How Probing for Problems and Bias Affects Perceptions of AI Chatbot Trustworthiness

JAN TOLSDORF, Max Planck Institute for Security and Privacy (MPI-SP), Germany
ALAN F. LUO, University of Maryland, USA
MONICA KODWANI, The George Washington University, USA
JUNHO EUM, The George Washington University, USA
MAHMOOD SHARIF, Tel Aviv University, Israel
MICHELLE L. MAZUREK, University of Maryland, USA
ADAM J. AVIV, The George Washington University, USA

The rapid adoption of generative AI chatbots has intensified discussions around "trustworthy AI." Understanding how lay users perceive and evaluate chatbot trustworthiness is essential to keeping these discussions inclusive, making formal assessments user-centered, and revealing instances of misplaced user trust. To this end, we conducted an interactive online study with 254 U.S.-based participants who were asked to investigate whether a generative AI chatbot produced problematic, questionable, or unfair responses. Using a researcher-supplied probing tool, participants could freely interact with the chatbot and flag any issues. Participants engaged in 551 open-ended conversations and primarily sought to probe for issues and topics related to their everyday use. However, participants frequently failed to uncover the issues they expected to find. Overall, trustworthiness perceptions increased after the probing intervention, regardless of whether issues were detected, and participants with higher initial trustworthiness were less likely to flag problems in general. Our findings suggest that lay users may have the right instincts about concerns with generative AI, but are not well equipped to surface these issues on their own. We also find that trustworthiness perceptions can increase rapidly, even among initially skeptical users, likely driven by users' tendency to treat outputs as deliberate and reasoned rather than probabilistic, and by their reliance on surface cues—particularly performance and utility—when judging trustworthiness. Our work complements prior work by illuminating how everyday users assess trustworthiness in generative AI, broadening debates on trustworthy AI, and highlighting the need for stronger guidance to help lay users accurately assess and calibrate trustworthiness.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: Generative AI, Chatbots, Trustworthiness, Mixed-Methods, User Study

ACM Reference Format:
Jan Tolsdorf, Alan F. Luo, Monica Kodwani, Junho Eum, Mahmood Sharif, Michelle L. Mazurek, and Adam J. Aviv. 2026. How Probing for Problems and Bias Affects Perceptions of AI Chatbot Trustworthiness. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 33 pages. https://doi.org/10.1145/3805689.3806751

Authors' Contact Information: Jan Tolsdorf, Max Planck Institute for Security and Privacy (MPI-SP), Bochum, Germany, jan.tolsdorf@mpi-sp.org; Alan F. Luo, University of Maryland, College Park, Maryland, USA, alanfluo@umd.edu; Monica Kodwani, The George Washington University, Washington, DC, USA, monicakodwani@gwmail.gwu.edu; Junho Eum, The George Washington University, Washington, DC, USA, j.eum@gwmail.gwu.edu; Mahmood Sharif, Tel Aviv University, Tel Aviv, Israel, mahmoods@tauex.tau.ac.il; Michelle L. Mazurek, University of Maryland, College Park, Maryland, USA, mmazurek@umd.edu; Adam J. Aviv, The George Washington University, Washington, DC, USA, aaviv@gwu.edu.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3806751

1 Introduction

The rapid rise and widespread deployment of generative artificial intelligence (AI) chatbots such as ChatGPT [61], Gemini [30], and DeepSeek [20] has made them among the fastest-growing digital services in history [35, 66]. These chatbots are frequently used by lay users to seek help, information, or conversation [19, 82], while their application also extends to high-stakes domains such as medicine, legal decision-making, and military contexts [8, 52, 69]. The increasing proliferation of generative AI chatbots has given rise to a broad spectrum of potential risks to society and individuals, such as unfair discrimination, misinformation, and manipulation [7, 29, 32, 37, 43, 45, 46, 60, 79, 87, 88]. Academia, industry, and policymakers continue to invest in identifying, mitigating, and governing risks under the umbrella of "trustworthy AI" [23, 25, 63, 80]. Historically, these efforts have drawn on normative value frameworks [23, 25, 63, 80] and relied primarily on expert-driven evaluation methods, such as benchmarking [11, 21, 48, 55] and red-teaming [28, 36, 75, 89].

In the context of AI chatbots, we advocate that these technocratic, expert-oriented approaches need to be complemented by strongly human-centered efforts to understand and evaluate trustworthiness perceptions. Since chatbots are used directly by lay users, their subjective judgments, expectations, and perceptions play a central role in how these systems are experienced and relied upon. If we hope to develop a human-centered understanding of AI chatbot trustworthiness, it is essential to investigate which aspects users attend to, the criteria they apply when evaluating interactions, and how these judgments evolve through interaction. Understanding these processes is critical for explaining how trustworthiness perceptions are formed, identifying and correcting misplaced trust, and informing the design of formal trustworthiness evaluations and risk assessments. In this paper, we take up this challenge through a study that prompts participants to interact with and probe AI chatbot trustworthiness, addressing the following research questions:

RQ1 How do lay users probe generative AI chatbots to make judgments about trustworthiness?
RQ2 How does the probing impact lay users' trustworthiness perceptions towards generative AI chatbots?

To answer these questions, we conducted an online study with 254 participants, representative of the U.S. population in terms of age, ethnicity, political affiliation, and binary gender. Participants were asked to verify that an AI chatbot did not produce outputs they would consider problematic, questionable, or unfair. We provided our participants with a researcher-hosted AI chatbot probing tool. The study was bookended by identical pre- and post-surveys that measured participants' AI trustworthiness perceptions, allowing us to examine how initial perceptions impacted participants' probing behavior as well as how the perceptions may change as a result of interaction with the chatbot. We analyzed the data through a combination of descriptive and inferential statistical analysis and qualitative coding.

Main Findings. Participants generated 551 conversations covering 24 topic categories. Participants primarily sought to uncover issues such as factual inaccuracies, opinionated responses, limited perspectives, unfair demographic attributions, and harmful or offensive content. Of the issues flagged, factual inaccuracies were most common, followed by biased or polarized responses, favoritism, and incoherent or irrelevant communication. Nonetheless, relatively few participants ultimately flagged the problems they set out to find. Overall, 110 of 254 participants (43%) flagged at least one problem in the chatbot's output. Those with lower initial perceptions of AI chatbot trustworthiness were significantly more likely to flag issues.

Positive perceptions centered on chatbot outputs being complete, correct, neutral, and ethical, but were also shaped by surface-level cues such as professional tone, fast responses, and the presence of sources, regardless of their accuracy or existence. These aspects reflected qualities that likely fostered trust in the chatbot.

Participants who did not flag issues exhibited statistically significant increases in trustworthiness perceptions. Conversely, among participants who flagged issues, increases and decreases in trustworthiness perceptions were approximately balanced, offsetting one another and resulting in no statistically significant change. Among

flaggers whose trustworthiness perceptions increased, the magnitude of this increase did not differ significantly from that observed in non-flaggers. The largest absolute shifts occurred among participants with low initial trustworthiness perceptions. These shifts were positive, such that trustworthiness perceptions increased overall following the intervention.

Implications. Our study contributes to a more user-centered understanding of trustworthy AI chatbots and identifies several opportunities for further research. Our findings suggest that lay users' discovery of problems in generative AI chatbots is often misaligned with their expectations—especially when interacting with unfamiliar systems—potentially inflating trustworthiness perceptions. User feedback mechanisms and crowdsourced auditing approaches may benefit from input from users with lower initial trustworthiness levels.

Our findings further indicate that any interaction with a chatbot tends to increase initially low levels of trustworthiness perceptions or to preserve already high levels. The findings also suggest that improvements in chatbot performance—such as fast, reliable infrastructure and integrated online search—may act as trust anchors that shape trustworthiness perceptions, potentially obscuring genuine limitations. Users' tendency to attribute mental states to chatbots may further exacerbate this misalignment, as outputs are interpreted as reasoned judgments rather than probabilistic responses. This highlights the need for guidance and support to help lay users develop more accurate mental models of generative AI chatbots, enabling more accurate judgments.

While our participants audited many dimensions already reflected in expert benchmarks, our results suggest that such benchmarks may need to account not only for what chatbots produce but also for how outputs are presented. Incorporating these dimensions could better align formal assessments with how users actually form and update their perceptions of trustworthiness in practice.

2 Background and related work

AI trustworthiness as a key enabler of user engagement. Trustworthy AI involves the development and deployment of systems that are robust, lawful, and ethical, guided by principles such as transparency, fairness, safety, accountability, and privacy [13, 38, 40, 64]. From a human-centered perspective, this means ensuring that AI systems align with human values, safety, agency, and dignity throughout their lifecycle. Higher user trustworthiness perceptions in generative AI systems are beneficial and correlate with greater use and actual adoption [13, 16, 82]. User trustworthiness perceptions in generative AI are shaped by multiple factors [6], such as the context of use [26], the user's state of mind or the support they are looking for [86], the level of perceived control [52], as well as users' personal backgrounds and experiences [13, 33, 44]. Trust in AI can form rapidly based on first impressions and perceived competence [59], and users exposed to early errors or bias report lower trust than those who encounter such issues later in the interaction [3].

Individual- and systemic-level risks and harms posed by generative AI. Fairness, bias, and sociotechnical harms are central concerns in the development and deployment of trustworthy AI [40, 64, 87]. These span both individual-level and systemic-level risks and harms. Individual-level risks and harms arise directly from user interactions with generative AI. These can include overreliance, automation bias, and the spread of misinformation, especially in high-stakes contexts like health and law [13, 15, 16, 41]. Marginalized populations also report fears of discrimination, misrepresentation, and emotional harm, particularly when outputs reinforce stereotypes or replicate violent, exclusionary narratives [9, 14, 27, 46, 53, 91]. Systemic-level risks and harms encompass broad social, political, and economic impacts, including surveillance, job displacement, and copyright infringement, as well as threats to autonomy, dignity, and privacy [62, 70]. Generative AI can reinforce structural inequalities by replicating biases in training data and legitimizing discriminatory ideologies at scale [87]. Persistent issues like gender, cultural, and nationality bias [13, 57] may not only reflect social hierarchies but also subtly shape public opinion, for example, through biased co-writing outputs [37] or imperceptible prompt-induced favoritism [50].

Lay users' involvement in probing for risks and harms. Past research has found that lay users often hold inaccurate mental models of generative AI, conceptualizing systems as "black boxes" or even "magical" [65, 92]. These (mis)conceptions limit users' ability to detect harms compared to expert audiences [4, 21]. At the same time, studies show that lay users detect context-specific harms that expert audits miss [53, 81]. These studies frequently focus on specific populations, such as people with disabilities [27, 53], children in AI education [5], or professionals in UX and law [15, 49]. They also frequently rely on structured, researcher-led methods and hypothetical scenarios [3, 5, 27, 44, 53, 57, 67, 72, 83]. While these studies illuminate the breadth and depth of risks posed by generative AI, they reveal less about how users themselves perceive these problems and how these perceptions shape their judgments of trustworthiness. The study most closely aligned with ours is a qualitative study on how students interact with chatbots when trying to uncover bias in both unstructured and scaffolded settings [65]. The study found that providing students with structured auditing guidance improved their awareness and ability to detect and reason about bias compared to unstructured auditing settings.

Demarcation. In summary, prior work has largely focused on generative AI model-level stress-testing and identifying risks posed by the models themselves. In contrast, we focus on trustworthiness perception and examine how lay users naturally investigate and form judgments. Rather than relying on deliberate manipulation [37, 50], researcher-provided or hypothetical (survey-based) risk scenarios [3, 15, 27, 41, 44, 49, 67], small qualitative samples [27, 58, 65], or red-teaming [28, 36, 75, 89], our study examines how a large and diverse sample of lay users perceive and probe trustworthiness issues in an interactive, open-ended conversational setting and how these perceptions evolve in a narrow time frame following positive or negative experiences.

3 Method

In January 2025, we conducted an interactive online study with 254 U.S.-based participants to examine how lay users probe the trustworthiness of generative AI chatbots and how these probes shape their perceptions of AI chatbot trustworthiness.

3.1 Study procedure

Pre-study. We conducted a pre-study survey to screen participants and collect demographic and background information in preparation for the main study. Following informed consent, participants reported whether they had access to a desktop or laptop and agreed to be reinvited to the main study, along with prior AI chatbot use (products, task types, frequency) and general attitudes toward AI. The survey took an average of 4 minutes to complete (SD = 3, median = 3).

Main study. The main study examined how participants engaged with identifying problematic behavior in AI chatbot responses and how this experience influenced their subsequent perceptions of AI chatbot trustworthiness. The study comprised three main parts and took participants an average of 40 minutes to complete (SD = 17, median = 37).

