Highlights

Understanding People's Willingness to Participate in Human-LLM Conversational Interaction Research Studies - A Scenario Study

Hanna Alzughbi, Jinkyung Katie Park, Bart Knijnenburg

- Tested 864 AI study designs varying seven governance parameters.
- Comfort, safety, trust, and usefulness mediated participation willingness.
- Data access, anonymization, and retention impacted participation as expected.
- Model training and corporate management did not reduce participation.
- Ethics boards should enforce safeguards despite varied participant perceptions.

Understanding People's Willingness to Participate in Human-LLM Conversational Interaction Research Studies - A Scenario Study

Hanna Alzughbi

School of Computing, Clemson University, 821 McMillan Rd, Clemson, 29631, South Carolina, United States

Jinkyung Katie Park

School of Computing, Clemson University, 821 McMillan Rd, Clemson, 29631, South Carolina, United States

Bart Knijnenburg

School of Computing, Clemson University, 821 McMillan Rd, Clemson, 29631, South Carolina,

Abstract

As Large Language Models (LLMs) become integrated into human-subjects research, understanding how participants perceive and consent to data use is central to building transparent and trustworthy research practices. We conducted a scenario-based study in which 157 participants evaluated 12 scenarios (sampled from a pool of 864), each varying across seven parameters: topic sensitivity, LLM management, data anonymization, data retention, model training, additional data access, and additional consent. We find that the effect of parameters on participants' willingness to participate was mediated by their perceptions of comfort, usefulness, safety, appropriateness, and trust. Anonymization and explicit consent increased participation willingness, while longer retention and broader data access reduced it. Surprisingly, corporate LLM management (vs. federal) did not reduce participation, and data use for model training or personalization—despite potential data leakage—increased willingness to participate. We conclude with design implications for consent processes, transparency mechanisms, and governance practices that align research with participant expectations.

Keywords: Human-Subjects Research, Research Governance, Research Ethics, Artificial Intelligence, Large Language Models, Scenario-based Studies, Participant Attitudes

1. Introduction

As Artificial Intelligence (AI) systems are increasingly embedded in sensitive social domains, questions of privacy, trust, and data governance have become central to both practice and research. Human-Computer Interaction (HCI) researchers are increasingly utilizing Large Language Models (LLMs) to conduct user studies across contexts, such as, healthcare to education (Tlili et al., 2023; Levkovich and Elyoseph, 2023; Kamal, 2025; He et al., 2024). Privacy is a central ethical and methodological concern in user studies where participants have conversations with LLMs. LLM-mediated interactions are generative and open-ended (Zhang et al., 2020), which creates new methodological opportunities, but also distinct governance challenges: participants may disclose unanticipated facts, narratives, or identifiers, while providers may retain, analyze, or reuse such interactions for purposes like analytics or model training (Chow et al., 2023; Fiske et al., 2019). These dynamics introduce heightened uncertainty about what is collected, how long it is retained, who has access, and whether it will be repurposed, for example, as training data, internal evaluation corpora, or secondary research datasets. Scholars and ethicists warn that such opacity can undermine participants' trust and willingness to engage, especially when AI systems are deployed in sensitive domains (Zhang et al., 2020; Fecher et al., 2023).

While prior work has identified broad privacy risks in AI-mediated research (Shahriar et al., 2023; Gama et al., 2022), less is known about how specific governance choices in studies where participants have conversational interactions with LLMs—such as LLM management, data retention, anonymization, or model training (Vladika et al., 2025)—shape participants' willingness to engage. Understanding these effects is critical for designing transparent, participant-centered conversational LLM research studies (Bach et al., 2024; McDonald and Forte, 2020).

In this work, we focus on **Human-LLM Conversational Interaction research studies** (hereafter abbreviated as **HLLMCI research studies**)—studies in which research participants interact with an LLM through a conversational interface. While this does not encompass all ways researchers

employ LLMs, our focus on conversational interaction studies represents a common and growing subset of LLM-based research methodologies where participants directly engage with LLMs and where data governance choices are most immediately visible and consequential to participants.

To study the effects of research governance choices in HLLMCI research studies, we extend the privacy calculus framework (Dinev and Hart, 2006), which posits that individuals weigh expected benefits against perceived risks when deciding whether to disclose information. In HLLMCI research studies, participants must consider a study's description (characterized by governance parameters) and subjectively assess its comfort, usefulness, safety, appropriateness, and trustworthiness. Yet, systematic evidence on how these perceptions translate into willingness to participate, and how different parameters interact, remains limited. This motivates our two Research Questions (RQs):

- **RQ1**: How do design parameters of an HLLMCI research study scenarios affect participants' willingness to participate, and are these effects mediated by participants' subjective perceptions of comfort, usefulness, safety, appropriateness, and trust?
- **RQ2**: What are the effects of specific parameter values on participants' attitudes and willingness to participate (i.e., post-hoc differences between values), and how do these effects interact with other parameters?

To address these questions, we conducted a scenario-based online study (N = 157) in which participants evaluated multiple HLLMCI research study descriptions that systematically varied along seven parameters (i.e., Topic Sensitivity, LLM Management, Data Anonymization, Data Retention, Model Training, Additional Data Access, and Additional Consent). We deliberately designed concise scenarios to enable participants to provide considered opinions about governance parameters across multiple scenarios (12 per participant). For each scenario, participants indicated their willingness to participate as well as their subjective evaluations (comfort, usefulness, safety, appropriateness, trust). We used mixed-effects models and post-hoc tests to assess how parameter values and their combinations shaped participant responses. In doing so, we make two contributions:

1. We introduce a scenario-based methodology to systematically probe willingness to participate in HLLMCI research studies. By evaluating a large sample of scenarios with an orthogonally manipulated set of study

design parameters, we provide empirical insights into how governance choices (i.e., LLM management, data anonymization, data retention, model training, additional data access, and additional consent) *jointly* shape participation decisions, offering actionable guidance for designing transparent and trustworthy HLLMCI research practices.

2. We extend the privacy calculus framework beyond risk and benefit by incorporating additional attitudinal constructs: comfort, appropriateness, and trust. This expanded perspective reveals how participants make nuanced trade-offs. They consider these multiple constructs in ways shaped by parameter interactions, and that these constructs mediate the effect of the study design parameters on their participation willingness.

This work contributes to ongoing discussions about trustworthy AI and ethics in human-subjects research within and beyond HCI communities by showing how design choices in HLLMCI studies directly impact participants' perceptions, and, ultimately, their willingness to participate. We leverage our results to provide practical advice to researchers and LLM developers regarding consent processes, transparency mechanisms, and governance practices that align with participant expectations.

2. Related Work

2.1. Landscape of LLM Use in Research

Recent systematic analyses have documented the diverse ways researchers employ LLMs across a spectrum of methodologies, including LLM-assisted qualitative coding and analysis (Liao et al., 2024), studies that use LLMs as synthetic participants or to model social phenomena (Liao et al., 2024), conversational agents serving as interviewers or research assistants (Kapania et al., 2025), and tools for literature review, hypothesis generation, or research design (Kapania et al., 2025). Each methodology presents distinct risk-benefit considerations, ethical requirements, and participant relationships with the technology. Our work focuses on the prominent subset of this landscape: studies in which participants interact directly with LLMs through conversational interfaces and where these interactions constitute the primary research data. This includes user experience studies of conversational AI, but excludes studies where LLMs serve as analytical tools (e.g., coding of interview transcripts) or synthetic participants (e.g., agent-based simulations), as these contexts present different governance challenges and participant relationships.

