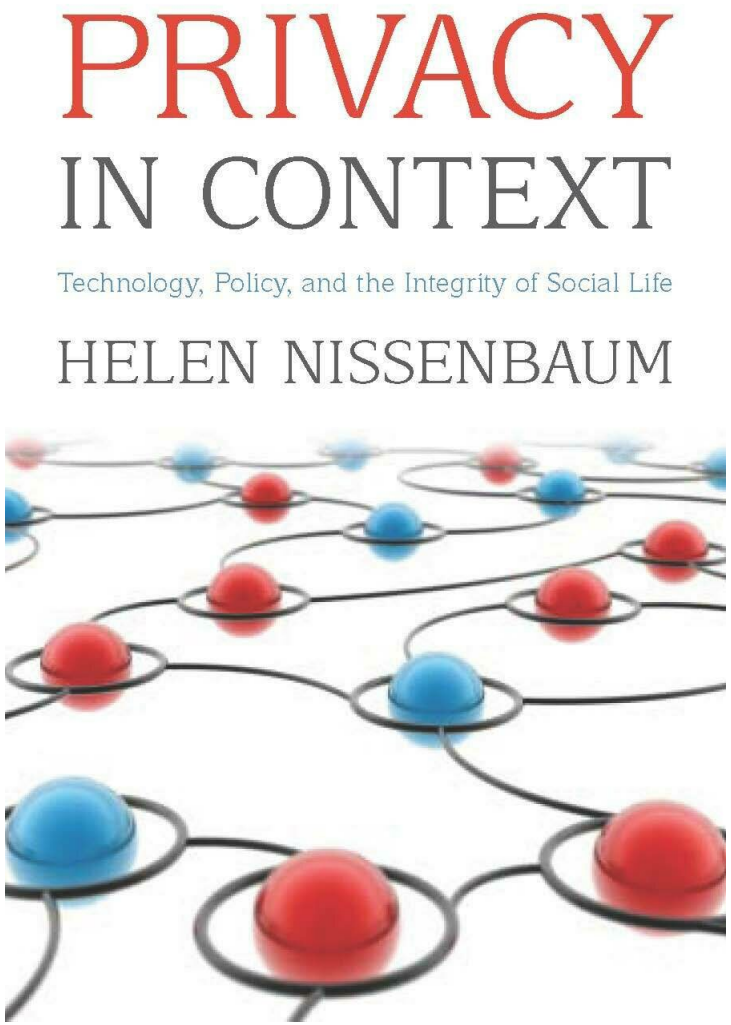
Privacy in Conversation: What People Expect and What ChatGPT Reveals

Mehedi Zaman

PhD student in Information Science at Rutgers

Privacy in Conversation: What People Expect and What ChatGPT Reveals



Appropriate flow of personal information within a conversational context

Appropriate flow of personal information within a conversational context

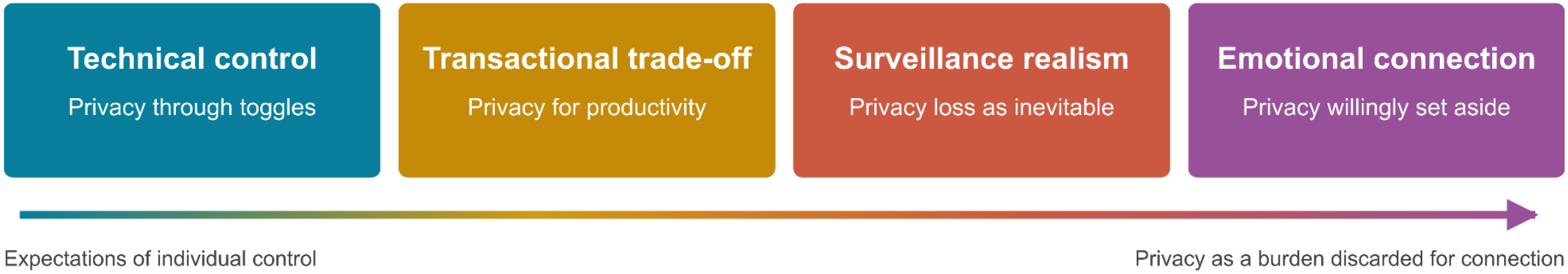


- 3 subreddits – r/ChatGPT, r/OpenAI, r/privacy
  - Data from Nov 30, 2022 – Sep 30, 2025
- After filtering, around 23k posts & 100k comments
- 4 themes after BERTopic & qualitative analysis



- 3 subreddits – r/ChatGPT, r/OpenAI, r/privacy
  - Data from Nov 30, 2022 – Sep 30, 2025
- After filtering, around 23k posts & 100k comments
- 4 themes after BERTopic & qualitative analysis

Four themes as a spectrum of Agency



Inferential Privacy Leakage in Anonymized ChatGPT Logs

- Over a thousand users' full ChatGPT conversation history
- Users from Global South countries: Pakistan, India, Nigeria, and Brazil
- Data from Nov 2022 to Feb 2026
- Ground truth of age, gender, and country

Dataset summary

	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 logs, Google Search, YouTube Search, YouTube Watch
Demographics	Age bracket, gender, country	Age, gender, country, religion, caste, education, income, voting preference

Inferential Privacy Leakage in Anonymized ChatGPT Logs

- NER filtering
- LLM-based filtering
- Manual validation

- Llama 3.3-70B 4bit
- Gender and country predicted at higher accuracy

Attribute	XGBoost F1	LLM F1
Age bracket	0.48	**0.51**
Gender	0.64	**0.76**
Country of residence	0.52	**0.86**

- Llama 3.3-70B 4bit
- Gender and country predicted at higher accuracy

Attribute	XGBoost F1	LLM F1
Age bracket	0.48	**0.51**
Gender	0.64	**0.76**
Country of residence	0.52	**0.86**

Chat history % (Context)

Attribute / group	Mean	Median
Age	53.8%	70%
Gender (overall)	24.6%	5%
Female	36.7%	20%
Male	21.1%	5%
Country (overall)	19.8%	5%
Brazil	22.2%	5%
India	17.6%	5%
Nigeria	20.0%	5%
Pakistan	25.1%	10%

- Llama 3.3-70B 4bit
- Gender and country predicted at higher accuracy

- Lower context needed for gender and country

Attribute	XGBoost F1	LLM F1
Age bracket	0.48	**0.51**
Gender	0.64	**0.76**
Country of residence	0.52	**0.86**

Chat history % (Context)

Attribute / group	Mean	Median
Age	53.8%	70%
Gender (overall)	24.6%	5%
Female	36.7%	20%
Male	21.1%	5%
Country (overall)	19.8%	5%
Brazil	22.2%	5%
India	17.6%	5%
Nigeria	20.0%	5%
Pakistan	25.1%	10%

Patterns in the model’s reasoning

- Tech ≡ male: Technical questions suggest male gender
- Tech ≡ Western: Nigerian and Pakistani tech users receive US/UK predictions
- English fluency ≡ US: Fluent English defaults to the US, while Portuguese anchors Brazil
- Contemporary content ≡ young: Older tech users receive younger age predictions

How does it compare with Google and Youtube

How does it compare with Google and Youtube

Attribute	ChatGPT logs	Google Search	YouTube Search	YouTube Watch
Age bracket	**0.54**	0.44	0.49	0.39
Gender	0.88	**0.90**	0.89	0.86
Religion	0.71	0.75	**0.83**	0.72
Education level	**0.78**	0.73	0.68	0.67
Monthly income	0.56	**0.62**	0.55	0.51
Voting preference	**0.52**	0.46	0.43	0.29

How is the LLM classifying

Message	Label	Explanation
“I am a Bird. I am a fruit. I also am a person. What am I?”	No Personal Disclosure	First-person statements form part of a riddle.
“Idk... i don’t feel like i have that much money”	Wealth Details Salary	Describes perceived financial circumstances.
“Ele é meu namorado estamos juntos a algum tempo”	Family Life and Relationship	Discloses an ongoing romantic relationship.
“More i learn about ai agent and cybersecurity more i feel detached to the world”	Mental State and Personality Mood	Describes a feeling of detachment.
“I drink milk tea in the early morning then I feel hypoxic every day why?”	Physical Health Diagnosis	Reports a perceived recurring physical symptom.
“I am pursuing my B.Tech in Computer Science and Engineering of Data Science.”	Job and Education	Explicitly states the author’s current degree program.
“all the tracker and ad blocking that i do on firefox”	Lifestyle and Habits	Describes the author’s browser privacy practices.
“Eu tô com desidratação severa agora”	Physical Health Diagnosis	Self-reports a current physical condition.
“It’s just that my baby needs more care. My spouse wanted it maybe when she’s 18 months”	Family Life and Relationship	Discloses family circumstances involving a baby and a spouse.
“No I want to get up now I am awake the fear is freezing me”	Mental State and Personality Mood	Describes current fear.

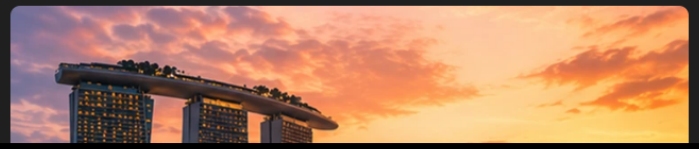
What does ChatGPT save in its Memory?

What does ChatGPT save in its Memory?

Without memory

Help me plan a trip to Singapore. I'm going to be there for a work trip in July and will have a few days free at the end.

Great add-on city for a work trip: Singapore is compact, safe, excellent for solo exploring, and easy to do in 2–4 free days. July is hot and humid; expect daytime highs around 31–33°C / 88–91°F, warm nights, and sudden rain, so plan outdoor time early morning or after 5 p.m. and keep midday for museums, malls, cafés, or hotel downtime. Singapore's official weather service notes typical daily highs of 31–33°C and lows of 23–25°C. (Singapore Weather Service)

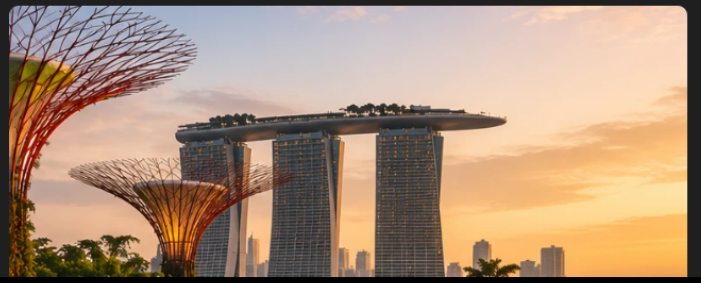


The model produces a generic response that is more touristy, doesn't help with hotel booking, and largely ignores the user's interests.

With memory

Help me plan a trip to Singapore. I'm going to be there for a work trip in July and will have a few days free at the end.

I'll build this around your known constraints: a few free days after work, hot/humid July weather, wildlife/photography/nature bias, strong AC needs for sleep, and good but not overly social dining. I'll verify current July logistics and then give a practical short itinerary rather than a generic Singapore list.



The model produces a response that is personalized to the user's interests in wildlife photography, quiet dinners, and their priorities when booking a hotel.