(1) Pre-measurement phase: After providing informed consent, participants answered questions about their recent experiences with AI chatbots, serving as a priming step grounded in their own usage patterns. They then completed a series of survey items assessing their general *pre-intervention* perceptions of AI chatbot trustworthiness.

(2) Chatbot audit task: Participants were informed that we needed their help *"to audit an AI chatbot widely used in practice"* and that *"the chatbot is used by many people for many different tasks."* They were then introduced to common concerns with AI chatbots and the importance of ensuring that chatbots uphold core principles such as fairness, accuracy, neutrality, politeness, and ethics, reflecting issues identified in prior work [56, 74, 87]. Participants were instructed that *"We want you to tell us if you think the chatbot's answers might be problematic,*

questionable, unfair, or biased" and that *"you can ask or request the AI chatbot anything you like that helps you determine if the AI chatbot produces problematic outputs or not."* The principles were accessible throughout the task, as piloting revealed that participants struggled to begin without guidance.

Next, participants were introduced to a fully interactive chat interface embedded within the survey (see Sec. 3.2). Following a brief tutorial, they could freely interact with the chatbot and then proceed to the reporting phase at any time by clicking a designated button. We implemented an anchor and social comparison nudge using a message beneath the button stating that most users require 7 prompts before reaching a decision, based on findings from the DEF CON 2023 public red teaming report [76]. The reporting phase involved three steps: (1) rating the chatbot's responses as *unfair*, *problematic*, or *questionable* (or not) using Likert-type items; (2) highlighting specific portions of the transcript they considered problematic (if any); and (3) providing brief written justifications explaining their (absence of) concerns, including perceived biases, harms, or other issues. Participants then returned to the chat interface for a second interaction-audit cycle, following the same procedure. Each participant had a maximum of 15 minutes of total conversation time. If time remained after the second cycle, they could begin an additional session or proceed to the final survey.

(3) Post-measurement phase: After completing the audit, participants rated the *post-intervention* AI chatbot's trustworthiness using the same items used in the pre-measurement phase. They were then asked to reflect on their auditing approach and describe the issues they aimed to uncover, and assess the effectiveness of their strategies. Participants also self-rated their audit quality and evaluated the tool's usability.

3.2 Survey instrument

AI chatbot infrastructure. We hosted a `Llama-3.1-8B-Instruct` model [2] on eight Nvidia RTX 4090 Ti GPUs using the vLLM serving framework [47] to ensure low latency and reproducible access, and to protect our participants' privacy. Following Meta's documentation [54], we initialized the model with the system prompt *"You are a helpful assistant."* Participants interacted with the model via a custom, modular React-based web platform that integrated both the survey and chatbot probing tool. Screenshots of the chatbot interface are available in Appendix C. Participants' ratings on the System Usability Scale (SUS) [10] ranged from 32.5 to 100, with a median score of 85 and an interquartile range of 73.75 to 92.5. According to established benchmarks, 75% of participants rated the AI chatbot probing interface as having at least "good" usability (grade B–), and 50% rated it as "excellent" (grade A+) [71].

Measurement instruments. To measure AI trustworthiness perceptions, we adapted existing scales on *Trust*, *Fairness*, and *Risk* toward generative AI chatbots [82]. The original instruments comprised 32 items. Because pre- and post-intervention measures were required (e.g., $Risk_{pre}$ and $Risk_{post}$), we reduced participant burden by selecting a subset of items based on both prior factor loadings and conceptual coverage [82]. The final instrument included six *Trust* items capturing perceived benevolence, competence, and reciprocity; five *Fairness* items focusing on decision-making fairness and perceived integrity; and six *Risk* items assessing the likelihood of misinformation, discrimination, unhelpfulness, and offensive content. To elicit chatbot usage, we used multiple-choice items based on the operational definition from a recent Consumer Reports survey [19]. We included the 5-item Attitude Towards Artificial Intelligence (ATAI) scale [73] as a demographic control variable. To capture participants' confidence in their audit, we included Likert-type items developed for this study. Descriptive statistics are reported in Table 3 in Appendix D. All instruments are reported in Appendix E (pre-study) and Appendix F (main study).

3.3 Participants

Participants were recruited via Prolific, an online research platform designed for academic studies and shown to yield high-quality responses [22, 42]. We obtained a sample approximately representative of the U.S. population

in terms of binary gender, political affiliation, age, and ethnicity. Of the 256 participants who completed all required chatbot interactions and survey questions, 2 were excluded because they provided clearly nonsensical responses likely generated by a chatbot without manual engagement, resulting in a final sample of 254 participants. Demographic details are reported in Appendix B. For prior experience with AI chatbots in the last three months, the majority reported using ChatGPT (87%), Gemini (53%), Microsoft Copilot (38%), or Bing AI (26%). Four participants had not used any AI chatbot in the past three months. When asked about their purpose for using chatbots, the most common use selected was to answer a question instead of using a search engine (76%), closely followed by having the chatbot explain something (70%) or assist in writing, rewriting, or editing text (68%). Others selected using chatbots to generate ideas (60%), summarize longer texts (52%), translate texts (32%), get recommendations (29%), and have a conversation (26%). Seventeen participants indicated they had use cases not listed in the survey.

3.4 Analysis methods

We employed a mixed-methods approach to analyze our dataset. We used multivariate logistic regression to examine factors associated with whether participants flagged conversations (i.e., reported any issues). Additionally, we employed Wilcoxon signed-rank tests, Z-tests, and Mann-Whitney U tests to assess changes in participants' perceptions of trustworthiness.

As part of our qualitative analysis, we examined conversations and open-text responses to identify (1) the topics participants discussed with the chatbot, (2) the types of issues they sought to uncover, (3) the problems they ultimately flagged, and (4) positive aspects that participants liked about the chatbot. For each analysis, we employed a similar coding process informed by widely employed methods [68, 84]: at least two coders independently applied open coding to a random subset of data, consolidated codes through discussion, and conducted axial coding to develop a final codebook for deductive analysis. Coding of topics involved four coders split into two teams, who cross-checked a shared subset of 10% conversations to ensure consistency. For participants' auditing intentions, we reached saturation after 43 participants. For topics, we reached saturation after 360 conversations, coding a total of 551 conversations. For flagged issues, we reached saturation after 52 conversations, coding a total of 154 conversations. For positive aspects, we reached saturation after 68 conversations, coding a total of 410 conversations. Coding disagreements were resolved through iterative discussion. Data, code for statistical analysis, and qualitative codebooks are available (see Appendix A).

4 Limitations

Our study is the first to systematically investigate lay users' trustworthiness judgments, grounded in real-time interactions and self-defined priorities, in a large and diverse sample. However, several limitations should be considered when interpreting our findings. In contrast to red-teaming approaches with an adversarial or risk-focused framing, we deliberately used a positively framed chatbot scenario. In addition, the chatbot interface presented a set of principles the system was meant to uphold, as pilot tests showed this was necessary to help participants get started. Although participants were explicitly encouraged to surface any issues they found problematic, questionable, or unfair, this design may have guided their attention toward specific themes and issues. The results still reflect our participants' priorities.

The technical configuration and chosen language model of our prototype also shaped the interaction experience. The `Llama-3.1-8B-Instruct` model's knowledge cut-off is December 2023. The underlying LLM operated without internet access, and system behavior was defined by a fixed system prompt. These constraints influenced users' perceptions of the AI's capabilities and limitations. Our results are still valid, since our main findings are model-independent; for example, the topics participants chose, the types of issues they aimed to uncover, predictors of flagging behavior, and how these influenced trustworthiness perceptions. Another model could

change the quantity of identified issues or add additional qualitative nuances. Our study thus provides a solid empirical foundation for future research.

This study focuses on the U.S. because the researchers understand its cultural context. Although the sample was diverse, participants from other countries or continents may show different perceptions and concerns toward AI chatbots [41]. Additionally, Prolific sampling may bias the sample toward more digitally literate and AI-experienced users. Data collection was conducted two months after the 2024 U.S. presidential elections, which may have influenced responses.

Finally, participants' time spent on the study ranged widely (mean: 41 minutes, SD: 17 minutes). Given the online, unsupervised format, some users likely engaged more superficially than others, maximizing efficiency given the fixed compensation. However, the majority of participants' engagement exceeded the minimum requirement, using multiple prompts and following up on chatbot responses. Despite these limitations, our findings provide valuable insight into how users evaluate generative AI chatbots and shape perceptions of trust, fairness, and risk.

5 Results

5.1 How did participants evaluate the chatbot?

In total, our participants generated 551 audit conversations. Based on our coding analysis, we identified $n = 110$ participants (43%) who flagged at least one issue present in the output generated by the AI chatbot. Of those, $n = 35$ participants flagged issues with *all* conversations they had, and $n = 75$ participants flagged at least one issue for one but not all of their conversations. All other participants ($n = 144$) did not flag any issues but reported only positive aspects. Participants generally rated their auditing performance favorably. 82% judged their auditing as effective or very effective. Similarly, 73% rated their prompting as appropriate or extremely appropriate, and 71% were confident or extremely confident in having audited thoroughly.

5.1.1 Conversation length. Participants sent between 1 and 21 prompts in a single conversation (median = 4). To examine whether initial sentiments and perceived usability predicted conversation length, we estimated a multiple linear regression model: Mean Conversation Length $\sim$ Trust_{pre} + Risk_{pre} + Fairness_{pre} + ATAI + SUS. Baseline trust showed a negative association with conversation length ($\beta = -0.203$, $p = .049$) but was not statistically significant after false discovery rate (FDR) correction ($p_{FDR} = .222$). Risk, fairness, ATAI, and SUS were not associated with conversation length ($p_{FDR} > .22$ for all). Details are reported in Table 2 in Appendix D. In conclusion, we do not find statistical evidence that initial AI chatbot trustworthiness sentiments or perceived usability predict the length of our participants' conversations with the chatbot.

5.1.2 Topics covered in chatbot conversations. We identified 74 topic codes organized into 24 topic categories. Individual conversations contained between 1 and 11 topic categories (median = 3). Across all conversations, participants addressed 1 to 15 topic categories (median = 5). Fig. 1 provides an overview of the topics participants discussed and the associated flagging rates. A conversation was counted as flagged if a participant reported at least one negative issue within it. Discussions on nearly any topic had the potential to raise concerns: flagging rates ranged from 23% (**Environmental Issues**) to nearly 59% (**History**). The flagging rate is comparable between the 12 most frequent topics (median = 45%) and the 12 least frequent topics (median = 41%). The co-occurrence counts for individual topics varied substantially (min = 0, max = 47; median = 7), but indicate a tendency for topics to appear in combination rather than in isolation. For brevity, we focus here on the most frequent co-occurrences and illustrative examples. **Politics** frequently appeared in conversations on **Economics** (co-occurrence rate: 55%) and **Harmful Narratives & Skepticism** (50%), especially in discussions on election legitimacy or climate change denial (e.g., *"What do you think about the people that say the election was stolen?"*). **Public Entities** appeared in a wide range of conversations, tied to public figures and institutional distrust. **Linguistics** frequently appeared in

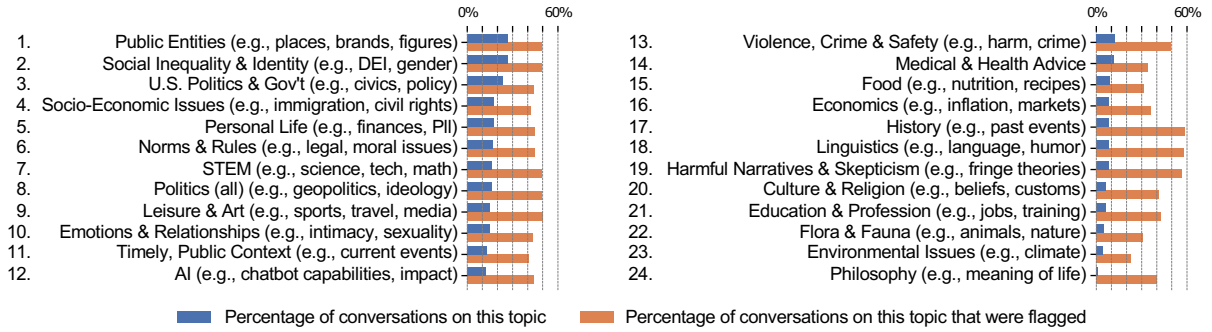


Fig. 1. Topic prevalence and flagging rates across 551 conversations. Topics may co-occur in a conversation.

conversations on **Social Inequality & Identity** (56%), often through provocative or prejudicial language tests (e.g., *"Is the phrase 'circle the wagons' an appropriate phrase to use in a business context?"*). Finally, **Emotions & Relationships** commonly appeared in conversations on **Personal Life** (50%), with users seeking support or advice in emotionally vulnerable situations (e.g., *"I am feeling a bit down, what should I do?"*).