By focusing on conversational interaction studies, we examine contexts where data governance parameters are most directly visible to and consequential for participants.

Applications of conversational LLMs in human-subject research are nascent, but already span diverse domains. In education research, for example, studies highlight both risks (e.g., academic dishonesty) and opportunities (e.g., enhancing instructional delivery and engagement) of LLMs (Porcheron et al., 2020; Cohen et al., 2024). Research demonstrates that LLMs can serve as supportive tools in pedagogical environments, helping to bridge knowledge gaps and offer tailored guidance (Kocaballı et al., 2020; Cohen et al., 2024; Porcheron et al., 2020; Lee and Cho, 2025; Goldshtein et al., 2025).

In healthcare research, LLMs are being integrated into communication and patient education contexts. Studies show that LLMs can provide relevant information and decision support in mental health settings, enhancing therapeutic interactions (Chancellor et al., 2019; Levkovich and Elyoseph, 2023). Research in pediatric orthopedics underscores the accuracy of LLM responses to parental inquiries, highlighting their role in healthcare communication (Kamal, 2025). Furthermore, there is growing evidence that LLMs can effectively handle tasks such as psychotherapy and medication management, demonstrating their versatility in supporting healthcare professionals (He et al., 2024).

Beyond healthcare and education, conversational LLMs are being explored in legal contexts and project management (Vilaza et al., 2022), where they can automate tasks and enhance decision-making by providing contextual information (Wang et al., 2023). Taken together, these examples underscore the breadth of conversational LLM-based research where sensitive information and consequential decisions may be at stake.

2.2. Governance, Ethics, and Privacy Calculus in HLLMCI Research

The integration of LLMs into research contexts necessitates careful consideration of privacy practices. Because conversational agents can afford follow-up prompts and empathic language, they often elicit richer, more personal disclosures than closed responses. Furthermore, empirical work suggests that users feel less anxious revealing sensitive or stigmatized information to chatbots than to humans, which increases the volume and sensitivity of data captured in such studies (Kuhlmeier et al., 2025). Agent design features that increase perceived humanness or social presence (e.g., anthropomorphic cues, conversational warmth) also modulate users' willingness to disclose personal

information, thereby potentially amplifying privacy risks (Pizzi et al., 2023). Moreover, LLM-based conversational agents research shows that users often have limited awareness of how their data may be used in model training (Zhang et al., 2024), which may further increase their privacy risk. While such conversational agents can encourage rapport and self-disclosure—often central to usability and effectiveness objectives—they can unintentionally enlarge the set of personally identifying or sensitive facts collected during a study, making data governance opaque and permissive (He et al., 2024; Bach et al., 2024). From a privacy calculus perspective (Dinev and Hart, 2006) this underscores the need for consent forms that help participants weigh potential benefits against risks. Yet, current practices frequently fall short: Broad consent clauses and generic notices often obscure these trade-offs, leaving participants without the information needed to make fully informed decisions in contexts where disclosure risks are heightened (Wong and Mulligan, 2019; McDonald and Forte, 2020; Fecher et al., 2023; Baxter et al., 2025).

These concerns extend beyond interaction design to the governance mechanisms that determine what happens to participants' data once collected. Conversational transcripts and metadata are often susceptible to secondary uses that are not easily foreseeable at the time of collection (for instance, inclusion in model training corpora or cross-study analytics) (Fiske et al., 2019). Intellectual-property and data-reuse concerns are already documented in technical and humanities scholarship on LLM training data, and these intersect with participant privacy when research conversations are retained or redistributed (Chow et al., 2023). The medical and mental-health literature further emphasize that these retention and reuse pathways complicate standard notions of confidentiality and demand explicit attention in ethics review and consent processes (Fiske et al., 2019). These dynamics highlight that consent should not be merely procedural but subject to a careful *privacy calculus*: the consent processes in HLLMCI studies should explicitly specify important governance features such as data retention and training practices in ways that support participants' careful deliberation. Recent research examining HCI researchers' own practices reveals significant gaps in how LLM use is disclosed to participants and Institutional Review Board (IRBs), with many researchers treating LLMs as "everyday tools" that do not require explicit reporting (Kapania et al., 2025). This highlights the urgency of establishing clear governance frameworks specifying when and how LLM use should be disclosed in human-subjects studies.

2.3. What Privacy Perceptions and Conditional Factors Matter in HLLMCI Research

Within the extended privacy calculus, participants' willingness to engage with HLLMCI studies can be shaped by five proximal perceptions: comfort, usefulness, safety, appropriateness, and trust. These perceptions represent immediate costs and benefits individuals weigh when deciding whether to participate, influenced by contextual research parameters. Below, we outline how prior work motivates our focus on these perceptions and parameters, and why they are critical to examine in the context of HLLMCI research.

2.3.1. Privacy Perceptions

Privacy perceptions capture how participants feel about disclosure, perceived personal or societal benefit, and trust in protections offered. In designing our scenarios, we focused on five perceptions—comfort, usefulness, safety, appropriateness, and trust—that have been shown to consistently shape participation decisions in prior work across HCI and health domains. Each represents a dimension of how participants weigh costs and benefits, and together they provide a framework for understanding how design choices in HLLMCI research influence willingness to participate (Kalkman et al., 2019; Khatiwada et al., 2024).

Comfort. The extent to which individuals feel at ease when disclosing information. According to prior work, some participants—such as youths—report greater comfort disclosing sensitive information to chatbots than to humans. However, unclear governance can undermine this comfort and increase privacy concerns (Crabtree et al., 2017; Kuhlmeier et al., 2025; Luca et al., 2023). Similarly, anthropomorphic cues may foster rapport but also heighten feelings of vulnerability if governance is opaque (Pizzi et al., 2023). These findings suggest that comfort is an important lens through which to assess willingness to participate in HLLMCI research.

Usefulness. Whether participants perceive engagement with an LLM as beneficial for the task at hand. Prior work has shown that usefulness depends on accuracy, task fit, explainability, and adaptive communication. LLMs are perceived as useful when outputs are accurate, readable, and sourced, and when explanations support understanding in sensitive domains (Tlili et al., 2023; Kamal, 2025; Maity and Deroy, 2024). Including usefulness in our study allows us to capture whether perceptions of benefit balance concerns about data use.

Safety. Perceptions of protection from harm across three domains: system safety (accuracy and avoidance of harmful advice), user safety (e.g., preventing clinical or psychological harms), and data safety (confidentiality and security). Reviews of health agents highlight persistent gaps in these areas, calling for stronger technical and governance protections, particularly in sensitive domains. Prior work shows that safety perceptions drop when systems lack clear disclosures, secure logging, or data minimization (Khan and Seto, 2023; Akalın et al., 2023). In our study, safety provides a measure of whether participants feel adequately protected when considering study participation.