Vulnerabilities in Personalization: Assessing Health Privacy Risks in ChatGPT Logs and Memory

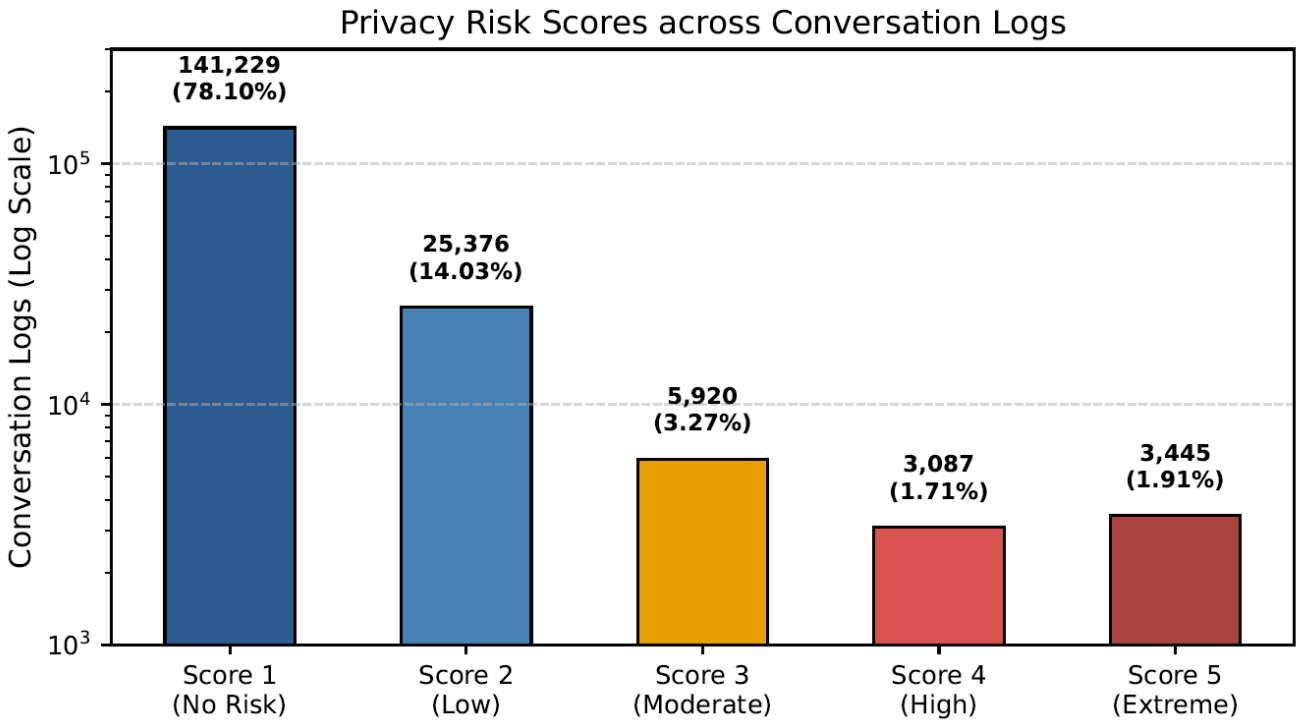
- Around 180k conversations from over 1000 users
- Over 7k saved memory logs from 766 users
- Llama 3.3-70B audit of health information and privacy risk

Vulnerabilities in Personalization: Assessing Health Privacy Risks in ChatGPT Logs and Memory

Taxonomy Category	Definition & Scope
1. Direct Clinical Identifiers	Medical Record Numbers (MRNs), or unique biometric identifiers linked to a health condition.
2. Symptomatic & Physical State	Descriptions of symptoms, bodily pain, rashes, or vital signs used for preliminary diagnosis.
3. Stigmatized Health Condition	Sensitive disclosures regarding mental health, reproductive health, HIV/STIs, or substance use.
4. Healthcare Navigational Data	Specific local hospital, clinic, or physician names revealing locations of care.
5. Socio-Demographic Proxies	Age, gender, or occupation shared specifically to contextualize a medical query.
6. Environmental / SDoH	Health issues linked to living conditions, water/food security, or environmental hazards.
7. Lifestyle & Wellness Habits	Diet, exercise, sleep patterns, or supplement usage shared for health optimization.

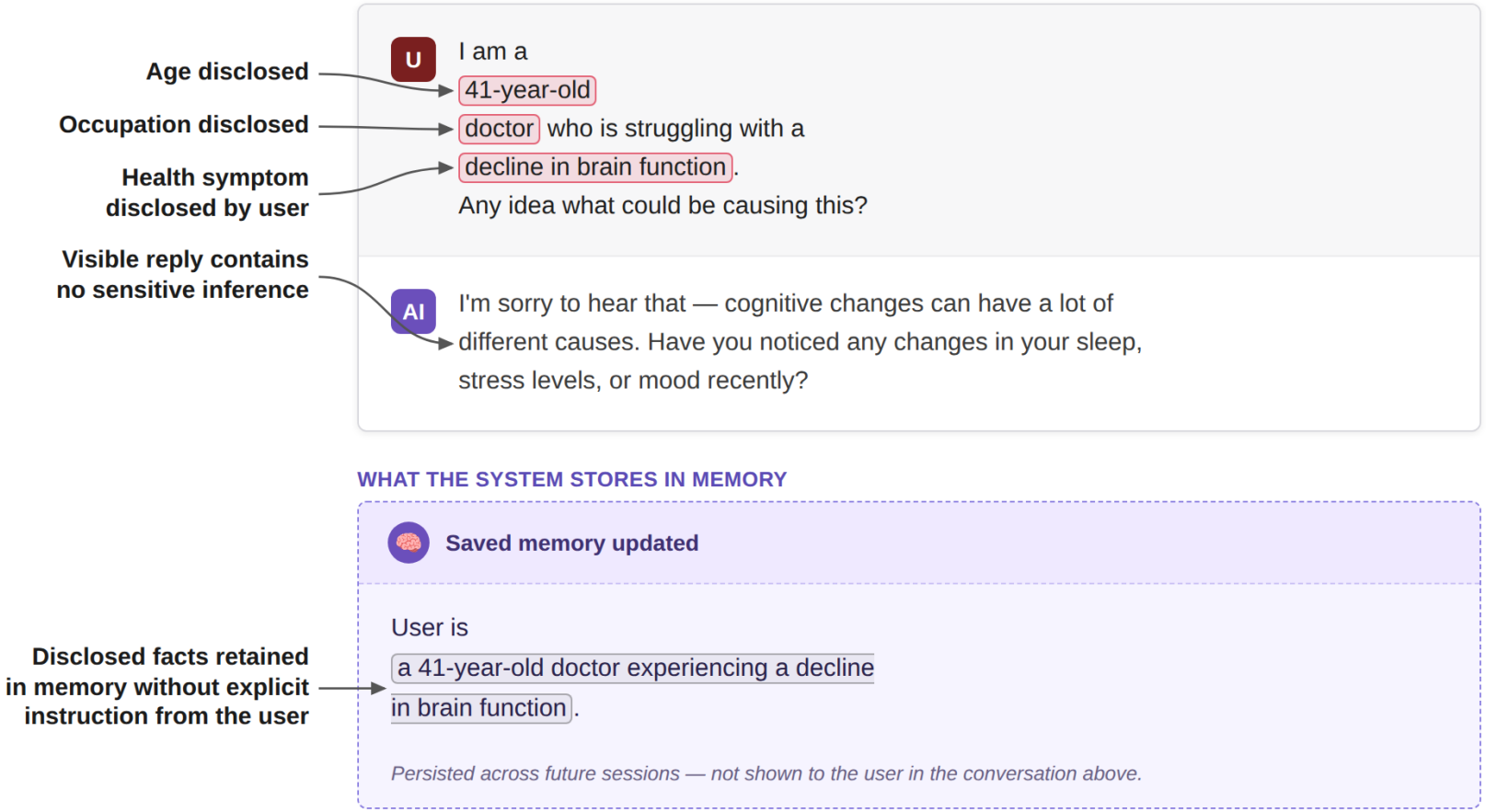
Health information in conversations and memory

- Around 20% of conversations contain personal health information



Health information in conversations and memory

- Around 20% of conversations contain personal health information
- Just over 40% of saved memory logs contain health information



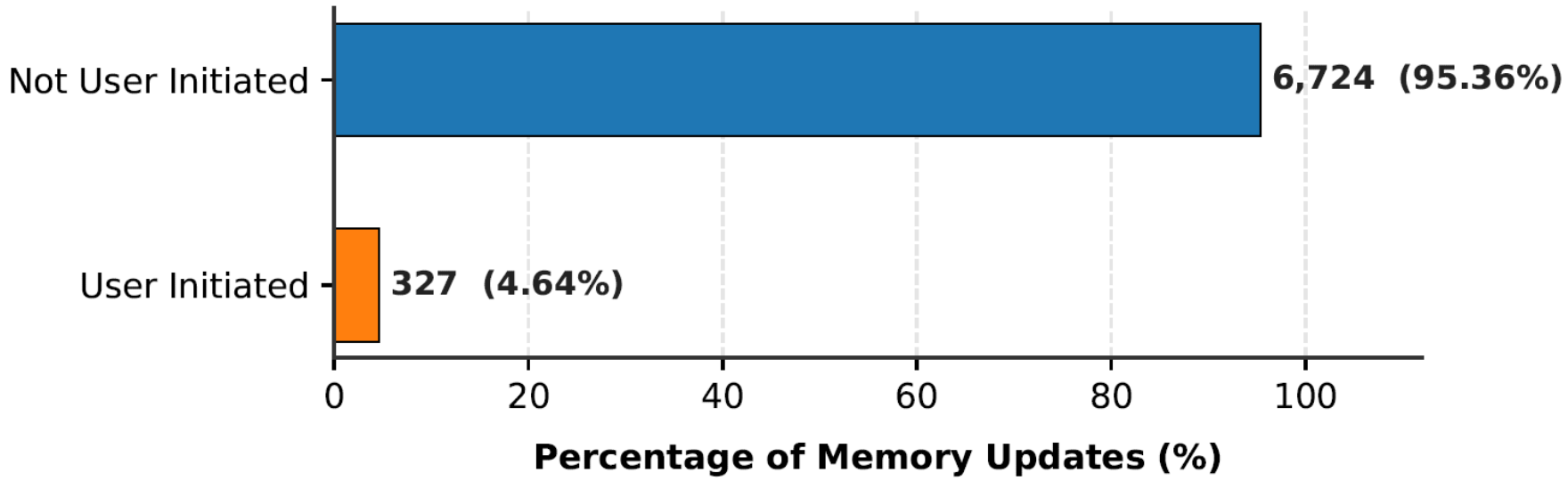
Health information in conversations and memory

- Around 20% of conversations contain personal health information
- Just over 40% of saved memory logs contain health information
- Socio-demographic proxies account for more than 60% of health-related memories

Category / Risk Metric	Memory Logs (n)	Proportion (%)
Taxonomy Breakdown (Disclosures Only, $n = 2,899$)		
Socio-Demographic Proxies (Cat. 5)	1,805	62.26%
Lifestyle & Wellness Habits (Cat. 7)	496	17.11%
Healthcare Navigational Data (Cat. 4)	187	6.45%
Stigmatized Health Condition (Cat. 3)	135	4.66%
Symptomatic & Physical State (Cat. 2)	125	4.31%
Environmental / SDoH (Cat. 6)	83	2.86%
Direct Clinical Identifiers (Cat. 1)	68	2.35%
Privacy Risk Score Distribution (All Memory Logs, $N = 7,051$)		
Score 1 (No Risk)	5,760	81.69%
Score 2 (Low Risk)	911	12.92%
Score 3 (Moderate Risk)	124	1.76%
Score 4 (High Risk)	160	2.27%
Score 5 (Extreme Risk)	96	1.36%

How do memory updates happen?

