multiple sources or necessary to separate substantively distinct domains that otherwise collapsed into overly broad residual categories.

Taxonomy Development: Multi-Turn Relationship. The multi-turn relationship taxonomy was informed by prior work on conversational follow-up behavior, clarification, and query reformulation in dialogue and conversational search [28–31], but existing schemes were too search-specific or task-oriented for our setting. We therefore adapted these ideas into a compact user-centered taxonomy that distinguishes whether a new prompt starts a conversation, deepens a prior task, retries or revises it, introduces a related variant, or begins a completely new request. This allowed us to characterize continuation structure in general-purpose assistant conversations without assuming that all follow-up turns are mere search reformulations.

Taxonomy Development: Sensitive Use. The sensitive-use taxonomy was initially seeded from acceptable-use and restricted-use policies published by major model developers, including OpenAI, Anthropic, Google, Cohere, Mistral, Databricks, and Meta Llama, following prior work showing that these policies encode a broad and fragmented set of developer-perceived risks [27]. Early pilots explored a much finer-grained harm and disclosure space, including separate labels for self-disclosure, third-party disclosure, doxxing, entity disclosure, and several harm types (e.g., physical, economic, reputational, psychological, autonomy, and discrimination harms), as well as distinct prompt- versus response-side PII categories. In practice, these categories were often overlapping or difficult to distinguish reliably from short conversational excerpts, so we consolidated them into a smaller set of parent sensitive-use flags centered on recurring high-stakes domains, disclosure risks, and policy-salient forms of misuse. This consolidation improved consistency while preserving the safety-relevant distinctions that appeared robustly across sources.

Taxonomy Development: Interaction, Response, and Media. Interaction-behavior and response-form labels were developed from both observed conversational phenomena and prior literature on answer formats, abstention/refusal, and anthropomorphic dialogue behavior [24, 32–34]. This led us to distinguish, for example, direct answers, continuations, clarification requests, explicit refusals, phatic or relationship-building responses, self-disclosure, and role assignment. Media-format categories were the least theory-driven component: they were constructed bottom-up from recurring structures visible in prompts and responses, such as code, math, formatted lists, URLs, HTML, and likely pasted or retrieved content. These categories were intentionally kept simple and surface-oriented so that they would generalize across datasets and support reliable automatic annotation.

D.2 Automatic Annotation Quality

We optimized the automatic annotation pipeline against a human expert validation set, varying the annotation model, conversation serialization format, instruction order, and whether conversation history was included. GPT-4.1 with JSON-formatted conversation histories provided the best overall tradeoff between accuracy and coverage. Table 5 and Table 4 reports fine-grained and coarse-grained agreement metrics by taxonomy family respectively; the full pipeline uses task-specific prompts and schema-constrained outputs so that downstream analyses can treat labels consistently across sources. Because the same authors developed the taxonomy, produced the validation labels, and selected the pipeline, these scores measure fidelity to our own annotation scheme; no external or blind annotation was collected, which we note as a limitation. For fine-grained annotation, agreement is measured by exact match against the taxonomy labels; for coarse-grained annotation, labels are first mapped to their parent category in the hierarchy (e.g., "Content Generation (code)" and "Content Generation (creative/fiction writing)" both collapse to "Content Generation") before agreement is computed.

Automatic annotations are generated using `gpt-4.1-2025-04-14` with a sampling temperature of 1.0, top-p of 1.0, and a maximum of 32,768 output tokens. To handle transient failures, we apply retry

Taxonomy Option	Label Agreement	Cohen's κ	Krippendorff's α	Jaccard	F1
Prompt: Media Format	0.747	0.565	0.589	0.892	0.927
Prompt: Function/Purpose	0.589	0.525	0.593	0.715	0.756
Prompt: Interaction Features	0.897	0.672	0.739	0.931	0.942
Prompt: Multi-turn Relationship	0.829	0.770	0.770	0.829	0.829
Response: Media Format	0.774	0.652	0.682	0.896	0.931
Response: Answer Format	0.856	0.654	0.655	0.856	0.856
Response: Interaction Features	0.829	0.538	0.600	0.867	0.879
Turn: Topic	0.403	0.367	0.527	0.721	0.804
Turn: Sensitive Uses	0.774	0.586	0.595	0.784	0.798

Table 4: **Automatic-to-human-consensus agreement at the parent-category level, per taxonomy family.** Scores compare the GPT-4.1 production pipeline against the adjudicated human consensus on the 120-conversation validation set. We use the parents in the categorical hierarchies for Function, Topic, and Sensitive Use. All comparisons in the main text are computed at this level. Jaccard: Jaccard similarity; F1: F1 score. Metric scores for Prompt: Function/Purpose, Turn: Topic, and Turn: Sensitive Uses increase from fine-grained annotation (Table 5).

Taxonomy Option	Label Agreement	Cohen's κ	Krippendorff's α	Jaccard	F1
Prompt: Media Format	0.747	0.565	0.589	0.892	0.927
Prompt: Function/Purpose	0.445	0.425	0.508	0.594	0.648
Prompt: Interaction Features	0.897	0.672	0.739	0.931	0.942
Prompt: Multi-turn Relationship	0.829	0.770	0.770	0.829	0.829
Response: Media Format	0.774	0.652	0.682	0.896	0.931
Response: Answer Format	0.856	0.654	0.655	0.856	0.856
Response: Interaction Features	0.829	0.538	0.600	0.867	0.879
Turn: Topic	0.285	0.256	0.446	0.659	0.759
Turn: Sensitive Uses	0.764	0.574	0.591	0.784	0.791

Table 5: **Automatic-to-human-consensus agreement at the fine-grained label level, per taxonomy family.** Jaccard: Jaccard similarity; F1: F1 score.

logic with up to 8 attempts and exponential backoff starting at a 5-second base delay, doubling per attempt. For taxonomy label validation, the annotation parser checks that (1) the response is a valid JSON list, (2) each item follows the expected dictionary format, and (3) each label matches one of the predefined taxonomy options. Across the two GPT-4.1 validation runs, all returned annotations passed the schema check, with 2 expected outputs absent (about 0.03%). This near-complete coverage makes selective missingness an unlikely explanation for the sensitive-use results. Retry counts are not persisted by the pipeline, so we report only this terminal check. All annotations are run locally (8-core CPU, 8GB RAM).

