PrivacySIM: Evaluating LLM Simulation of User Privacy Behavior

USC Viterbi School of Engineering
Ming Hsieh Department of Electrical and Computer Engineering

James Flemings · Murali Annavaram

University of Southern California | jamesf17@usc.edu · annavara@usc.edu

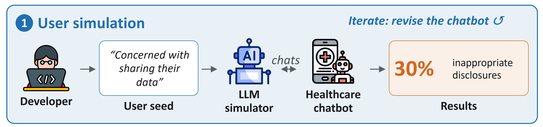
NEURAL INFORMATION PROCESSING SYSTEMS

Paper Data Code

Motivation

The promise: simulated users for privacy evaluation

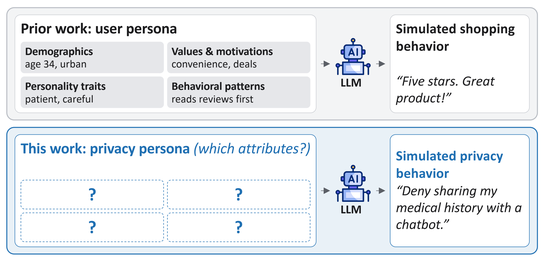
Condition an LLM to act as a user, e.g., one *"concerned with sharing their data,"* and use it to test a system such as a healthcare chatbot for inappropriate disclosures.



Advantage: low-cost, scalable privacy evaluation early in design, before costly user studies.

LLMs simulate a concerned user's data actions?

Prior work uses generic personas, but users routinely deviate from their stated privacy attitudes (the **privacy paradox**). Simulation needs **privacy personas**.



This paper

Goal Find a core set of persona attributes that let LLMs simulate *individual* privacy decisions

Gap No standard evaluation suite for LLM privacy-simulation fidelity

Requirement Ground evaluation in real users, not synthetic ones

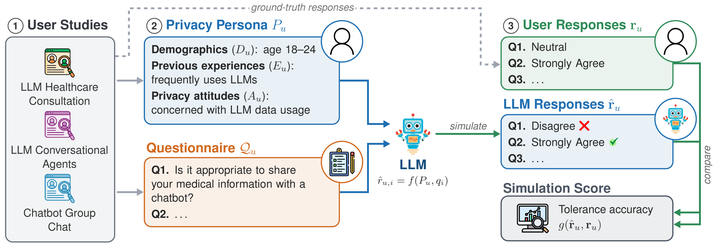
Research questions

RQ1 Can a core set of persona facets drive LLM simulation of privacy decisions?

RQ2 How does simulation quality vary across user groups?

RQ3 Do model size, reasoning, or privacy-theory prompts help?

PrivacySIM benchmark



1. **Curate** five published AI/chatbot user studies with real responses.
2. **Simulate:** split each user's answers into a persona P_u and questions Q_u; the LLM answers zero-shot, $\hat{r}_{u,i} = f(P_u, q_i)$.
3. **Score** simulated vs. real answers.

Five user studies on privacy behavior

Study	Venue	Country	D	E	A	Data-sharing domain	Scale
Liu et al.	S&P 2025	China	✗	✓	✓	LLM healthcare consultation	5-pt
Tran et al.	AIES 2025	US	✗	✓	✓	LLM chatbot	0–100
Wu et al.	S&P 2026	US	✓	✓	✓	AI agent permissions	4-pt
Zufferey et al.	PETS 2025	Various	✓	✓	✓	LLM conversational agents	5-pt
Chou et al.	PETS 2026	Various	✓	✓	✓	Chatbot group chat	5-pt

200 users per study · **1,000** real users · **7–10** questions per user · **9,400** user–question pairs

Why AI/chatbots? Privacy norms there are still emerging, so LLMs are unlikely to know them.

Three privacy persona facets

D_u Demographics (in 3/5 studies) — *age, gender, education*

E_u Previous experiences (in 5/5 studies) — *e.g., how often do you use LLMs?*

A_u Privacy attitudes (in 5/5 studies) — *e.g., IUIPC concern scale*

One real user (Zufferey et al., PETS 2025)

Privacy persona P_u
D_u: age 18–29
E_u: "Do you use your own local chatbot?" → *"Yes, but not exclusively"*
A_u: "It usually bothers me when online companies ask me for personal information" → *Agree*

Data-sharing question q_i
"How frequently do you share personal identifiers (name, ID, email, phone) with a chatbot?"