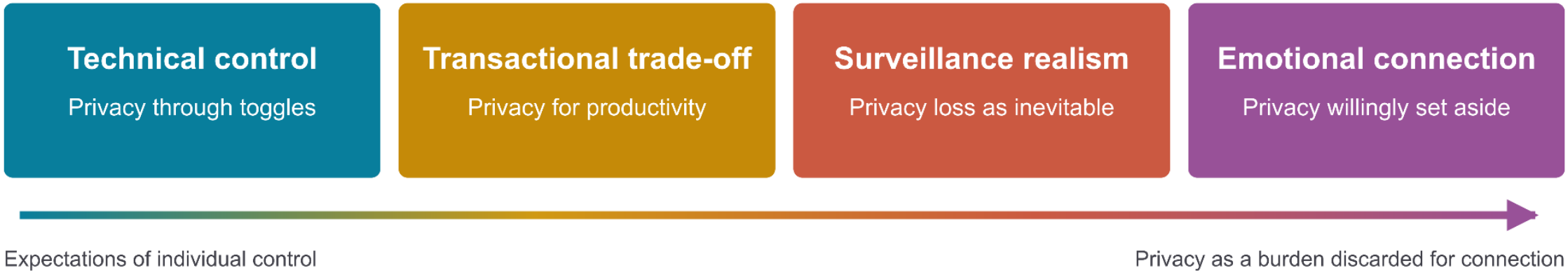
- More than 95% had no detected memory command in the immediately preceding user message
- Around 5% followed an explicit memory command



When a conversation becomes memory

- Temporary health disclosures can persist as profile attributes
- Memory can combine health details with demographic context
- Contextual integrity depends on the conditions of retention and reuse

Takeaways


Technical control
Privacy through toggles
Transactional trade-off
Privacy for productivity
Surveillance realism
Privacy loss as inevitable
Emotional connection
Privacy willingly set aside
Expectations of individual control
Privacy as a burden discarded for connection

Takeaways

Technical control
Privacy through toggles

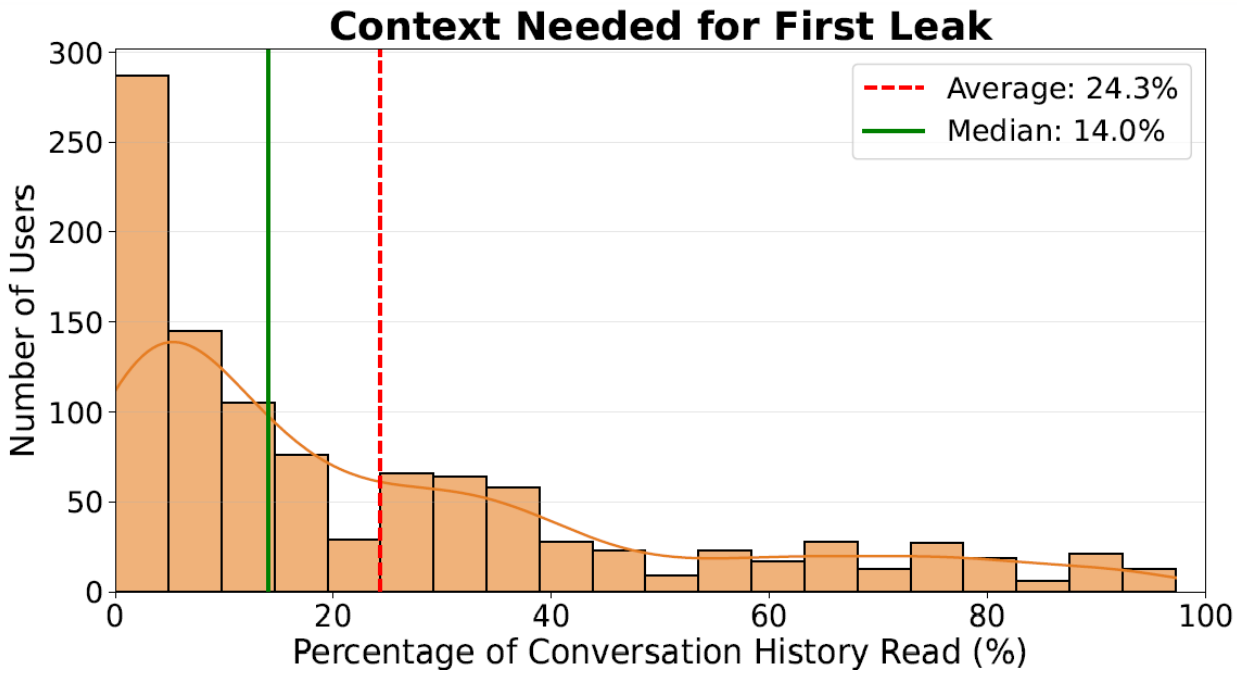
Transactional trade-off
Privacy for productivity

Surveillance realism
Privacy loss as inevitable

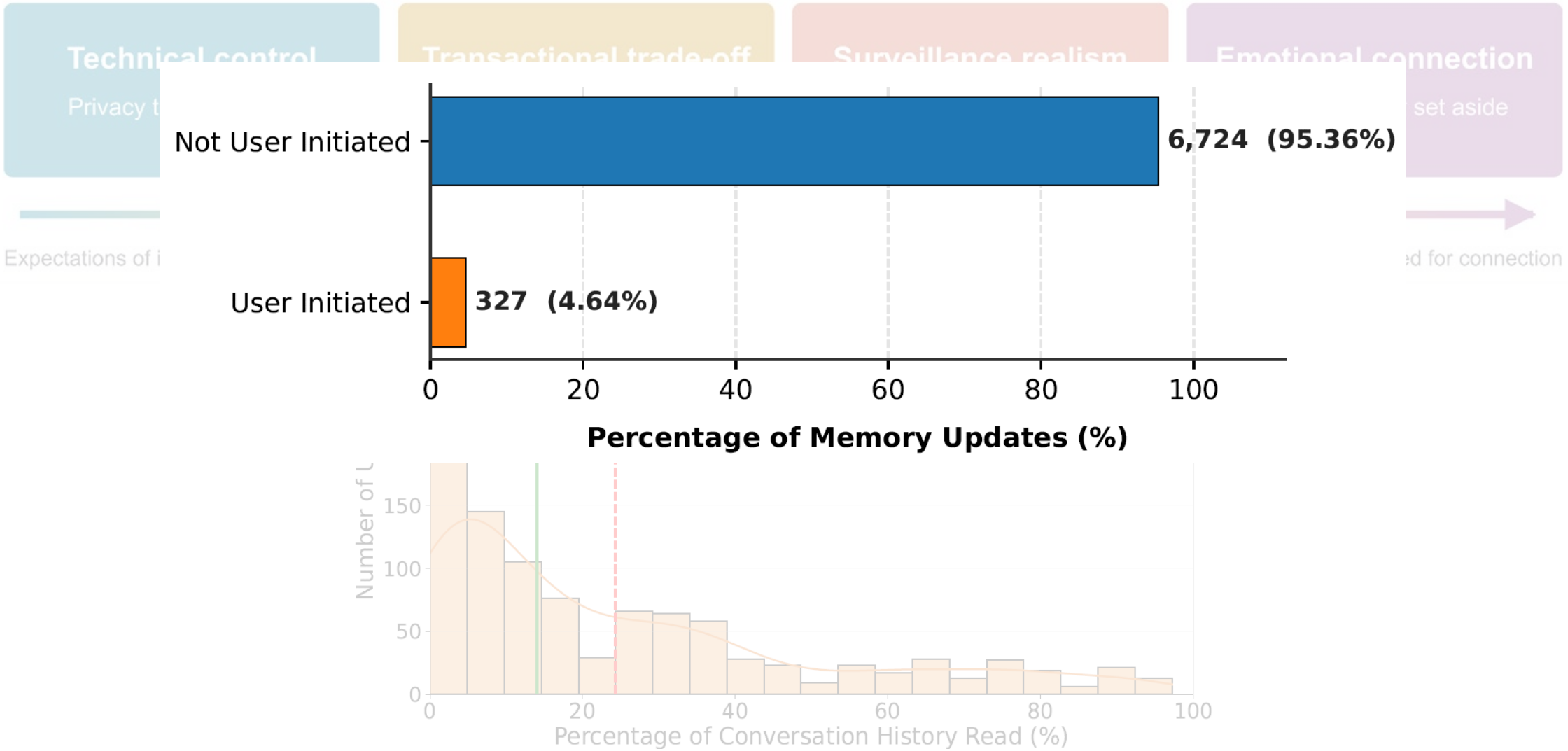
Emotional connection
Privacy willingly set aside

Expectations of individual control

Privacy as a burden discarded for connection



Takeaways


Not User Initiated
6,724 (95.36%)
User Initiated
327 (4.64%)
0
20
40
60
80
100
Percentage of Memory Updates (%)
Number of
150
100
50
0
0
20
40
60
80
100
Percentage of Conversation History Read (%)

Recommendations

- Make inference and memory visible and controllable
- Minimize exposure by design
- Protect conversational data beyond anonymization
- Restrict harmful profiling and audit unequal effects
- Treat privacy as collective governance

Limitations

- Open-weights testing possible with the sensitive data
- The users are skewed towards young, tech-literate population
- We didn’t test downstream application of inference
- Only ChatGPT’s data used
- Legacy memory architecture used but newer versions are here

Papers

- Unpacking Privacy in the Age of ChatGPT: A Critical Analysis of Public Discourse on Reddit (Working paper; short poster version linked)
- Inferential Privacy Leakage in Anonymized Conversational AI Logs
- Vulnerabilities in Personalization: Assessing Health Privacy Risks in ChatGPT Logs and Memory



Collaborators



Dr. Kiran Garimella



Dr. Britt Paris



Saubhagya Joshi



Yiyi Wu



Md. Mozammel Hoque

How I use AI

- I use Codex
- Streamlining my coding and environment pipelines
- I try not to use it to make slides; but appendix sure!
- Searching for relevant papers and people
- For visualizing and running quick analysis on new data
- Updating CV, building quick dashboards/websites

- Not use it for initial draft
- Not take AI's words at face value; verify twice
- Not write with it; bullet points are fine

Thanks Lazer Lab!

Appendix

- Reddit discourse: slides 34–41
- Demographic inference: slides 42–57
- Health and memory: slides 58–75

Reddit: why these communities?