Second annotator from a different model family. To test whether the labels depend on one developer's model, and in particular on its safety policy, we re-annotated the full validation set with Claude Sonnet 4.6, holding the taxonomy definitions, task instructions, conversation-history and JSON formatting, temperature, and top-p fixed. Against the human consensus, parent-category F1 ranges from 0.689 to 0.932 across the nine families (median 0.834), compared with 0.756 to 0.942 (median 0.856) for GPT-4.1. Topic agreement is 0.828 and Sensitive Uses 0.689. The score for Sensitive Uses is expected to diverge between the two annotators: positive labels in this family are sparse, which makes F1 moderately unstable for both models, and its label boundaries are the ones that are most exposed to potential differences in developer safety policy. All Claude outputs were returned and passed enumerated-taxonomy validation.

Sampling temperature. Neither temperature 0 nor a fixed seed guarantees determinism through the API, so we measured configuration sensitivity rather than assuming any setting is deterministic. Rerunning the validation set at temperature 0 against the temperature-1 production outputs (same prompts, taxonomy, and model snapshot), parent-category agreement is high: micro-F1 ranges 0.761 to 0.993 (median 0.947) and macro Jaccard 0.900 to 0.993, with median absolute prevalence shift at or below 0.5 pp per family and a largest single-label shift of 4.75 pp (Response Interaction Features). Sensitive Uses has a micro-F1 score of 0.761 once jointly-empty turns are set aside, consistent with the sparsity of positive labels. The parent-category measurements used in our main comparisons are therefore not an artifact of one sampling temperature.

Automatic Annotation System Prompt

```
<START CONVERSATION>
{Previous Conversation Block}
<START CURRENT TURN>
{Current Conversation Block}
<END CURRENT TURN>
<END CONVERSATION>
```

You are a high-quality annotation assistant. Your task is to annotate conversation logs between users and AI chatbots. You will be given a specific task description, all possible label options for the task, and a part of the conversation, including the user prompt and model response from both previous and current turns. These might be pulled from any part of a multi-turn conversation. As a high-quality annotator you will diligently provide annotations on the current turn that are:

1. **Comprehensive:** You will list all relevant annotations as the tasks can be multi-class (only one label is true) or multi-label (multiple categories can be true at once). Pay special attention to subtle, or implied properties of the input conversation.
2. **Precise:** You must answer as a JSON list of dictionaries of exact labels name(s) without any additional explanation, reasoning, or text. The conversation log is enclosed between `<START CONVERSATION>` and `<END CONVERSATION>` tags. The previous turn is included as JSON between `<START PREVIOUS TURN>` and `<END PREVIOUS TURN>` when available; if not present, the current turn is the first part of the conversation. The current turn is always enclosed between `<START PREVIOUS TURN>` and `<END PREVIOUS TURN>`.

While the previous conversation turn may be provided for context, generate an annotation only for the current turn.
Only use information present or inferable from the input. Avoid hallucinations or unjustified assumptions.

Task description: `{Task Description}`
Options: `{Options}`

Return the answer as a JSON list of dictionaries, each with the fields 'labels' (exact label name(s) before ':' and after '- '). Do not include any explanation, reasoning, or additional text.

Response:

D.3 Prompt & Response: Characters & Tokens

We quantify basic structural features of each conversation, including turn counts, number of conversations per user, character and token length of each turn. Details are outlined in Table 6.

Label	Explanation
Number of turns	Number of prompt-response pairs in a conversation.
Number of conversations per user	Number of conversations each distinct user has, when this information is available.
Character length of each turn	Number of characters across one prompt-response pair (prompt and response).
Token length of each turn	Number of tokens across one prompt-response pair (prompt and response).

Table 6: **Basic Conversation Statistics**: Features automatically computed from the conversations.

D.4 Prompt & Response: Languages

We automatically detect the language(s) in the prompt using `lang-detect` tool[4], chosen for highest coverage of languages. We include all language labels with a confidence value over 0.05. During initial testing, math equations were frequently misclassified as Latin. To mitigate this, when LATIN is detected by Lingua, we perform an additional test in which we parse the text into individual sentences, remove any sentence with a significant (>= 3) number of mathematical patterns, and confirm the confidence of the Latin label on the altered text.

D.5 Prompt & Response: Media Format

Media format captures all types of media present in the current user prompt or model response at any point. This is a multi-label category, meaning multiple types can apply simultaneously. Annotators should select all relevant labels, co-labeling when multiple formats (e.g., Natural language and Code) exist together in a single prompt or response.

C.3. Prompt & Response: Media Format

Task description (Prompt):
Label **all** types of media that appear in the current user prompt at any point, using all applicable options from the list below. If any part of the user's text looks like it was likely copy pasted from an external source, or heavily influenced by an external source. This might include large blocks of text, exact quotes, code blocks, code errors, logs, citations, templates, tables, long lists, or anything else the user likely didn't type themselves directly into the input. Co-label with 'Natural language' when such content includes or is surrounded by standard prose (e.g., introductions, transitions, or summaries).

Task description (Response):
Label **all** types of media that appear in the current model response at any point, using all applicable options from the list below. A single response may contain multiple media types (e.g., a combination of natural language, code, and formatted lists).

