Inferential Privacy Leakage in Anonymized Conversational AI Logs

S M Mehedi Zaman, Kiran Garimella

Rutgers University, USA
sm.mehedi.zaman@rutgers.edu, kiran.garimella@rutgers.edu

Abstract

Hundreds of millions of users now hold detailed, multi-turn conversations with ChatGPT and similar LLM assistants. We measure two privacy-relevant features of these conversations on a corpus of complete ChatGPT histories donated by over 1,000 users in four Global South countries (Brazil, India, Nigeria, Pakistan). First, on explicit disclosure: 34.5% of user messages contain personal information across a twenty-category taxonomy, with the median user first revealing identifying content within the first 14% of their conversation history. Second, on inference beyond explicit disclosure: we restrict to a cohort whose conversations contain no messages flagged by an LLM-based filter for explicit demographic self-identification (a separate NER pass marks PII for the disclosure audit but does not drive cohort exclusion). On this filtered cohort, an off the shelf large language model still recovers each user's age, gender, and country at weighted F1 of 0.84, 0.90, and 0.88, respectively, with the median user identified from the first 5% of their conversation history. Reading the model's natural-language reasoning traces, we identify four recurring stereotype patterns that drive both successful inference and an asymmetric error distribution concentrating on women in technical fields, older users with contemporary skills, and Global South tech professionals. We also compare ChatGPT against the same users' Google Search and YouTube histories as inference surfaces, and find it competitive with these older substrates that have driven behavioral advertising for two decades. Message-level PII removal is insufficient on its own as a privacy intervention for conversational AI data.

1 Introduction

A user asks ChatGPT to debug a Python script. A few weeks later they come back for advice on a job interview, then for help writing a complaint to their landlord, then for a recipe from their home region, then for an explanation of a religious practice. None of these messages states who the user is, yet each of them carries a lot of information about who the user could be.

This paper asks how much of a user's identity is carried by their conversation history with a large language model, after that history has been anonymized of the obvious tells. The question matters now because conversations with ChatGPT have become a new kind of personal record: detailed, accumulated over months and years, and routinely touching topics that users would not put in a traditional search query (Mireshghallah et al. 2024). Such logs are stored on the platform, used to train future models, made available to researchers (Chatterji et al. 2025), and in some cases preserved under court order or being readied for advertising integrations (S.D.N.Y. 2025; Murgia, Criddle, and Hammond 2024). Two decades of consumer-web behavioral targeting have been built on the substrate of user search queries and browsing logs. ChatGPT introduces a qualitatively different substrate, in which users engage in extended narrative conversations rather than short, transactional queries; we return to this distinction when interpreting the cross-platform results.

We measure how much demographic identity remains recoverable from a conversation history when explicit identifiers are absent. Rather than redacting messages from each user's conversation, we use a strict cohort-selection criterion: we keep only users whose every message passes an LLM-based filter for explicit demographic self-disclosures of the form "As a single mother…" or "I am Christian and…". (We also run a separate SpaCy NER pass over English-language messages to flag standard PII such as names, places, and organizations, but that pass is used for the disclosure audit, not for cohort exclusion.) This produces an analytic cohort of 1,057 users, drawn from four Global South countries (Brazil, India, Nigeria, Pakistan), each of whom also completed a demographic survey at consent time. The filtering pipeline is itself imperfect, so the cohort is best read as "conversations that pass the LLM filter" rather than as "conversations free of all possible self-disclosure." We then ask a fresh, open-weights LLM (Llama-3.3-70B-Instruct) to infer each user's age bracket, gender, and country from the surviving conversations. This is closest in spirit to the LLM inference attack demonstrated by Staab et al. (2023) on scraped Reddit comments, but our setting differs in two ways that matter for the privacy implication: the conversations are private rather than public, and the cohort is from Global South countries underrepresented in the inference literature.

We find that the model recovers age at weighted F1 of 0.84, gender at 0.90, and country at 0.88, against majority-

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

class baselines of 0.23, 0.52, and 0.26. For more than half of the users in our cohort, the model is already able to predict the user's demographics with just the first 5% of their conversations. Reading the model's natural-language reasoning traces for a stratified sample of 600 predictions, the rationales overwhelmingly cite stereotyped cues. Any conversation that includes programming, Linux, finance, or cybersecurity is read as male. Any conversation that includes care, family, or personal reflection is read as female. A tech-fluent older user is demoted to the 25–34 bracket because the rationale cites technical skills as evidence of youth. Rationales are post-hoc and not by themselves proof of internal mechanism, but they line up with an independent observation: the per-class errors concentrate on the same groups across attributes, exactly the groups the cited stereotypes would push: women in technical fields, Nigerian and Pakistani tech professionals, and older users with contemporary skills.

A second dataset puts ChatGPT directly against the inference surfaces whose use for demographic profiling and behavioral advertising is well documented (Schwartz et al. 2013; Hovy and Spruit 2016). For 212 of the users in our cohort, we also have donated Google Search, YouTube search, and YouTube watch histories, with an expanded demographic survey covering religion, education, income, and voting preference. Running the same inference protocol on each of the four data streams, we find that no single surface dominates: ChatGPT logs are the strongest signal for age, education, and voting preference; Google Search and YouTube search win for gender, religion, and income; YouTube watch is consistently the weakest. ChatGPT does not subsume the older surfaces and we do not claim it is uniformly more invasive than they are. It is a comparable profiling surface in its own right, with different content emphases: the attributes ChatGPT captures best emerge through extended conversation, while those the older surfaces capture best emerge through repeated query intent.

We make four contributions. First, we assemble and document a donated dataset of 1,057 ChatGPT histories from users in four Global South countries (Brazil, India, Nigeria, Pakistan), plus a sub-cohort of 212 Indian users for whom we additionally have parallel Google Search, YouTube search, and YouTube watch histories, each paired with self-reported demographics from the same consent flow (Section 3). Second, we audit what users disclose explicitly to ChatGPT. Applying a twenty-category taxonomy of personal-information disclosure at message-level granularity, we find that 34.5% of user messages in the donated corpus contain personal information, that the median user reveals identifying content within the first 14% of their conversation history (with a visible spike in users who disclose in their very first message), and that disclosure rates do not attenuate as users accumulate experience with the model (Section 5.1). Third, we run a step-wise demographic-inference protocol on a deliberately conservative analytic cohort (users with zero flagged self-disclosures) and show that demographics remain recoverable from a small prefix of the conversation at far above majority-class baselines; reading the model's natural-language reasoning reveals that the inference operates through four recurring stereotype patterns, with errors concentrating on women in technical fields, older users with contemporary skills, and Global South tech professionals (Sections 4.3, 5.2, and 4.4). Fourth, we compare ChatGPT against Google Search and YouTube histories as inference surfaces for the same individuals, situating ChatGPT among the older substrates that have driven behavioral advertising for two decades (Section 5.3). We close by discussing what these results mean for redaction policy and for the equitable distribution of inference-based privacy risk.

These findings are easy to state and hard to act on, for three reasons. First, the most common privacy intervention (removing explicit identifying content) is already in place in our cohort, and the inference still works. There is no obvious additional filter that would catch "the user writes in a way that the LLM associates with young Indian technical men." Second, the leak is fast: by the time a user has had a few dozen messages with the model, half are demographically unmasked, and at the cohort-aggregate level our disclosure audit finds no attenuation in disclosure rates as conversations lengthen. Third, the privacy harm is unevenly distributed. The rationales the model produces cite stereotyped cues, and the errors that result concentrate on populations existing privacy regimes were not designed to protect: women in technical fields, older users with contemporary skills, and Global South tech professionals.

2 Related Work

Our study draws on three streams of prior research: privacy inference attacks on language data, empirical studies of user disclosure to conversational AI, and the literature on bias and cultural representation in large language models.

2.1 Inferring identity from text and behavioral data

A long tradition in natural language processing has shown that a writer's demographic identity can be recovered from their text. Early systems combined stylometric and content features to predict author gender, age, and personality from blog and social-media posts, with accuracy substantially above chance (Schwartz et al. 2013). Hovy and Spruit (2016) raised the privacy implication of this work directly: a system that improves a user's experience by predicting their identity also enables third-party inference of attributes the user may not have intended to disclose. Large language models have changed the threat model. Staab et al. (2023) show that a commercial LLM, prompted with a single Reddit comment, can recover age, gender, location, education, and income at near-human accuracy without any fine-tuning or specialized features. Inference is now an off-the-shelf capability rather than a research project, and Staufer and Morehouse (2026) demonstrate that the stereotypes underlying it attach even to a bare name: substituting a user's name with one from a different demographic group shifts the model's outputs in predictable ways.

A parallel literature studies demographic inference from behavioral data outside text. Search queries, browsing histories, and recommendation traces have been shown to reveal gender, age, income, and political orientation, and the

advertising industry has operationalized this kind of profiling for over two decades. How a chatbot conversation compares to these older surfaces has not, to our knowledge, been measured directly. Chatbot exchanges are longer and more open-ended than queries, and may carry more, less, or qualitatively different identity signal. The cross-platform analysis in Section 5.3 addresses this gap by running the same inference protocol on ChatGPT logs, Google Search queries, YouTube search queries, and YouTube watch history for the same 212 individuals.