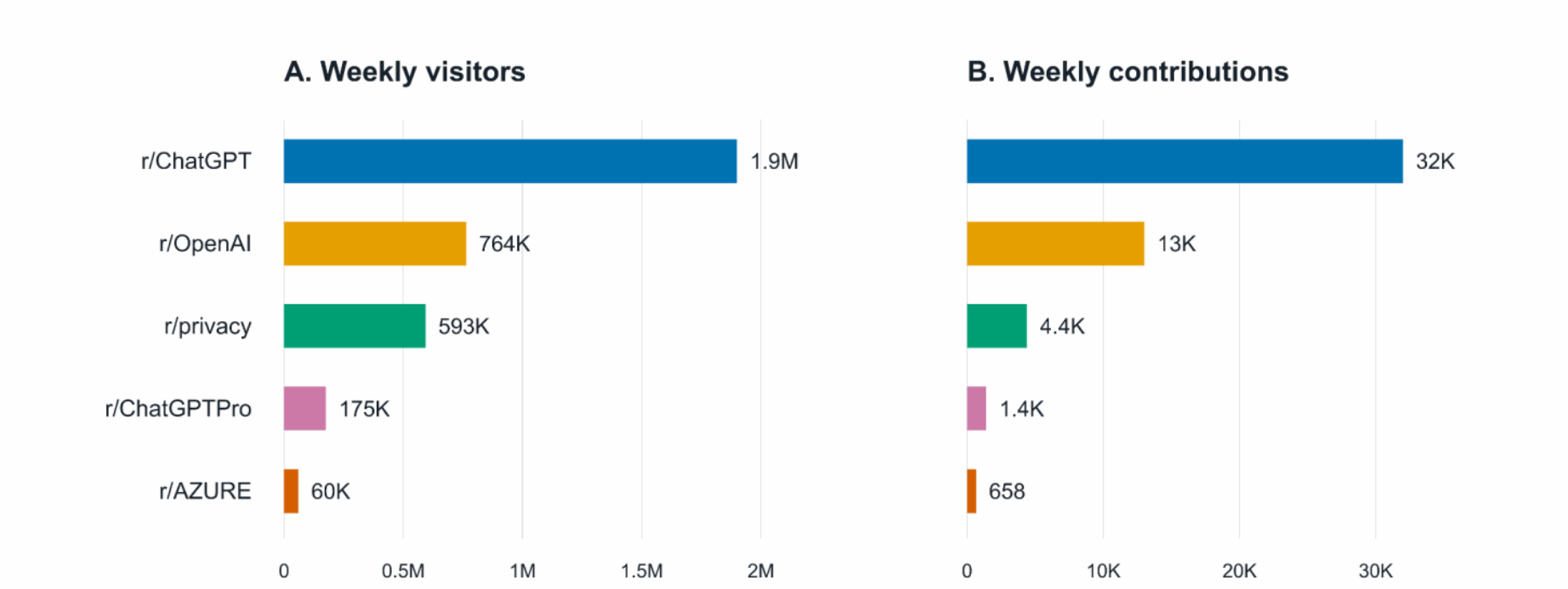


Figure 1. Weekly activity across the five shortlisted subreddits. Panel A shows weekly visitors, and Panel B shows weekly contributions (non-removed posts and comments). Values retain Reddit's displayed rounding; K denotes thousands and M denotes millions.

Reddit: community selection

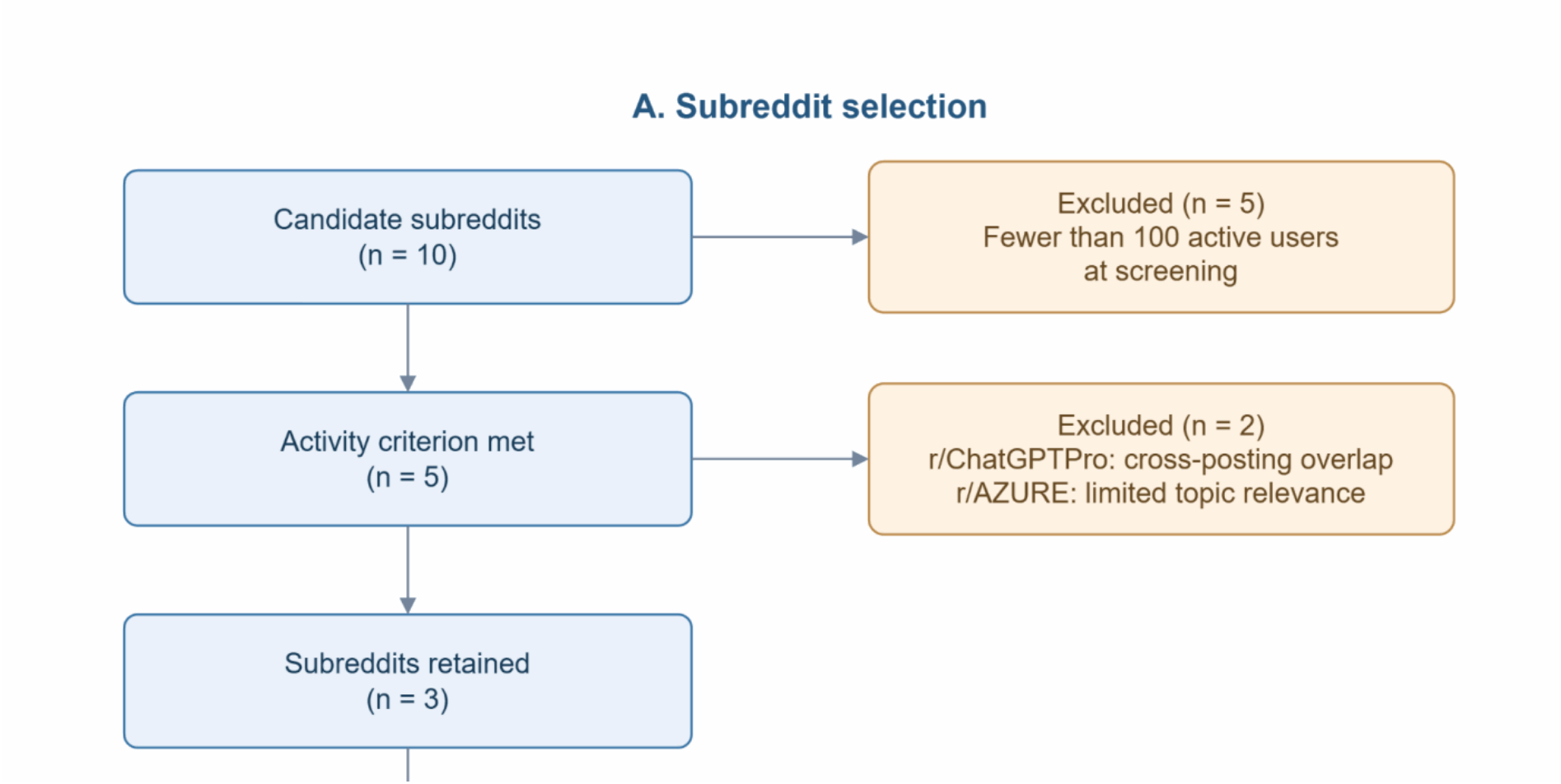


Figure 2A: selecting communities for the preliminary study.

Reddit: preliminary post screening

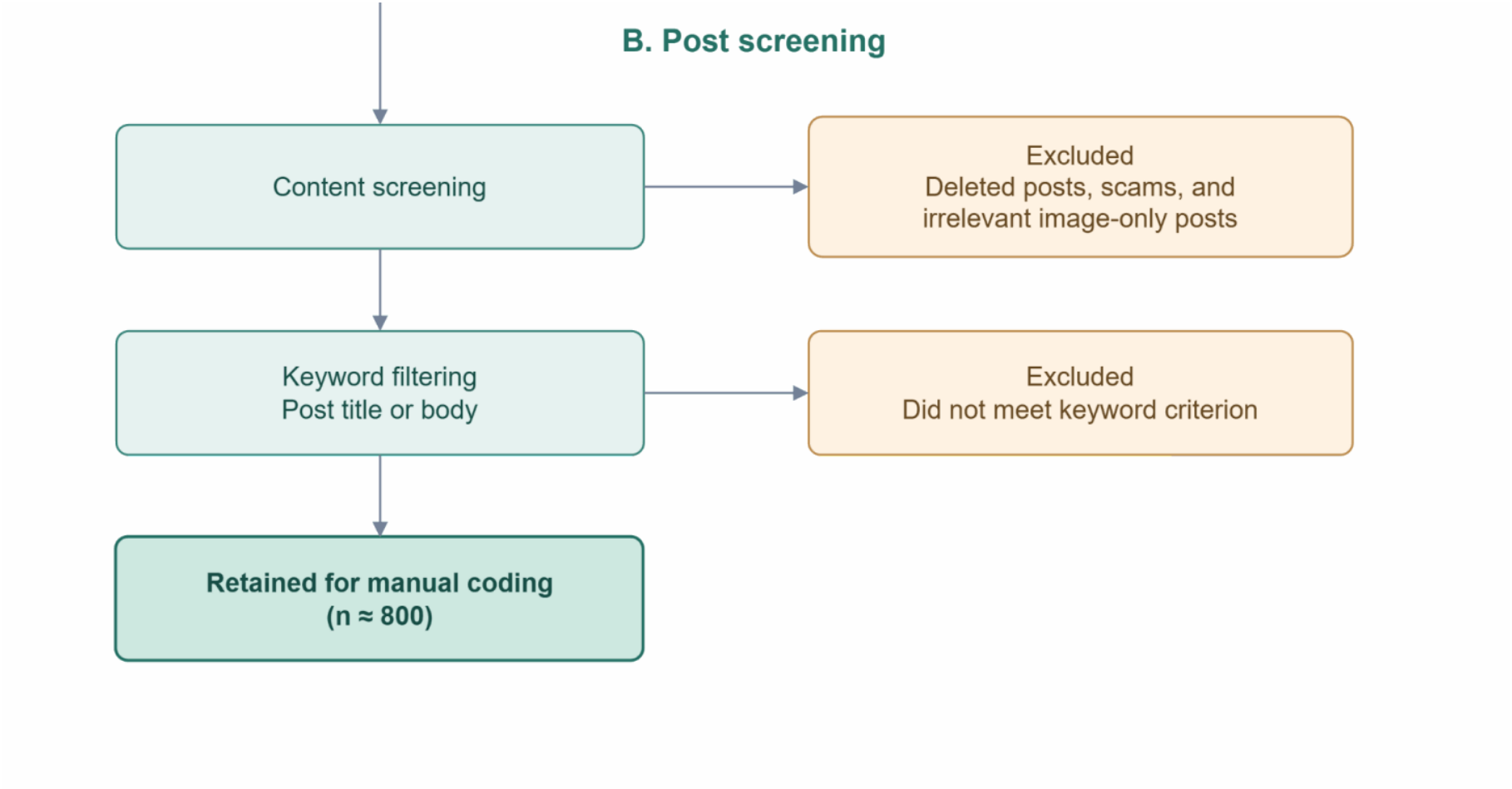


Figure 2B concerns the preliminary pool, not the final 23,017-post corpus.

Reddit: developing the keywords

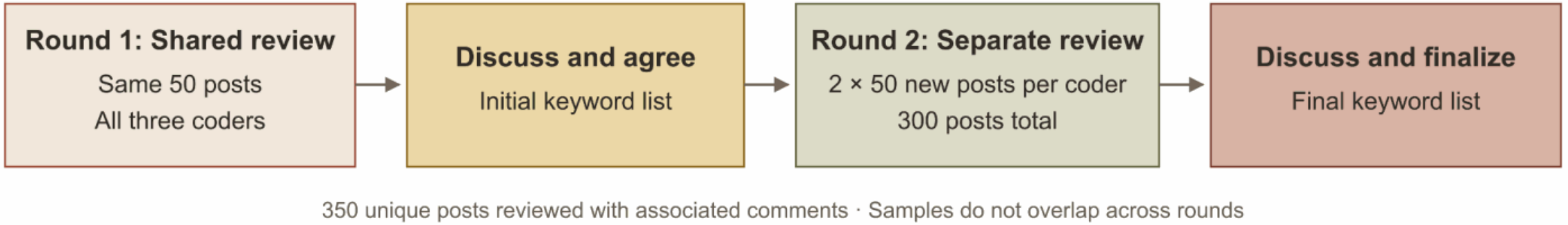


Figure 3. Two-round keyword development through manual coding.