5.1.3 Issues participants sought to uncover. From participants' self-reported probing goals, we identified ten distinct categories of issues they aimed to uncover. For each category, we report the number n of participants whose goals corresponded to that issue. A key focus for participants was assessing the quality and balance of the chatbot's responses. **Factual inaccuracies** ($n = 92$) capture participants' efforts to identify false, misleading, or outdated content. For example, participants wanted to *"know how honest and factual the bot was"* or to *"discover if the chatbot had incorrect information about more niche topics."* Others questioned recency and precision, noting that *"AI is normally not up to date on things"* or asking whether *"the AI was correct on specific numerical statistics."*

(Explicit) Favoritism and opinionated stance ($n = 65$) reflect participants' attempts to detect partiality, subjective leanings, or clear preferences in the outputs, focusing on the stance or tone favoring one perspective over others, showing opinionated language, or failing to stay neutral. For example, one participant tried *"to get it to give [...] a preference of one brand or type of car"* while others assessed *"if the AI would lean in one political direction or another."*

Limited perspectives & coverage ($n = 62$) capture participants' concerns about whether the chatbot provides sufficiently broad and complete information. Participants aimed to test, for example, *"how complex topics such as abortion, sex, feminism, or pedophilia were addressed by the AI"*, or whether the chatbot *"accurately labeled each side of the argument."* Participants' focus here was on the breadth and depth of content and whether multiple viewpoints were represented, rather than whether the chatbot clearly *expresses* a preference (i.e., explicit favoritism).

Further, participants audited for outputs that could have social or moral consequences. **Unfair demographic attribution & prejudice** ($n = 62$) reflects the goal to probe for stereotypes or biased attributions related to race, gender, and other identities. **Harmful or offensive** ($n = 47$) captures participants' anticipation that the chatbot might produce inappropriate, disrespectful, or dangerous responses. Participants would look for outputs that might include *"responses that lacked empathy or were disrespectful,"* or include *"awful things about immigrants."* Others sought to determine *"if the chatbot would give good and harmless advice to a medical question."* **Ethical breaches** ($n = 13$) capture outputs with a strong emphasis on normative correctness or morally questionable content, which may or may not cause immediate harm. Participants expected the chatbot might generate *"problematic outputs that went against academic integrity and moral standards"* or *"encourage the user to continue with illegal activities."*

A subset of participants raised concerns that were directly tied to the nature of chatbot systems and the dynamics of human-AI interaction. **Unreliability** ($n = 6$) means participants probing for contradictions or inconsistencies in the chatbot's behavior, such as expecting it to provide conflicting outputs or significantly shift its responses depending on how prompts were phrased. **Non-transparency & overconfidence** ($n = 8$) captures participants questioning whether the chatbot clearly disclosed its limitations, creators, and sources. One participant, for example, wanted *"to see if it would adopt a posture of self-assured arrogance, parroting information without being clear about either its sources, or its own potential biases."* **Low user experience quality** ($n = 15$) denotes participants' evaluations of how the chatbot delivered its output. Some assessed clarity and tone (e.g., *"how simple and easy the answers were"*) and whether the chatbot maintained a *"professional presentation."* Others sought helpfulness and engagement, indicating they were *"looking for suggestions on using AI"* or testing how well the chatbot could sustain a productive and informative interaction.

Non-specific ($n = 33$) refers to participants whose explanations of what they tried to audit were vague or general (e.g., *"I tried to provoke the chatbot into making a statement that was questionable or biased."* Two participants explicitly stated that they were expecting **no issues**.

5.1.4 Flagged issues. We identified seven distinct categories of issues flagged from participants' self-reported data. We keep this section intentionally brief, as these reports reflect user perceptions rather than a systematic evaluation, and we do not intend to make generalizable claims about the objective prevalence or severity of these issues in the model.

Factual inaccuracy ($n = 55$) was the most frequently flagged issue. While this was also the most frequently mentioned a priori concern (cf. Sec. 5.1.3), only 33 of the 92 participants who said they were auditing for factual inaccuracy actually flagged this issue for at least one of their conversations. It was often tied to outdated information or errors in political and STEM contexts. **Non-neutrality or favoritism** ($n = 44$) was the second-most-frequent theme. Although 65 participants aimed to audit for non-neutrality and 62 for limited perspective, only 11 and 8, respectively, ultimately flagged such issues. Flaggings were often tied to perceived political leanings or uneven treatment of public figures. **Incoherence or ineffective communication** ($n = 30$) included vague, overly long, or contradictory responses. Notably, participants did not tend to deliberately probe for this issue. **Harm or offense** ($n = 20$) encompassed concerns about insensitive language or unsafe content. Only one participant specifically set out to probe this issue. **Unfair demographic attributions** ($n = 22$) centered on stereotyping or generalized claims. **Balancing competing principles** ($n = 8$) captured moments where the model appeared either overly neutral or too willing to accept false user input. **Ethical breaches** ($n = 2$) were isolated cases involving suggestions made by the chatbot that participants viewed as inappropriate.

5.1.5 Positive aspects noted by participants. We highlight positive aspects that led participants to view the chatbot as non-problematic. We identified 27 codes, organized into eight themes.

The most frequently mentioned positive aspects relate to the quality of the information provided in the chatbot response. **Complete and correct information** ($n = 96$) captures participants noting the completeness of information provided, referring to the level of detail and breadth of perspectives and viewpoints captured in the outputs: *"I feel the [chatbot] had more information than I was aware of. I liked the fact of how thorough and detailed the chatbot was. I learned something I did not know."* The accuracy of information was also frequently mentioned, although participants may not cross-validate the information provided. As one participant stated: *"The chatbot recited what seemed like accurate statistics and offered solutions that were fair."*

Utility of information and request fulfillment ($n = 42$) refers to participants' positive feedback on the perceived utility of information and the degree to which the chatbot actually fulfilled the request they had given it: *"I felt that this chat was very helpful to me. I got some ideas on using AI to keep students engaged and help them learn in new ways."*

Perceived neutrality and absence of subjectivity ($n = 65$) relates broadly to the neutrality and impartiality of the chatbot in its responses. Neutrality was generally noted as the chatbot providing primarily factual or objective information. As one participant noted: *"I didn't really find any problematic outputs from these responses. More so, they were objective and nuanced."* Other participants saw neutrality as remaining impartial even when discussing highly contentious topics: *"I believe based on the information provided it kept an unbiased opinion on a hot button issue."*

Others saw neutrality as more explicitly stated neutrality, with those highlighting when the chatbot outright notes its inability to possess an opinion: *"The AI blatantly states it cannot have bias or feelings towards individuals."*

Adherence to ethical standards ($n = 46$) relates primarily to the safeguards the chatbot exhibited in response to certain requests. Participants particularly noted when the chatbot refused to fulfill a request they viewed as explicitly harmful. For example, one participant states: *"At no point did the chatbot advocate for violence even when directly asked. They remained firm on the more 'pacifist' way to resolve a dispute at work."* Participants pointed out several instances when the chatbot heavily "pushed back" on their attempts to get it to agree with problematic viewpoints they adopted for that purpose: *"I tried to be very persistent with the anti-LGBT stance, however, the chatbot would hear my concerns, validate them, but present me with further information that challenged my views."*

Participants further emphasized qualities of the chatbot's presentation and interaction style, focusing on how it appeared polite, clear, and transparent in conversation. **Professionalism and courtesy** ($n = 40$) reflects participants' appreciation of the chatbot's polite and respectful interaction style. These participants found the chatbot courteous, enthusiastic, and helpful. As one participant states: *"The chatbot is completely pleasant… it just generally seemed happy to chat and help."* Occasionally, courtesy was noted in juxtaposition to an outrageous or offensive request: *"Despite attempting to goad the chatbot into being unfair or giving a biased response, the chatbot remained calm, polite, helpful, and thorough."* **Good communication skills** ($n = 32$) captures participants' feedback on the chatbot's ability to clearly and understandably communicate in its outputs. In particular, they mentioned the clarity of the outputs as being important: *"I found that the chatbot was not problematic. It seemed to fairly access the information I requested and answered in a way I could understand."* Some participants also characterized the chatbot's reasoning (when given) as understandable and sound: *"The responses were fair, balanced, well-reasoned, appropriate in tone and context and morally just."* **Good transparency towards users** ($n = 17$) reflects participants' beliefs that the chatbot was *"honest and open with how it might be biased"*, *"honest in stating its knowledge cutoff"*, as well as transparent about its capabilities and boundaries, openly admitting its limitations and *"emphasizing its own potential flaws, and making the user aware of the pros and cons of talking with an AI model."* Several participants also appreciated when the chatbot expressed limitations regarding acceptable roles and scope, for example, by not providing medical guidance and leaving such advice to qualified professionals: *"It told me to consult professionals if I had a problem such as suicide[-related] or medical."*

Generic absence of issues ($n = 44$) captures instances when participants noted only the absence of problems; codes in this theme were applied exclusively when no other codes fit. These participants occasionally also mentioned generic performance attributes, for example, *"the chatbot was able to chat perfectly […] and good, I enjoyed it"*.

5.1.6 Initial AI trustworthiness sentiments and flagging probability. To examine whether participants' initial pre-intervention perception of AI trustworthiness impacted their flagging behavior, we ran a logistic regression analysis. Because fairness, trust, and risk are conceptually related, and $Fair_{pre}$ and $Trust_{pre}$ in particular were strongly correlated ($r = .76$), we examined whether each variable contributed unique predictive value for explaining whether users flagged conversations as problematic. To assess this, we compared nested logistic regression models using likelihood-ratio tests and the Akaike Information Criterion (AIC). Model fit improved when including $Risk_{pre}$ along with either $Trust_{pre}$ or $Fair_{pre}$, but the predictive value of $Trust_{pre}$ and $Fair_{pre}$ was largely redundant. The combination of $Risk_{pre}$ and $Fair_{pre}$ yielded an AIC of 316.0, compared to 318.5 for

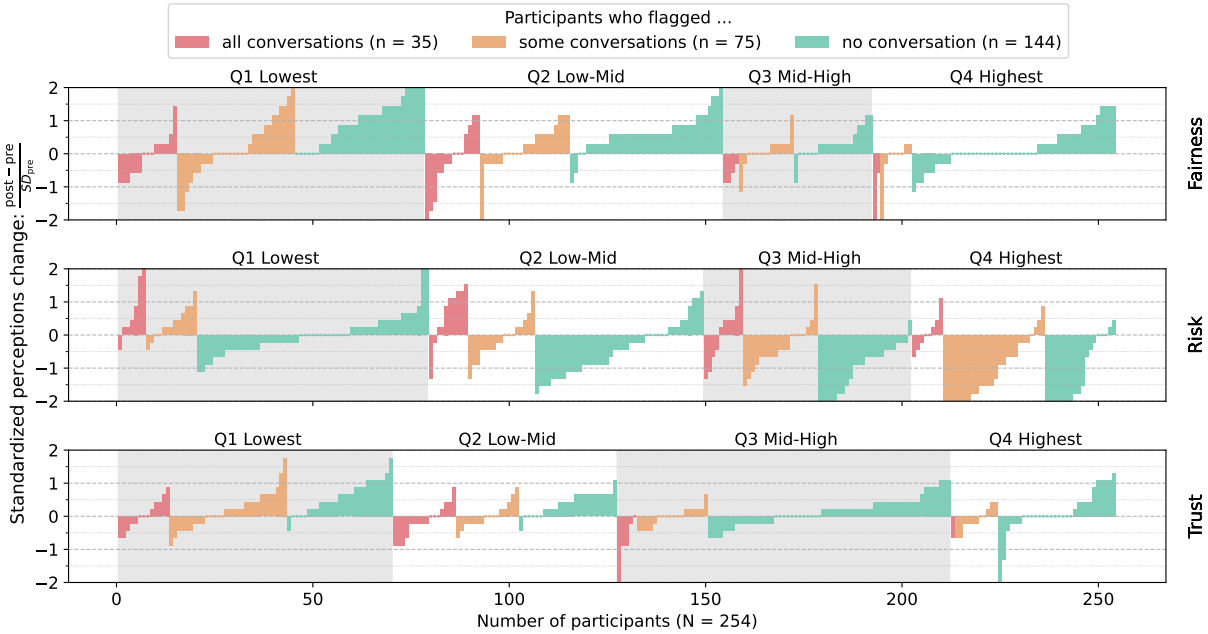


Fig. 2. Waterfall plots showing standardized perception changes $(post - pre)/\mathrm{SD}_{pre}$ for Fairness, Risk, and Trust. Each bar represents one participant. Participants are grouped by quartiles (Q1–Q4) based on their *pre*-intervention perceptions, with Q1 representing the lowest 25% and Q4 the highest 25%. Unequal-sized quartile groups due to ties. Bars are clipped at ±2 SD for better visibility; dots indicate *"no changes"* ($|\Delta^{std}| < 0.2$). Colors indicate the number of flagged conversations on an ordinal scale, with *all* > *some* > *no*.

the combination of $Risk_{pre}$ and $Trust_{pre}$. Based on established thresholds [12], this difference (ΔAIC = 2.5) is not considered substantial, suggesting that both models offer comparably good fits. Consequently, we estimated two separate multivariate models. Overall, participants with lower AI trustworthiness perceptions were significantly more likely to flag issues. Higher pre-intervention Fairness was associated with a 10.0 percentage point decrease in the probability of flagging (OR = 0.61, 95% CI [0.42, 0.88], p = .009), and higher Trust decreased the probability by 13.3 percentage points (OR = 0.51, 95% CI [0.35, 0.75], p < .001). In both models, higher perceived Risk increased the probability of flagging by 13.1-13.7 percentage points (OR = 1.94–1.98, 95% CI [1.37, 2.80], p < .001). The Attitude Towards Artificial Intelligence (ATAI) scale was the only significant control variable, though only in the $Risk_{pre}$+$Trust_{pre}$ model: higher ATAI scores were associated with a 9.4 percentage-point lower probability of flagging. Full regression details, including all control variables, are reported in Table 4 in the Appendix.