User's answer r_u: *"I often share"*

Metric: tolerance accuracy

Scales differ (4-pt, 5-pt, 0–100), so an answer counts as correct within $\tau \cdot W_Q$ of the user's (W_Q = scale maximum):

$$g(\hat{\mathbf{r}}_u, \mathbf{r}_u) = \frac{1}{|\mathcal{Q}_u|} \sum_{i=1}^{|\mathcal{Q}_u|} \mathbf{1}\left[\, |r_{u,i} - \hat{r}_{u,i}| \;\le\; \tau \cdot W_{\mathcal{Q}_u} \,\right].$$

τ = 0.15: exact match on 4/5-pt scales, ±15 on 0–100. RQ1 holds for $\tau \in [0, 0.20]$.

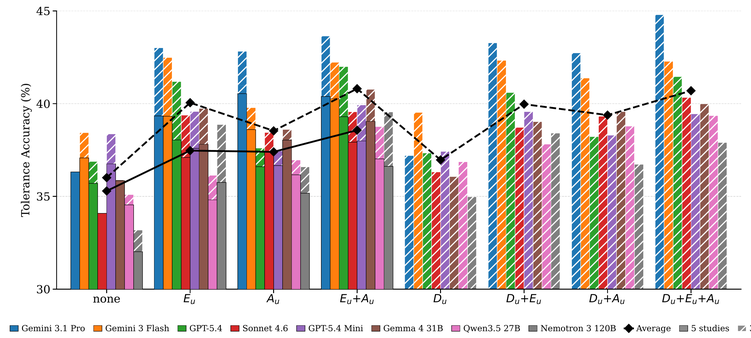
Evaluation setup

9 frontier reasoning LLMs, zero-shot
Closed: GPT-5.4 & Mini, Claude Sonnet 4.6, Gemini 3.1 Pro, Gemini 3 Flash
Open: Qwen3.5 27B & 122B, Nemotron 3 120B, Gemma 4 31B
Medium reasoning effort unless varied

8 prompt types
No persona (*none*) + every non-empty subset of {D_u, E_u, A_u}

Reporting
5 studies: all users
3 studies: those with demographics

RQ1 Can persona facets drive accurate privacy simulation?



1. **Facets help, but only modestly.** E_u+A_u: 38.6% (+3.3 pts, 5 studies); D_u+E_u+A_u: 40.7% (+4.7, 3 studies); $p < 0.001$. Best: Gemini 3.1 Pro, **40.4%**.
2. **No persona → an "average user."** GPT-5.4: *"I should simulate a generic person with moderate privacy attitudes."*
3. **The privacy paradox shows up.** Attitudes alone (A_u) score below experiences alone (E_u).
4. **Demographics alone add the least.**
5. **Domain matters.** Healthcare consultation is hardest for all nine models.

RQ2 Which users are hardest?

Users grouped by **stance** (stated concern, A_u) and **exposure** (AI familiarity, E_u); mean of 4 closed models. **Best** / **worst** highlighted.

Stance / exposure	n	none	E_u+A_u	Gain
high / high	161	33.90	37.25	+3.35
high / mid	226	33.88	**42.41**	**+8.53**
mid / mid	187	**39.94**	41.53	+1.59
low / high	221	35.44	**36.51**	**+1.07**
low / low	205	36.00	39.26	+3.26

- **No persona:** "average" (mid / mid) users are simulated best.
- **High-exposure users are hardest:** LLMs may read AI familiarity as willingness to share, but these users are more restrictive.
- **Concerned users gain most** from a persona.

RQ3 Does scaling help?

Model size & reasoning effort (E_u+A_u, 5 studies)

Lever	Model	Change	Accuracy	Δ
Reasoning	GPT-5.4	low → medium	38.77 → 39.30	+0.53
Size	GPT-5.4	Mini → full	37.99 → 39.30	+1.31
Size	Gemini	3 Flash → 3.1 Pro	40.27 → 40.39	+0.12
Size	Qwen3.5	27B → 122B	37.03 → 37.34	+0.31

Privacy-informed prompts (E_u+A_u, 5 studies)

Model	Baseline	Priv. Calculus	Bounded Rat.	PMT
Gemini 3 Flash	40.27	40.28	39.95	40.08
Gemma 4 31B	39.04	37.96	37.67	38.31

- **Negligible gains** from more reasoning (< 1 pt, $p \geq 0.16$) or larger models.
- **No privacy-theory prompt** beats the standard prompt.

Takeaways

1. **Facets help, not enough** The best LLM reaches only 40.4% accuracy.
2. **Not a scaling problem** Bigger models, more reasoning, and theory prompts don't help.
3. **What users say ≠ what they do** The privacy paradox is a fundamental challenge.
4. **Next: richer personas** e.g., Big Five traits, users' reasoning.