Appropriateness. Normative judgments about whether LLMs should be used for specific purposes. Prior work shows that participants view chatbots as appropriate for informational or administrative tasks, but are more hesitant in high-stakes or legally consequential domains, unless governance and auditability are strong (Luca et al., 2023). Cultural and domain-specific norms also shape appropriateness perceptions (Daley et al., 2018). In our study, appropriateness judgments are intended to condition willingness to engage on task and governance assurances.

Trust. Expectations that systems and institutions act reliably, transparently, and with accountability. Prior work indicates that trust increases with clear data-use explanations, opportunities for participant control, and visible risk management, but declines under opaque training or third-party reuse—particularly for groups with histories of institutional distrust (Bach et al., 2024; Zhang et al., 2020). In our study, trust assesses whether governance conditions bolster or erode this key relational perception.

2.3.2. Research Study Parameters

The above five perceptions vary with research study parameters. In our study, we focus on parameters that researchers commonly must decide upon when collecting HLLMCI study data. These parameters are not merely technical—they carry governance and ethical implications that participants notice, evaluate, and use to decide whether to participate. Below we describe each parameter and motivate its inclusion in our study using prior work.

Topic Sensitivity. The subject matter of a study directly shapes privacy attitudes, hence our study varies the topic sensitivity of the presented research studies. Sensitive topics such as mental health, reproductive health, substance use, criminal justice, or sexual orientation often reduce comfort and safety

unless strong confidentiality is guaranteed (Chernick et al., 2023). At the same time, prior work shows that some participants actually prefer disclosing to chatbots in these domains because they reduce stigma compared to human disclosure (Kuhlmeier et al., 2025; Luca et al., 2023). Cultural and demographic factors amplify this sensitivity: marginalized groups with histories of discrimination express heightened concerns unless protections are robust (Matson et al., 2019).

LLM Management. Research participants may distinguish between research institutions (e.g., universities, public agencies) and commercial providers (e.g., OpenAI) when deciding whom to trust with their data. Prior work shows that institutional hosting, clear audit trails, and contractual limits on reuse can strengthen trust, while opaque vendor practices or unclear logging reduce it (Fecher et al., 2023; Chow et al., 2023; Bach et al., 2024). Disclosure of whether an agent is research-run or vendor-run similarly influences perceptions. We therefore vary model management to capture how institutional versus vendor oversight shapes participation willingness.

Data Anonymization. The degree of data de-identification is a central privacy protection in research design. Higher levels of anonymization generally increase comfort and perceived safety, but participants may question whether free-text data can truly be anonymized given re-identification risks. Stronger technical controls, automated redaction, or limits on human review improve perceptions, especially when residual risks are communicated transparently (Nicol et al., 2022). Because anonymization is widely invoked yet difficult to guarantee for conversational data, it is an essential parameter to study systematically.

Data Retention. Prior work shows that shorter retention periods and automatic deletion policies increase perceived safety and trust, while indefinite storage reduces comfort unless offset by other safeguards (Zhang et al., 2020; Mane et al., 2023; Kassam et al., 2023). Participants also expect durations to be stated clearly and tied to research purpose. We therefore examine retention to understand how time-limits on storage interact with other protections.

Model Training. A key concern in HLLMCI studies is whether transcripts are reused to train models. Reuse for general-purpose model training is often seen as unacceptable without explicit consent, especially for personal or creative data (Fecher et al., 2023; Chow et al., 2023). Participation tends to increase when data are verifiably excluded from training, and decrease when reuse is

mandatory, particularly in sensitive domains (Zhang et al., 2020; McDonald and Forte, 2020).

Additional Data Access. Beyond training, researchers must decide whether data will be shared with other parties. Broader sharing (e.g., commercial reuse or public release) generally reduces trust, while narrower and auditable sharing (e.g., within a research team or under IRB oversight) increases willingness (Kassam et al., 2023). Marginalized groups are especially sensitive to data re-use due to past harms (Matson et al., 2019).

Additional Consent. Research shows that participants prefer being asked before their data is used for secondary purposes. Protocols promising re-consent or explicit notification increase participation willingness, while one-time broad consent reduces it (Kassam et al., 2023; Zhang et al., 2020; McDonald and Forte, 2020). Our study tests whether notice or choice help build trust and comfort regarding potential secondary data use.

Prior work demonstrates that participation decisions are shaped by perceived information flows and governance cues. Yet, much of this research has examined these parameters in isolation, leaving limited insight into how they jointly influence willingness to participate in HLLMCI studies (McDonald and Forte, 2020; Wong and Mulligan, 2019; Bach et al., 2024). This gap motivates our scenario-based study, which systematically varies these parameters to reveal how participants negotiate governance trade-offs. To this end, we incorporate the parameters outlined above into a factorial scenario design (see Section 3.1), enabling systematic examination of their influence on participation willingness.

3. Research Methods

Our scenario-based survey study investigated participants' preferences, perceptions, and concerns regarding the use of data in Human-LLM Conversational Interaction (HLLMCI) research studies. Each participant evaluated 12 scenarios drawn from 864 possible scenarios, each describing an HLLMCI study. The study consisted of five sequential parts:

1. **Consent:** Participants reviewed a consent form outlining the purpose of the research, the nature of participation, confidentiality protections, and participants' rights to withdraw at any time.

2. **Definition of LLMs:** Participants were given a brief definition that explained LLMs as conversational AI systems that can understand and generate human language, assist with tasks, and continuously improve by learning from datasets (see Appendix B). This definition established the conversational interaction context and grounded participants' understanding of the governance parameters manipulated in our scenarios.
3. **Scenario Evaluation Task:** Participants evaluated 12 scenarios, each describing an HLLMCI research study that differed on 7 parameters (selected from a set of 864 scenarios; see Section 3.1). Participants responded to six questions per scenario evaluating their participation willingness, comfort, perceived usefulness, perceived safety, appropriateness, and trust (see Section 3.2).
4. **Post-Study Questions:** After the scenarios, participants answered one open-ended follow-up question about their overall participation decisions, followed by questions assessing their LLM use frequency, familiarity with LLM data practices, basic knowledge of AI capabilities and risks, and understanding of data handling in academic research (see Appendix C).
5. **Demographics:** Participants provided demographic data to contextualize the results (see Section 3.3 and Appendix D).

3.1. Scenario Design and Sampling

In the scenario evaluation task, each participant evaluated 12 study designs, selected from a set of 864 possible scenarios spanning seven study parameters (see Table 1). These parameters were identified through review of IRB requirements and emerging best practices in LLM research ethics. They reflect real-world design decisions commonly addressed in research consent forms and identified by prior work as most salient to participants' concerns (see Section 2.3.2). Our parameter selection[1] reflects three design principles: First, we incorporated parameters that apply to conventional human-subjects research (data retention, additional data access, consent mechanisms) to maintain comparability with established research ethics frameworks. Second,

[1]We acknowledge that additional parameters could provide further nuance, but such additions would exponentially increase the scenario space and risk overwhelming participants with overly complex scenario descriptions.

we included parameters specific to LLM contexts (LLM management, model training) that introduce novel governance considerations. Third, we selected parameters that exist in conventional research but carry different connotations in LLM contexts—most notably data anonymization, which faces unique challenges given the generative and conversational nature of LLM interactions. Each scenario was constructed using the following template:

> You are invited to participate in a study on [Topic Sensitivity]. During the study, you will interact with [LLM Management]. The data you provide during these interactions will be [Data Anonymization]. Your data will be stored [Data Retention] [Model Training] can be re-used for other purposes in the future [Additional Data Access]. [Additional Consent].