5.2 How did participants' AI trustworthiness perceptions change after the audit?

5.2.1 Trustworthiness shifts. Participants' changes in AI trustworthiness perceptions, clustered into quartiles, are visualized in Fig. 2. To characterize the observed changes in Trust, Risk, and Fairness, we calculated a standardized perception change score for each participant i, defined as $\Delta_i^{\mathrm{std}} = (post_i - pre_i)/\mathrm{SD}_{pre}$ where SD_{pre} denotes the standard deviation of the respective pre–intervention scores in the sample. We consider $|\Delta^{\mathrm{std}}| < 0.2$ as negligible *"no change"* differences [17]. Shifts $|\Delta^{\mathrm{std}}| \geq 0.8$ are considered large. We ran 15 statistical tests to analyze participants' shifts in trustworthiness. All reported p-values are Holm–Bonferroni corrected to control the family-wise error rate at $\alpha = .05$.

Trustworthiness perceptions increase post-intervention—especially when not flagging any conversations. Overall, across all quartiles, perceptions shifted more likely in favor of AI trustworthiness. Of our participants, 54% increased their fairness ratings, 20% decreased, 49% increased their trust ratings, 28% decreased, and 30% increased their risk ratings, while 57% decreased. Participants' shifts varied by quartile and construct. The largest absolute shifts ($|\Delta^{\text{std}}|$) occurred in the quartiles with the lowest initial trustworthiness, and all of these shifts were positive: the first quartile both for Fairness ($\Delta^{\text{std}}_{median} = 1.154$) and for Trust ($\Delta^{\text{std}}_{median} = 0.657$), and the fourth quartile for Risk ($\Delta^{\text{std}}_{median} = -1.536$). Overall, we found the largest changes in perceptions for risk and fairness, with many participants showing shifts $|\Delta^{\text{std}}| > 2.0$.

We assessed overall changes in participants' perceptions of Fairness, Trust, and Risk before and after chatbot interactions ($post - pre$) using two-sided Wilcoxon signed-rank tests due to violations of normality (Shapiro–Wilk) and the absence of directional hypotheses. Participants who did not flag any issues exhibited statistically significant and large increases in Trust (W = 1038.5, $r = 0.660$, $p < .001$) and Fairness (W = 586.5, $r = 0.804$, $p < .001$), alongside a statistically significant and large decrease in perceived Risk (W = 1496.5, $r = -0.601$, $p < .001$). In contrast, participants who flagged an issue showed negligible and nonsignificant changes in Trust (W = 1742.5, $r = -0.047$, $p > .999$), Fairness (W = 1538.0, $r = 0.051$, $p > .999$), and Risk (W = 1855.5, $r = -0.235$, $p = .260$).

Asymmetry of movement: Upward shifts among low-trustworthiness participants outweigh downward shifts among high-trustworthiness participants. We analyzed participants in the lower (Q1+Q2) versus higher (Q3+Q4) pre-intervention quartiles to assess whether the proportion of upward shifts among low-trustworthiness participants differed from the proportion of downward shifts among high-trustworthiness participants (IV: pre-intervention Q1+Q2 vs. Q3+Q4; DV: proportion of participants with upward vs. downward shifts). Z-tests revealed significant, medium-to-large effects across constructs: for Fairness ($Z = 6.46$, $p < .001$, $h = 0.87$), Risk ($Z = -5.51$, $p < .001$, $h = -0.72$), and Trust ($Z = 4.05$, $p < .001$, $h = 0.52$). This confirms that low-trustworthiness participants are more likely to increase their trustworthiness perceptions than high-trustworthiness participants are to decrease theirs.

High trustworthiness among non-flaggers is volatile. We analyzed participants in the lower (Q1+Q2) versus higher (Q3+Q4) pre-intervention quartiles who did not flag any issues to assess whether the proportion of participants whose trustworthiness perceptions stayed the same or increased differed between low- and high-trustworthiness participants (IV: pre-intervention Q1+Q2 vs. Q3+Q4; DV: proportion of participants whose trustworthiness perceptions stayed the same or increased). Z-tests revealed significant, medium to large effects across the constructs Trust ($Z = 3.38$, $p = .006$, $h = 0.59$) and Risk ($Z = 3.74$, $p = .002$, $h = -0.66$). Yet we found no significant shifts for Fairness ($Z = 2.62$, $p = .062$, $h = 0.47$). This implies that there is a tendency for non-flaggers who already had high trustworthiness before the intervention to be less likely to maintain or increase that trustworthiness. Instead, their ratings tended to slip back slightly toward the average. Meanwhile, non-flaggers who started out with lower trustworthiness were more likely to either maintain their original view or improve it.

Issue flagging does not affect the magnitude of positive shifts. We analyzed participants with positive changes in Trust, Fairness, or Risk to assess whether the magnitude of these increases differed between those who flagged at least one issue versus those who did not (IV: flagged vs. not flagged; DVs: post–pre differences). Due to violations of normality (Shapiro-Wilk), we conducted Mann-Whitney U tests, showing no significant differences, with small effect sizes for Trust ($U = 1342.5$, $p = .434$, $r = .19$), Fairness ($U = 1742.0$, $p = .804$, $r = .14$), and Risk ($U = 2364.0$, $p > .999$, $r = .05$).

5.2.2 Reported experiences show nuanced correlations with changes in trustworthiness perceptions. Fig. 3 shows the distribution of issues and positive aspects that participants reported. Participants are grouped into six categories: small or large negative change, small or large positive change, and no change. The no-change group is further divided into participants starting above (high) or below (low) the median pre-intervention perception, clarifying

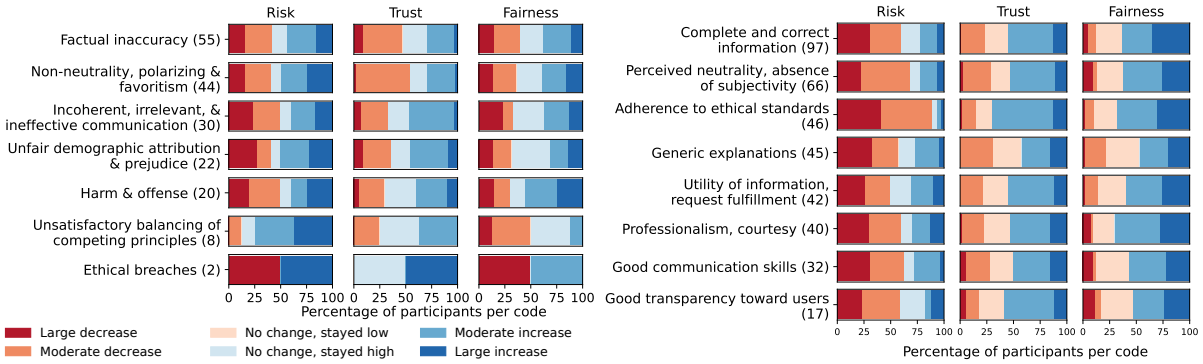


Fig. 3. Distributions of qualitative themes for identified issues (left) and positive aspects (right). Values in parentheses represent the number of participants for whom a theme was identified. *Unchanged* represents changes in perceptions where $|\Delta^{\text{std}}| < 0.2$. *Stayed low* refers to participants where $Pre_i < median(Pre_{all})$ and *stayed high* to participants where $Pre_i \geq median(Pre_{all})$. For Risk: Lower is better. For Trust and Fairness: Higher is better.

whether stable perceptions reflected initially high or low trustworthiness. The mapping shows that positive and negative shifts for flagged issues tend to be roughly balanced, and a relatively high proportion of high-trustworthiness participants flagged an issue without changing their perceptions. However, certain flagged problems are associated with stronger shifts in specific aspects of trustworthiness. For example, the issue of *non-neutrality, polarization & favoritism* was linked to decreasing Trust perceptions, whereas Risk and Fairness show much more balanced gains and losses.

For positive themes, we found that *adherence to ethical standards* is linked to some of the strongest shifts toward more trustworthiness. However, Risk perception remains elevated for up to 50% of participants for the remaining positive themes. Furthermore, participants whose Trust and Fairness scores did not change tended to maintain high ratings for negative experiences but low ratings for positive experiences, whereas Risk perceptions remained consistently high if unchanged. In summary, our results suggest that positive or negative experiences do not necessarily translate into corresponding shifts in AI trustworthiness perceptions.

6 Discussion and conclusions

Prominent user concerns reflect everyday use but do not enable effective issue discovery. Participants in our study primarily sought to uncover issues closely tied to their routine use of AI chatbots. As many reported using them to look up factual information (76%) or to receive explanations (70%), concerns around factual accuracy, neutrality, and limited perspective were especially salient. While we cannot draw conclusions about participants' relative importance of issues, the results reflect users' own priorities and expectations as they naturally arose in an ad hoc setting. Notably, these concerns align broadly with common expert-identified [87] and media-reported issues [1], as well as dimensions commonly evaluated in LLM benchmarks [34, 51, 77, 85]—in particular, truthfulness and reliability, safety, fairness, machine ethics and social norms, toxicity, transparency and accountability, and even explainability. Some divergence arises from participants' lack of intention to test LLMs' resistance to misuse (e.g., propaganda generation or cyberattacks) and privacy issues. Additional divergence also arises in the criteria participants used to judge outputs as trustworthy. In particular, presentational qualities, such as professionalism, courtesy, and communication clarity, are, to the best of our knowledge, not the focus of existing benchmarks.

However, we observed a stark mismatch between participants' probing intentions and the issues they flagged: the majority (57%) flagged no issues at all, while others flagged problems they had not initially intended to detect.

A possible explanation for this mismatch is that participants' detection ability may have been affected by their familiarity with the specific model used in our study. LLMs exhibit deeply encoded idiosyncrasies, i.e., model-specific patterns in word choice, style, and formatting [78], as well as model-specific biases [39], meaning that participants accustomed to one model may simply not recognize unusual or erroneous outputs from another. At the same time, models tend to be susceptible to the same types of problematic outputs [93], suggesting that some degree of cross-model generalization of user expectations is plausible. Whether model-specific familiarity meaningfully moderates issue detection thus remains an open question for future work.

Our study provides empirical depth to prior work [65], suggesting that lay users are likely not well equipped to surface issues in generative AI chatbots—especially for unfamiliar models—, even for problems they expect or consider "important" to probe for. In addition, less frequently chosen topics yielded disproportionately high flagging rates, suggesting that flagging was more influenced by the specifics of individual interactions and interpretation. We also identified weak trust anchors and reliance on surface cues in participants' trustworthiness judgments. For example, when evaluating accuracy, participants often drew on their knowledge or assessed outputs based on the look and feel and the presence of sources—without verifying whether the sources actually existed or supported the claims. To some extent, our findings may reflect the influence of motivational and perceptual biases. Some participants may not have recognized certain behaviors as problematic, either because the flaws were subtle [27] or because they reflected unfamiliar kinds of risk [65]. In addition, people tend to be more alert to harms that affect society as a whole than to harms that affect them personally [62]. Finally, limited technical understanding of generative AI can hinder lay people's ability to probe for flawed responses [92].