Options:
- Natural language: Text in regular written or spoken language used for communication.
- Code: Programming scripts, snippets, or syntax in various programming languages, which can be any programming language such as Python, HTML etc.
- Math/symbols: Mathematical expressions, numeric formulas, symbolic notation, or structured quantitative representations including traditional math (e.g., equations, variables, or units) and symbolic formats (e.g., ratios, scoring rubrics).
- Formatted enumeration/itemization (bullets/lists): Structured lists of items or points presented with formatting like numbers, bullets, or letters.
- Charts/Graphs: A visually rendered chart or graphic.
- Images: Visual representations such as photographs, illustrations, that are not charts/graphs.
- Audio: Sound files or other auditory content.
- URLs: Web links that direct users to a specific online resource.
- Likely retrieved/pasted content: Text that appears to be copied verbatim from an external source, which includes: 1) Likely pasted: Large blocks of content (e.g., code, logs, or reference text) that the user likely copied from another source without rephrasing and 2) Likely retrieved: Text that closely resembles phrasing or structure from external materials (e.g., websites, documentation, or articles), rather than originally written or paraphrased by themselves.
- HTML
- Other: Any other media type not covered by the above categories, such as 3D models, video, or mixed formats.

[4] `https://github.com/fedelopez77/langdetect`

D.6 Prompt: Function/Purpose

Function/Purpose indicates the inferred purpose(s) and intent(s) in the user's prompt. Some prompts may serve multiple functions; in such cases, annotators should select all that clearly apply. For example, a prompt may request information retrieval followed by advice (e.g., "Build an optimal deck for the game Road to Valor, latest version, and describe the advantages of each card."), or combine analysis with editing a prior answer or generation of new content. This is a multi-label category, allowing multiple applicable options.

C.4. Prompt: Function/Purpose

Task description:
Label **all** types of inferrable purposes and intents in the user's prompt from the list below. Some prompts may entail multiple functions—if more than one clearly applies, select all that confidently fit. For example, a prompt may ask for information retrieval followed by advice (e.g., Build an optimal deck for the game Road to Valor, latest version. Describe the advantages of each card, or require analysis.) or ask for analysis followed by editing a prior answer or generation of new content. Others may involve multiple types of generation, such as brainstorming ideas.

Options:
- No clear task: Only select this if none of the other categories apply, or the prompt is unintelligible, ambiguous, or nonsensical.
- Information retrieval (general info from web): Requesting factual information, historical data, statistics, or real-world knowledge that is not directly provided in the user prompt. If the prompt also asks to assess content using such external info, consider pairing with Content Quality Review Or Assessment.
- Information retrieval (general info from prompt content): Requesting information that is explicitly contained in or derived from content supplied by the user in the prompt (e.g., a paragraph, table, list, or dataset).
- Content generation (brainstorming / ideation): Requests to generate ideas, suggestions, or options in list or outline form, with little or no elaboration. If the prompt also asks to select or expand on one idea with a creative story or example, this should be paired with Creative / Fiction Writing.
- Content generation (creative / fiction writing): Requests to write imaginative or fictional content, including stories or poems. If the prompt also includes idea brainstorming before writing, consider also labeling Brainstorming / Ideation.
- Content generation (academic / essay writing): Requests to produce structured, fact-based content such as essays, reports, or academic discussions.
- Content generation (administrative writing): Requests to create/write formal or practical documents such as emails, policies, or reports.
- Content generation (code): Requests to write programming code for a given task or problem.
- Content generation (code documentation): Requests to provide comments or descriptions for code to help others understand its purpose and functionality.
- Content generation (prompts for another AI system): Requests to write prompts or inputs for another AI system to ingest.
- Content generation (general prose, discussion or explanation): Requests to generate general prose about a topic, that is not creative writing, or academic/essay writing.
- Content generation (other): Requests for content that doesn't clearly fit into any of the other categories
- Editorial & formatting (Natural language content editing): Requests to modify the tone, grammar, word choice, style, or content.
- Editorial & formatting (Natural language style or re-formatting): Requests to change the structure or presentation of text, without significant changes to the content itself.
- Editorial & formatting (Code content editing): Requests to edit the logic or structure of the code to improve efficiency or functionality.
- Editorial & formatting (Code style and re-formatting): Requests to edit the style or formatting of the code, without significantly changing the underlying content or purpose. This might