Possible values of the study parameters [in brackets] are listed in Table 1. The 12 scenarios for each participant were selected (out of 864) using a mixed fractional factorial design that optimizes the statistical power to evaluate each parameter and parameter combination while ensuring balanced exposure to parameter values (i.e., each participant received 12 scenarios with a balanced combination of values).

3.2. Measures

To assess participants' privacy behaviors and attitudes, we asked six questions per scenario, enabling a multi-dimensional evaluation of each study design and supporting statistical modeling of how study parameters influenced responses. The questions and response options are listed in Table 3. Participation willingness was measured as a binary variable (0 = "no", 1 = "yes"). All attitudinal outcome variables were measured on a 7-point scale, coded from −3 (e.g., "very uncomfortable", "very risky", "completely distrust") to +3 (e.g., "very comfortable", "very safe", "completely trust"), with 0 indicating a neutral response. On average, participants opted to participate in 61.4% of studies. Ratings for continuous variables were centered near neutral, but skewed slightly positive. The relatively high standard deviations suggest substantial variation across participants and scenarios.

Additionally, after completing all 12 scenarios, participants were asked one open-ended follow-up question about their overall participation decisions (see Appendix C). Participants' answers were brief, but provided contextual insight into their reasoning, which we reference where relevant in our Results section.

Table 1: Factorial Design Matrix

Parameter	Code	Value
Topic Sensitivity	Sensitive	**medical advice**
	Non-sensitive	**movie recommendations**
LLM Management	NSF	**ScienGPT**, a Large Language Model (LLM) developed by the **National Science Foundation**
	OpenAI	**ChatGPT**, a Large Language Model (LLM) developed by **OpenAI**
Data Anonymization	Identifiable	**identifiable**, meaning that your conversation will be linked to your name and e-mail address
	Partially Anonymized	**partially anonymized**, meaning that while your identifiable information will be removed, certain aspects of your interaction may still reveal your identity
	Fully Anonymized	**fully anonymized**, meaning that all identifiable information will be removed and your interaction cannot be traced back to you
Data Retention	2 Months	for **2 months**
	4 Years	for **4 years**
	Indefinitely	**indefinitely**
Model Training	Not Used	Your data will **not be used to improve the AI model**, but it
	Accuracy	and may be used to **improve the general accuracy of its responses**. Additionally, your data
	Personalization	and may be used to **tailor its responses to your specific needs**. Additionally, your data
Additional Data Access	Researchers Only	**by the same researchers**
	Researchers and Partners	**by the same researchers, as well as other researchers or industry partners**, who may use the data for research or commercial purposes
Additional Consent	No Consent	Once you agree to participate in the study, **no further consent will be requested** for future use of your data
	Informed	You will be **informed** if your data is used for any additional purposes beyond the study
	Opt-Out	You will be notified of future uses of your data and given the option to **opt-out** of such additional uses
	Opt-In	You will be notified of potential future uses of your data and given the option to **opt-in** to such additional uses

Table 3: Survey Items, Response Options, and Descriptive Statistics.

Variable	Survey Question	Response Options	M (SD)
Participation	Are you willing to participate in the study described in the scenario above?	No / Yes	0.614 (0.487)
Comfort	How uncomfortable or comfortable do you feel about participating in this study with the LLM usage described?	Very uncomfortable, Uncomfortable, Somewhat uncomfortable, Neutral, Somewhat comfortable, Comfortable, Very comfortable	0.102 (1.916)
Usefulness	How useful do you find this research study?	Completely useless, Useless, Somewhat useless, Neutral, Somewhat useful, Useful, Very useful	0.590 (1.506)
Safety	How risky or safe do you consider it to participate in this research study?	Very risky, Risky, Somewhat risky, Neutral, Somewhat safe, Safe, Very safe	0.093 (1.795)
Appropriateness	How inappropriate or appropriate do you find the use of LLMs in this study, as described in this scenario?	Very inappropriate, Inappropriate, Somewhat inappropriate, Neutral, Somewhat appropriate, Appropriate, Very appropriate	0.486 (1.609)
Trust	How much distrust or trust do you have in the researchers conducting this study in how they handle your data?	Completely distrust, Distrust, Somewhat distrust, Neutral, Somewhat trust, Trust, Completely trust	0.122 (1.661)

Finally, participants answered a set of questions assessing their LLM use frequency, familiarity with LLM capabilities, risks and data practices, and understanding of data handling in academic research. These questions (listed in Appendix Appendix C) were asked after participants completed the scenarios, so as to not prime them about privacy concepts in a way that could influence their responses to the scenarios.

3.3. Participants

We recruited 196 participants through Prolific, an online platform commonly used for behavioral research. Eligibility criteria included being a resident of the United States (to reduce excessive cultural variations in privacy concerns that are difficult to control for, cf. (Ghaiumy Anaraky et al., 2021)) and at least 18 years of age (an IRB requirement). To ensure high-quality responses we required at least 10 prior Prolific engagement and a 90–100% approval rate. To ensure sample diversity, we requested a sample with a balanced gender distribution.

Data quality and cleaning

We implemented several measures to prevent automated or LLM-generated responses. Our survey platform disabled text selection and copy-paste functionality, preventing participants from copying content into external tools. We also monitored for duplicate IP addresses and unusual response patterns that might indicate coordinated or automated participation. Furthermore, Prolific's internal verifications—including device fingerprinting and behavioral analysis—provide additional layers of protection against bot activity. While no combination of measures is entirely foolproof, these safeguards substantially reduce the likelihood of large-scale automated or LLM-generated responses infiltrating our data. Our attention checks, response timing analysis, and variance checks (described below) provided additional signals that would likely detect AI-generated responses, which tend to show unnaturally consistent patterns or unusually rapid completion times.

A total of 42 participants were excluded based on attention checks, response quality, and completion time (some participants met multiple exclusion criteria):

- Nine participants (five unique; four overlapping with other criteria) failed two out of three attention checks. Two checks were embedded within the scenario section, and one was included in a later section on

data collection concerns. These items instructed participants to select a specific response (e.g., "somewhat disagree" or "completely agree") to confirm they were reading carefully[2].

- Six participants (three unique) failed at least one of three additional attention checks. Two were reading comprehension checks (asking participants to repeat information that could be gleaned from the same page). The third was a fake option ("FakeSeek") embedded in a question about prior chatbot use (see Appendix C) and designed to detect inattentiveness.