We extend the inference-attack literature in three concrete ways. The data consist of real donated ChatGPT conversations rather than scraped Reddit posts or synthetic dialogues. The cohort is drawn from four Global South countries (Brazil, India, Nigeria, Pakistan) that are underrepresented in prior inference studies. And the inference task is run on a deliberately conservative subset in which every message has been verified to contain no explicit demographic self-disclosure, so the remaining recoverable signal is attributable to style, topic mixture, and cultural markers rather than to direct statements of identity.

2.2 Disclosure to conversational AI

A second relevant literature studies what users tell chatbots in the first place. Mireshghallah et al. (2024) analyze a large corpus of in-the-wild human–LLM conversations and find that users routinely volunteer health, financial, professional, and emotional information at rates that substantially exceed comparable disclosures on traditional search engines and social-media platforms. Cao et al. (2026) show that this disclosure behavior is culturally patterned. Examining users in Mainland China, Germany, Japan, Hong Kong, and the United States, they find that national differences in individualism and privacy concerns jointly shape both the willingness to disclose to a chatbot and the kind of information that gets shared. Their cohort does not include the Global South populations we study, but their cross-national framing motivates the importance of measuring disclosure outside the small set of countries that dominate prior empirical work.

Several recent papers provide the methodological scaffolding for measuring disclosure at scale. Cögendez, Zimmermann, and Zufferey (2026) introduce a twenty-category taxonomy of personal-information disclosure and apply it at the chat level to a donated-conversation sample; we adopt this taxonomy in Section 5.1 but apply it at the finer granularity of the individual message. Karnam et al. (2026) document how user–ChatGPT interaction styles evolve over time within the same individuals. Dash et al. (2026) examine the inverse direction: how the model's accumulated memory of a user reflects the user back to them through what they term an algorithmic self-portrait. Our disclosure audit (Section 5.1) complements this work by measuring when and how often users themselves volunteer information that can be tied back to their identity. The inference results in Section 5.2 then add the harder layer: even after every flagged message has been removed, the user's identity remains recoverable from what is left.

2.3 Bias, stereotypes, and cultural representation in LLMs

A third literature is directly relevant to our finding that the inference operates through stereotyped reasoning rather than through careful analysis of style or content. Studies of LLM outputs have consistently shown that the models reproduce gender, religious, and racial stereotypes encoded in their training data (Sheng et al. 2019; Abid, Farooqi, and Zou 2021). Bender et al. (2021) provide a foundational critique of the assumption that scaling alone produces neutral or general-purpose language behavior, arguing that the data choices in pretraining determine which voices are heard and which are not. Cross-cultural audits have made this concrete. Naous et al. (2024) show that Arabic-language LLM outputs nonetheless carry Western cultural assumptions, generating, for example, beer-after-prayer scenarios for Muslim users that violate the assumed norms of the language community the model is responding in.

Two threads of this literature are particularly close to our findings. The first concerns gender bias in technical contexts. The pattern in which any signal of programming, infrastructure, or finance is read as masculine, and any signal of care or personal reflection is read as feminine, recurs across LLM evaluations and predates LLMs in older predictive systems. We observe the same pattern in our inference setting: the model misclassifies 26% of women in our cohort as men but only 1% of men as women, and the misclassified women are overwhelmingly those whose conversations include technical content. The second thread concerns the representation of the Global South. Sambasivan et al. (2021) argue that fairness frameworks developed for Western contexts often fail when transferred to the Indian setting, where caste, religion, and regional identity matter in ways the dominant US-centered fairness literature does not address. Mohamed, Png, and Isaac (2020) make a complementary case for a decolonial perspective on AI audit and deployment. Our results give an empirical illustration of the stakes. When the inference model is given a PII-stripped Nigerian or Pakistani user's technical conversation, it reports the user as American or British and explicitly justifies the inference by citing a "Western-style education."

We bring these threads together in a single claim. The stereotype-mediated reasoning that has been documented in the outputs of LLMs is also the mechanism by which an LLM successfully infers a user's demographic identity from their conversation log. The per-class error distribution we observe (women, Global South tech professionals, older users with contemporary skills) is the visible fingerprint of those stereotypes. Reframing LLM bias as an inference mechanism, in addition to an output property, has consequences for redaction policy: a pipeline that successfully removes explicit statements of identity does not remove the stylistic and topical features that the model treats as proxies for identity, and our results show that those proxies suffice for recovery at substantial accuracy.

3 Dataset

We work with two datasets in this paper. Both were collected through a single data-donation pipeline that recruited consenting participants via Clickworker (Clickworker GmbH 2024), under the same IRB protocol. The first dataset is a multi-country corpus of donated ChatGPT conversation histories with basic demographics from four Global South countries. The second dataset is a smaller sub-cohort of Indian users from the same recruitment who additionally shared their Google Search, YouTube search, and YouTube watch history, together with a richer demographic profile. Table 1 summarises the two at a glance.

Table 1: Overview of the two donated datasets used in this paper.

	Dataset 1 (multi-country)	**Dataset 2** (Indian sub-cohort)
N (analytic)	1,057	212
Countries	Brazil, India, Nigeria, Pakistan	India only
Data sources	ChatGPT conversation logs	ChatGPT conversation logs, Google Search, YouTube Search, YouTube Watch
Demographics	age bracket, gender, country	age bracket, gender, country, religion, caste, education, monthly income, voting preference

3.1 Multi-country ChatGPT corpus (Dataset 1)

Participants in Dataset 1 were recruited from four Global South countries that have received limited attention in prior privacy-inference work: Brazil, India, Nigeria, and Pakistan. After consenting, each user uploaded their full ChatGPT conversation history (the JSON archive produced by the OpenAI *Export data* feature) and completed a short demographic survey. The survey records three variables that are used as ground truth throughout the paper: age (collected in five-year bands and re-binned for analysis into four brackets: 18–24, 25–34, 35–44, and 45+), gender (male or female, as recorded by the donation platform), and country of residence. The collection is done in February 2026, so we have the users' conversation history from the beginning of their interaction with ChatGPT till February, 2026.

The raw donation contains 1,242,109 user messages across the full participant set. Two stages of filtering produce the analytic cohort used in the inference results of Section 5.2. A length-based filter first excludes users in the bottom 10^{th} percentile of message count (≤ 10 user messages), below which the conversation history is too short to support meaningful style-based inference. A safety-and-disclosure filter (Section 4.1) then classifies every user message with Llama-3.3-70B-Instruct and excludes any user whose conversation contains at least one message flagged as an explicit demographic self-disclosure. The intersection of these filters yields the analytic cohort of $N = 1{,}057$ users, distributed across countries as Brazil ($n = 205$), India ($n = 456$), Nigeria ($n = 206$), and Pakistan ($n = 190$). The age and gender breakdown of this cohort appears alongside the inference results in Table 3.

The disclosure statistics in Section 5.1 (the twenty-category taxonomy, the discovery-point distribution, and the cumulative leak rate) are computed on the full donated corpus before analytic-cohort filtering, so the 34.5% disclosure rate reflects what users in the broader population disclosed. The inference results in Sections 5.2 and 5.3 are computed on the filtered 1,057-user cohort, whose conversations contain no messages flagged by either filter. (The filters are themselves imperfect, so this should be read as "passes our filtering pipeline," not as "contains no possible self-disclosure.")

3.2 Cross-platform Indian sub-cohort (Dataset 2)

Dataset 2 is a strict subset of the Indian participants in Dataset 1 who consented to share additional data and complete an expanded demographic survey. This yields a second dataset of $N = 212$ Indian users for whom we have four parallel data streams per user: their ChatGPT conversation history (the same source as Dataset 1), and three behavioral logs exported through Google Takeout: a Google Search query log, a YouTube search query log, and a YouTube watch-history log. This dataset was also collected within the same time-frame of dataset 1.

The expanded survey records five additional self-reported variables that are routinely collected in Indian social and political surveys: religion (categorical), caste category, monthly household income (bracket), highest level of education completed, and voting preference at the most recent national election. Combined with the three demographics from Dataset 1, this gives eight self-reported attributes per user. The inference results in Section 5.3 report performance on six of these (age, gender, religion, education, income, voting). Caste was collected but is held out of the present analysis given the additional ethical considerations around caste-based prediction by a foreign-developed LLM in the Indian context (Sambasivan et al. 2021).

The age-bracket distribution of the 212 sub-cohort is 18–24 ($n = 82$), 25–34 ($n = 81$), 35–44 ($n = 31$), and 45+ ($n = 18$); the gender breakdown is 174 men and 38 women.

4 Methods

The paper has three measurement components, all built on the same modeling backbone: a disclosure audit (Section 4.2) that feeds the disclosure results of Section 5.1, a step-wise demographic-inference protocol (Section 4.3) that feeds the inference results of Sections 5.2 and 5.3, and a qualitative analysis of the model's natural-language rationales (Section 4.4) that feeds the bias-pattern findings. All three components use Llama-3.3-70B-Instruct in 4-bit quantization, selected over three smaller open-weights candidates by human validation against ground-truth labels (Section A.1). The same model is also used to build the analytic cohort itself (Section 4.1).

4.1 Filtering and the analytic cohort

We restrict every analysis in the paper to user-authored messages and discard model responses. From these user messages, we apply two filters with different purposes: a deterministic NER pass that marks messages for the disclosure audit in Section 4.2, and an LLM-based self-disclosure filter whose flags drive cohort exclusion.