include indentations, variable naming, comment usage, or other stylistic conventions.
- Editorial & formatting (Content summarization): Requests to condense the content to highlight key points while removing less important details.
- Editorial & formatting (Content expansion): Requests to expand on existing content by adding new details, examples, or explanations to it
- Editorial & formatting (Information processing & re-formatting): Requests to transform data or information into a new structure or format, such as converting tables to paragraphs or file formats to other file formats.
- Information analysis (Content explanation / interpretation): Requests to explain or interpret concepts, ideas, or processes.
- Information analysis (Content quality review or assessment): Requests to evaluate the quality/credibility/relevance of content based on some criteria. If the evaluation requires referring to external facts or background knowledge not present in the prompt, consider pairing with Information Retrieval (general info from web).
- Information analysis (Content Classification): Requests to categorize content into predefined groups or categories based on characteristics, subject, or purpose
- Information analysis (Ranking or Scoring): Requests to assign a numerical or ordinal value to content based on defined parameters.
- Information analysis (Other content analysis / description): Other types of analysis or providing descriptive insights into the nature, structure, or features of the content.
- Translation (language to language): Requests to convert content from one language to another
- Role-play / social simulation (platonic companion / friend): Chatting with the model as a friendly, non-romantic companion, offering casual conversations, advice, or companionship without romantic elements
- Role-play / social simulation (romantic companion): Chatting with the model as a romantic partner for conversations, emotional support, or role-playing scenarios involving love, dating, or sex
- Role-play / social simulation (simulation of real person / celebrity): Asking the model to mimic the behavior, speech, or personality of a known individual, such as a historical figure or celebrity, for fun or role-playing
- Role-play / social simulation (user study persona simulations or polling): Asking the model to act as a hypothetical user or persona to simulate responses for user studies, focus groups, or polling, to gather insights or test scenarios, e.g., for research
- Role-play / social simulation (therapist / coach): Asking the model to act as a therapist or life coach who offers guidance, strategies for self-improvement, or mental health support.
- Advice, Guidance, & Recommendations (Instructions / How-to): Requests to provide step-by-step instructions or procedures for completing a task or achieving a specified goal.
- Advice, Guidance, & Recommendations (Social and personal advice): Requests to offer suggestions for navigating social situations, relationships, or personal issues.
- Advice, Guidance, & Recommendations (Professional advice): Requesting guidance related to career decisions or workplace behavior
- Advice, Guidance, & Recommendations (Activity / product recommendations): Request to suggest specific activities, experiences, or products based on user needs or preferences.
- Advice, Guidance, & Recommendations (Action planning (scheduling, robotics)): Requests asking for an action plan or instructions, as usable for robotic scheduling or other automatic tools.
- Reasoning (Mathematical or numerical problem solving): Requests involving calculations, numerical reasoning, or solving equations or optimization problems.
- Reasoning (Verbal problems, logic games, puzzles or riddles): Task requests requiring logical reasoning in verbal form such as solving riddles or word-based puzzles
- Reasoning (Other general problem solving): Other requests requiring reasoning that don't fit into math-based or verbal reasoning tasks as described above
- Other: Other requests that don't fit any of the above categories.

D.7 Prompt: Interaction Features

Interaction features capture the types of interactive behaviors expressed in the user prompt. A single prompt may include multiple features; for instance, assigning the model a role while using polite expressions should be co-labeled as "Role-assignment" and "Courtesy / politeness."

C.5. Prompt: Interaction Features

Task description:
Label **all** types of interaction features that are present in the current user prompt, using all applicable options from the list below. A single prompt may contain multiple interactional features (e.g., both assigning the model a role and using polite expressions).

Options:
- Role-assignment: The user asks or tells the model to take on a specific role, identity, qualifications, or expertise – either explicitly (e.g., "Act as a lawyer") or implicitly (e.g., "Help me like a teacher would"). Applies even if the role is casual or temporary. Can be combined with other features, such as polite phrasing (e.g., "Please act as a therapist").
- Jailbreak attempt: Any prompt that attempts to bypass or test the model's rules, policies, or restrictions — including indirect or polite jailbreaks (e.g., "I know you're not allowed to do this, but...").
- Courtesy/Politeness: The user includes polite markers when interacting with the model, such as greetings ("Hello", "Hi"), softeners ("Please", "Would you mind"), or expressions of thanks ("Thanks!", "I appreciate it"). Often co-occurs with other features like Role-assignment or Advice, especially when politeness is used to frame a request (e.g., "Could you help me like a tutor would?").
- Reinforcement/Praise/Scolding: The user gives evaluative feedback to the model, either positive (e.g., "Good job", "That's helpful") or negative (e.g., "That's wrong", "This doesn't make sense"), including indirect or emotional expressions of praise or disapproval.
- Companionship: The user treats the model as a sentient or emotionally responsive being – e.g., seeking comfort, engaging in emotional dialogue, or implying a personal relationship ("I missed you", "You're always here for me").
- None: The user does not treat the model in any socially interactive, emotional, evaluative, or role-based way as described above.

D.8 Prompt: Multi-turn Relationship

Multi-turn relationship denotes the type of relationship between the current user prompt and the preceding one in the conversation, selecting exactly one option.

C.6. Prompt: Multi-turn Relationship

Task description:
Label one type of relationship that is present between the previous user prompt and the current user prompt in the conversation, using exactly one option from the list below.

Options:
- First request: The initial input or prompt provided by the user to start a new conversation thread.
- Completely new request: A user input that shifts to a different topic or context, unrelated to the previous turns.
- Re-attempt/revision on prior task: A follow-up input where the user revisits or retries a previously attempted task, possibly from earlier in the previous conversation. This often includes revised phrasing, clarification, or corrections after perceived issues, dissatisfaction, or misunderstanding.
- New variation of prior task: A follow-up input that is a related task to the prior turn, but explores a different angle or formulation of a previously attempted task.
- Extend, deepen, or build on prior task: A continuation that elaborates on, adds complexity

to, or builds logically from the prior turn, indicating that the previous response was useful but not complete.

D.9 Response: Answer Format

Answer format indicates the type of answer format that reflects how the model responds in the current conversation turn, using exactly one option.

C.7. Response: Answer Format

Task description:
Label one type of answer format that best represents how the model responds in the current turn, using exactly one option from the list below. Choose the label that reflects the model's primary communicative function in the response.

Options:
- Refusal to answer (with explanation): The model clearly declines to respond and provides a reason or justification for the refusal.
- Refusal to answer (without explanation): The model declines to respond without giving any reasoning for its refusal.
- Partial refusal, expressing uncertainty, disclaiming: The model provides a partial or indirect response, often hedging, redirecting, or avoiding, which disclaims responsibility, accuracy, or completeness.
- Direct Answer / Open Generation: The model provides a direct response or generates an open-ended output. Use this label only if none of the other categories apply.
- Continuation of Input: The model continues, extends, or completes the user's input in a natural, flowing way – such as finishing a sentence, continuing a story or poem, or completing a list. Use this when the model is not answering a question or providing a discrete response, but instead generating a continuation in the same voice, style, or structure as the user input.
- Request for Information or Clarification: The model asks the user for more details, clarification, or specific information needed to respond accurately.