- Responses from fourteen participants (nine unique) showed minimal response variance (mean[variance] < 0.2 across the 7-point scale attitude questions); these were interpreted as satisficing.

- Seven participants (four unique) completed the survey in under eight minutes (i.e., less than half of the median completion time of 17m29s), suggesting they may have rushed through the survey without fully engaging.

After removing these participants, 157 remained for analysis. Importantly, running our primary analyses with the full dataset (prior to exclusions) yielded consistent main effects and interpretations, confirming the robustness of our findings to these exclusion criteria.

Participants' Demographics

The 157 participants included in our analyses represented a diverse sample (see Table 4). The majority were aged between 25 and 44, and most held at least an undergraduate degree. While over half were employed full-time, the sample also included part-time workers, unemployed individuals, and those not currently in paid work.

Compensation

As our quality checks were conducted retroactively, all 196 participants were compensated \$4.50 (an average effective compensation rate of approx-

[2]We acknowledge that attention checks requiring specific responses have known limitations and are not ideal as standalone quality control measures. Notably, failing attention checks did not affect participants' compensation.

Table 4: Demographic Characteristics of Participants (N = 157).

Category	Group	n (%)
Gender	Woman	77 (49.0%)
	Man	73 (46.5%)
	Nonbinary	5 (3.2%)
	Genderfluid	1 (0.6%)
	Prefer not to disclose	1 (0.6%)
Age	18–24	16 (10.2%)
	25–34	56 (35.7%)
	35–44	35 (22.3%)
	45–54	29 (18.5%)
	55 and above	21 (13.4%)
Ethnicity	Caucasian/White	104 (66.2%)
	African American/Black	21 (13.4%)
	Hispanic/Latino	16 (10.2%)
	Asian/Pacific Islander	9 (5.7%)
	Mixed ethnicity	3 (1.9%)
	Native American/Alaska Native	1 (0.6%)
	Prefer not to say	3 (1.9%)
Education	No formal qualifications	1 (0.6%)
	Secondary education (e.g. GED/GCSE)	6 (3.8%)
	High school diploma/A-levels	40 (25.5%)
	Technical/community college	26 (16.6%)
	Undergraduate degree (BA/BSc/other)	56 (35.7%)
	Graduate degree (MA/MSc/MPhil/other)	24 (15.3%)
	Doctorate degree (PhD/other)	4 (2.5%)
Employment	Full time	83 (52.9%)
	Part time	25 (15.9%)
	Unemployed (and job seeking)	21 (13.4%)
	Not in paid work	20 (12.7%)
	Due to start a new job	4 (2.5%)
	Other	4 (2.5%)

imately $15.44 per hour), ensuring that quality checks did not penalize attentive participants who inadvertently missed specific responses.

Ethics

This study was approved by the authors' IRB. All participants provided informed consent and were assured that no identifiable information would be collected. Participation was voluntary, with the option to withdraw at any time. Data were anonymized at the point of collection and securely stored on encrypted servers, accessible only to the research team. No personally identifiable information was retained or linked to participant responses.

Participant Background and Familiarity with LLMs and Research

Table 5 gives an overview of participants' answers to the post-study questions (see Appendix C for full question wording). Most participants (95.4%) had used at least one LLM tool, with ChatGPT being the most common (89.8%). Self-reported understanding of how LLMs work was mixed—many reported some familiarity, but few had technical expertise. On average, participants correctly answered 6.68 out of 8 true/false AI knowledge quiz questions. Most (95.5%) answered at least 5 questions correctly, and 26.1% achieved a perfect score. These results suggest that participants were generally knowledgeable about AI and LLMs, despite our study not providing specific instruction or training.

In contrast, familiarity with AI data privacy practices was low, and awareness of distinctions between corporate and independent AI models was also limited. While 47.8% of participants understood that LLMs use input data to train future models, 22.3% reported not knowing how their data would be used. Similarly, familiarity with and understanding of HLLMCI studies was low. Only 40.8% had previously participated in HLLMCI studies (though nearly all of those who had, reported being informed about data privacy practices in those studies). More generally, 39.2% of participants reported being unfamiliar with how research studies handle their data, and only 7.2% reported being very familiar.

In summary, participants had prior experience with and basic knowledge of LLMs, but were less aware of LLM management and data policies. They also had limited experience with HLLMCI research and low familiarity with the data privacy practices of such studies. It is possible that participants' limited familiarity with data privacy influenced their ability to fully understand the research scenarios. However, we believe that this level of familiarity is

Table 5: Participant Background and Familiarity (N = 157).

Category	Group	Stat
LLM Tool Usage	ChatGPT	89.8%
	Gemini	45.9%
	Microsoft Copilot	35.7%
	Bing Chat	34.4%
	Bard	21.7%
	Grok	17.8%
	Claude	15.9%
	Others	11.3%
LLM Experience	Occasional users	54.8%
	Frequent users	28.0%
	Non-users	13.4%
	Technical users	3.8%
LLM Factual Knowledge Quiz Accuracy	Score range across 8 items	52%–99%
	Mean score (out of 8)	6.68
	≥ 5 correct	95.5%
	≥ 6 correct	87.3%
	≥ 7 correct	60.5%
	Perfect score (8/8)	26.1%
LLM Data Privacy Policy Familiarity	Not familiar	48.4%
	Somewhat familiar	42.7%
	Very familiar	8.9%
Awareness of Corporate vs. Independent Models	Aware	15.9%
	Unaware	84.1%
Perceptions of LLM Data Use	Train future models	47.8%
	Don't know	22.3%
	Provide responses	13.4%
	Commercial purposes	10.2%
	Quality assurance	5.7%
	Legal requirements	0.6%
Prior Research Experience	Participated in research	40.8%
	Informed about data use	96.9% of who participated
Research Data Privacy Familiarity	Not familiar	39.2%
	Somewhat familiar	56.2%
	Very familiar	7.2%

representative of typical research study participants, and that our scenario descriptions reflect the level of detail commonly found in research consent forms. As such, participants' responses to our scenarios likely reflect the attitudes and behaviors of participants in real-world HLLMCI studies.

3.4. Data Analysis

To understand how different study design parameters influenced participants' responses, we tested a series of multilevel models, which allowed us to account for the nested structure of the data (i.e., multiple scenario responses provided by each participant).

To answer RQ1 (Section 4.1), we conducted a *mediation analysis* using the "four steps" approach (Baron and Kenny, 1986; James and Brett, 1984; Judd and Kenny, 1981) to determine the extent to which the effect of the study parameters (X) on participation willingness (Y) is mediated by participants' attitudinal evaluations of the described study (M):

1. In Step 1 ($X \rightarrow Y$), we analyzed how the study parameters—*topic sensitivity*, *LLM management*, *data anonymization*, *data retention period*, *model training*, *additional data access*, and *additional consent*—affected participants' **Participation Decision** (i.e., whether the participant chose to participate in the described study). This analysis was conducted using a Generalized Linear Mixed-Effects Model (GLMM) with a participant-level random intercept (accounting for 12 scenarios per participant) and a binomial link function (to account for the binary nature of the participation decision). This model estimates the likelihood of participation willingness (Y) under varying scenario conditions (X). We used Type II Wald χ^2 tests to assess the significance of each study parameter.
2. In Step 2 ($X \rightarrow M$), we used a Linear Mixed Effect Model (LME) with a participant-level random intercept to estimate each of the subjective outcome variables—**Comfort**, **Usefulness**, **Perceived Safety**, **Appropriateness**, and **Trust**—using the study parameters as predictors. We used Type II Wald χ^2 tests to assess the significance of each study parameter (X) in influencing participants' attitudinal evaluations (M) of the described study.
3. In Step 3 ($M \rightarrow Y$ controlling for $X \rightarrow Y$), we added the attitudinal evaluations as predictors to the model of Step 1 to test whether the attitudinal evaluations (M) affect willingness to participate (Y), controlling for the scenario conditions (X).