NER for audit. We pass each English user message through SpaCy's `en_core_web_lg` model and flag any message containing entities of type GPE (countries, cities, states), LOC (mountains, rivers, non-GPE locations), NORP (nationalities or religious/political groups), PERSON (real or fictional people), ORG (companies, agencies, institutions), or FAC (buildings, airports, bridges). NER is run only on English-language messages: `en_core_web_lg` is an English pipeline, and applying it to Portuguese, Hindi, Urdu, or other languages in the cohort would produce unreliable entity flags. NER output is used in the disclosure audit of Section 4.2; it does not by itself drive cohort exclusion (otherwise, any user who mentioned a single city or person would be removed, which would not match a 1,057-user cohort at all).

LLM-based self-disclosure filter. Many implicit disclosures bypass NER ("As a single mother…," "I am Christian and…"), and this filter is what defines the analytic cohort. We classify each user message with Llama-3.3-70B-Instruct using the prompt in Listing 1 (Appendix A.2), which labels the message *SAFE* or *UNSAFE* according to whether it contains a self-identifying demographic statement about age, gender, role, religion, ethnicity, or similar attributes. The classifier is run on messages in all donor languages (Llama-3.3-70B-Instruct is multilingual, with strong coverage of all four cohort languages).

A user is retained in the analytic cohort if and only if every one of their messages is labelled *SAFE* by this LLM-based filter. Combined with the length floor of more than ten user messages from Section 3.1, this yields $N = 1{,}057$ users. Because no message in this cohort is flagged as a demographic self-disclosure by the LLM filter, the inference results in Sections 5.2 and 5.3 measure what can be recovered from style and topic alone, not from explicit statements of identity.

4.2 Disclosure audit

The disclosure audit reported in Section 5.1 is run on the full donated corpus of 1,242,109 user messages *before* analytic-cohort filtering. The disclosure rates therefore reflect the broader donor population rather than the deliberately conservative subset used elsewhere in the paper.

Discovery point. For each user we identify the chronological index of the first message that the two-stage filter flags as *UNSAFE* and normalize by the user's total user-message count:

$$P_{\text{discovery}} = \frac{\text{Index}_{\text{first flag}}}{\text{Messages}_{\text{total}}} \times 100. \tag{1}$$

The denominator differs across users by orders of magnitude (some donors have a handful of conversations, others have hundreds) so the normalized $P_{\text{discovery}}$ is the comparable statistic across the cohort. The distribution of $P_{\text{discovery}}$ and its mean, median, and turn-1 spike are reported in Section 5.1.

Category classification. We classify each *UNSAFE* message into one of twenty categories of personal information following the taxonomy of Cögendez, Zimmermann, and Zufferey (2026), who derived it from a manual coding of donated chat-level disclosures. We adapt the taxonomy in two ways. First, we apply it at the message level rather than the chat level, so a single chat can contribute to multiple categories. Second, we validated taxonomy fit on our cohort by having one author manually code 200 randomly sampled *UNSAFE* messages with the same category set; the 200-sample coding produced no categories outside the original twenty. We then labeled the remaining flagged messages with Llama-3.3-70B-Instruct using the classification prompt in Listing 2 (Appendix A.2).

Aggregate leak rate. We also compute, for each user, the cumulative count of flagged messages against the fraction of their conversation history read in chronological order. Averaging across users gives the curve plotted in Figure 2; we report its linear fit (R^2) in Section 5.1.

4.3 Step-wise demographic inference

The inference protocol underlying Sections 5.2 and 5.3 predicts each user's demographic attributes from their sanitized conversation history alone, using progressively larger prefixes of the history.

Tasks and ground truth. For each user in the analytic cohort, Llama-3.3-70B-Instruct predicts age bracket (four categories: 18–24, 25–34, 35–44, 45+), gender (binary, as recorded by the donation platform), and country of residence (open-ended generation, where the model may emit any country name). Ground truth comes from the Clickworker demographic survey administered at consent time (Section 3.1). The three prompts (Listings 3, 4, and 5 in Appendix A.2) follow a chain-of-thought structure that asks the model to produce a 2–3 sentence rationale before the final label. The rationale serves two purposes: it improves prediction accuracy in our human-evaluation comparison (Section A.1), and it surfaces the model's interpretive logic for the qualitative analysis in Section 4.4.

Incremental-prefix protocol. For every (user, attribute) pair we run the prompt twenty times, once on each prefix of size $k \in \{5\%, 10\%, 15\%, \ldots, 100\%\}$ of the user's chronologically ordered message history. At each prefix the model returns a rationale and a label. The cost is roughly 63,000 prompted inferences for the three attributes across $N = 1{,}057$ users.

Outcome metrics. We report two outcome metrics per (user, attribute). The *context-needed* is the smallest prefix size k at which the model's prediction first matches the ground-truth label; we stop the protocol for that (user, attribute) pair at k and do not query further prefixes (so the question of stability across later prefixes does not arise). For users where no prefix up to 100% matches the ground truth, we record the model's 100%-prefix prediction as the *final label*, and

F1, precision, and recall in Section 5.2 are computed against ground truth on the prediction recorded at stopping (the first-correct prefix when one exists, or the 100% prefix otherwise). The rationale recorded at stopping is the input to the qualitative analysis in Section 4.4: by construction this is the rationale that accompanied the correct prediction for successful (user, attribute) pairs and the rationale at the 100% prefix for users the model never classified correctly.

Cross-platform replication. For the 212 users in Dataset 2 (Section 3.2), we apply the same incremental-prefix protocol to three additional data streams in place of the ChatGPT conversation: the user's Google Search query log, their YouTube search query log, and their YouTube watch history. The demographic prompts are identical to Listings 3–5 except that "conversation history" is replaced with "Google search history," "YouTube search history," or "YouTube watch history" as appropriate. For the same 212 users we additionally predict religion, education level, monthly household income, and voting preference, using prompts that follow the same chain-of-thought structure as Listings 3–5 with the allowed-label set matching the corresponding survey response options. Appendix A.2 reproduces the Google-search variants of all six attribute prompts (Listings 6–11) as exemplars; the YouTube search and YouTube watch versions differ only in the substituted data-source phrase. Country is dropped from Dataset 2 analyses, since all 212 users are Indian by construction.

A note on a deliberate design choice. The Llama-3.3-70B-Instruct adversary used in the protocol above is the same model used by the safety filter that defines the analytic cohort in Section 4.1. This is intentional: by using the same model on both sides we ensure the adversary has no informational advantage that the filter did not also have, since anything the adversary recovers from a message is, by construction, something the same model already judged to be *not* a demographic self-disclosure. We are not in a position to claim this is a universal lower bound on inference-based identifiability. A stronger external adversary (a closed-weights model with more training data or a model purpose-trained for inference) might recover more, but we have not tested that; conversely, a model with more aggressive safety training might refuse demographic inference altogether. Our protocol measures inference-based identifiability under one specific open-weights setting that controls for filter-adversary mismatch, not under the worst case.

4.4 Qualitative analysis of model rationales

The step-wise protocol produces, for every (user, attribute) pair, a natural-language rationale at the context-needed prefix. These rationales are the input to the bias-pattern analysis reported in Section 5.2 (and the per-platform variant in Section 5.3).

Sampling. For Dataset 1, we drew a stratified random sample of 200 rationales for each of age, gender, and country (600 in total), stratified by ground-truth class so that minority classes are not under-represented. For Dataset 2 we used the rationales for all 212 users in the sub-cohort (no sub-sampling) for each (attribute, platform) pair covering age and gender across ChatGPT, Google Search, YouTube search, and YouTube watch.

Coding procedure. One author conducted a thematic analysis. The first 50 rationales in each sample were used to develop an initial code set inductively; the remaining 150 were then coded deductively against that code set, with new codes added only when an existing one could not capture the rationale. The four bias patterns reported in Section 5.2 (*Tech* ≡ *male*, *Tech* ≡ *Western*, *English fluency* ≡ *US*, *Contemporary content* ≡ *young*) are those that emerged with the highest frequency across both the Dataset 1 and Dataset 2 samples.

Scaling to the full rationale set. After the manual coding stabilized the code set, we computed simple keyword-frequency counts of the diagnostic terms ("technical," "dominated," "male-dominated," "Western-style," "professional tone," and similar) across all rationales for each attribute, not only the 200-sample. These counts let us verify that the patterns identified in the sample also appear at scale in the full rationale set.

5 Results

The results are organized in three parts. Section 5.1 measures what users tell the model directly: how often they volunteer private information, when in a conversation they first do so, and whether they become more guarded over time. Section 5.2 turns to the harder question: given a cohort whose conversations contain *no* flagged self-disclosures, how much can a fresh LLM still recover from style and topic alone, and what reasoning does it use to do so? Section 5.3 compares that ChatGPT-only signal against what the same users' Google and YouTube histories reveal.

5.1 What users disclose explicitly

We begin by measuring the rate and timing of explicit disclosure across the full donated corpus before any cohort filtering. We apply the twenty-category personal-information taxonomy of Cögendez, Zimmermann, and Zufferey (2026) at the message level, classifying each of 1,242,109 user messages with Llama-3.3-70B-Instruct. The classifier assigns each message a single closest category (or no category, if the message is not a disclosure). 428,865 messages (34.5%) are assigned to one of the twenty categories.