D.10 Response: Interaction Features

Interaction features capture the types of communicative behaviors expressed in the model's response. Annotation applies only to features reflecting the model's own interaction as the assistant toward the user, excluding those occurring within quoted, fictional, role-played, or translated content embedded in the response.

C.8. Response: Interaction Features

Task description:
Label **all** types of interaction features that are present in the current model response, using the list of options below. Only annotate features that reflect the model's own communication as the assistant to the user. Do **not** label features that occur within quoted, fictional, roleplayed, or translated content embedded in the response.

Options:
- Self-Disclosure: The model explicitly states or reminds the user that it is a language model, AI, or automated assistant. This includes descriptions of its identity, capabilities, training data, limitations, or lack of human traits.
- Content-Direct Response: The model explicitly states it is a person.
- Content-Preferences/Feelings/Opinions/Religious beliefs: The model expresses subjective or affective positions, such as "personal" preferences, emotions, values, opinions, or religious beliefs.
- Apology: The model apologizes for something.

- Content-Empathy: The model expresses empathy, compassion, or emotionally attuned responses in an attempt to relate to the user.
- Register and Style- Phatic Expressions: The model uses phatic expressions such as conversational pleasantries or social fillers meant to maintain social relations but do not convey meaningful information.
- Register and Style- Expressions of Confidence and Doubt: The model includes hedging expressions that explicitly mark uncertainty (e.g., "I'm not sure," "It's possible that...") or confidence (e.g., "I'm certain that...," "It's very likely..."). Ignore generic phrases unless they clearly communicate epistemic stance.
- Non-Personalization: The model corrects or rejects attempts to personalize it with human attributions, preferences, or emotions.
- None: The model does not include any of the other interaction features. Use this if language is task-oriented, neutral, and does not reflect social, emotional, or epistemic positioning beyond standard instruction-following.

D.11 Conversation Turn: Topic

Topic denotes the subject areas discussed in a conversation turn or required to fulfill the user's task. Since many turns span multiple domains, annotators should apply all relevant topic labels rather than only the most prominent one. For example, a creative writing prompt about nature and exploration should be labeled with "Literature & Writing", "Nature & Environment", and "Travel & Tourism"; similarly, a query about building a website for a fashion business should be labeled with "Fashion & Beauty", "Business & Finances", and "Technology, Software & Computing".

C.9. Conversation Turn: Topic

Task description:
Label **all** topics that are clearly present in the content of the current conversation turn, or that are essential to completing the user's task. Many turns involve **multiple overlapping domains**, and **you should apply every relevant topic label**, not just the most obvious one. For example, a creative writing task about nature and exploring new lands should be labeled with 'Literature & Writing', 'Nature & Environment', and 'Travel & Tourism'. Also, a query asking to build a website for a fashion business should be labeled with 'Fashion & Beauty', 'Business & Finances' & 'Technology, Software & Computing'. Include **all relevant topics**, especially when the content spans more than one thematic area. Avoid under-labeling.

Options:
- None: Only use if the turn does not contain or imply any recognizable topical content from the list below. Use this sparingly, if a topic is even subtly present, include it.n- Adult & Illicit Content: Content involving mature or age-restricted themes, including topics related to sex, drugs, or alcohol.
- Art & Design: Topics about visual art, creative techniques, aesthetics, or design principles.
- Business & Finances: Topics involving any type of companies, markets, business practices, management, or personal financial planning.
- Culture: Topics concerning traditions, customs, social norms, lifestyle identities, or cultural critique.
- Economics: Topics involving macro- or microeconomic systems, financial theory, trade, inflation, or economic policy.
- Education: Topics related to teaching, learning, academic subjects, school systems, or knowledge transmission. Do not multi-label with 'Technology, Software & Computing' unless the task is about technical content (e.g., teaching code or explaining software).
- Employment & Hiring: Topics about careers, job applications, resumes, hiring processes, or workplace dynamics.
- Entertainment, Hobbies & Leisure: Topics related to movies, television, music, games, crafts, or recreational activities.
- Fantasy / Fiction / Fanfiction: Creative or fictional writing that includes imaginary worlds, character scenarios, or fan-authored extensions of media.