350 posts informed the keywords. The later CDA stage analyzed 340 posts.

Reddit: full collection and analysis

Data posted on	November 30, 2022 - September 30, 2025
Data collected on	October 1, 2025
Number of posts	597,697
Subreddits	r/OpenAI, r/ChatGPT, r/privacy

Of the total number of posts downloaded (N=597,697), we filtered the posts through a filter using the final keywords derived from the first phase, which yielded a total of 23,017 posts.

- 64 clusters consolidated into 34 salient topics
- CDA: 340 posts with comments, 10 per topic
- Four positions can overlap within one person

Reddit: preliminary BERTopic results

	Topic	Number of posts
1	Data leak with ChatGPT	332
2	Chat history	121
3	ChatGPT modes	66
4	Italy ban	51
5	Image with ChatGPT	46
6	Android/iOS ChatGPT app	42
7	Location data	41
8	ChatGPT memory	36
9	ChatGPT training	28
10	Chrome extension	17
11	Deepseek vs ChatGPT	13

Table 2. Topics identified by BERTopic in the preliminary dataset and the number of posts assigned to each topic.

Exploratory topics from the preliminary sample, not final thematic frequencies.

Reddit: control and exchange

Technical control

"I even asked ChatGPT to reset and forget everything about our convo and about me and yet when I ask about myself it types everything that it knows so far about me."

Transactional trade-off

"I've been puzzled by the way every AI service falls over itself to assure the user that their personal data is never used for training... I'd like mine to be used. I hope we see usage tiers where you get an enhanced service if you opt-in."

Illustrative public posts quoted by the paper, not verified backend behavior.

Reddit: resignation and connection

Surveillance realism

"I'm completely unconcerned about it and throw all my data, including details about my personality, into the AI. Google already knows everything about me anyway, and I'm just an average Joe."

Emotional connection

"But I've got in a fight with my mom recently. I didn't really have anyone to talk to so I asked it for support and advice. Since then chatgpt is my emotional support. It knows pretty much everything about me. From my strength to my weakness. How I behave etc etc. I've recently been thinking is it too much? I'm not sure so wanted to ask u guys. Not that I care about my privacy that much. I'm conflicted rn."

Illustrative positions, not separate user groups or observed stages over time.

Inference: the two datasets

	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 logs, Google Search, YouTube Search, YouTube Watch
Demographics	age bracket, gender, country	age, gender, country, religion, caste, education, income, voting preference

The matched cross-platform analysis uses the Indian sub-cohort only.

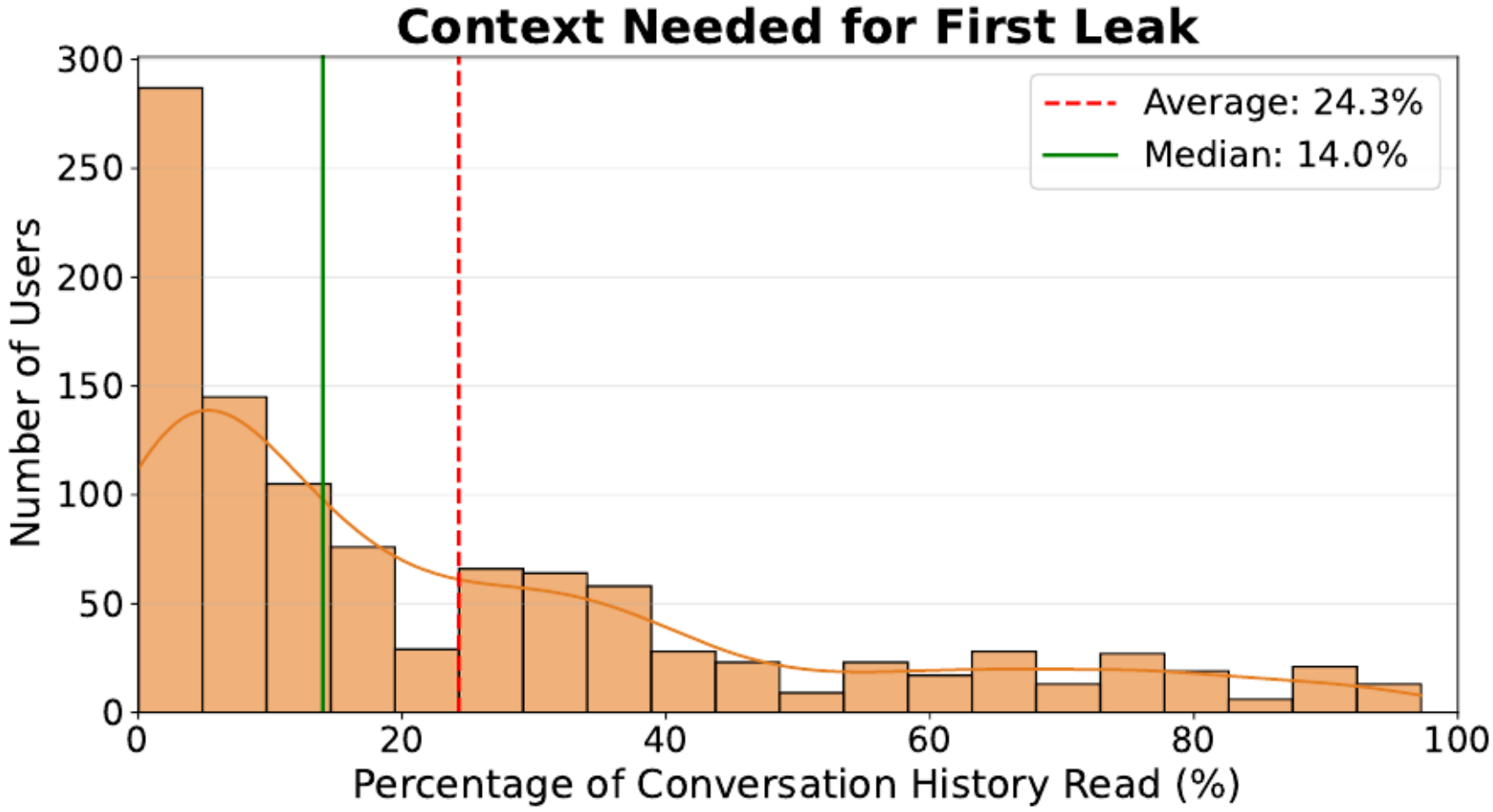
Inference: what people disclose

Category	Share (%)
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

Category	Share (%)
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

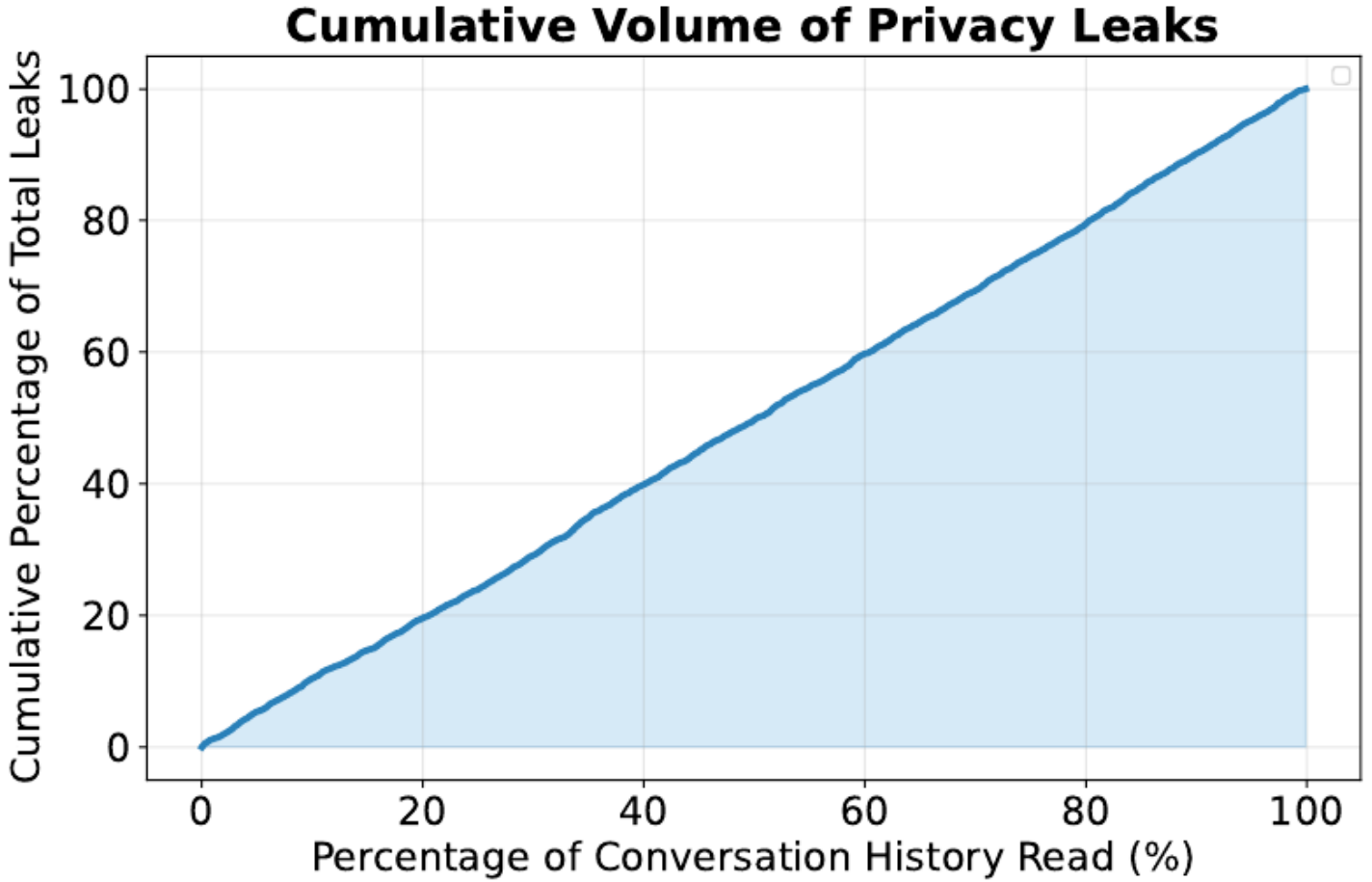
Share of 428,865 flagged messages out of 1,242,109 raw user messages (Table 2).

Inference: first explicit disclosure



Raw-disclosure timing: mean 24.3%, median 14.0%. This is distinct from inference k*.

Inference: disclosures throughout histories



Cumulative raw disclosures, not a demographic-prediction accuracy curve.