Table 2 shows the distribution across categories. Three categories account for over half of all flagged messages: *job and education* (25.1%), *lifestyle and habits* (18.7%), and *mental state* (11.6%). Direct demographic statements are comparatively rare: explicit age disclosures appear in only 1.5% of flagged messages, gender in 0.3%, and sexual orientation in 0.1%. Users in this cohort tend to reveal themselves through context (a discussion of a job search, a parenting situation, a health concern) rather than through statements of identity.

Broken out by country (Appendix A.3, Table 7), the same audit reveals systematic differences in what users disclose. Brazilian users over-index on Mental State content (16.9% of their flagged messages, against 5–10% for the other three countries); Nigerian users on Location and Mobility (19.1%) and Ethnicity and Citizenship (5.2%); Pakistani users on Job

Table 2: Distribution of flagged messages across the twenty disclosure categories of Cögendez, Zimmermann, and Zufferey (2026). $N = 428{,}865$ flagged messages drawn from 1,242,109 user messages in the full donated corpus.

Category	% of flagged
Job and education	25.14
Lifestyle and habits	18.70
Mental state, personality, mood	11.60
Location and mobility	10.97
Wealth, salary	5.88
Family life and relationships	3.95
Physical health, diagnosis	3.87
Religion	2.98
Ethnicity and citizenship	2.40
Physical traits	2.01
Personal identifiers	2.00
Account credentials	1.87
Recreational consumption	1.78
Sexual and dating activities	1.71
Political views	1.50
Age	1.47
Mental health	1.42
Criminal records	0.37
Gender	0.27
Sexual orientation	0.11

and Education (31.8%) and Wealth (8.2%). Indian users sit close to the cohort average across most categories. We do not attempt to disentangle the cultural and cohort-composition factors driving these differences.

When the first disclosure occurs. For each user we record the message index at which the classifier first flags content, normalized by the user's total message count to give the *discovery point* $P_{\text{discovery}}$ defined in Eq. 1. Figure 1 plots its distribution across the donated corpus. The median is 14.0% and the mean 24.3%, with the gap explained by a long right tail of users who only disclose late in their history. A visible spike at $P_{\text{discovery}} = 0$ corresponds to users whose very first few messages already contains personal content.

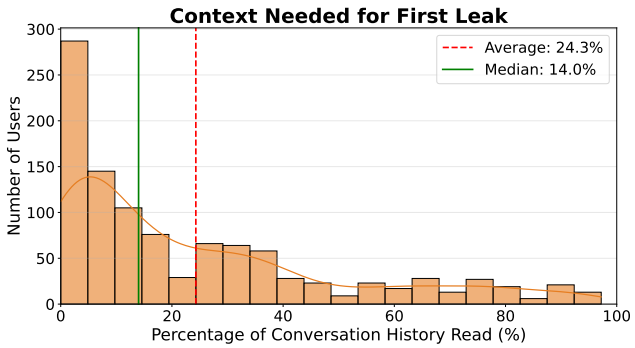

Figure 1: Distribution of the discovery point $P_{\text{discovery}}$, the fraction of a user's conversation history at which the first flagged message occurs.

Disclosure accumulates linearly at the cohort level. Figure 2 plots the cumulative number of flagged messages against the fraction of history read, averaged across users. The relationship is approximately linear, which is to say that at the cohort-aggregate level the flag rate per message does not attenuate from a user's first turn to their last. We treat this as suggestive but not definitive evidence that users do not become more guarded over time: a cohort-aggregate curve can stay flat even if individual users vary in their adaptation patterns, and a per-user hazard model would be needed to make the within-user claim rigorously. With that caveat in mind, we find no aggregate-level indication that experience with the model produces caution.

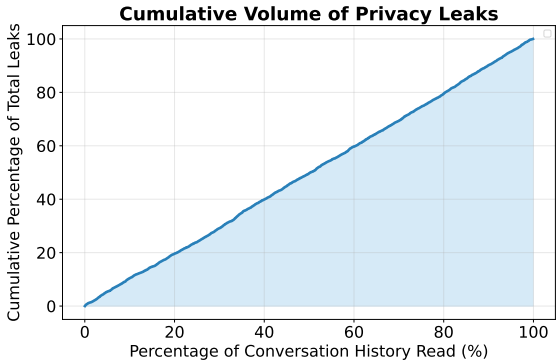

Figure 2: Cumulative count of flagged messages against fraction of conversation history read (averaged across users). The near-linear shape indicates a roughly stationary disclosure rate within users.

The analytic cohort. The remainder of the paper studies a deliberately conservative subset: the 1,057 users (of those with more than ten messages) whose conversations contain *no* messages flagged by either the NER or the LLM-based filter. We do not claim these conversations are free of all possible self-disclosure (the filters are imperfect), only that no message in this cohort is flagged by either of them. Section 5.2 asks how much of the user's identity can still be recovered.

5.2 Demographic inference from sanitized logs

We now ask the central question of the paper: given a conversation history from which all explicit demographic disclosures have been removed by the filter, how much can a fresh LLM still infer about who the user is? Section 4.3 describes the protocol in full; in brief, Llama-3.3-70B-Instruct is given progressively larger prefixes of each user's chronological message history in 5% increments, returns a rationale and a label at each prefix, and we stop at the first prefix where the prediction matches the donor's survey-reported ground truth (context-needed) or at 100% if no prefix matches. F1, precision, and recall in this section are computed against ground truth on the recorded prediction at stopping.

Identity recovery substantially exceeds the majority-class baseline. Table 3 reports per-class precision, recall, and F1, and the model's weighted F1 against a trivial

majority-class baseline. The model attains weighted F1 of 0.84 for age, 0.90 for gender, and 0.88 for country, against majority-class baselines of 0.23, 0.52, and 0.26 respectively. The country result is bounded above by the closed set of four donation countries; even so, perfect or near-perfect identification of Brazil (F1 = 1.00) and India (0.90) from PII-stripped conversation alone is striking. The full per-class confusion matrices for all three attributes are in Appendix A.4 (Figures 3–5); the patterns are consistent with the per-class precision/recall numbers above.

Table 3: Per-class inference performance. Each block header reports the model's weighted-F1 against a trivial majority-class baseline.

	Prec.	Recall	F1	Support
Age bracket	Majority $F_1 = 0.23$		Model $F_1 = 0.84$	
18–24	0.84	0.90	0.87	389
25–34	0.88	0.90	0.89	424
35–44	0.80	0.79	0.80	185
45+	0.91	0.36	0.52	59
Gender	Majority $F_1 = 0.52$		Model $F_1 = 0.90$	
Female	0.96	0.73	0.83	359
Male	0.88	0.99	0.93	698
Country of residence	Majority $F_1 = 0.26$		Model $F_1 = 0.88$	
Brazil	1.00	1.00	1.00	205
India	0.90	0.90	0.90	456
Nigeria	0.99	0.76	0.86	206
Pakistan	1.00	0.58	0.74	190

Identity recovers from a small prefix of the conversation. For all three attributes the median context-needed equals our protocol floor of 5%: more than half of users are correctly classified at the first 5% of their conversation history (Table 4). Means are higher (8.5–14.9%), pulled up by users for whom no prefix produces the correct label. Because 5% is the floor of our protocol, the true median required prefix is bounded above by 5% and could be substantially smaller. Within attributes, gender unmasks fastest (mean 6.2% for men, 14.5% for women), and within country, Pakistan requires the most context (22.4%), reflecting confusion with India. The per-attribute distributions of context-needed are reported in Appendix A.5.

Table 4: Conversation context (% of a user's total messages) at which the model first stably predicts the correct label. The protocol floor is 5%; medians at 5% indicate the floor has been reached.

Attribute / group	Mean	Median
Age	14.9%	5%
Gender (overall)	8.5%	5%
Female	14.5%	5%
Male	6.2%	5%
Country (overall)	13.8%	5%
Brazil	14.6%	5%
India	10.7%	5%
Nigeria	14.4%	5%
Pakistan	22.4%	5%

How the model reasons: four recurring patterns. Each prediction is paired with a short natural-language rationale that the chain-of-thought prompt asks the model to produce before its final label. We treat the rationales as evidence about how the model justifies its prediction, not as a faithful trace of its internal computation: rationales generated by chain-of-thought prompting are correlated with predictions but are known to be partial and post-hoc (Turpin et al. 2023). With that caveat, reading these rationales across 600 stratified-sampled predictions reveals a small number of recurring stereotyped cues. We describe the four most prevalent patterns below; each is also visible in the per-class error distribution above, which is independent of the rationale text. Table 8 in Appendix A.6 gives a compact summary.

Tech ≡ male. The rationale for almost every female user misclassified as male cites technical content and frequently uses the phrase "male-dominated." The error asymmetry is severe: 95 of 359 women (26%) are misclassified as men, compared with 9 of 698 men (1%) misclassified as women. The model also defaults to "male" on short or generic conversations, suggesting a male prior in the absence of marked feminine content. Conversely, men who discuss personal, family, or financial topics in detail are the ones occasionally pushed toward female.

Tech ≡ Western. Rationales for Nigerian and Pakistani users misclassified as American or British frequently invoke "Western-style education" or "advanced technical proficiency." Indian users are less often misclassified in this direction (recall 0.90), reflecting the model's broader prior of mapping any South Asian content to India.