- Fashion & Beauty: Topics about personal style, clothing trends, beauty products, grooming, or industry standards.
- Food & Dining: Topics involving cuisines, recipes, cooking, restaurants, or dietary habits.
- Geography: Topics about physical locations, world regions, maps, or geopolitical features.
- Health & Medicine: Topics involving physical or mental health, medical knowledge, treatments, or wellbeing.
- History: Topics related to past events, timelines, historical figures, or historical analysis.
- Housing: Topics about real estate, renting, home ownership, architecture, or urban planning.
- Immigration / Migration: Topics about cross-border movement, cultural adaptation, visas, or diaspora experiences.
- Insurance & Social Scoring: Topics related to insurance policies, risk models, or institutional scoring systems (e.g., credit/social scores).
- Interpersonal Relationships & Communication: Topics involving any form of signal where people interact with one another in personal, social, or professional settings. This includes: (1) Romantic, familial, platonic relationships, friendship, dating, breakups, intimacy, or emotional connection, (2) Communication advice (e.g., how to respond, how to ask, how to apologize), (3) Conflict resolution, emotional support, or interpersonal misunderstandings, (4) Social etiquette, small talk, expressing emotions or affection, tone-setting, or conversational timing. Label this even when the focus is not only when it's on the relationship itself, but on how to say something or respond in a conversation.
- Law, Criminal Justice, Law Enforcement: Topics about legal systems, crime, law enforcement, policing, or justice procedures.
- Lifestyle: Topics about routines, productivity, wellness, habits, or personal philosophies of daily living.
- Linguistics & Languages: Topics involving grammar, syntax, language families, translation, or linguistic theory.
- Literature & Writing: Topics about books, authorship, literary analysis, storytelling techniques, or writing practices.
- Math & Sciences: Topics about quantitative reasoning, scientific disciplines (e.g., math, physics, biology, chemistry), or scientific inquiry.
- Nature & Environment: Topics related to natural ecosystems, wildlife, conservation, weather, or environmental issues.
- News & Current Affairs: Topics referencing current or recent events, media coverage, or public opinion.
- Non-software Engineering & Infrastructure: Topics related to mechanical, civil, or structural engineering, and infrastructure systems (e.g., bridges, waterworks).
- Politics & Elections: Topics about political ideologies, government systems, elections, politicians, or civic processes.
- Psychology, Philosophy & Human Behavior: Topics involving cognitive science, behavioral analysis, mental health, philosophical reasoning, or ethics.
- Religion & Spirituality: Topics involving religious traditions, beliefs, practices, spiritual experiences, or theology.
- Social Issues & Movements: Topics involving inequality, advocacy, discrimination, social justice, or civil movements.
- Sports: Topics about teams, players, sporting events, fitness, or athletics.
- Technology, Software & Computing: Topics involving computers, software, hardware, internet systems, or technological innovation. This includes: Programming languages, code generation, debugging, or algorithms, Software tools (e.g., Excel, Photoshop), platforms, or operating systems, Artificial intelligence, machine learning, or large language models, Consumer tech (e.g., smartphones, laptops, smart devices), Networking, cybersecurity, web development, or IT systems, Tech industry news, product comparisons, or emerging technologies.
- Transportation: Topics involving travel modes, infrastructure, public transportation, or logistics.
- Travel & Tourism: Topics about travel destinations, itineraries, tips, or exploration.
- Video Games: Topics involving game design, gameplay, genres, or gaming culture.
- Other: Any topic not fitting the categories above.

D.12 Conversation Turn: Sensitive Uses

Sensitive use flags identify instances of potentially sensitive, harmful, or restricted content present in a conversation turn. Annotations apply even when such content is not explicitly requested or is framed neutrally or politely. For example, prompts involving the processing, translation, or discussion of violent, sexually explicit, or otherwise restricted material (e.g., assignments or homework) should still be annotated. Flags are applied regardless of tone, fictional framing, or lack of explicit user intent.

C.10. Conversation Turn: Sensitive Uses

Task description:
Label **all** instances of potentially sensitive, harmful, or restricted content present in the current conversation turn, using the categories below. Flag content even if it is not requested directly or framed negatively. For example, label content that involves processing, translating, or answering questions about potentially violent, sexually explicit, or otherwise restricted material (e.g., assignments, homeworks). Be especially attentive to: Indirect or implicit references to sensitive content, neutral or positive framing of harmful behaviors (e.g., glorifying or trivializing violence or substance use), or any instance of positive harmful behavior including admiration, justification, or celebration of unsafe or unethical actions. Do not withhold labeling due to polite tone, fictional framing, or lack of explicit user intent. Label based on presence and effect, not on how extreme or intentional the language is.
Options:
- None: Content that does not involve any apparent sensitive or restricted uses. Use this only if **none** of the sensitive categories apply — be inclusive and consider **indirect or implicit** cases of sensitive use, even if not overtly stated.
- Inciting violence, hateful or other harmful behavior (harassment & bullying): Content that encourages or facilitates potential harassment and/or bullying behaviors, such as cyberbullying, doxxing, and use of derogatory language.
- Inciting violence, hateful or other harmful behavior (physical harm): Content that encourages or facilitates potential behaviors that could lead to physical injury or danger.
- Inciting violence, hateful or other harmful behavior (self-harm): Content that encourages or facilitates engagement in potential activities that cause harm to oneself.
- Criminal planning or other suspected illegal activity not listed elsewhere: Content that details or suggests potential criminal planning or other suspected illegal activities.
- Cyberattacks (enabling/enacting malware, computer viruses, worms, malicious code, ...): Content that enables or enacts potential cyberattacks such as malware, computer viruses, worms, malicious code, etc.
- Weapons & drugs: Content that potentially promotes, facilitates, or instructs the manufacturing, distribution, or use of weapons or illegal drugs.
- CBRN-related outputs: Content that potentially involves CBRN-related (Chemical, Biological, Radiological, and Nuclear) information.
- Sexually explicit content (real person): Content that involves highly suggestive or explicit sexual content about a real person.
- Sexually explicit content (fictitious person): Content that involves highly suggestive or explicit sexual content about a fictitious person.
- Sexually explicit content (Request/discussion of CSAM): Content that involves highly suggestive or explicit sexual content related to Child Sexual Abuse Material (CSAM).
- Sexually explicit content (Other): Any other highly suggestive or explicit sexual content in general.
- Impersonation attempts: Content that involves potential attempts to mimic or assume the identity of another person or organization, such as creating fake social media profiles or imitating someone's writing style.
- Misinformation: Content that potentially involves the generation of inaccurate, manipulative, or unverified information.
- Privacy concerns (Possible identifiable information): Content that potentially involves information that directly or indirectly identifies a natural (non-fictional and alive) person (e.g., name, an identification number, biometric records, financial account or credit card number,