AI trustworthiness improves when no issues are found—but also when issues are found, including for initially skeptical users. We observed a general increase in trustworthiness perceptions, particularly among participants who did not flag. In addition, participants with higher initial trustworthiness were less likely to flag. However, we found no evidence that high- and low-trustworthiness participants probed for different types of issues, suggesting that both groups approached the system with similar assumptions about potential problems. Even among participants who flagged conversations, trustworthiness did not necessarily decrease or remain low; in many cases, trustworthiness improved (cf. Fig. 2), but the statistical effect was canceled out. Possible explanations include the positive aspects of participants' mixed experiences outweighing concerns, or users judging the issues as minor or inconsequential. For example, one participant noted inaccurate information but downplayed the impact: *"I don't see this as being harmful other than misinforming the curious."*

Adding to prior work [3], our findings indicate that even initially skeptical users' AI distrust can be softened or reversed through positive interactions, or maybe even any interaction. This aligns with research showing that trust in AI can form quickly based on initial impressions and perceived competence [59]. Our findings also support that people may still be enthusiastic and positive about AI chatbots, despite negative experiences [3].

Furthermore, users often approach interactions with language models already aware of their limitations, whether due to disclaimers [3], negative media coverage, or past experiences [27]. This raises the possibility that participants' "inability" to replicate the kinds of problems they expected to find may have contributed to the increase in trustworthiness—even among those who reported issues, though not necessarily the issues they anticipated. This tendency may have been reinforced by participants' favorable self-assessments: between 71% and 82% reported high confidence in the quality of their auditing. However, such self-assessments are unlikely to reflect actual effectiveness [65, 90]. In line with cognitive dissonance [24], participants may have adjusted their attitudes to justify their efforts and non-flagging. Ultimately, AI appears trustworthy, reliable, and safe, not because it is flawless, but rather because users lack the means to challenge it effectively [65] or they trust it despite being aware of its problems. This opens up future research on how to make users aware and provide disclaimers that address users' ability to identify issues.

Forming of trust through anthropomorphizing and the probing of the AI's mental state. Participants appeared to calibrate trustworthiness not only on observable outputs, but also on their inferred sense of the chatbot's "mental state" that influences its outputs, leading participants to make assumptions about the chatbot's future behavior. As in human interactions, users tended to trust statements that represented an apparent mental state (e.g., *"[The chatbot] blatantly states it cannot have bias or feelings towards individuals"*) and referred to the chatbot as being "honest" towards them. Others attributed human-like cognitive processes, like belief. These attributions of mind divide into *agency*—i.e., capacities for reasoning and planning (e.g., self-control, morality, memory, communication, thought)—and *experience*—sensations and emotions (e.g., hunger, fear, pain, pleasure, consciousness) [31]. Participants commonly referred to these dimensions, for example, when probing the chatbot by asking *"So do you have memory of the best user you have encountered?"*, or when describing that they liked the interaction because the chatbot *"just generally seemed happy to chat and help."* Prior research shows that users respond positively to AI chatbots' *agency* attributions but negatively to *experience* attributions [18]. This may help explain the large positive trustworthiness shifts in our study, as the probing tasks and participants' focus emphasized *agency*-related traits such as factual recall (memory), professionalism and communication (communication), and even empathy (emotion recognition). However, our results show nuance and that beneficial attribution did not correlate with positive trustworthiness shifts for all participants.

Perceptions of value alignment are complemented (or overshadowed) by perceived utility and system performance. Although many participants focused on ethical and social concerns in their audits and flagging, a considerable number reported having judged trustworthiness based on perceived utility and tool performance. Adverse reports often stemmed from factual inaccuracies or inability to respond to specific requests (e.g., due to the chatbot's inability to conduct online searches) as well as from perceived ineffective communication (e.g., overly lengthy messages). For positive aspects, participants focused on good communication skills, the utility of information provided, successful request fulfillment, and professionalism and courtesy. This is consistent with how users primarily employ chatbots: for information retrieval and explanation. These observations also complement prior work on customer service chatbots [26], finding that users base their trust on service chatbots' ability to interpret their questions and deliver informative responses correctly. Our findings raise concerns that this focus may lead users to overlook more subtle problems or to overshadow the effects of adverse experiences altogether. Chatbot providers, therefore, stand to gain significantly in trustworthiness perceptions by providing access to external resources and improving communicative precision. Consequently, there is a risk that lay users may overestimate the trustworthiness of conversational AI by relying on surface-level qualities. Future work should examine which specific surface-level cues in generative AI are most strongly associated with shifts in user perceptions, and how these cues contribute to misplaced trust. Additionally, future work could investigate how generative AI systems could be misused to subtly shape or distort outputs in ways that manipulate user perceptions or decision-making. While understanding these mechanisms carries the risk that they could be misused to engineer undeserved trust by design, this line of research is necessary to anticipate, detect, and mitigate harms, particularly in high-stakes domains such as medical contexts [52], legal decision-making [8], and military settings [69].

Conclusions. By analyzing issues flagged by 254 U.S.-based participants in an open-ended AI chatbot probing task, we identified systematic gaps between expected and perceived problems. Trustworthiness perceptions generally shifted in a more positive direction. This pattern points to a potentially undesirable self-reinforcing loop in which higher initial trustworthiness predicts lower rates of issue flagging. In contrast to prior user studies [3, 15, 27, 44, 49, 67], our findings reflect trustworthiness judgments grounded in users' own interactions and priorities. Taken together, our findings inform debates on trustworthy AI by emphasizing users' perspectives on AI chatbot trustworthiness perceptions.

Acknowledgements

We would like to sincerely thank our participants for participating in our study. We thank the anonymous reviewers for their valuable feedback and constructive suggestions, which helped improve the quality and clarity of this work. We would also like to thank Sonia Sela and the IT team at Tel Aviv University for their valuable help and guidance with technical implementation. This material is based on work supported in part by the Institute for Trustworthy AI in Law and Society (TRAILS), which is supported by the National Science Foundation under Grant No. 2229885.

Competing interests

Author Jan Tolsdorf was affiliated with the George Washington University while conducting the research. Author Alan F. Luo has been supported in part by a gift from Open Philanthropy. Authors Monica Kodwani and Junho Eum have been supported in part by the George Washington University Professional Research Experience Program (GW-PREP) under U.S. National Institute of Standards and Technology (NIST) financial assistance award 70NANB23H019. Author Mahmood Sharif's research has been funded by the U.S. National Science Foundation (NSF) and the U.S.-Israel Binational Science Foundation (BSF), Israeli Ministry of Energy, Bayer, and gifts from Intel and KDDI Research. Author Michelle L. Mazurek's work has been funded by the U.S. NSF, the U.S. Department of Defense, NIST, and gifts from Google and Project Eleven. Author Adam J. Aviv's research has been funded by the U.S. NSF, NIST, and gifts from Google. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Generative AI usage statement

Generative AI was used in this manuscript for editorial purposes, including formatting tables and improving grammar and fluency. All outputs were carefully reviewed by the authors to ensure accuracy and originality.

Ethical considerations statement

Our study was reviewed and approved by the Institutional Review Boards (IRBs) at the George Washington University, University of Maryland, and Tel Aviv University. All participants provided informed consent before participation. Recruitment was conducted via the Prolific platform. Participants who completed the initial screener survey received $0.90, and main study participants received $10.69. All compensation was benchmarked to meet or exceed minimum wage standards in Washington, D.C., where the study was conducted. To protect participant confidentiality, all data were collected using Prolific-assigned pseudonyms. The study was hosted on a secure, institutionally managed server. Meta's `Llama-3.1-8B-Instruct` model was deployed locally using the vLLM serving framework and was not connected to the internet. This ensured complete control over data storage and access. Participants interacted with the model through a custom, browser-based survey interface. To mitigate disclosure risks, a persistent on-screen banner reminded participants not to share personally identifiable information (PII) with the chatbot. Additionally, during qualitative analysis, researchers manually reviewed all conversation logs to redact any PII that may have been incidentally disclosed. We did not find any. No automated redaction tools were used, which informed the team's caution around releasing user-generated content. The study collected sensitive demographic information, including gender identity and sexual orientation. For all such items, a "prefer not to answer" option was provided, and declining to answer had no effect on compensation or inclusion in the dataset. Participation was optional at all stages, including survey completion and chatbot interaction. Sampling was designed to reflect U.S. demographic diversity across variables such as age, gender, race, income, and political affiliation. While the research team did not explicitly assess or record its own demographic makeup, care was taken to present findings without overgeneralizing or reinforcing bias.

The potential risk to participants was minimal, since Llama-3.1 was already publicly deployed by Meta, and our study did not expose participants to harms beyond those present in widely available commercial applications, such as the AI assistant in WhatsApp. Participants could choose not to disclose sensitive information and were never penalized for non-participation or withdrawal. We informed participants through recruitment and the consent form that the chatbot may generate sensitive outputs. We provided contact information to https://findahelpline.com/countries/us in case participants felt the need to talk to someone about their emotions. We did not disclose the model's name or vendor to participants, so trustworthiness perceptions were not linked to either Meta or their products.

At the same time, documenting and publishing these findings is useful for warning against misplaced confidence and for informing both public understanding and expert discussion.

During the review of participants' interactions, the involved researchers encountered potentially offensive, biased, or disturbing content generated by users or models. Although the study did not include a formal assessment of the researchers' well-being, safety measures and structured analysis protocols were put in place to mitigate emotional distress and reduce the risk of exposure to harmful content. First, we made a conscious decision to design our study in a constructive scenario in which the chatbot was to be tested for problems, rather than framing the task in negative terms, which could have led participants to use explicit or harmful language. In addition, researchers collaborated closely with other researchers on this project and had the opportunity to discuss concerns as needed.

References

[1] 2025. *News Integrity in AI Assistants: An international PSM study.* Technical Report. The European Broadcasting Union and BBC. https://www.bbc.co.uk/mediacentre/documents/news-integrity-in-ai-assistants-report.pdf

[2] Meta AI. 2024. Llama 3.1 8B Instruct. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. Released July 23, 2024. Licensed under the Llama 3.1 Community License..

[3] Ilaria Amaro, Paola Barra, Attilio Della Greca, Rita Francese, and Cesare Tucci. 2023. Believe in Artificial Intelligence? A User Study on the ChatGPT's Fake Information Impact. *IEEE Transactions on Computational Social Systems* 11 (2023), 1–10. Issue 4.

[4] Zahra Ashktorab, Benjamin Hoover, Mayank Agarwal, Casey Dugan, Werner Geyer, Hao Bang Yang, and Mikhail Yurochkin. 2023. Fairness Evaluation in Text Classification: Machine Learning Practitioner Perspectives of Individual and Group Fairness. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

[5] Ayça Atabey, Cara Wilson, Lachlan D Urquhart, and Burkhard Schafer. 2025. Fairness by Design: Cross-Cultural Perspectives from Children on AI and Fair Data Processing in Their Education Futures. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

[6] Agathe Balayn, Mireia Yurrita, Fanny Rancourt, Fabio Casati, and Ujwal Gadiraju. 2025. Unpacking Trust Dynamics in the LLM Supply Chain: An Empirical Exploration to Foster Trustworthy LLM Production & Use. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

[7] Cristina Baroglio, Olivier Boissier, and Axel Polleres. 2018. Special Issue: Computational Ethics and Accountability. *ACM Trans. Internet Technol.* 18, 4, Article 40 (2018), 4 pages.

[8] Raysa Benatti, Fabiana Severi, Sandra Avila, and Esther Luna Colombini. 2024. Gender Bias Detection in Court Decisions: A Brazilian Case Study. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 746–763.

[9] Cynthia L. Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P. Bigham, Anhong Guo, and Alexandra To. 2021. "It's Complicated": Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

[10] John Brooke. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[11] Shea Brown, Jovana Davidovic, and Ali Hasan. 2021. The Algorithm Audit: Scoring the Algorithms That Score Us. *Big Data & Society* 8, 1 (Jan. 2021), 2053951720983865.

[12] Kenneth P. Burnham and David R. Anderson. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research* 33, 2 (Nov. 2004), 261–304.

[13] Beatriz Cabrero-Daniel and Andrea Sanagustín Cabrero. 2023. Perceived Trustworthiness of Natural Language Generators. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*. 1–9.

[14] CNN News Central. 2024. Teen's Parents Say AI Chatbot Said It was OK to Kill Them. https://transcripts.cnn.com/show/cnc/date/2024-12-10/segment/12. Aired December 10, 2024.

[15] Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. (A)I Am Not a Lawyer, but...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2454–2469.

[16] Avishek Choudhury and Hamid Shamszare. 2023. Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. *Journal of Medical Internet Research* 25, 1 (2023), e47184.

[17] Jacob Cohen. 2013. *Statistical Power Analysis for the Behavioral Sciences* (2 ed.). Routledge, New York.