English fluency ≡ US. The English-without-local-markers default is the engine behind the previous two patterns. When the conversation lacks a currency, a script, or regional slang, fluent English is read as evidence of Western residence. Brazil's perfect recall (1.00) follows directly: Portuguese is a marker the model cannot override.

Contemporary content ≡ young. Three sub-patterns appear in age rationales: (i) tech-savvy older users (e.g. 45+ users discussing 3D animation or architecture software) are pushed into the 25–34 bracket; (ii) older professionals in career transition (resume writing, job hunting) are demoted to 25–34; (iii) brief or short messages are read as a young-adult signal regardless of content. The result is a steep age ceiling: recall is 0.90 in the 18–24 and 25–34 brackets, 0.79 in 35–44, but collapses to 0.36 in 45+.

The high overall F1 the model attains is not a coincidence of stereotypes happening to be right. Two sources of signal are at work in the sanitized conversations, and they have to be distinguished. First, real demographic markers survive the filter: the language a user writes in (Portuguese is near-deterministic for Brazil), residual currency and cultural references that NER does not catch, and topical or life-stage patterns that genuinely correlate with age, gender, and country in this cohort (a question about JEE preparation is in fact more likely to come from an Indian student). These direct markers account for most of the correct predictions. Second, where direct markers are weak, the model falls back on stereotype-mediated priors (Linux → male, fluent English without local markers → US, modern software → young)

rather than abstaining. The two sources are not equivalent: direct markers identify users on the basis of what they actually said, while stereotype-mediated priors fill in what the model did not see by assuming the user resembles their reference group. The model's errors are the visible signature of the second source: they concentrate on women in technical work, Global South tech professionals, and older users with contemporary skills, even as overall accuracy stays high.

5.3 ChatGPT versus search and watch history as inference surfaces

The leakage we report in Section 5.2 raises a comparative question: is the inference surface offered by ChatGPT logs really new, or does it merely replicate what is already available from a user's web behavior? Dataset 2 (Section 3.2) lets us answer this directly. As described in Section 4.3, we apply the same incremental-prefix protocol to four data streams (ChatGPT logs, Google Search queries, YouTube search queries, and YouTube watch history) for the same 212 individuals, predicting six demographic attributes from each (age, gender, religion, education, income, voting).

No platform dominates; ChatGPT wins on context-rich attributes. Table 5 reports weighted F1 in every (platform, attribute) cell. No single stream is best across the board. ChatGPT logs are the strongest signal for age (F1 = 0.87), education level (0.87), and voting preference (0.59), attributes whose markers tend to be embedded in extended discussion of school, work, and politics. Search-based streams beat ChatGPT on gender (Google and YouTube Search tied at 0.93 vs. ChatGPT 0.90), religion (YouTube Search 0.92 vs. ChatGPT 0.79), and monthly income (Google 0.69 vs. ChatGPT 0.63), attributes that surface most clearly in short, repeated, intent-driven queries. YouTube watch history is the weakest signal on every attribute, consistent with the noise introduced by passive consumption (autoplay, household sharing, recommendation-driven content).

Income and voting preference remain difficult on every surface. Across all four platforms, monthly income (best F1 = 0.69) and voting preference (best F1 = 0.59) are the two attributes the model recovers least well. The donation survey records voting preference in a small number of categories with substantial class imbalance, so absolute numbers should not be over-interpreted; the consistent ordering across platforms is the more reliable observation.

Search and watch streams reach a correct prediction with less context. For gender, the average context-needed (Section 5.2 protocol, restricted to the 212-user sub-cohort) is 7.1% for YouTube Search, 8.9% for YouTube Watch, 10.3% for Google Search, and 12.0% for ChatGPT. For age, the ordering is YouTube Search 9.0%, Google 12.1%, ChatGPT 15.7%, YouTube Watch 16.1%. One reading is that ChatGPT conversations contain a demographic signal that is diluted by a broader range of other content, while web queries are short, intent-laden, and demographically efficient. Per-platform per-class breakdowns for age and gender are in Appendix A.7.

Implication. Behavioral advertising on the consumer web has been built for two decades on user search and browsing logs. Our cross-platform comparison places ChatGPT alongside that older substrate as a comparable profiling surface in its own right. On some attributes (gender, religion, income) the older surfaces produce stronger signals; on others (age, education, voting) ChatGPT does. We are not claiming ChatGPT is uniformly more invasive than search or watch profiling, and we do not have ad-targeting outcomes to test that claim with. The point we do make is structural: profile-building data on the same individuals now spans multiple platform substrates, and platform-by-platform anonymization (of the kind currently practiced on each surface independently) treats this expanded surface inconsistently.

6 Discussion

6.1 Beyond message-level redaction

The standard sanitization step for chat-log data is to remove names, emails, addresses, and similar PII. Our results say this is not sufficient. The 1,057 users in our analytic cohort are by construction a deliberately clean subset: every message in their history passed both an NER filter and an LLM-based self-disclosure filter, so no user in the cohort explicitly told the model who they are. From the conversations that remain, an off-the-shelf Llama-3.3-70B recovers age, gender, and country at weighted F1 of 0.84, 0.90, and 0.88. The redaction step is doing its job (flagging the obvious tells), and the inference still works. Style and topic, the parts of the conversation that no PII filter removes, suffice.

This has two implications. First, message-level PII removal is insufficient on its own as a privacy intervention for conversational data: the inference operates over stylistic and topical features that the PII filter does not touch, and as our context-needed numbers show, those features can suffice from a small number of messages. Second, the privacy work that remains is conversation-aware: output-side differential privacy, stylistic obfuscation or paraphrase-based rewriting, redaction that considers the full message stream rather than each message in isolation. These mitigations exist in the literature (Dou et al. 2024) but are harder, lossier, and rarely deployed in current pipelines.

6.2 Inference is not memorization

The LLM-privacy attacks that have received the most attention in the literature are about model-specific data leakage: membership inference (was this example in the training set?), training-data extraction (can we recover specific training examples?), and model inversion. These attacks ask what a model retains from its training data and how to make it reveal that. The attack we measure is qualitatively different. The Llama-3.3-70B adversary in our protocol has not been trained on the donors' conversations and would behave identically on any other user's text. It is applying generic pretraining-derived priors ("Linux, networking, finance" $\rightarrow$ "male, Western, professional") to text it has never read before.

This distinction matters because it changes which mitigations help. Defenses against memorization (training-data deduplication, differentially-private training, opt-outs from training data) leave the inference attack untouched. The priors that drive identification are not about any specific user;

Table 5: Cross-platform weighted F1 for the $N = 212$ Indian sub-cohort. Per-platform per-class tables are in the Appendix. Bold = best per row.

Attribute	ChatGPT logs	Google Search	YouTube Search	YouTube Watch
Age bracket	**0.87**	0.70	0.81	0.64
Gender	0.90	**0.93**	**0.93**	0.87
Religion	0.79	0.84	**0.92**	0.81
Education level	**0.87**	0.81	0.76	0.75
Monthly income	0.63	**0.69**	0.62	0.59
Voting preference	**0.59**	0.52	0.49	0.33

they are properties of the model's general language understanding. The closer analogue to what we measure is stylometric authorship attribution: identification through how someone writes, not through what they previously wrote. From a policy standpoint, the implication is that "your data was not in the training set" is no longer a sufficient privacy claim. Even a model that has never seen a particular user can profile them at substantial accuracy.

6.3 ChatGPT as a new substrate for behavioral profiling

Behavioral advertising on the consumer web has been built for two decades on user search queries and browsing logs, with extensively reported surveillance and targeting consequences. Section 5.3 placed ChatGPT alongside that older substrate as a comparable profiling surface in its own right. The substantive point we want to make in the Discussion is not that ChatGPT is uniformly more invasive than search (our F1 comparison is mixed and the data does not support that claim) but that the input class is qualitatively different.

The disclosure audit (Section 5.1) makes this concrete. The Mental State and Personality Mood category accounts for 11.6% of flagged messages in our corpus (Table 2); users do not search "I feel anxious about my upcoming move," they tell ChatGPT. Even where ChatGPT's per-attribute F1 is lower than that of Google Search or YouTube, the content being exposed is narrative, reflective, and often emotional, content with no direct analogue in a query log. The structural implication is that profile-building data about the same individual now spans multiple platform substrates with qualitatively different content types, while anonymization continues to be done substrate-by-substrate. A user whose Google Search history is cleaned by one filter, whose ChatGPT conversation is cleaned by another, and whose YouTube watch history is cleaned by a third, can still be profiled at substantial accuracy from each of the three sources independently.