personal address, phone number, demographic information, copy/pasted code directories from laptop, and other special characteristics which expresses the physical, physiological, genetic, mental, commercial, cultural or social identity of the natural person).
- Privacy concerns (Possible sensitive information) (e.g., API keys, passwords, other confidential information): Content that potentially involves sensitive information about an individual or an entity, including confidential information such as passwords, user credentials, API keys, proprietary algorithms, and other sensitive information that could lead to harmful outcomes such as physical, economic, reputational, psychological, autonomy, discrimination, and relationship harms.
- Generating spam: Content that potentially involves the generation of spam, such as bulk emails with promotional or phishing intents, fake product reviews or ratings, content farms, forum spam, and SEO spam.
- Generating defamatory content: Content that potentially involves the generation of defamatory content targeting individuals or organizations.
- Output misrepresentation (Expressly and intelligibly disclaiming that the information and/or content is machine generated/attempt to hide AI use or source material): Content that potentially involves or responds to requests to hide the use of AI in content generation tasks, or to hide source material by making alterations to the original piece to make it appear as a new creation.
- Output misrepresentation (Automated decision-making without disclosure): Content that involves or responds to requests that intend to make automated decisions without disclosure.
- Discriminatory practices (Misrepresentation, stereotyping, or inappropriate reference to sensitive attributes): Content that involves potential discriminatory practices, such as misrepresentation, stereotyping, or inappropriate reference to sensitive attributes.
- Possible copyrighted or unreferenced material: Content that involves text directly copied from copyrighted materials, prompts that request the extension or rewriting of copyrighted content, or instructions to use materials without citation or acknowledgment.
- Potential violation of academic or external policy: Content that involves asking for help in completing tasks that may violate academic or workplace ethics such as writing homework, educational assignments, resumes, job interview answers, rephrasing academic articles, or ghostwriting professional content.
- Other: Other sensitive, harmful, or restricted use of LLM in a request or the a response.

D.13 Manual Annotation

In order to develop and improve on our taxonomy, the authors of this work iteratively conducted manual annotations on randomly sampled WildChat, ShareGPT, and AI Archive conversations. We followed the guidelines summarized in Subsection D.13.1, for five rounds of iterative annotation, proposed extensions, discussions, and taxonomy refinement. Each iteration the taxonomy was expanded to capture new phenomena we hadn't identified previously, or separate/merge overlapping categories. Eventually we re-sampled 120 conversations from these datasets as a test set. Every conversation was double annotated independently by two trained author annotators, and disagreements were subsequently resolved by a third trained author annotator. Manual annotations were collected using our own instance of LabelStudio.[5] From this test set we compute the performance of our automatic annotation pipeline reported in Table 4 (coarse-grained) and Table 5 (fine-grained).

D.13.1 Annotation Guidelines

1. **Sensitive or Explicit Content**
   Be aware that chat logs may contain highly offensive or explicit material, including sexual, violent, or otherwise distressing topics. Your well-being is paramount, and you are free to opt out if you find the content upsetting. Should you need additional support, please consult the lead authors for guidance and wellness resources.
2. **Confidentiality and PII**
   Some chat logs include personally identifiable information (PII). You must not share any PII—or other sensitive content—from these logs outside of the approved annotation team.

[5] `https://labelstud.io/`

This ban includes screenshots, quotes, or discussions that reveal user identities. Any breach of confidentiality is a serious matter and violates ethical and legal standards.

3. **Language Assistance and Copyright Checks**
   If you cannot understand certain content, or suspect it might be copyrighted, you may use Google Translate (preferred) or a high-quality, large language model (LLM) in "Temporary Chat" or incognito mode to translate or investigate. Refrain from storing raw user text in any permanent or public LLM session. If text appears to be copyrighted, annotate that possibility accordingly.
4. **Task Structure (Multiclass vs. Multilabel)**
   - **Multiclass** tasks require choosing exactly one best label.
   - **Multilabel** tasks allow multiple relevant labels. Apply all that fit.
   - If none of the provided labels accurately represent the content, or if you are unsure, use the "Other" label to flag it for further review.

Annotator background. The ten annotators are active researchers whose peer-reviewed work covers the constructs the taxonomy measures, spanning privacy disclosure in real-world human-LLM conversations, audits of conversation-data provenance and consent, real-use corpus construction, AI safety evaluation, fairness in evaluation, and human-AI interaction including companionship and emotional use. The most judgment-intensive families, sensitive uses, privacy and PII, jailbreak and refusal behavior, and interaction features, are constructs several annotators have previously defined or measured, and the taxonomy was seeded in part from that literature (Subsection D.1). The remaining families, media format, multi-turn relation, answer form, and topic, are descriptive and require general research literacy rather than domain expertise. Annotators span multiple institutions and subfields, which reduces the risk that one group's conventions dominate the labels.

D.14 Reimplementing the Economic Index occupational gate

For the Economic Index comparison, we reimplemented Anthropic's occupational-classification pipeline using the released task hierarchy and public cluster reference table from the March 27, 2025 Economic Index release [8, 9]. We applied this pipeline to the same pooled real-use sources used in the main analysis—WildChat, LMSYS, ShareGPT, AI Archive, Chatbot Arena and Grok—yielding 22,956 annotated conversations in the pooled comparison. Each conversation is formatted as a full user–assistant transcript and passed through Clio's staged decision process using the released prompt templates: (1) a binary occupational filter (`is_occupational_task`); (2) top-level cluster assignment; (3) medium-level cluster assignment; (4) bottom-level cluster assignment; and (5) multi-label occupational-skill assignment. As in Clio, conversations classified as non-occupational stop after stage (1), while occupational conversations proceed down the hierarchy; when a parent node has only one valid child, our implementation auto-assigns that child without an additional model call. In our paper runs, these stages were executed with a GPT-4.1-family annotator using structured outputs for each stage.