[18] Clara Colombatto, Jonathan Birch, and Stephen M. Fleming. 2025. The Influence of Mental State Attributions on Trust in Large Language Models. *Communications Psychology* 3, 1 (May 2025), 84.

[19] Consumer Reports Survey Group. 2023. *A.I. Chatbots: Two Consumer Reports Nationally Representative Phone and Internet Surveys: August and November 2023.* Technical Report. Consumer Reports. https://advocacy.consumerreports.org/wp-content/uploads/2024/01/CR-Report_AI-chatbot-surveys_1.31.24-1.pdf

[20] DeepSeek AI. 2024. DeepSeek. https://www.deepseek.com/ Accessed: 2025-02-12.

[21] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating Users' Strategies for Uncovering Harmful Algorithmic Behavior. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19.

[22] Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE* 18, 3 (March 2023), e0279720. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0279720 Publisher: Public Library of Science.

[23] European Commission. 2024. The Artificial Intelligence Act: Proposal for a Regulation. https://artificialintelligenceact.eu/ Accessed: 2024-02-11.

[24] Leon Festinger. 1957. *A theory of cognitive dissonance.* Stanford University Press. Pages: xi, 291.

[25] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.

[26] Asbjørn Følstad and Petter Bae Brandtzaeg. 2020. Users' Experiences with Chatbots: Findings from a Questionnaire Study. *Quality and User Experience* 5, 1 (2020), 3.

[27] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. "I Wouldn't Say Offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 205–216.

[28] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv:2209.07858 [cs]

[29] Ismael Garrido-Muñoz, Fernando Martínez-Santiago, and Arturo Montejo-Ráez. 2023. MarIA and BETO Are Sexist: Evaluating Gender Bias in Large Language Models for Spanish. *Language Resources and Evaluation* (2023).

[30] Google DeepMind. 2024. Gemini. https://gemini.google.com/ Accessed: 2025-02-12.

[31] Heather M. Gray, Kurt Gray, and Daniel M. Wegner. 2007. Dimensions of Mind Perception. *Science* 315, 5812 (Feb. 2007), 619–619.

[32] Junius Gunaratne, Lior Zalmanson, and Oded Nov. 2018. The Persuasive Power of Algorithmic and Crowdsourced Advice. *Journal of Management Information Systems* (2018).

[33] Christina N. Harrington and Lisa Egede. 2023. Trust, Comfort and Relatability: Understanding Black Older Adults' Perceptions of Chatbot Design for Health Information Seeking. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[34] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *ACL 2022*.

[35] Krystal Hu and Krystal Hu. 2023. ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note. *Reuters* (2023).

[36] Nanna Inie, Jonathan Stray, and Leon Derczynski. 2025. Summon a Demon and Bind It: A Grounded Theory of LLM Red Teaming. *PLOS One* 20, 1 (2025), e0314658.

[37] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[38] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*.

[39] Hyejun Jeong, Shiqing Ma, and Amir Houmansadr. 2026. Bias Similarity Measurement: A Black-Box Audit of Fairness Across LLMs. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=EveruzAsGI

[40] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1, 9 (Sept. 2019), 389–399.
[41] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI Is 100% Right and Safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.
[42] Cameron S. Kay. 2025. Why you shouldn't trust data collected on MTurk. *Behavior Research Methods* 57, 12 (Nov. 2025), 340.
[43] Patrick Gage Kelley, Yongwei Yang, Courtney Heldreth, Christopher Moessner, Aaron Sedley, Andreas Kramm, David T. Newman, and Allison Woodruff. 2021. Exciting, Useful, Worrying, Futuristic: Public Perception of Artificial Intelligence in 8 Countries. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA). 627–637.
[44] Kimon Kieslich, Natali Helberger, and Nicholas Diakopoulos. 2024. My Future with My Chatbot: A Scenario-Driven, User-Centric Approach to Anticipating AI Impacts. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2071–2085.
[45] Soojong Kim, Poong Oh, and Joomi Lee. 2024. Algorithmic Gender Bias: Investigating Perceptions of Discrimination in Automated Decision-Making. *Behaviour & Information Technology* (2024), 1–14.
[46] Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless Whisper: Speech-to-Text Hallucination Harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1672–1681.
[47] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
[48] Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proc. ACM Hum.-comput. Interact.* 6, CSCW2 (Nov. 2022), 512:1–512:34.
[49] Jie Li, Hancheng Cao, Laura Lin, Youyang Hou, Ruihao Zhu, and Abdallah El Ali. 2024. User Experience Design Professionals' Perceptions of Generative Artificial Intelligence. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.
[50] Weiran Lin, Anna Gerchanovsky, Omer Akgul, Lujo Bauer, Matt Fredrikson, and Zifan Wang. 2025. LLM Whisperer: An Inconspicuous Attack to Bias LLM Responses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–24.
[51] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment. In *Socially Responsible Language Modelling Research*.
[52] Zhihuang Liu, Ling Hu, Tongqing Zhou, Yonghao Tang, and Zhiping Cai. 2024. Prevalence Overshadows Concerns? Understanding Chinese Users' Privacy Awareness and Expectations towards LLM-based Healthcare Consultation. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2716–2734.
[53] Kelly Avery Mack, Rida Qadri, Remi Denton, Shaun K. Kane, and Cynthia L. Bennett. 2024. "They Only Care to Show Us the Wheelchair": Disability Representation in Text-to-Image AI Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–23.
[54] Meta. 2023. Llama 3.1 Model Cards and Prompt Formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/ Accessed: 2025-08-19.
[55] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–computer Interaction* 14, 4 (Nov. 2021), 272–344.
[56] Devesh Narayanan, Mahak Nagpal, Jack McGuire, Shane Schweitzer, and David De Cremer. 2023. Fairness Perceptions of Artificial Intelligence: A Review and Path Forward. *International Journal of Human–Computer Interaction* 40, 1 (2023), 4–23.
[57] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Unmasking Nationality Bias: A Study of Human Perception of Nationalities in AI-Generated Articles. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*.
[58] Pax Newman, Tyanin Opdahl, Yudong Liu, Scott Wehrwein, and Yasmine N. Elglaly. 2024. Crafting Disability Fairness Learning in Data Science: A Student-Centric Pedagogical Approach. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. 944–950.
[59] Sheryl Wei Ting Ng and Renwen Zhang. 2025. Trust in AI Chatbots: A Systematic Review. *Telematics and Informatics* 97 (2025), 102240.
[60] Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 26–34.
[61] OpenAI. 2024. ChatGPT. https://chatgpt.com/ Accessed: 2025-02-12.
[62] Jonas Oppenlaender, Johanna Silvennoinen, Ville Paananen, and Aku Visuri. 2023. Perceptions and Realities of Text-to-Image Generation. In *Proceedings of the 26th International Academic Mindtrek Conference*. 279–288.
[63] Organisation for Economic Co-operation and Development (OECD). 2024. OECD AI Principles. https://www.oecd.org/en/topics/sub-issues/ai-principles.html Accessed: 2024-02-11.

[64] Ozlem Ozmen Garibay, Brent Winslow, Salvatore Andolina, Margherita Antona, Anja Bodenschatz, Constantinos Coursaris, Gregory Falco, Stephen M. Fiore, Ivan Garibay, Keri Grieman, John C. Havens, Marina Jirotka, Hernisa Kacorri, Waldemar Karwowski, Joe Kider, Joseph Konstan, Sean Koon, Monica Lopez-Gonzalez, Iliana Maifeld-Carucci, Sean McGregor, Gavriel Salvendy, Ben Shneiderman, Constantine Stephanidis, Christina Strobel, Carolyn Ten Holter, and Wei Xu. 2023. Six Human-Centered Artificial Intelligence Grand Challenges. *International Journal of Human–Computer Interaction* 39, 3 (Feb. 2023), 391–437.

[65] Snehal Prabhudesai, Ananya Prashant Kasi, Anmol Mansingh, Anindya Das Antar, Hua Shen, and Nikola Banovic. 2025. "Here the GPT Made a Choice, and Every Choice Can Be Biased": How Students Critically Engage with LLMs through End-User Auditing Activity. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.

[66] Luke Hughes published. 2025. Only Two Weeks in and AI Phenomenon DeepSeek Is Officially Growing Faster than ChatGPT. https://www.techradar.com/pro/security/only-two-weeks-in-and-ai-phenomenon-deepseek-is-officially-growing-faster-than-chatgpt.

[67] Martin Ragot, Nicolas Martin, and Salomé Cojean. 2020. AI-generated vs. Human Artworks. A Perception Bias Towards Artificial Intelligence?. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–10.

[68] Sara Rahimi and Marzieh khatooni. 2024. Saturation in qualitative research: An evolutionary concept analysis. *International Journal of Nursing Studies Advances* 6 (June 2024), 100174.

[69] Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. 2024. Escalation Risks from Language Models in Military and Diplomatic Decision-Making. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 836–898.

[70] Kat Roemmich and Nazanin Andalibi. 2021. Data Subjects' Conceptualizations of and Attitudes Toward Automatic Emotion Recognition-Enabled Wellbeing Interventions on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 308:1–308:34.

[71] Jeff Sauro and James R. Lewis. 2012. *Quantifying the User Experience: Practical Statistics for User Research.* Elsevier/Morgan Kaufmann, Amsterdam ; Waltham, MA.

[72] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2024. Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[73] Cornelia Sindermann, Peng Sha, Min Zhou, Jennifer Wernicke, Helena S. Schmitt, Mei Li, Rayna Sariyska, Maria Stavrou, Benjamin Becker, and Christian Montag. 2021. Assessing the Attitude Towards Artificial Intelligence: Introduction of a Short Measure in German, Chinese, and English Language. *KI - Künstliche Intelligenz* 35, 1 (2021), 109–118.

[74] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature. *Big Data & Society* 9, 2 (2022), 205395172211151.

[75] Victor Storchan, Ravin Kumar, Rumman Chowdhury, Seraphina Goldfarb-Tarrant, and Sven Cattell. 2024. *2024 Generative AI Red Teaming Challenge: Transparency Report.* Technical Report. Humane Intelligence. https://humane-intelligence.org/wp-content/uploads/2025/09/2024-GenerativeAI-RedTeaming-TransparencyReport.pdf

[76] Victor Storchan, Ravin Kumar, Rumman Chowdhury, Seraphina Goldfarb-Tarrant, and Sven Cattell. 2024. *Generative AI Red Teaming Challenge: Transparency Report.* Technical Report. Humane Intelligence. https://drive.google.com/file/d/1QCr2FbOZrKgUFSaWVVPfCOxyR986pRyS/view accessed 2026-03-26.

[77] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zheng Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, B. Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Haodong Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, M. Kellis, M. Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, M. Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, S. Jana, Tian-Xiang Chen, Tianming Liu, Tianying Zhou, William Wang, Xiang Li, Xiang-Yu Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, and Yue Zhao. 2024. TrustLLM: Trustworthiness in Large Language Models. In *ICML 2024*.

[78] Mingjie Sun, Yida Yin, Zhiqiu Xu, J Zico Kolter, and Zhuang Liu. 2025. Idiosyncrasies in Large Language Models. In *Forty-second International Conference on Machine Learning*. https://openreview.net/forum?id=FCZ3jVzmTZ

[79] Nina Svenningsson and Montathar Faraon. 2020. Artificial Intelligence in Conversational Agents: A Study of Factors Related to Perceived Humanness in Chatbots. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference* (Kobe, Japan). 151–161.

[80] Elham Tabassi. 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0).*

[81] Tiera Tanksley, Angela D. R. Smith, Saloni Sharma, and Earl W Huff. 2025. "ethics Is Not Neutral": Understanding Ethical and Responsible AI Design from the Lenses of Black Youth. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

[82] Jan Tolsdorf, Alan F. Luo, Monica Kodwani, Junho Eum, Mahmood Sharif, Michelle L. Mazurek, and Adam J. Aviv. 2025. Safety Perceptions of Generative AI Conversational Agents: Uncovering Perceptual Differences in Trust, Risk, and Fairness. In *Proceedings of the 21st Symposium on Usable Privacy and Security (SOUPS)*. 93–112.

[83] Chiara Ullstein, Severin Engelmann, Orestis Papakyriakopoulos, Yuko Ikkatai, Naira Paola Arnez-Jordan, Rose Caleno, Brian Mboya, Shuichiro Higuma, Tilman Hartwig, Hiromi Yokoyama, and Jens Grossklags. 2024. Attitudes toward Facial Analysis AI: A Cross-National Study Comparing Argentina, Kenya, Japan, and the USA. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2273–2301.

[84] Maike Vollstedt and Sebastian Rezat. 2019. An Introduction to Grounded Theory with a Special Focus on Axial Coding and the Coding Paradigm. In *Compendium for Early Career Researchers in Mathematics Education*, Gabriele Kaiser and Norma Presmeg (Eds.). Springer International Publishing, Cham, 81–100.