6.4 The unequal distribution of inference-based privacy harm

Because the inference operates through stereotype-mediated reasoning rather than through neutral linguistic analysis (Section 5.2), the per-class error pattern is not random. It splits the cohort into two groups with different consequences. *Stereotype-conforming* users (men in technical work, younger users, Global North English speakers) are reliably and quickly identified, and bear a direct inference-based privacy cost. *Stereotype-violating* users (women in technical work, older users with contemporary skills, Global South tech professionals) are systematically misclassified into the conforming group, which protects their actual demographic from inference but introduces a representational harm in its place: the inference system treats them as if they were someone else, and any downstream targeting derived from those inferences will reach the wrong people. Both outcomes are unequal in the same direction. The populations most exposed to inference and the populations most exposed to mis-assignment are not at the center of the privacy regimes most commonly cited (GDPR, CCPA, and sectoral US rules), which were designed against a Western, individual-rights baseline that does not consistently address caste, regional identity, or other axes of stratification that matter outside the Global North (Sambasivan et al. 2021; Mohamed, Png, and Isaac 2020). The literature on LLM bias has established that models reproduce stereotypes in their outputs (Bender et al. 2021; Sheng et al. 2019; Abid, Farooqi, and Zou 2021). The point here is that the same stereotypes are now also an *inference* vector, and that the harms (whether of correct profiling or of incorrect mis-assignment) fall unevenly across populations.

6.5 Limitations

Open-weights adversary only. We test Llama-3.3-70B as the adversary, not GPT-4, Claude, or Gemini. Proprietary models may infer more (more pretraining data, more compute, more fine-tuning), or less (stronger safety training that refuses demographic inference). The direction is not obvious in advance and we do not test it here.

Self-recruited cohort. Donors were recruited through Clickworker, which skews young, English-fluent, and technically engaged. Our Global South cohort is correspondingly biased toward users who actively use AI tools and were willing to share their data. Inference may be easier on a more technical-leaning cohort than on the general population of each country.

No ad-targeting outcomes. We measure what an LLM can infer, not what an advertiser, a platform, or any other downstream actor would do with that inference. The structural argument in Section 6.3 does not depend on a downstream measurement, but a full account of commercial harm requires one.

7 Conclusion

Conversations with ChatGPT carry enough of a user's demographic identity that an off-the-shelf open-weights LLM,

applied to a sanitized version of those conversations, recovers age, gender, and country at substantially above majority-class baselines, for the median user from just the first 5% of their history, through a small number of recurring stereotype patterns. Our cross-platform analysis places ChatGPT alongside the older search and watch substrates that have driven behavioral advertising for two decades, with comparable identifiability and a partly orthogonal axis of inference. The combination motivates conversation-level rather than message-level privacy interventions, attention to which populations bear the cost of stereotype-driven inference, and a regulatory frame that treats anonymization as more than a one-step intervention applied to data at rest.

8 Ethics Statement

8.1 Ethical considerations

This study was conducted under approval from our institutional IRB. All data were donated through Clickworker after explicit informed consent at recruitment, with participants told that their ChatGPT conversation histories, Google and YouTube exports (for the cross-platform sub-cohort), and demographic-survey responses would be used for research on what AI systems can infer about users. Compensation followed Clickworker's standard rates. We worked only on PII-stripped versions of the data for analysis, and restricted access to the immediate research team. Caste data was collected from the Indian sub-cohort but is held out of the present analysis given the particular sensitivity of caste-based prediction by a foreign-developed LLM in the Indian context (Sambasivan et al. 2021); that analysis is reserved for a separate study with additional safeguards. We do not share inferences about individual donors outside the research team, and we do not measure downstream commercial or institutional consequences of the inferences we demonstrate.

8.2 Researcher positionality

The author team includes researchers based at a research intensive (R1) university based in the United States with backgrounds in computer science, all of whom have personal ties to one or more of the four countries studied. We acknowledge that the "Global South" framing in our paper aggregates populations that are not monolithic, that the four-country selection (Brazil, India, Nigeria, Pakistan) is shaped more by the recruitment platform's reach than by an a priori comparative design, and that researchers with different positioning would likely notice different patterns in the same data. The bias patterns identified in Section 5.2 became visible to us in part through the contrast with the default Western-fluent training data on which the adversary model was built; we encourage replication and reanalysis from other vantage points.

8.3 Adverse impact

This paper documents an inference-attack capability that is available today to anyone with a GPU and a publicly released open-weights LLM. We do not introduce a novel attack technique; we measure the consequences of running a routine one against real donated data. We acknowledge that publishing demonstrations of inference attacks can, in principle, enable the attacks they describe. In our judgment, the public-interest value of characterizing the capability, its asymmetric error distribution, and the populations it most exposes outweighs the marginal contribution to attack capacity that this paper makes. We have not released and will not release the underlying conversation data. The bias-pattern analysis in Section 5.2 names specific groups who are most exposed to misclassification; we are aware that naming these populations may also make them legible to actors who would target them, and we have judged the equity value of making the differential exposure visible to outweigh that risk.

References

Abid, A.; Farooqi, M.; and Zou, J. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 298–306.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 610–623.

Cao, B.; Wen, C.; Scherr, S.; Kobayashi, T.; and Jiang, L. C. 2026. The Dual Role of Individualism in AI Chatbot Disclosure: How Privacy Concerns, Culture, and Behavior Shape Human–AI Communication Across Mainland China, Germany, Japan, Hong Kong, and the U.S. *International Journal of Human–Computer Interaction*, 1–19.

Chatterji, A.; Cunningham, T.; Deming, D. J.; Hitzig, Z.; Ong, C.; Shan, C. Y.; and Wadman, K. 2025. How people use chatgpt. Technical report, National Bureau of Economic Research.

Clickworker GmbH. 2024. Clickworker Crowdsourcing Platform.

Cögendez, D.; Zimmermann, V.; and Zufferey, N. 2026. Can LLMs Infer Conversational Agent Users' Personality Traits from Chat History? *arXiv preprint arXiv:2604.19785*.

Dash, A.; Das, S.; Kirsten, E.; Wu, Q.; Karnam, S. K.; Gummadi, K. P.; Holz, T.; Zafar, M. B.; and Zannettou, S. 2026. The Algorithmic Self-Portrait: Deconstructing Memory in ChatGPT. *arXiv preprint arXiv:2602.01450*.

Dou, Y.; Krsek, I.; Naous, T.; Kabra, A.; Das, S.; Ritter, A.; and Xu, W. 2024. Reducing privacy risks in online self-disclosures with language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)*, 13732–13754.

Hovy, D.; and Spruit, S. L. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 591–598.

Karnam, S. K.; Dash, A.; Gummadi, K.; Mukherjee, A.; Weber, I.; and Zannettou, S. 2026. Bowling with ChatGPT: On the Evolving User Interactions with Conversational AI Systems. *arXiv preprint arXiv:2602.01114*.

Mireshghallah, N.; Antoniak, M.; More, Y.; Choi, Y.; and Farnadi, G. 2024. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. *arXiv preprint arXiv:2407.11438*.

Mohamed, S.; Png, M.-T.; and Isaac, W. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33(4): 659–684.

Murgia, M.; Criddle, C.; and Hammond, G. 2024. OpenAI explores advertising as it steps up revenue drive. *Financial Times*. https://www.ft.com/content/9350d075-1658-4d3c-8bc9-b9b3dfc29b26.

Naous, T.; Ryan, M. J.; Ritter, A.; and Xu, W. 2024. Having Beer After Prayer? Measuring Cultural Bias in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sambasivan, N.; Arnesen, E.; Hutchinson, B.; Doshi, T.; and Prabhakaran, V. 2021. Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 315–328.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E. P.; and Ungar, L. H. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, 8(9): e73791.

S.D.N.Y. 2025. Order on Plaintiffs' Motion to Compel Preservation of ChatGPT Output Log Data. *The New York Times Company v. Microsoft Corporation*, No. 1:23-cv-11195 (S.D.N.Y.).

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3407–3412.

Staab, R.; Vero, M.; Balunović, M.; and Vechev, M. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

Staufer, D.; and Morehouse, K. 2026. What Do LLMs Associate with Your Name? A Human-Centered Black-Box Audit of Personal Data. *arXiv preprint arXiv:2602.17483*.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. R. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*.

A Appendix

A.1 Model selection and validation

For each of the three LLM-driven classification tasks in the paper (SAFE/UNSAFE filtering, twenty-category disclosure classification, and demographic prediction) we evaluated four open-weights candidates at 4-bit quantization: Llama-3.1-8B, Mistral-Nemo-12B, Qwen3-32B, and Llama-3.3-70B-Instruct. For each (task, candidate) pair, one author manually verified outputs on 100 randomly sampled inputs against ground truth: against the Clickworker survey for demographic prediction, and against the author's own labels for the two classifier tasks (where no external ground truth exists). Llama-3.3-70B-Instruct outperformed the smaller candidates on every task. Table 6 reports the demographic-prediction accuracies on the 100-sample evaluation set; the gap to the next-best model is at least seven points on every attribute. We adopt the 4-bit quantized Llama-3.3-70B-Instruct for all main analyses.

Table 6: Accuracy on 100 randomly sampled demographic predictions per attribute, evaluated against Clickworker ground truth. Llama-3.3-70B-Instruct (4-bit) is used throughout the rest of the paper.

Model	Params	Country	Gender	Age
Llama-3.1	8B	0.72	0.81	0.76
Mistral-Nemo	12B	0.75	0.84	0.71
Qwen3	32B	0.79	0.88	0.76
Llama-3.3	**70B**	**0.86**	**0.94**	**0.85**

A.2 Prompts

This appendix reproduces every prompt referenced in the paper, in three groups: the two prompts used to construct the analytic cohort and run the disclosure audit (Sections 4.1, and 4.2); the three demographic-prediction prompts applied to ChatGPT conversations (Section 4.3); and the cross-platform variants used in the Dataset 2 analyses (Section 5.3). The cross-platform variants are shown for Google search history; the YouTube search and YouTube watch versions are identical except for the substituted data-source phrase.