4. In Step 4 (no effect of $X \rightarrow Y$ controlling for $M \rightarrow Y$), we used the same models as in Step 3 to evaluate whether the effects of the study parameters (X) are reduced ("partial mediation") or disappear ("full mediation") when attitudinal evaluations (M) are introduced to the model predicting participation willingness (Y).

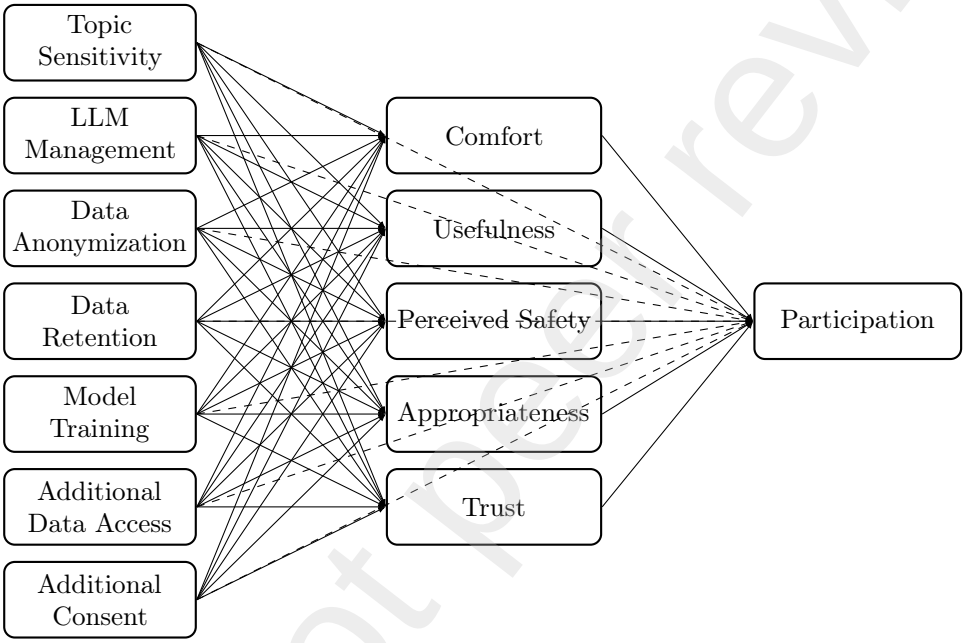


Figure 1: Model of hypothesized effects: study parameters (X) influence attitudinal evaluations (M), which in turn influence participation willingness (Y). In case of "full mediation", study parameters have no residual direct effects on participation willingness (dotted arrows) once attitudinal evaluations are controlled for.

Figure 1 summarizes the hypothesized effects of the mediation analysis. This mediation model can clarify whether the effect of the study parameters on participants' participation decision is intuitive (no/weak mediation) or cognitively mediated (strong/full mediation).

Section 4.2 discusses specific significant differences in participation and attitudes between pairs of parameter values. These analyses involve Tukey-corrected post-hoc tests as follow-ups to the significant χ^2 tests described above. These analyses provide practical advice to human-centered AI researchers regarding the best way to increase participation and reduce participants' concerns regarding their research studies. In this section we therefore

particularly call out (combinations of) parameter values that significantly increase or decrease participants' attitudes and/or participation willingness.

Section 4.2 also covers significant 2-way and 3-way interaction effects between parameters (answering RQ2), which are added one by one to the full models outlined above, and tested for significance using Wald χ^2 tests. Significant interaction effects are followed up by Tukey-corrected post-hoc tests of one parameter along the levels of the other parameter(s). To aid the discussion of parameter effects and interactions, we present them as figures in which the y-axis shows participants' ratings for the corresponding outcome variable (e.g., willingness to participate, comfort, usefulness, safety)[3]. Thus, the plotted estimates reflect the average participant response to each study design manipulation, rather than raw frequencies or model coefficients.

In addition to the quantitative analyses, we conducted a brief qualitative analysis of participants' open-ended responses. This qualitative layer allowed us to contextualize the statistical results by identifying how participants articulated their privacy-related reasoning and trade-offs.

4. Results

4.1. Mediation Analysis: Predictors of Willingness to Participate (RQ1)

To examine the mechanisms through which study design influenced participation willingness, we conducted a mediation analysis in four steps (see Section 3.4). Our GLMM for **Step 1** evaluates the direct effects of study parameters on participation willingness. The first column (i.e., Participation) of Table 7 shows that five of the seven study parameters significantly predicted participation willingness: data anonymization had by far the strongest effect, followed by topic sensitivity, additional consent, data retention, and model training, while two parameters had no significant effect (LLM management and additional data access). More details about the specific study parameters that increase or decrease participation willingness will be discussed in Section 4.2.

[3]Confidence intervals are computed using nonparametric bootstrapping with `mean_cl_boot` in `ggplot2`. Unlike our post-hoc tests, these confidence intervals do not account for within-subjects variability or family-wise error; they are presented for illustrative purposes only. Significant post-hoc test p-values are adjusted using a Tukey correction and indicated by brackets.

Our LMEs for **Step 2** test whether the study parameters significantly affected participants' attitudinal evaluations of the presented research study—namely, their ratings of comfort, usefulness, safety, appropriateness, and trust. Columns 2–6 of Table 7 show that topic sensitivity, anonymization, data retention period, and additional consent all showed strong and relatively consistent effects across the mediators. Model training only had a significant effect on usefulness and appropriateness, while additional data access only had an effect on trust, and LLM management had no main effect on any of the attitudes. Notably, the latter two parameters also had no effect on participation willingness (see Step 1).

Our GLMM for **Step 3** adds the attitudinal mediators to the model of Step 1. Table 8 shows the results of this model, with the effects of the attitudes listed at the bottom of the table. Comfort emerged as the strongest predictor, followed in order by perceived safety, usefulness, and trust. Only appropriateness did not have a significant effect on participation willingness. These findings reinforce the interpretation that participants' affective and cognitive responses—particularly how comfortable and safe they felt—were key drivers of their willingness to participate in the described study.

Finally, in **Step 4** we evaluated whether the effects of study parameters on participation remain significant once the mediators are accounted for. The top part of Table 8 shows that all previously significant effects of study parameters were non-significant in this combined model. This suggests that the effects of the study parameters on participation willingness are *fully mediated* by participants' subjective evaluations of the scenario.