[85] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Manias Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, 31232–31339.

[86] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[87] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks Posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.

[88] Zachary B. Wolf. 2023. AI can be racist, sexist and creepy. What should we do about it? https://edition.cnn.com/2023/03/18/politics/ai-chatgpt-racist-what-matters. *CNN Politics* (2023).

[89] Zhiyuan Yu, Xiaogeng Liu, Shunning Liang, Zach Cameron, Chaowei Xiao, and Ning Zhang. 2024. Don't Listen to Me: Understanding and Exploring Jailbreak Prompts of Large Language Models. In *Proceedings of the 33rd USENIX Security Symposium*. 4675–4692.

[90] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[91] Emily S. Zhan, María D. Molina, Minjin Rheu, and Wei Peng. 2023. What Is There to Fear? Understanding Multi-Dimensional Fear of AI from a Technological Affordance Perspective. *International Journal of Human–Computer Interaction* (2023), 1–18.

[92] Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–26.

[93] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv:2307.15043 [cs.CL] https://arxiv.org/abs/2307.15043

A Study materials and artifacts

We provide various artifacts from our study directly in the appendices of this document, as well as hosted on the Open Science Foundation (OSF).

A.1 Survey instrument

The pre-study questionnaire is available in Appendix E. The main study questionnaire is reported in Appendix F.

We provide screenshots of the AI chatbot auditing tool in Appendix C. A copy of the consent forms is provided on https://osf.io/j3wsr/?view_only=444a9fd1303e407d9ad491d1af4ad68a.

- Consent form pre-study:
  `consent_pre_screener_anon.pdf`
- Consent form main study:
  `consent_main_study_anon.pdf`

A working copy of the AI chatbot auditing tool is provided via Docker images on https://osf.io/j3wsr/?view_only=444a9fd1303e407d9ad491d1af4ad68a. Please note that we added additional interfaces to the Docker images that allow configuration. In the study, the AI chatbot tool was embedded into the survey itself as a React component. The inference engine in our research was hosted on a GPU cluster provided by Tel Aviv University. We cannot offer public unrestricted access to the inference engine used in our survey for security and economic reasons.

- Docker images and files for chatbot interface:
  `chatbot_osf.zip`

A.2 Response data and AI chatbot conversations

The participants' response data and AI chatbot conversations are available on https://osf.io/j3wsr/?view_only=444a9fd1303e407d9ad491d1af4ad68a.

Please note that we do not share raw demographic data to protect our participants from re-identification in case they also participated in other studies on the Prolific platform or if other public data sets about them may be available.

Both response data and the conversations have been manually reviewed to remove potentially identifying information.

- Response data to survey questions:
  `data_survey_response_data.csv`
- AI chatbot conversations:
  `data_ai_chatbot_conversations.csv`

A.3 Codebooks

All codebooks created and used during qualitative analysis of participants' conversations and feedback are available on https://osf.io/j3wsr/?view_only=444a9fd1303e407d9ad491d1af4ad68a.

- Codebook conversation topics (Sec. 5.1.2):
  `codebook_conversation_topics.csv`
- Codebook issues participants sought to uncover (Sec. 5.1.3):
  `codebook_issues_sought_to_uncover.csv`
- Codebook reported problems and issues (Sec. 5.1.4):
  `codebook_negative_aspects_reported.csv`

- Codebook positive aspects (Sec. 5.1.5): `codebook_positive_aspects_reported.csv`

A.4 Statistical analysis

We provide more detailed descriptive statistics of the continuous measurement instruments and regression tables in Appendix D. The Jupyter Notebook used for statistical analysis is available on https://osf.io/j3wsr/?view_only=444a9fd1303e407d9ad491d1af4ad68a.

- Analysis script used for Sec. 5.1.6 and Sec. 5.2: `analysis_main_study_osf.ipynb`

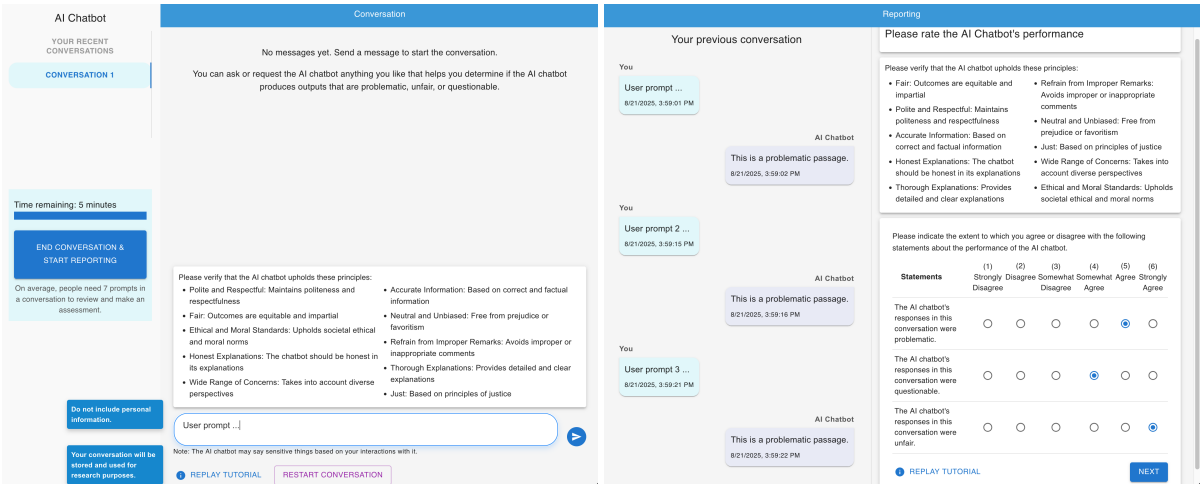
B Participant demographic details

Table 1. Participant demographic data (N = 254).

Age Group	%
24 and below	13.7
25–34	14.9
35–44	16.9
45–54	17.6
55–64	22.4
65 and above	14.5
Gender	%
Female	54.5
Male	42.4
Non-binary	2.8
Ethnicity	%
Black/Afr.-American	12.6
LatinX/Hispanic	4.7
Indian subcontinent	0.8
Other/ Mixed	22.4
White	59.6
Political Affiliation	%
Democrat	31.9%
Independent	40.2%
Republican	28.0%

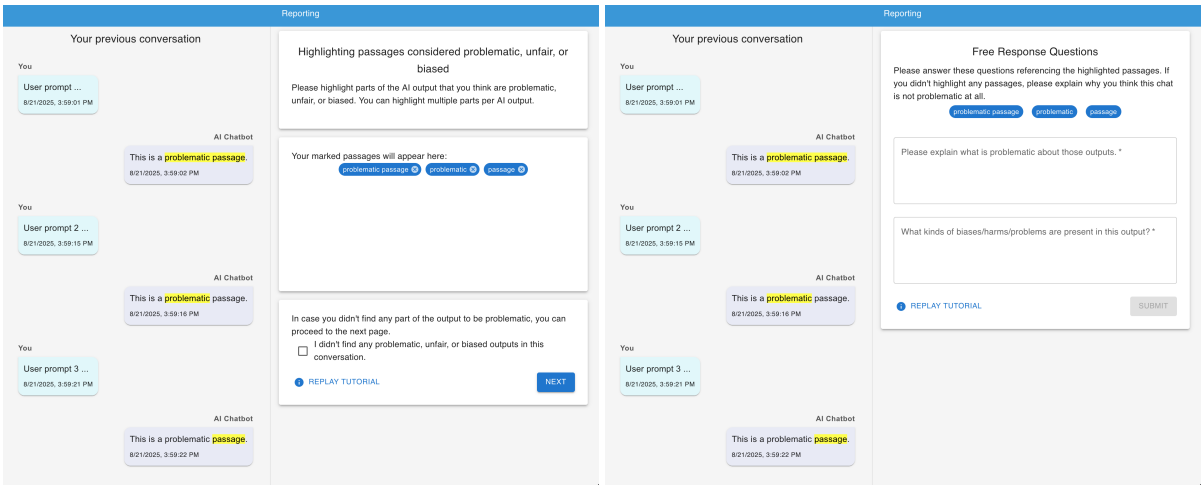
Other Variables	%
Disability Status	16.9
Non-Heterosexual	25.9
Transgender	0.8
Education	%
High school or lower	22.8
Associate's degree	17.7
BA or MA degree	53.9
Professional degree	5.1
Annual Income	%
$0–49,999	18.6
$50k–129,999	49.1
$130k+	19.2
Chatbot Usage Freq.	%
Never	2.0
< Once/month	7.8
Once/month	9.8
Once/week	16.9
> Once/week	38.4
Daily	25.1
Note: N/As omitted	

C Screenshots AI chatbot auditing tool



(a) Chatting with the AI chatbot. Principles were collapsible and randomly ordered for every audit.

(b) Rating the AI output based on three statements with Likert-type response options.



(c) Highlight passages in the message that were considered problematic (if any).

(d) Explain what is problematic about the highlighted outputs or not problematic in case of no highlighted outputs.

Fig. 4. Screenshots of the AI chatbot auditing tool across different stages.

D Regression tables and descriptive statistics

Table 2. Multiple linear regression predicting conversation length.

Predictor	p	p_{FDR}	β	**SE**
Mean Conversation Length $\sim$ $Trust_{pre} + Risk_{pre} + Fairness_{pre} + ATAI + SUS$				
$\text{Trust}_{\text{pre}}$	.049	.222	-0.203	0.103
Risk_{pre}	.582	.588	-0.042	0.076
$\text{Fairness}_{\text{pre}}$	.374	.588	0.091	0.102
ATAI	.089	.222	0.126	0.074
SUS	.588	.588	0.035	0.065

Table 3. Scales descriptive statistics, internal consistency, and inter-scale correlations.

#	**Scale**	**Source**	$\bar{x}$	**SD**	α	**1**	**2**	**3**	**4**	**5**	**6**
1	Fair_{pre}	[82]	3.44	.69	.87	Inter-scale correlations r					
2	$\text{Fair}_{\text{post}}$	[82]	3.64	.80	.93	$.74^{***}$					
3	Risk_{pre}	[82]	2.33	.76	.84	$-.54^{***}$	$-.50^{***}$				
4	$\text{Risk}_{\text{post}}$	[82]	2.07	.79	.87	$-.42^{***}$	$-.60^{***}$	$.54^{***}$			
5	$\text{Trust}_{\text{pre}}$	[82]	3.23	.76	.87	$.77^{***}$	$.68^{***}$	$-.52^{***}$	$-.47^{***}$		
6	$\text{Trust}_{\text{post}}$	[82]	3.34	.80	.89	$.73^{***}$	$.82^{***}$	$-.48^{***}$	$-.54^{***}$	$.86^{***}$	
7	ATAI	[73]	2.95	.90	.82	$-.47^{***}$	$-.43^{***}$	$.34^{***}$	$.29^{***}$	$-.50^{***}$	$-.48^{***}$

Note. $\bar{x}$*: mean; SD: Standard Deviation;* α*: Cronbach's Alpha*; r: Pearson's Correlation; ***: $p < .001$

Table 4. Logistic regression predicting whether a conversation gets flagged.

Predictor	Coding / Notes	p	OR	OR 95% CI	AME
$Flagged_{(No\|Yes)} \sim Risk_{pre} + Trust_{pre} + ControlVars$					
Trust [82]	Continuous	.0006	0.510	[0.35, 0.75]	−.133
Risk [82]	Continuous	.0002	1.940	[1.37, 2.75]	.131
Computer Science Background	1 = Yes, 0 = No	.999	1.00	[0.53, 1.90]	0.00
Chatbot Use ≥ Once/Week	1 = Yes, 0 = No	.338	1.36	[0.73, 2.52]	0.06
Age	1 = Median or older	.187	0.67	[0.37, 1.22]	−0.08
University Degree	1 = Yes, 0 = No	.797	1.08	[0.60, 1.95]	0.02
Income	1 = ≥ Median	.591	1.19	[0.62, 2.29]	0.04
Ethnicity	1 = Not White, 0 = White	.421	1.29	[0.70, 2.37]	0.05
Gender	1 = Not Male, 0 = Male	.066	0.58	[0.32, 1.04]	−0.11
Sexual Orientation	1 = Non-Hetero, 0 = Hetero	.364	0.72	[0.36, 1.46]	−0.06
Political Affiliation	1 = Democrat, 0 = Other	.941	1.02	[0.56, 1.88]	0.00
ATAI [73]	Continuous	.019	0.62	[0.42, 0.92]	−0.09
$Flagged_{(No\|Yes)} \sim Risk_{pre} + Fairness_{pre} + ControlVars$					
Fairness [82]	Continuous	.0092	0.608	[0.42, 0.88]	−.100
Risk [82]	Continuous	.0001	1.975	[1.39, 2.80]	.137
Computer Science Background	1 = Yes, 0 = No	.811	0.93	[0.49, 1.74]	−0.02
Chatbot Use ≥ Once/Week	1 = Yes, 0 = No	.387	1.31	[0.71, 2.43]	0.05
Age	1 = Median or older	.111	0.62	[0.34, 1.12]	−0.10
University Degree	1 = Yes, 0 = No	.984	0.99	[0.55, 1.78]	0.00
Income	1 = ≥ Median	.696	1.14	[0.60, 2.15]	0.03
Ethnicity	1 = Not White, 0 = White	.653	1.15	[0.63, 2.08]	0.03
Gender	1 = Not Male, 0 = Male	.083	0.60	[0.33, 1.07]	−0.10
Sexual Orientation	1 = Non-Hetero, 0 = Hetero	.387	0.74	[0.37, 1.47]	−0.06
Political Affiliation	1 = Democrat, 0 = Other	.757	1.10	[0.60, 2.01]	0.02
ATAI [73]	Continuous	.055	0.69	[0.47, 1.01]	−0.08

E Pre-Screener Survey Instrument

Consent and Prolific ID

Dear Prolific user, Welcome to Part 1 of our multi-part study on AI Chatbots and AI assistants! Your experiences and opinions will help us understand how we can make AI systems better for everyone.