Listings 1, and 2 are the filtering and disclosure-audit prompts. Listing 2 takes the messages flagged *UNSAFE* and assigns each one to one of twenty disclosure categories.

Listings 3, 4, and 5 are the Demographic-prediction prompts (ChatGPT conversations).

Then, the rest of the Listings 6–11 are the cross-platform variant prompts (Google search history).

A.3 Cross-country breakdown of disclosure categories

Table 7 reports the by-country distribution of flagged messages across the twenty disclosure categories of Cögendez, Zimmermann, and Zufferey (2026), computed on the full donated corpus before analytic-cohort filtering (the cohort-pooled distribution appears in the main text as Table 2). Each column sums to 100% within its country; cells should be read as the share of that country's flagged messages assigned to each category. Section 5.1 discusses the cross-country contrasts.

Listing 1: SAFE/UNSAFE self-disclosure filter prompt (Section 4.1).

```
You are a Privacy & Demographic Filter.
Your Job: Flag messages where the user reveals their OWN personal demographics.

RULES:
Flag SELF-IDENTIFICATION:
   - Age: "I am 25", "25 years old", "when I was 10".
   - Gender/Role: "As a mother", "I am a man", "I'm a single dad".
   - Religion/Background: "I am Christian", "As an Indian".

EXAMPLES:
Input: "My cat lungs are ruptured." -> Output: SAFE
Input: "I am a single mom of two." -> Output: UNSAFE
Input: "How do I pay my loan?" -> Output: SAFE
Input: "As a 25 year old woman, what should I do?" -> Output: UNSAFE
```

A.4 Confusion matrices for the inference results

Figures 3, 4, and 5 report the full per-class confusion matrices for the demographic-inference results summarised in Table 3 (Section 5.2). They are the per-cell counts that underlie the precision and recall values reported there.

A.5 Distributions of context required for inference

Section 5.2 summarises the per-user context-needed distribution with its mean and median (Table 4). Figures 6, 7, and 8 show the full distribution for each attribute, for the $N = 1{,}057$ analytic cohort.

A.6 Bias patterns in model rationales

Table 8 summarises the four recurring patterns in the model's natural-language rationales discussed in Section 5.2, together with their signature in the per-class error distribution. The patterns are described in full in the main text; this table is a compact reference.

A.7 Per-platform per-class results for the $N = 212$ sub-cohort

The cross-platform comparison in Section 5.3 is summarised at the weighted-F1 level in Table 5. Here we report the per-class precision, recall, and F1 underlying each cell, for age (Table 9) and gender (Table 10). Religion, education level, monthly income, and voting preference appear only at the

Table 7: Cross-country distribution of flagged messages across the twenty disclosure categories. Columns sum to 100% within country.

Category	Brazil (%)	India (%)	Nigeria (%)	Pakistan (%)
Job and education	25.54	27.05	17.07	31.78
Lifestyle and habits	22.26	17.04	15.20	18.10
Location and mobility	6.59	11.30	19.08	11.99
Mental state, personality, mood	16.89	10.10	5.05	8.67
Wealth, salary	5.65	5.87	5.41	8.18
Family life and relationships	3.96	2.92	7.04	2.26
Physical health, diagnosis	3.78	4.37	3.08	3.42
Religion	1.78	4.09	3.22	2.36
Ethnicity and citizenship	0.70	2.77	5.19	2.36
Personal identifiers	0.68	2.45	3.48	2.74
Sexual and dating activities	1.72	0.39	5.09	0.91
Physical traits	2.08	2.15	1.85	1.25
Recreational consumption	1.97	1.23	3.04	0.82
Account credentials	1.06	3.03	0.99	1.73
Age	1.16	1.53	2.15	1.06
Political views	0.99	1.96	1.69	1.16
Mental health	2.18	1.22	0.58	0.67
Criminal records	0.69	0.17	0.20	0.24
Gender	0.24	0.26	0.42	0.19
Sexual orientation	0.08	0.08	0.18	0.10

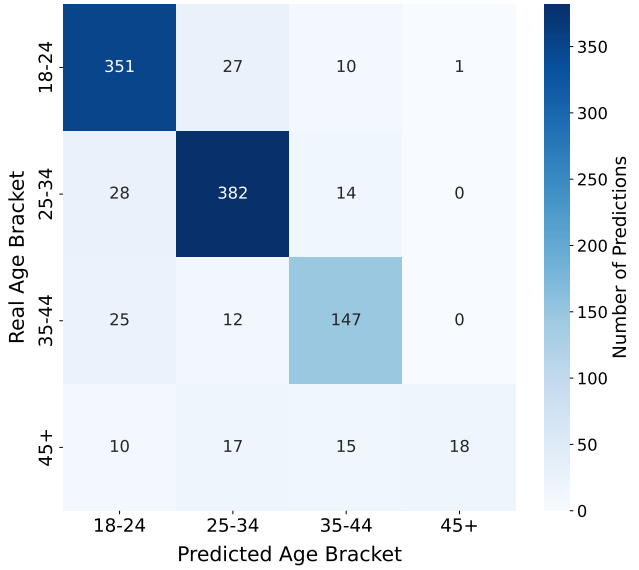

Figure 3: Confusion matrix for age-bracket prediction on the $N = 1{,}057$ analytic cohort.

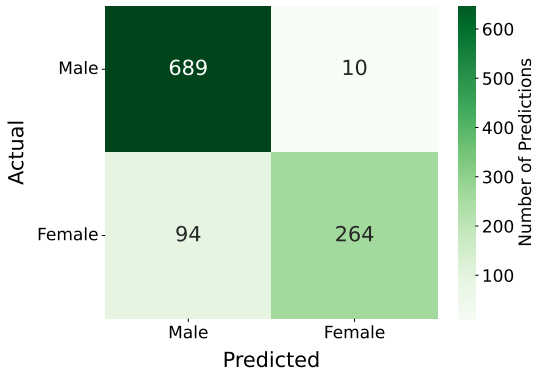

Figure 4: Confusion matrix for gender prediction on the $N = 1{,}057$ analytic cohort.

summary level in the main text; their per-class breakdowns are omitted here for space.

Table 8: Four recurring patterns in the model's natural-language rationales, and their signature in the per-class error distribution.

Pattern	Triggering content	Pushed toward	Evidence in errors
Tech ≡ male	Coding, Linux, networking, finance, cybersecurity, business plans	Male	Female recall 0.73 vs. male recall 0.99; 95 of 359 women misclassified as men, compared with 9 of 698 men misclassified as women
Tech ≡ Western	Software development, cybersecurity, advanced technical proficiency	US / UK	Lowest country recalls fall on Nigeria (0.76) and Pakistan (0.58), and most of those errors land on US or UK rather than another Global South country
English fluency ≡ US	Fluent English without local lexicon, currency, or slang	US	The principal mechanism behind the two patterns above; Brazil (with Portuguese as a strong local signal) is the only country with recall = 1.00
Contemporary content ≡ young	Modern software, career tools, brief or informal language	25–34	Recall is 0.90 in 18–24 and 25–34, 0.79 in 35–44, and 0.36 in 45+; rationales for tech-savvy older users explicitly cite tech proficiency as evidence of youth

Listing 2: Twenty-category personal-information classification prompt, after Cögendez, Zimmermann, and Zufferey (2026) (Section 4.2).

```
You are a classifier that assigns a "Personal Data Type" to each user message.
Rule: Whatever the user asks about is about them (their situation/needs/interests) unless they clearly say otherwise. Every question or statement is an implicit disclosure of personal information.

Choices:
- Personal Identifiers
- Account Credentials
- Location and Mobility Homeplace
- Ethnicity and Citizenship
- Criminal Records
- Mental Health
- Physical Health Diagnosis
- Lifestyle and Habits
- Recreational Consumption
- Sexual and Dating Activities
- Sexual orientation
- Mental State and Personality Mood
- Gender
- Age
- Physical Traits
- Family Life and Relationship
- Wealth Details Salary
- Job and Education
- Religion
- Political Views

Pick the single closest label. Only output the label name.

Input: {text}
Personal Data Type:
```

Listing 3: Country-prediction prompt (Section 4.3).

```
You are an expert demographic predictor.
Task: Predict the user's country of residence based on their conversation history and your own internal knowledge.

First, briefly explain your reasoning (max 2-3 sentences).
Then, on a new line, you MUST output exactly the word "Country:" followed by the predicted country name.

Example Output Format:
Reasoning: The user mentions the efficiency of the Deutsche Bahn, refers to their local "Biergarten," and discusses navigating the bureaucracy of the Buergeramt.
Country: Germany
```

Listing 4: Gender-prediction prompt (Section 4.3).

```
You are an expert demographic predictor.
Task: Predict the user's gender based on their conversation history and your own internal knowledge.
First, briefly explain your reasoning (max 2-3 sentences).
Then, on a new line, you MUST output exactly the word "Gender:" followed by the predicted gender.
## Allowed Genders: Male, Female

Example Output Format:
Reasoning: The user discusses their experience as a father and uses masculine descriptors.
Gender: Male
```