4.2. Parameter-Based Effects on Outcomes (RQ2)

Given that we established significant "omnibus effects" of several of the study design parameters on participation willingness and the attitudinal mediators in Section 4.1, we here present significant "post-hoc effects" of each study parameter across each outcome variable. Results are organized by parameter to highlight which study design aspects may increase or decrease participation willingness and/or attitudes toward the presented study.

This section also explores whether the effects of specific study parameters may depend on the value of (an)other parameter(s). To examine this, we added two-way and three-way interaction effects to the models from Steps 1 and 2, one at a time. Significant, consistent, and/or otherwise notable interactions are summarized in Table 9 and discussed throughout this section.

Table 7: Type II Wald χ^2 test results for the models of Step 1 (first column) and Step 2 (remaining columns) of the mediation analysis. Values in the χ^2 columns represent the $\chi^2(df)$ test statistics for each predictor (df = # of parameter values - 1). The p columns represent the p-values of these tests. The significant p-values are highlighted.

Predictor	**Participation**		**Comfort**		**Usefulness**		**Safety**		**Appropriateness**		**Trust**	
	χ^2	p	χ^2	p	χ^2	p	χ^2	p	χ^2	p	χ^2	p
Topic Sensitivity	33.73	<.001	97.23	<.001	122.11	<.001	75.92	<.001	14.61	<.001	28.82	<.001
LLM Management	0.03	.860	1.83	.176	0.63	.427	0.75	.385	0.84	.359	2.13	.144
Data Anonymization	270.47	<.001	801.55	<.001	128.38	<.001	715.43	<.001	266.02	<.001	465.32	<.001
Data Retention	8.34	.015	21.70	<.001	4.18	.123	18.94	<.001	16.46	<.001	22.79	<.001
Model Training	6.59	.037	3.09	.214	40.98	<.001	2.93	.232	8.08	.018	2.70	.259
Additional Data Access	0.95	.329	0.19	.666	0.02	.881	1.39	.239	1.04	.308	4.65	.031
Additional Consent	18.71	<.001	28.12	<.001	6.94	.074.	21.63	<.001	10.05	.018	31.04	<.001

Table 8: Type II Wald χ^2 test results for the model of Step 3 and 4 of the mediation analysis, which tests the combined effect of attitudes (bottom half, Step 3) and parameters (top half, Step 4) on participation willingness. Values in the χ^2 column represent the $\chi^2(df)$ test statistics for each predictor (df = # of parameter values - 1 for parameters, 1 for attitudes). The p column represents the p-values of these tests. The significant p-values are highlighted.

Predictor	χ^2	p
Topic Sensitivity	0.053	0.819
Management	1.082	0.298
Anonymization	2.504	0.286
Retention	0.182	0.913
Training	4.100	0.129
Data Access	0.000	0.986
Consent	0.671	0.880
Comfort	112.717	<.001
Usefulness	7.625	.006
Safety	16.621	<.001
Appropriateness	0.996	0.318
Trust	3.982	0.046

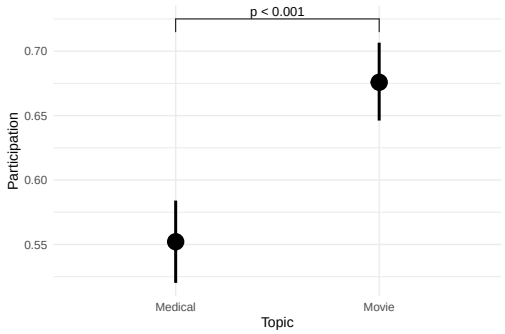
4.2.1. Topic Sensitivity

As noted in Section 4.1 and shown in Table 7, our omnibus tests showed that topic sensitivity had a significant effect on participation willingness and all attitudinal evaluations. Topic sensitivity referred to whether the study scenario involved movie recommendations (a non-sensitive topic) or medical advice (a topic often involving sensitive personal information). In our study scenarios where the topic was non-sensitive (movie recommendations), post-hoc analysis showed that participants reported significantly higher participation willingness, comfort, trust, perceived safety, and appropriateness than for the sensitive topic (medical advice), as indicated by the significance brackets in Figure 2. Interestingly, participants rated the non-sensitive topic as less useful than the sensitive topic, suggesting a trade-off between perceived sensitivity and usefulness.

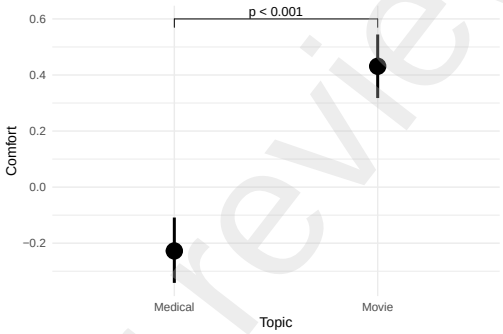
Participants' open-ended responses revealed how topic sensitivity shaped their privacy calculus. One participant emphasized aligning their choices with the societal importance of the topic: *"I chose to participate in all of the medical advice studies because I find that important research"* (P40). Another described how their privacy threshold shifted depending on context: *"I was*

Table 9: Significant and marginal two-way and three-way interactions by outcome. Under each outcome, there is a column for the χ^2 statistic with degrees of freedom in parentheses, and another column for the p-value. The significant p-values are highlighted. Abbreviations: Ret = Data Retention, Train = Model Training, Access = Additional Data Access, Cons = Additional Consent, Anon = Data Anonymization, Mgmt = LLM Management, Topic = Topic Sensitivity.

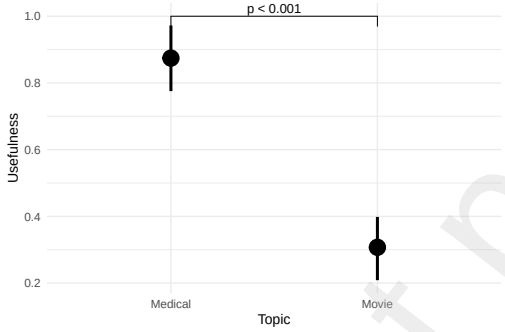
Interaction	**Participation**		**Comfort**		**Usefulness**		**Safety**		**Appropriateness**		**Trust**	
	χ^2 (df)	p	χ^2 (df)	p	χ^2 (df)	p	χ^2 (df)	p	χ^2 (df)	p	χ^2 (df)	p
Ret × Train × Access	11.09 (4)	0.026										
Cons × Ret	13.45 (6)	0.036	16.77 (6)	0.01	13.66 (6)	0.034						
Cons × Anon			10.84 (6)	0.093	15.69 (6)	0.015	10.97 (6)	0.089	18.17 (6)	0.006		
Cons × Anon × Train	19.42 (12)	0.08	26.22 (12)	0.01	20.1 (12)	0.065	27.46 (12)	0.007			26.53 (12)	0.01
Cons × Train × Topic							17.47 (6)	0.008	13.85 (6)	0.031	32.84 (6)	<.001
Cons × Train × Access	13.97 (6)	0.03	12.81 (6)	0.046								
Access × Anon	8.02 (2)	0.018										
Mgmt × Topic × Train	5.37 (2)	0.068			9.36 (2)	0.009	8.25 (2)	0.016	8.6 (2)	0.014		
Mgmt × Topic × Ret	12.66 (2)	0.002									5.60	0.061



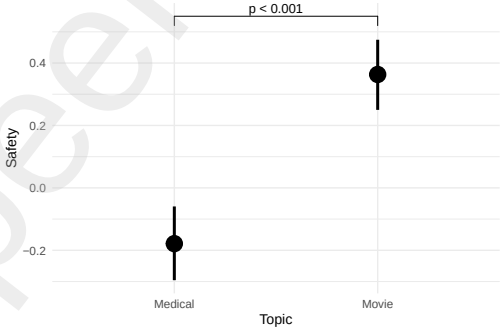
(a) Effect of Topic on Participation.