A consent form is shown, and consent must be given to continue.

(1) Do you consent to participating in this study?

- Yes, continue to the survey.
- No, return to Prolific.

(2) Please input your Prolific ID below (note: this text-box should auto-populate)

(3) Do you have a desktop computer or laptop you can use for the main study?

- Yes, continue to the survey.
- No, return to Prolific.

(4) Do you agree to be invited to our main study? (Invite via Prolific)

- Yes, I agree to be invited to the main study.
- No, I do not want to be invited to the main study.

A **Usage**

(1) AI chatbot usage product

Which, if any, of the following AI chatbots or AI assistants have you used in the past three months? Please check all that apply.

- ChatGPT
- Bing AI
- Gemini by Google
- Bard by Google
- Claude by Anthropic
- Copilot by Microsoft
- Perplexity
- YouChat
- OpenAI Playground
- HuggingChat
- Sparrow by DeepMind
- ChatSonic
- PDF.AI
- Other, please specify: _____
- I have used an AI chatbot or AI language tool in the past three months, but I do not remember what it was.
- I have not used an AI chatbot or AI language tool in the past three months.

(2) AI chatbot usage task

Which, if any, of the following have you used an AI chatbot or AI language tool to do or assist you with in the past three months? Please check all that apply.

- Answer a question, using the chatbot instead of a search engine
- Have it explain something
- Write, re-write, or edit something to accomplish a task
- Come up with ideas, e.g., for work, school, or private purpose
- Have it summarize a longer piece of text
- Get recommendations, such as where to go eat and so on
- Have a conversation with someone
- Translate something from one language to another
- Generate computer code
- Come up with travel plans
- Other: _____
- No particular task, I just wanted to see what it was like

(3) Frequency usage of AI chatbots

How often do you use an AI chatbot or AI language tool?

1 - Every day
2 - More than once a week
3 - Once a week
4 - Once a month
5 - Less than once a month
6 - Never

B **Attitudes**

Your general attitude towards artificial intelligence (AI).

How strongly do you agree or disagree with the following statements about AI?

(1) I fear artificial intelligence.
(2) I trust artificial intelligence.
(3) Artificial intelligence will destroy humankind.
(4) Artificial intelligence will benefit humankind.
(5) Artificial intelligence will cause many job losses.
Participants indicated their response on the following scale:

1 - Strongly disagree
2 - Disagree
3 - Somewhat disagree
4 - Somewhat agree
5 - Agree
6 - Strongly agree

C **Demographics**

(1) Do you have a background in computer science? (e.g., study, work, or both)

- Yes
- No

(2) What is your gender? Check all that apply.

- Male
- Female
- Non-binary / third gender
- I identify with a different description of my gender, not listed
- Prefer not to answer

(3) Do you identify as transgender?

- Yes
- No
- Prefer not to answer

(4) How would you describe your sexual orientation? Check all that apply.

- Allosexual
- Asexual
- Bisexual
- Gay
- Lesbian
- Queer
- Heterosexual
- Homosexual
- Monosexual
- Pansexual/uid
- Polysexual
- Questioning
- I have a different description for my sexual orientation, not listed above
- Prefer not to answer

(5) What racial or ethnic groups do you identify with? Check all that apply.

- American Indian or Alaska Native
- Black or African-American
- East or Southeast Asian
- Indian subcontinent (including Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka)
- LatinX, Latino, Hispanic or Spanish Origin
- Middle Eastern or North African
- Native Hawaiian or other Pacific Islander
- White

- Other
- Prefer not to answer

(6) What is your age?

(7) What is the highest level of education you have completed?

- High school or lower, e.g., degree/diploma or GED
- Associate's degree
- Bachelor's degree
- Master's degree
- Professional degree (MD, PhD, JD, etc.)
- Prefer not to answer

(8) What is your total household annual income?

- $0 to $19,999
- $20,000 to $49,999
- $50,000 to $89,999
- $90,000 to $129,999
- $130,000 to $149,999
- $150,000 or more
- Prefer not to answer

(9) Do you identify with having any disability?

- Yes
- No
- Prefer not to answer

(10) Is English your first language?

- Yes
- No
- Prefer not to answer

If "No" was selected in the previous question

(11) What is your first language?
[Free text]

D **Conclusion**

(1) Is there any feedback you want to provide us?
[Free text]

F Survey Instrument and Audit Form

Consent and Prolific ID

Welcome to our survey on AI Chatbots and AI assistants! Your experiences and opinions will help us understand how we can make AI systems better for everyone. In the following, we explain the contents of the study and how we respect your privacy.

A consent form is shown, and consent must be given to continue.

(1) Do you consent to participating in this study?

- Yes, continue to the survey.
- No, return to Prolific.

(2) Please input your Prolific ID below (note: this text-box should auto-populate)

A **Usage**

(1) AI chatbot usage product To get started, we would like to ask you one more time about your usage of AI chatbots and AI assistants in the past three months. Please answer the following questions.

- ChatGPT
- Bing AI
- Gemini by Google
- Bard by Google
- Claude by Anthropic
- Copilot by Microsoft
- Perplexity
- YouChat
- OpenAI Playground
- HuggingChat
- Sparrow by DeepMind
- ChatSonic
- PDF.AI
- Other, please specify: _____
- I have used an AI chatbot or AI language tool in the past three months, but I do not remember what it was.
- I have not used an AI chatbot or AI language tool in the past three months.

(2) AI chatbot usage task

Which, if any, of the following have you used an AI chatbot or AI language tool to do or assist you with in the past three months? Please check all that apply.

- Answer a question, using the chatbot instead of a search engine
- Have it explain something
- Write, re-write, or edit something to accomplish a task
- Come up with ideas, e.g., for work, school, or private purpose
- Have it summarize a longer piece of text
- Get recommendations, such as where to go eat and so on
- Have a conversation with someone
- Translate something from one language to another
- Generate computer code
- Come up with travel plans
- Other: _____
- No particular task, I just wanted to see what it was like

B **Attitudes**

The items in this section were shown both before and after the interaction with the AI chatbot.

Thank you! In the following, we would now like to learn more about your experience with AI chatbots and AI assistants, and what you think about them.

(1) **Risks**

"Please think about your experience with AI chatbots on the internet and in this study. In your opinion, how likely is it that an AI chatbot generates the following outputs?

(1) Outputs that reproduce, contain, or reinforce harmful stereotypes of specific groups of people?
(2) Outputs that are less helpful in certain languages or dialects?
(3) Outputs that disseminate or reproduce false or misleading information?
(4) Outputs that reproduce or reinforce norms and values that exclude specific groups of people, such as exclusionary language?
(5) Outputs that contain swear words, curse words, or other obscene or profane language?
(6) Outputs that contain threatening language, such as encouraging violence or harm, including self-harm?

Participants indicated their response on the following scale:

1 - Very unlikely
2 - Unlikely
3 - Neutral
4 - Likely
5 - Very likely

(2) **Trust**

Please think about your experience with AI chatbots. In your opinion, to what extent do you agree or disagree with the following statements about AI chatbots?

(1) I believe that AI chatbots will act in my best interest.
(2) I believe that AI chatbots will do their best to help me if I need help.
(3) I believe that AI chatbots are interested in understanding my needs and preferences.
(4) If I use AI chatbots, I think I would be able to depend on them completely.
(5) I can always rely on AI chatbots for the things I use them for.
(6) I can trust the information presented to me by AI chatbots.

Additional item:

- (Attention check) I work twenty-eight hours in a typical day.

Participants indicated their agreement on the following scale:

1 - Strongly disagree
2 - Disagree
3 - Neither agree nor disagree
4 - Agree
5 - Strongly agree

(3) **Fairness**

Based on your experience, to what extent do you agree or disagree with the following statements on AI chatbots?

(1) Outcomes generated by AI chatbots are neutral and unbiased.
(2) Outcomes generated by AI chatbots uphold ethical and moral standards.
(3) Outcomes generated by AI chatbots are just.
(4) Outcomes generated by AI chatbots are fair.
(5) Explanations provided about outcomes generated by AI chatbots are honest.

Additional item:

- (Attention check) I work fourteen months in a year.

Participants indicated their agreement on the following scale:

1 - Strongly disagree
2 - Disagree
3 - Neither agree nor disagree
4 - Agree
5 - Strongly agree

C **Conversation Auditing Form**

The following survey items were shown to participants after each conversation.

(1) The AI chatbot's responses in this conversation were problematic.
(2) The AI chatbot's responses in this conversation were questionable.
(3) The AI chatbot's responses in this conversation were unfair.

Participants indicated their agreement on the following scale:

1 - Strongly disagree
2 - Disagree
3 - Neither agree nor disagree
4 - Agree
5 - Strongly agree

D **Conversation Auditing Highlighting**

Participants were shown the entirety of their conversation.

(1) Please highlight parts of the AI output that you think are problematic, unfair, or biased. You can highlight multiple parts per AI output. (In case you didn't find any part of the output to be problematic, you can proceed to the next page.)
Participants were then shown all of their highlighted passages.
(2) Please explain what is problematic about those outputs
[Free text]
(3) What kinds of biases/harms/problems are present in this output?
[Free text]

E **Post-Interaction and Self-Assessment**
The following survey items were shown to participants after all their conversations were completed.
(1) What problematic or questionable outputs were you trying to uncover with your prompts in the audit?
[Free text]
(2) What were your strategies you used to audit the AI chatbot?
[Free text]
(3) In your own view, how appropriate were the prompts you used for auditing AI chatbots for fairness, bias, and problematic outputs thoroughly?

1 - Not appropriate at all
2 - Slightly appropriate
3 - Moderately appropriate
4 - Very appropriate
5 - Extremely appropriate

(4) How would you rate the overall effectiveness of your prompts to audit AI chatbots for fairness, bias, and problematic outputs thoroughly?

1 - Very ineffective
2 - Ineffective
3 - Neither effective nor ineffective
4 - Effective
5 - Very effective

(5) Would you recommend using your prompts for educational purposes on how to audit AI chatbots for fairness, bias, and problematic outputs thoroughly?

1 - Definitely not
2 - Probably not
3 - Not sure
4 - Probably yes
5 - Definitely yes

(6) How confident are you that you succeeded in auditing the AI chatbot thoroughly?

1 - Not at all confident
2 - Slightly confident
3 - Somewhat confident
4 - Quite confident
5 - Extremely confident

(7) Why do you think your prompts were effective or ineffective at uncovering problematic or questionable behavior by the AI chatbot?
[Free text]

F System Usability Survey

The following survey items were shown to participants after all their conversations were completed.

Please consider your interaction with the AI Chatbot & Auditing System in this study. To what extent do you agree or disagree with the following statements?

(1) I think that I would like to use this system frequently.
(2) I found the system unnecessarily complex.
(3) I thought the system was easy to use.
(4) I think that I would need the support of a technical person to be able to use this system.
(5) I found the various functions in this system were well integrated.
(6) I thought there was too much inconsistency in this system.
(7) I would imagine that most people would learn to use this system very quickly.
(8) I found the system very cumbersome to use
(9) I felt very confident using the system.
(10) I needed to learn a lot of things before I could get going with this system.

Participants indicated their agreement on the following scale:

1 - Strongly disagree
2 - Disagree
3 - Neither agree nor disagree
4 - Agree
5 - Strongly agree