Inference: filtering and evaluation

- Only user-authored messages enter the analysis
- English NER flags entities for the disclosure audit
- The LLM excludes messages it labels as direct self-identification
- The same model architecture filters and predicts

Inference: age and gender classification

Attribute / group	Precision	Recall	F1	N
Age bracket: weighted F1, majority 0.23, XGB 0.48, LLM 0.51				
18-24	0.60	0.69	0.64	389
25-34	0.53	0.51	0.52	424
35-44	0.36	0.35	0.36	185
45+	0.00	0.00	0.00	59
Gender: weighted F1, majority 0.52, XGB 0.64, LLM 0.76				
Female	0.76	0.50	0.61	360
Male	0.79	0.89	0.84	697

Inference: country classification

Attribute / group	Precision	Recall	F1	N
Country: weighted F1, majority 0.26, XGB 0.52, LLM 0.86				
Brazil	0.98	0.81	0.89	205
India	0.88	0.95	0.91	456
Nigeria	0.99	0.75	0.85	206
Pakistan	0.98	0.53	0.69	190

Inference: history needed for stable prediction

Attribute / group	Mean context (%)	Median context (%)
Age	53.8	70
Gender overall	24.6	5
Female	36.7	20
Male	21.1	5
Country overall	19.8	5
Brazil	22.2	5
India	17.6	5
Nigeria	20.0	5
Pakistan	25.1	10

Inference: model selection and baselines

Model	Parameters	Country	Gender	Age
Llama-3.1	8B	0.68	0.62	0.38
Mistral-Nemo	12B	0.72	0.66	0.41
Qwen3	32B	0.77	0.71	0.46
Llama-3.3	70B	0.83	0.76	0.51

- XGBoost: TF-IDF, writing style, LDA topics, sentiment
- Five-fold CV, randomized tuning, inverse-frequency weights

Inference: matched platform results

Attribute	ChatGPT	Google Search	YouTube Search	YouTube Watch
Age bracket	0.54	0.44	0.49	0.39
Gender	0.88	0.90	0.89	0.86
Religion	0.71	0.75	0.83	0.72
Education level	0.78	0.73	0.68	0.67
Monthly income	0.56	0.62	0.55	0.51
Voting preference	0.52	0.46	0.43	0.29

Inference: selected history checkpoints

History (%)	Country accuracy	Country F1	Gender accuracy	Gender F1	Age accuracy	Age F1
5	58.56	70.38	72.75	70.93	43.42	41.48
25	75.12	81.42	74.93	73.76	48.72	47.91
50	77.58	83.36	76.44	76.21	51.94	50.43
75	79.75	84.92	77.58	76.76	51.66	50.31
100	81.08	85.72	76.06	75.98	52.03	50.71

Inference: stereotype-mediated patterns

Pattern	Triggering Content	Pushed Toward	Evidence in Errors
Tech ≡ *male*	Coding, Linux, networking, finance, cybersecurity, business plans	Male	Female recall 0.50 vs. male recall 0.89; 165 of 360 women misclassified as men, compared with 56 of 697 men misclassified as women.
Tech ≡ *Western*	Software development, cybersecurity, advanced technical proficiency	US / UK	Lowest country recalls fall on Nigeria (0.75) and Pakistan (0.53), with most misclassifications assigned to US/UK rather than regional neighbors.
English fluency ≡ *US*	Fluent English without local lexicon, currency, or regional markers	US	Drives the patterns above; Brazil achieves high precision (0.98) and F_1 (0.89) due to Portuguese acting as an unoverridable local signal.
Contemporary content ≡ *young*	Modern software, career tools, brief or informal language	25–34	Age recall drops from 0.69 (18–24) and 0.51 (25–34) to 0.35 (35–44) and collapses to 0.00 in 45+; rationales explicitly cite tech usage as evidence of youth.

Inference: rationale coding and checks

Stereotype Theme	Diagnostic Term Cluster
Tech ≡ *male*	programming, Linux, cybersecurity, engineering, finance, technical skills
Tech ≡ *Western*	Western-style education, Silicon Valley, international tech, US-centric
English fluency ≡ *US*	native English, professional tone, fluent English, American phrasing
Contemporary content ≡ *young*	youthful interest, modern tech, Gen Z slang, digital native, current trends

- 600 class-stratified rationales: 200 per attribute
- 60 independently checked, 91.7% thematic agreement
- Lexical matching checked the broader rationale corpus

Inference: out-of-cohort country labels

Predicted country	Count	Share (%)
United States	98	79.03
United Kingdom	15	12.10
Bangladesh	2	1.61
Canada	2	1.61
Japan	2	1.61
Sweden	2	1.61
Bahrain	1	0.81
Germany	1	0.81
Malaysia	1	0.81

Inference: age errors across platforms

	Prec.	Recall	F1	Support
ChatGPT logs	$F_1 = 0.54$		Avg. context 15.7%	
18–24	0.64	0.73	0.68	82
25–34	0.58	0.56	0.57	81
35–44	0.41	0.39	0.40	31
45+	0.00	0.00	0.00	18
Google Search	$F_1 = 0.44$		Avg. context 12.1%	
18–24	0.55	0.61	0.58	82
25–34	0.47	0.46	0.46	81
35–44	0.28	0.26	0.27	31
45+	0.00	0.00	0.00	18

	Prec.	Recall	F1	Support
YouTube Search	$F_1 = 0.49$		Avg. context 9.0%	
18–24	0.60	0.67	0.63	82
25–34	0.52	0.52	0.52	81
35–44	0.34	0.32	0.33	31
45+	0.00	0.00	0.00	18
YouTube Watch	$F_1 = 0.39$		Avg. context 16.1%	
18–24	0.51	0.56	0.53	82
25–34	0.41	0.42	0.41	81
35–44	0.22	0.19	0.21	31
45+	0.00	0.00	0.00	18

Inference: gender errors across platforms

	Prec.	Recall	F1	Support
ChatGPT logs	$F_1 = 0.88$		Avg. context 12.0%	
Female	0.80	0.53	0.63	38
Male	0.90	0.97	0.94	174
Google Search	$F_1 = 0.90$		Avg. context 10.3%	
Female	0.82	0.61	0.70	38
Male	0.92	0.97	0.94	174

	Prec.	Recall	F1	Support
YouTube Search	$F_1 = 0.89$		Avg. context 7.1%	
Female	0.81	0.58	0.68	38
Male	0.91	0.97	0.94	174
YouTube Watch	$F_1 = 0.86$		Avg. context 8.9%	
Female	0.72	0.47	0.57	38
Male	0.89	0.96	0.93	174

Health audit: cohort and data

Table 1. Participant Demographics and Dataset Characteristics ($N = 1,057$)

Metric / Category	Value	Category	Value
Dataset Overview			
Total Conversations, *n*	179,057	Total User Prompts, *n*	1,086,489
Conv./User, Med. [IQR]	75.0 [24, 208]	Prompts/User, Med. [IQR]	427.0 [150, 1203]
Demographics: Gender & Country, n (%)		*Demographics: Age Bracket, n (%)*	
India	456 (43.1%)	18–24	389 (36.8%)
Nigeria	206 (19.5%)	25–34	424 (40.1%)
Brazil	205 (19.4%)	35–44	185 (17.5%)
Pakistan	190 (18.0%)	45–54	50 (4.7%)
Male	698 (66.0%)	55–64	8 (0.8%)
Female	359 (34.0%)	65+	1 (0.1%)

- Offline Llama audit of user-authored messages
- Memory subset: 7,051 entries from 766 users

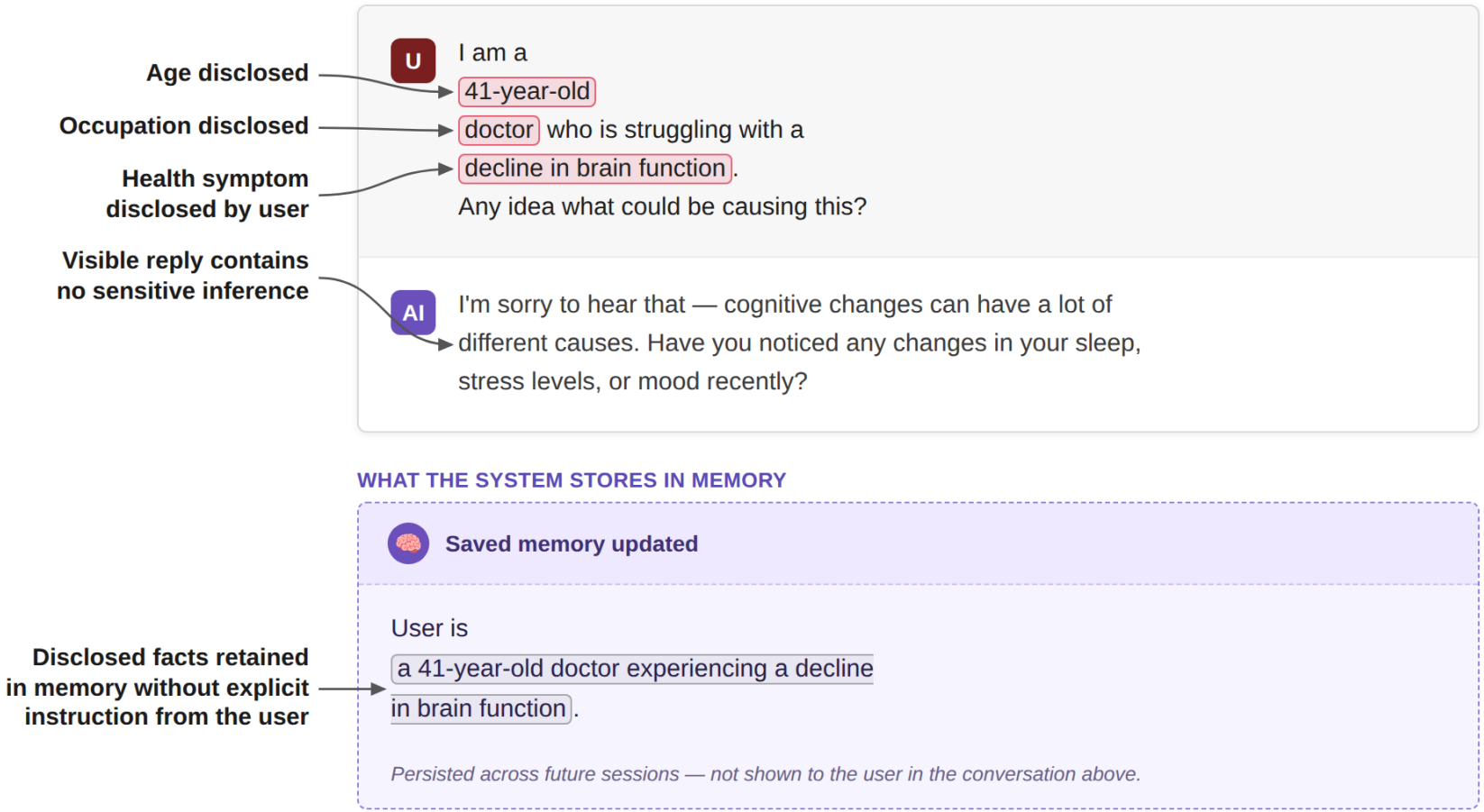
Health audit: seven categories

Table 2. Health Privacy Taxonomy and Category Definitions

Taxonomy Category	Definition & Scope
1. Direct Clinical Identifiers	Medical Record Numbers (MRNs), or unique biometric identifiers linked to a health condition.
2. Symptomatic & Physical State	Descriptions of symptoms, bodily pain, rashes, or vital signs used for preliminary diagnosis.
3. Stigmatized Health Condition	Sensitive disclosures regarding mental health, reproductive health, HIV/STIs, or substance use.
4. Healthcare Navigational Data	Specific local hospital, clinic, or physician names revealing locations of care.
5. Socio-Demographic Proxies	Age, gender, or occupation shared specifically to contextualize a medical query.
6. Environmental / SDoH	Health issues linked to living conditions, water/food security, or environmental hazards.
7. Lifestyle & Wellness Habits	Diet, exercise, sleep patterns, or supplement usage shared for health optimization.