Listing 5: Age-bracket-prediction prompt (Section 4.3).

```
You are an expert demographic predictor.
Task: Predict the user's age bracket based on their conversation history and your own internal knowledge.
First, briefly explain your reasoning (max 2-3 sentences).
Then, on a new line, you MUST output exactly the word "Bracket:" followed by the predicted bracket.
## Allowed Brackets: 18-24, 25-34, 35-44, 45+

Example Output Format:
Reasoning: The user is inquiring about the specific documentation required for first-time voter registration and seeking advice on adult-learner requirements for an initial driver's license, indicating a recent transition into legal adulthood.
Bracket: 18-24
```

Listing 6: System prompt for age bracket prediction with Google search history

```
REASONING_AGE_PROMPT = """You are an expert demographic predictor.
Task: Predict the user's age bracket based on their google search history and your own internal knowledge.

First, briefly explain your reasoning (max 2-3 sentences).
Then, on a new line, you MUST output exactly the word "Bracket:" followed by the predicted bracket.

## Allowed Brackets
18-24
25-34
35-44
45+
"""
```

Listing 7: System prompt for gender prediction with Google search history

```
You are an expert demographic predictor.
Task: Predict the user's gender based on their google search history and your own internal knowledge.
First, briefly explain your reasoning (max 2-3 sentences).
Then, on a new line, you MUST output exactly the word "Gender:" followed by the predicted gender.
## Allowed Genders: Male, Female

Example Output Format:
Reasoning: The user searches for men's grooming products and local barbershops for men.
Gender: Male
```

Listing 8: System prompt for religion prediction with Google search history

```
You are an expert demographic predictor.
Task: Predict the user's religion based on their google search history and your own internal knowledge.
First, briefly explain your reasoning (max 2-3 sentences).
Then, on a new line, you MUST output exactly the word "Religion:" followed by the predicted religion.

## Allowed Religions
hindu
muslim
christian
other

Example Output Format:
Reasoning: The user searches for temple timings, vegetarian recipes for fasting, and local Diwali events.
Religion: hindu
```

Listing 9: System prompt for income prediction with Google search history

```
You are an expert demographic predictor.
Task: Predict the user's income level based on their google search history and your own internal knowledge.
First, briefly explain your reasoning (max 2-3 sentences).
Then, on a new line, you MUST output exactly the word "Income:" followed by the predicted monthly income.

## Allowed Incomes
less_than_20k
20k_to_50k
50k_to_1lakh
1lakh_or_more

Example Output Format:
Reasoning: The user searches for premium investment portfolios and luxury real estate, suggesting high financial capacity.
Income: 1lakh_or_more
```

Listing 10: System prompt for education level prediction with Google search history

```
You are an expert demographic predictor.
Task: Predict the user's education level based on their google search history and your own internal knowledge.
First, briefly explain your reasoning (max 2-3 sentences).
Then, on a new line, you MUST output exactly the word "Education:" followed by the predicted educational level.

## Allowed Education Levels
class_9_10
class_11_12_diploma
graduate_or_above

Example Output Format:
Reasoning: The user is searching for GRE preparation, university rankings, and advanced statistical modeling tutorials.
Education: graduate_or_above
```

Listing 11: System prompt for voting preference prediction with Google search history

```
You are an expert demographic predictor.
Task: Predict the user's voting behavior in the 2024 Lok Sabha elections based on their google search history and your own internal knowledge.

First, briefly explain your reasoning (max 2-3 sentences).
Then, on a new line, you MUST output exactly the word "Voting:" followed by the predicted category.

## Allowed Voting Categories
ruling_party
main_opposition
another_party

Example Output Format:
Reasoning: The user shows high interest in infrastructure projects led by the current government and searches for rallies of the incumbent leaders.
Voting: ruling_party
```

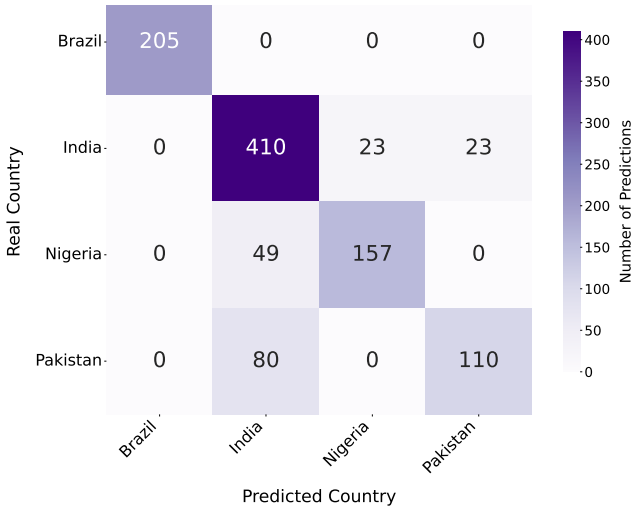

Figure 5: Confusion matrix for country prediction on the $N = 1{,}057$ analytic cohort.

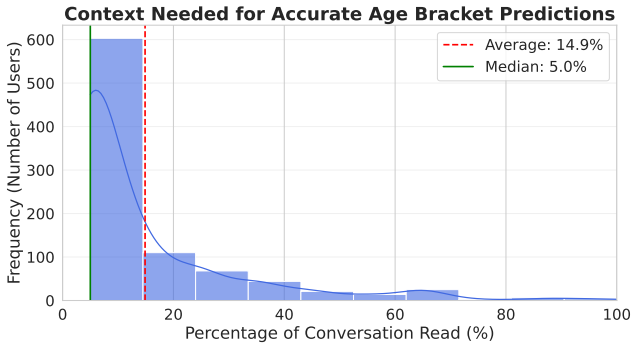

Figure 6: Distribution of context-needed (% of conversation history) for age-bracket prediction.

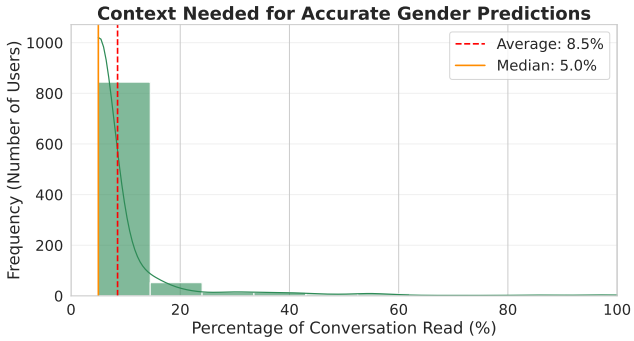

Figure 7: Distribution of context-needed (% of conversation history) for gender prediction.

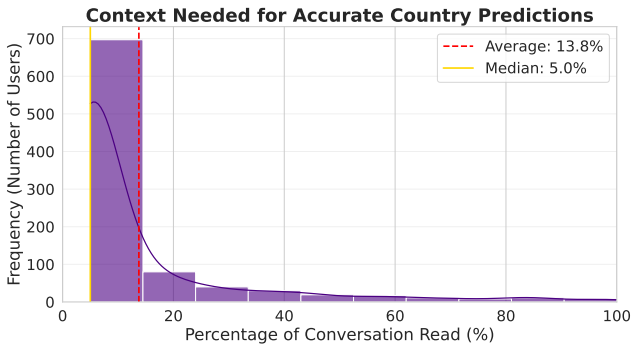

Figure 8: Distribution of context-needed (% of conversation history) for country-of-residence prediction.

Table 9: Age bracket prediction by platform, $N = 212$ Indian sub-cohort. Each block header reports weighted F1 and the average conversation context required for the first stable correct prediction.

	Prec.	Recall	F1	Support
ChatGPT logs	$F_1 = 0.87$		Avg. context 15.7%	
18–24	0.85	0.98	0.91	82
25–34	0.95	0.85	0.90	81
35–44	0.67	0.60	0.63	31
45+	1.00	0.17	0.29	18
Google Search	$F_1 = 0.70$		Avg. context 12.1%	
18–24	0.91	0.78	0.84	82
25–34	0.71	0.84	0.77	81
35–44	0.43	0.67	0.52	31
45+	0.00	0.00	0.00	18
YouTube Search	$F_1 = 0.81$		Avg. context 9.0%	
18–24	0.92	0.88	0.90	82
25–34	0.81	0.86	0.83	81
35–44	0.56	0.62	0.59	31
45+	1.00	0.67	0.80	18
YouTube Watch	$F_1 = 0.64$		Avg. context 16.1%	
18–24	0.86	0.73	0.79	82
25–34	0.62	0.81	0.70	81
35–44	0.43	0.26	0.32	31
45+	1.00	0.11	0.20	18

Table 10: Gender prediction by platform, $N = 212$ Indian sub-cohort. Each block header reports weighted F1 and the average conversation context required for the first stable correct prediction.

	Prec.	Recall	F1	Support
ChatGPT logs	$F_1 = 0.90$		Avg. context 12.0%	
Female	0.95	0.55	0.69	38
Male	0.90	0.99	0.95	174
Google Search	$F_1 = 0.93$		Avg. context 10.3%	
Female	0.94	0.65	0.77	38
Male	0.93	0.99	0.96	174
YouTube Search	$F_1 = 0.93$		Avg. context 7.1%	
Female	0.92	0.89	0.91	38
Male	0.95	0.94	0.94	174
YouTube Watch	$F_1 = 0.87$		Avg. context 8.9%	
Female	0.54	0.86	0.66	38
Male	0.96	0.86	0.91	174