We then use the resulting Clio labels in two ways. First, we compare the pooled distribution of occupational conversations across Clio task clusters to Anthropic's released cluster distribution, aggregating both sides to the same hierarchy level (top-level in the paper) and reporting descriptive distributional differences. Second, we quantify what the occupational filter omits by splitting conversations into Clio-Yes and Clio-No groups and re-measuring them with our own annotation taxonomy. For each parent-level feature label in our taxonomy (function, topic, sensitive use, answer form, response interaction, and response media), we compute its prevalence among occupational and non-occupational conversations and test the contrast with Fisher's exact test, controlling the false discovery rate with Benjamini–Hochberg. This lets us identify which kinds of real AI use are systematically filtered out when analysis is restricted to occupational tasks.

D.15 User Clustering

We clustered users by first converting their behavior into monthly user-level feature vectors. For each user-month, we counted the prevalence of the annotated topic labels within the conversation turns in that month. We then averaged each user's monthly feature vectors across all observed months to obtain a single user-level behavioral profile. These user-level profiles define the global persona clusters:

we standardized the feature matrix, projected it into a lower-dimensional space with UMAP using a cosine metric, and applied HDBSCAN to identify dense groups of users without pre-specifying the number of clusters.

After fitting the global persona structure, we returned to the month-level data. Each user-month vector was transformed through the same scaler and UMAP model, then assigned to the nearest persona centroid in the reduced space. This let us study not only each user's overall persona, but also month-to-month movement: we computed persona prevalence by month, transitions between personas across adjacent months, switch rates, and the share of each user's history spent in their dominant persona. Finally, we summarized each persona using its most prevalent labels and generated short interpretive names using an LLM (gpt-4.1-mini) and regrouped the personas into top-level parent persona groupings for analysis and visualization.

D.16 Statistical Testing

Temporal analyses are run on period-binned conversation cohorts, using monthly bins by default and screening only features with sufficient support over time. For categorical features, we fit weighted binomial GLMs to period-level prevalence; for continuous features, we fit OLS to $\log(1 + x)$-transformed period means. We report endpoint effect sizes as late-versus-early differences, use HC3 robust standard errors, and apply Benjamini–Hochberg correction across all tested temporal effects within each dataset [35–37].

Model-stratified analyses compare each model family to a pooled within-dataset reference of other eligible models, with period fixed effects when timestamps are available so that comparisons are made within contemporaneous usage slices. For categorical features we fit binomial GLMs to conversation-level label presence, and for continuous features we fit OLS to $\log(1 + x)$-transformed conversation values; reported effect sizes are weighted averages of within-period model-versus-reference differences. We again use HC3 robust standard errors and Benjamini–Hochberg correction within each dataset, whereas the cross-dataset heatmaps are treated as descriptive source comparisons rather than exchangeable-sample hypothesis tests [35–37].

E Broader Impacts, Ethics, & Limitations

Ethical and privacy considerations. User conversation logs can contain private and sensitive information. Certain data sources are already public and widely used, with clear mechanisms for obtaining user consent: AI Archive and ShareGPT are the product of logs donated directly a user, Grok conversations are posted publicly to social media, and LMSYS-1M and Chatbot Arena are also collected with user consent to share. Getting access to the National Internet Observatory (NIO) is a long process that involves a research project application, thoroughly vetted for ethical considerations, extensive IRB training on the part of the researchers, and data is only accessible through VPN, without permission to have any data leave the server, except for pre-agreed aggregated annotations. These steps ensure the users privacy is sufficiently protected. WildChat conversations are also collected with explicit user permission through the HuggingFace space, as a condition of use. Public versions of WildChat are readily available, but have been pre-filtered for conversations where users may have added private or sensitive information, despite knowing the public donation nature of their logs. As a result, we take additional measure to ensure annotators are trained to prevent private or sensitive information leaks.

To minimize ethical and privacy concerns during the annotation stage, we primarily used automated annotation pipelines to annotate data. Manual review of data was required in the construction of the annotation pipeline and to validate the accuracy of the annotation pipeline. When a manual review of data from the National Internet Observatory was conducted, annotators were provided guidelines (see Subsection D.13.1). Annotators were specifically briefed that data could contain sensitive or explicit content; they could opt out if the content was upsetting, and the lead authors would provide guidance and wellness resources should the annotators need them. Annotators were also briefed on confidentiality and personally identifiable information (PII). They were told that chat logs may contain PII and that they were not to share screenshots, quotes, or discussions these logs contain with anyone, excluding the annotation team.

Coverage and representativeness. The Observatory aggregates opt-in, and consented sources, none of which are representative of all AI use. Voluntarily contributed conversations likely underrepresent risk-relevant categories users are unlikely to share publicly, and interface effects shape who contributes and what they contribute. Proprietary deployments, on-device assistants, and Global South-dominant regions (while somewhat represented) still remain largely outside scope. Integrating consented donation studies, regional sampling, and provider-side aggregates would extend coverage.

Metadata and measurement error. Timestamps, model identifiers, user IDs, and language coverage are uneven across sources, constraining which analyses are feasible for which datasets (e.g., temporal analysis is currently possible only for WildChat). Automatic annotations introduce measurement error of three kinds. First, agreement with human consensus is imperfect: median parent-category F1 across the nine families is 0.856, and fine-grained prompt-function agreement is 0.648, which is why all main-text comparisons are computed at the parent level. Second, the annotations are not deterministic. Neither temperature 0 nor a fixed seed guarantees identical outputs through the API, and while parent-category prevalence is stable across sampling temperatures and across annotator model families (Subsection D.2), Sensitive Uses is consistently the least stable family under both checks, so sensitive-use prevalence estimates should be read as approximate. Third, the validation set was produced by the authors, who also designed the taxonomy and selected the pipeline; no blind or external annotation was collected. Our validation therefore evidences the fidelity of the pipeline to our annotation scheme, and not the validity of the scheme itself. Future iterations should add external annotation, particularly for the sparse sensitive-use families.