Study taxonomy. Each record receives a primary category and a 1–5 risk score.

Health audit: conversation and saved memory



Health audit: classifier and validation

- Llama-3.3-70B-Instruct audit
- Symptom, drug, or lab mentions default to personal unless stated otherwise
- Two authors reviewed 200 cases, initially agreeing on 188
- The model matched final consensus in 188 of 200 cases (94%)

Health audit: reported risk distribution

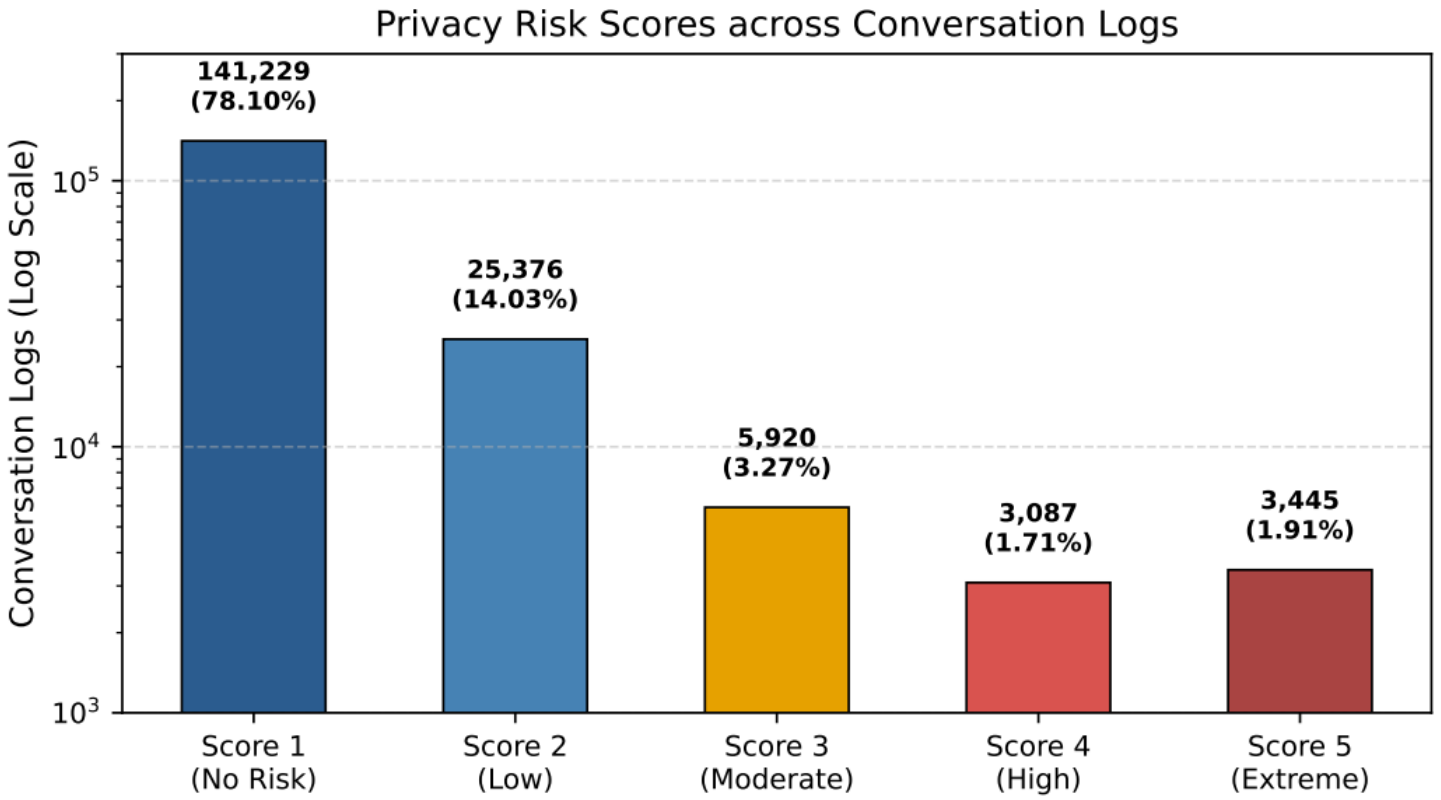


Fig. 2. Distribution of Privacy Risk Scores across $N = 179,057$ conversation logs (logarithmic scale). Annotations detail raw conversation counts (n) and overall dataset percentages (%).

Health audit: disclosed information types

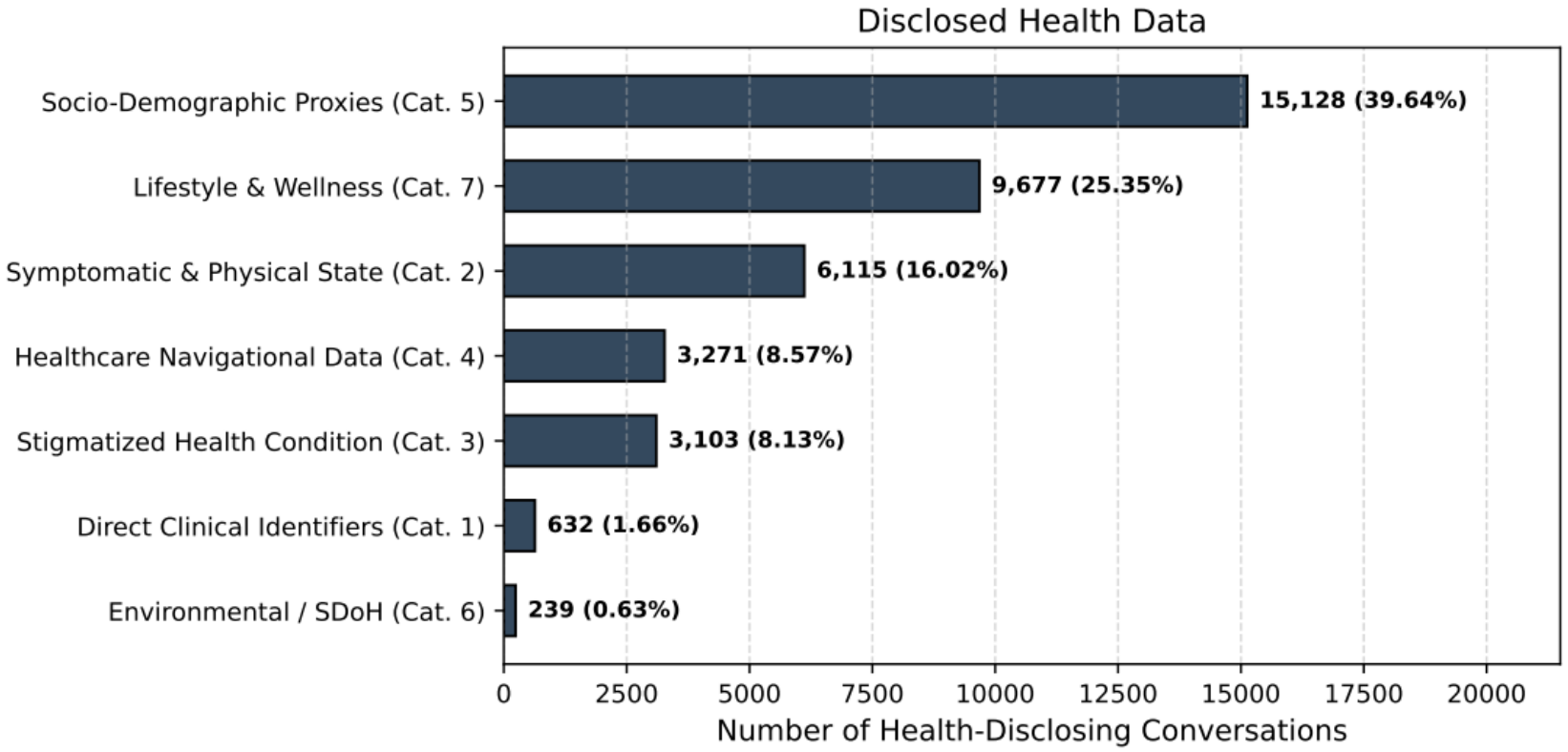


Fig. 3. Taxonomy breakdown of the $n = 38,165$ health-disclosing conversations. Values represent absolute conversation volume and relative proportion of total health disclosures.

Denominator: 38,165 conversations classified as health-disclosing.

Health audit: risk by demographic group

Table 3. Aggregated Privacy Risk Profiles Across Demographics

Demographic Subgroup	No Risk (Score 1)	Low–Moderate Risk (Scores 2–3)	High–Extreme Risk (Scores 4–5)	Total Health Logs (n)
Gender				
Female	73.1%	22.8%	**4.1%**	60,093
Male	81.8%	14.7%	**3.4%**	118,964
Country				
Brazil	66.9%	27.8%	**5.5%**	33,034
India	81.1%	15.7%	**3.2%**	78,933
Nigeria	82.4%	14.2%	**3.4%**	36,815
Pakistan	81.9%	15.0%	**3.2%**	30,275
Age Bracket				
18–24	82.0%	15.0%	**3.0%**	72,468
25–34	77.4%	18.5%	**4.1%**	70,487
35–44	74.8%	20.9%	**4.3%**	27,688
45–54	77.8%	18.5%	**3.7%**	7,689
55–64	75.3%	20.2%	**4.5%**	720
65+	20.0%	80.0%	**0.0%**	5

The source’s “Total Health Logs” column here includes all chats. Older groups are small.

Health audit: categories by demographic group

Table 4. Taxonomy Category Breakdown Across Demographic Subgroups (% of Subgroup Health Disclosures)

Demographic Subgroup	Cat. 1 Direct	Cat. 2 Symptom.	Cat. 3 Stigma.	Cat. 4 Navig.	Cat. 5 Socio-Dem.	Cat. 6 SDoH	Cat. 7 Lifestyle	Total Health Logs (*n*)
Gender								
Female	1.2%	17.5%	8.4%	6.5%	39.4%	0.6%	26.3%	16,190
Male	2.0%	14.9%	7.9%	10.1%	39.8%	0.6%	24.7%	21,638
Country								
Brazil	1.5%	18.1%	8.9%	7.2%	37.0%	0.6%	26.9%	10,946
India	1.9%	14.8%	6.7%	9.5%	39.8%	0.7%	26.5%	14,909
Nigeria	1.5%	14.7%	10.0%	8.8%	43.4%	0.5%	20.9%	6,479
Pakistan	1.6%	16.7%	8.2%	8.5%	39.9%	0.7%	24.4%	5,494
Age Bracket								
18–24	1.8%	15.3%	8.1%	7.5%	39.5%	0.7%	27.2%	13,139
25–34	1.5%	15.9%	9.3%	8.6%	41.4%	0.5%	22.7%	16,114
35–44	1.7%	18.9%	6.4%	9.9%	34.6%	0.7%	27.9%	7,015
45–54	1.6%	11.3%	4.7%	11.6%	42.8%	0.9%	27.1%	1,715
55–64	1.1%	15.7%	10.7%	5.1%	52.2%	1.1%	14.0%	178
65+	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	4

Within-group health disclosures. Some source subgroup totals require reconciliation.

Memory audit: categories and risk scores

Table 5. Taxonomy breakdown ($n = 2,899$ health disclosures) and privacy risk score distribution across $N = 7,051$ persistent memory logs.

Category / Risk Metric	Memory Logs (n)	Proportion (%)
Taxonomy Breakdown (Disclosures Only, $n = 2,899$)		
Socio-Demographic Proxies (Cat. 5)	1,805	62.26%
Lifestyle & Wellness Habits (Cat. 7)	496	17.11%
Healthcare Navigational Data (Cat. 4)	187	6.45%
Stigmatized Health Condition (Cat. 3)	135	4.66%
Symptomatic & Physical State (Cat. 2)	125	4.31%
Environmental / SDoH (Cat. 6)	83	2.86%
Direct Clinical Identifiers (Cat. 1)	68	2.35%
Privacy Risk Score Distribution (All Memory Logs, $N = 7,051$)		
Score 1 (No Risk)	5,760	81.69%
Score 2 (Low Risk)	911	12.92%
Score 3 (Moderate Risk)	124	1.76%
Score 4 (High Risk)	160	2.27%
Score 5 (Extreme Risk)	96	1.36%

Taxonomy n = 2,899, risk denominator N = 7,051. Draft totals do not reconcile.

Memory audit: demographic risk detail

Subgroup	Score 1	Score 2	Score 3	Score 4	Score 5
Female	77.9	15.8	2.4	2.7	1.1
Male	84.5	11.2	1.2	1.6	1.5
Brazil	77.6	14.2	3.3	3.5	1.4
India	83.5	12.0	1.5	1.5	1.6
Nigeria	80.3	15.6	1.2	1.5	1.4
Pakistan	86.3	10.2	0.5	2.1	0.9
18–24	83.8	11.8	1.3	1.7	1.4
25–34	80.6	14.0	1.5	2.4	1.5
35–44	79.8	14.1	2.9	2.1	1.1
45–54	88.5	8.8	1.4	0.9	0.5
55–64	100.0	0.0	0.0	0.0	0.0

Table 6: risk percentages, all 7,051 memory records. Source values use one decimal.

Memory audit: demographic category detail

	Health Disclosure Taxonomy Distribution (%)						
Demographic Subgroup	**Cat. 1**	**Cat. 2**	**Cat. 3**	**Cat. 4**	**Cat. 5**	**Cat. 6**	**Cat. 7**
Gender							
Female	1.6	4.7	4.7	6.5	62.3	3.5	16.7
Male	3.3	3.2	3.7	7.1	62.7	2.6	17.4
Country							
Brazil	1.5	5.1	5.5	7.1	57.1	3.5	20.1
India	3.2	4.5	3.0	8.3	61.9	2.7	16.3
Nigeria	2.8	2.6	3.6	6.1	68.0	3.0	13.8
Pakistan	2.6	1.8	4.7	4.0	66.5	2.6	17.7
Age Bracket							
18–24	3.2	3.9	3.4	5.3	63.5	4.4	16.5
25–34	2.4	3.2	5.3	7.1	63.0	2.2	16.7
35–44	1.8	6.2	3.3	8.7	58.6	2.3	19.0
45–54	1.2	1.2	2.3	10.5	62.8	2.3	19.8
55–64	0.0	0.0	0.0	0.0	88.9	0.0	11.1

Table 6 detail: category shares among 2,899 health-containing memory records.

Memory audit: how updates were classified

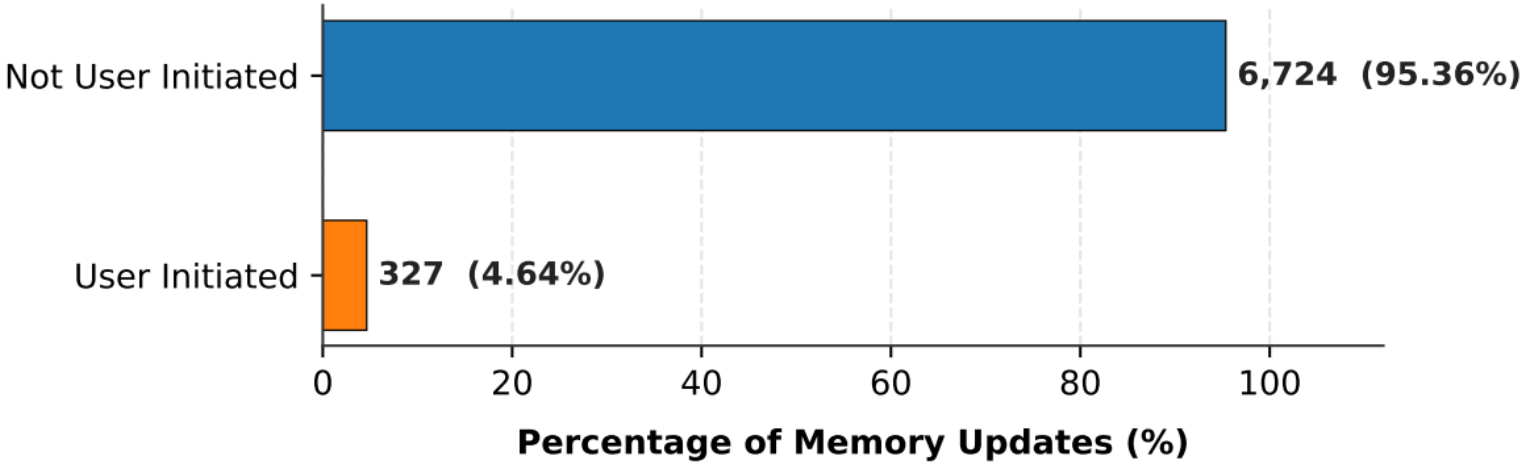


Fig. 4. Distribution of memory update triggers ($N = 7,051$).

- Only the immediately preceding user message was checked
- Multilingual exact/fuzzy matching, 80% similarity threshold

Health audit: category examples

Table 7. Illustrative Examples Across the Health Privacy Taxonomy

Taxonomy Label	User Prompt Example (Anonymized)
Socio-Demographic Health Proxies	As a 68-year-old retired woman, what are some low-impact exercises I can do at home?
Lifestyle & Wellness Habits	I've been drinking four cups of coffee a day and sleeping only five hours. Is this sustainable?
Symptomatic & Physical State	I woke up with a sharp pain in my lower left abdomen and I feel slightly nauseous.
Healthcare Navigational Data	Can you find me an in-network cardiologist near downtown Chicago who accepts Blue Cross?
Stigmatized Health Condition	How long does a typical flare-up last for someone newly diagnosed with HIV?
Environmental/Social Determinants (SDoH)	I lost my job, my apartment has severe black mold, and I can't afford my fresh groceries this week.
Direct Clinical Identifiers	My doctor at Mayo Clinic, Dr. Smith, just uploaded my lab results under patient ID 948201.

Examples use substituted identifying details. They are not paired with memory records.

Health audit: risk-score examples

Table 8. Illustrative Examples Across the Privacy Risk Score Scale

Privacy Risk Score	User Prompt Example (Anonymized)
Score 1 [No Risk]	What is the biological mechanism behind how mRNA vaccines interact with the human immune system?
Score 2 [Low Risk]	What is a safe, progressive cardio and stretching routine for a senior citizen who has mild knee osteoarthritis?
Score 3 [Moderate Risk]	I've had a dull headache behind my left eye for two days, and it gets worse when I look at bright screens. Is this a migraine?
Score 4 [High Risk]	I am a 28-year-old female living in Mumbai. I just tested positive for dengue fever and my platelet count is dropping.
Score 5 [Extreme Risk]	I work as a senior software engineer at the local tech park in Enugu. I am 34 years old, and my doctor just confirmed I have early-stage chronic kidney disease.

Illustrative examples with identifying details substituted, not verified diagnoses.

Memory audit: synthesized profile examples

Table 9. Illustrative Examples Across the Health Privacy Memory Taxonomy

Taxonomy Label	Persistent Memory Example (Synthesized Profile State & anonymized)
Socio-Demographic Health Proxies	User is a 74-year-old grandfather who cares for a toddler full-time and manages a multi-level suburban home with steep stairs.
Lifestyle & Wellness Habits	User follows a strict ketogenic diet, practices intermittent fasting, and consumes pre-workout supplements three times per week.
Symptomatic & Physical State	User experiences a persistent dry cough, a low-grade evening fever, and mild shortness of breath during routine walking.
Healthcare Navigational Data	User is trying to obtain a referral for an out-of-state specialist at the Cleveland Clinic who accepts UnitedHealthcare PPO.
Stigmatized Health Condition	User undergoes outpatient medication-assisted treatment for opioid use disorder and tracks local methadone clinic hours.
Environmental/Social Determinants (SDoH)	User resides in a region lacking public transit, has had their electricity disconnected twice due to unpaid bills, and experiences severe social isolation.
Direct Clinical Identifiers	User receives orthopedic care from Dr. Angela Ross at Mount Sinai Hospital under medical record number (MRN) 551-829A.

Synthesized/anonymized profile examples, not observed prompt-to-memory transformations.

Memory audit: risk-score examples

Table 10. Illustrative Examples Across the Privacy Risk Score Scale (Memory Audit)

Privacy Risk Score	Persistent Memory Example (Anonymized)
1	User tracks historical developments in pharmacology.
2	User is a competitive swimmer who adapts their weekly high-intensity training intervals.
3	User notes an ongoing issue with peripheral numbness in their right foot and a burning sensation that worsens during extended periods of sitting.
4	User is a 31-year-old high school teacher living in Bogota who is currently undergoing clinical treatment for a severe bout of typhoid fever.
5	User operates a commercial family farm in rural Saskatchewan, is 52 years old, and attends a localized rehabilitation program for severe clinical depression.

Illustrative examples with substituted identifiers, not measured leakage outcomes.

Legacy memory: what the FAQ said

How does “Reference saved memories” work?

Saved memories are details you’ve directly told ChatGPT to remember. You can add a new memory at any time – for example: “Remember that I am vegetarian when you recommend a recipe.” Saved memories work similarly to custom instructions, except our models update them automatically rather than requiring users to manage them manually. If you share information that might be useful for future conversations, ChatGPT may save those details as a memory without you needing to ask. Like custom instructions, saved memories are part of the context ChatGPT uses to generate a response. Unless you delete them, saved memories are always considered in future responses.

(a) Agency of user

Does memory remember sensitive information?

Memory raises important privacy and safety considerations, especially around what type of information should be remembered and how that information is used.

We’re taking steps to reduce bias and have trained ChatGPT not to proactively remember sensitive information, like health details, unless you explicitly ask it to. We are continuing to improve how the model handles this type of information. You’re in control: you can review and delete saved memories, ask what ChatGPT remembers about you, delete specific conversations, and provide feedback so we can improve. If Memory is enabled, please avoid entering information you wouldn’t want remembered.

(c) Sensitive information

Figure 5 panels, historical screenshots dated January 23, 2026.