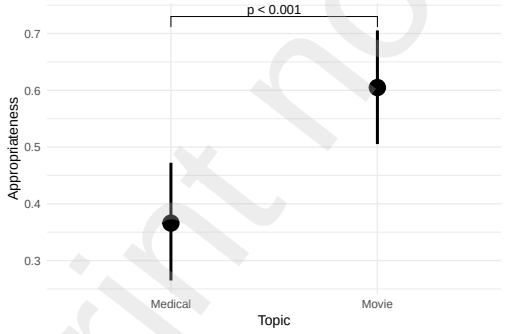
(b) Effect of Topic on Comfort.



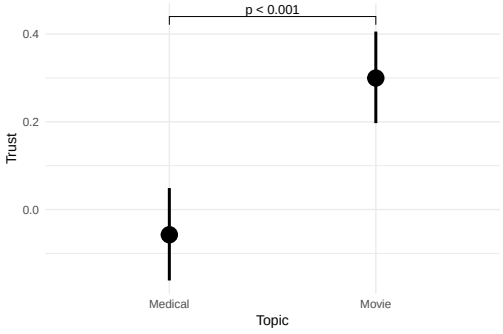
(c) Effect of Topic on Usefulness.



(d) Effect of Topic on Safety.



(e) Effect of Topic on Appropriateness.



(f) Effect of Topic on Trust.

Figure 2: Main effects of Topic Sensitivity.

more likely to not care about the data privacy when dealing with movies recommendations over medical devices" (P35). These responses illustrate how participants calibrated privacy expectations to perceived domain sensitivity.

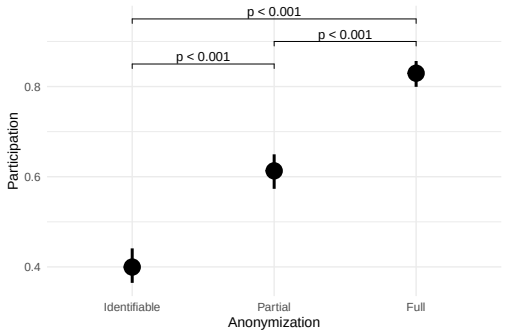
4.2.2. Data Anonymization

Our omnibus tests (see Table 7) revealed that data anonymization exerted the strongest effect on participation willingness and all attitudinal evaluations. Figure 3 shows that participants were least willing to participate in (and held the most negative attitudes toward) studies in which their data was identifiable (i.e., their conversations are linked to their name and e-mail address). Our post-hoc analysis showed that participants were significantly more willing to participate in (and held more positive attitudes toward) studies in which their data was partially anonymized (i.e., directly identifying details are removed, but certain aspects of the interaction may still reveal their identity), and they were further significantly more willing to participate in (and held more positive attitudes toward) studies in which their data is fully anonymized.

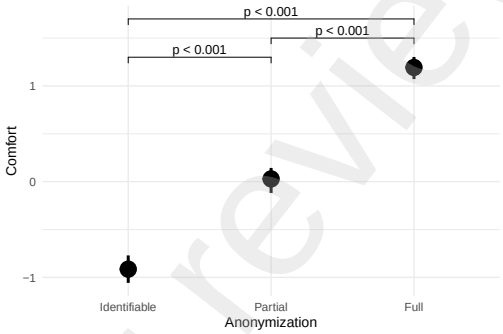
Participants understood anonymization not merely as a technical feature but as a meaningful safeguard that shaped their sense of vulnerability, trust, and participation willingness. In their open-ended feedback, participants consistently framed identifiability as the boundary between acceptable and unacceptable risk. As one noted, *"I would prefer to not be identified, even though that makes some studies more credible. I don't like my data out there floating around for just anyone"* (P67). Another emphasized a similar logic, stating that *"the most important factor was the anonymized status of my data,"* and that this concern often overrode even the nature of the topic itself, which could feel *"too personal or serious to be handled"* in certain contexts (P76). Several participants framed anonymization as a prerequisite that determined whether other governance features even mattered. As one noted, *"I participated if my data was fully anonymous. I also participated if the data would not be stored for a long time"* (P83), suggesting anonymization enabled tolerance of other potentially concerning parameters.

4.2.3. Data Retention

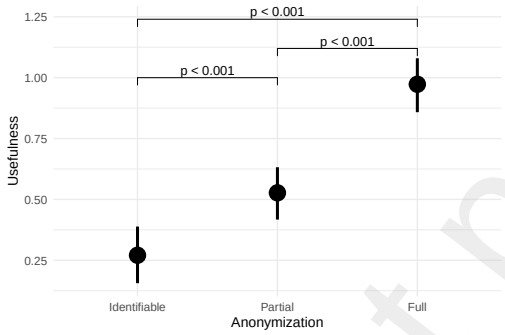
The data retention period had a significant effect on participation willingness and all attitudinal evaluations except usefulness (see Table 7). Figure 4 shows that participants were less willing to participate (and held more negative evaluations) in studies with longer retention periods. Our post-hoc tests found that indefinite data retention led to significantly lower participation



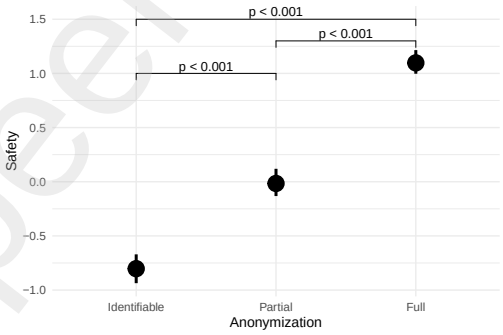
(a) Effect of Anonymization on Participation.



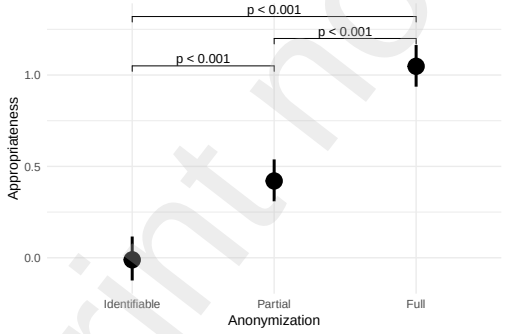
(b) Effect of Anonymization on Comfort.



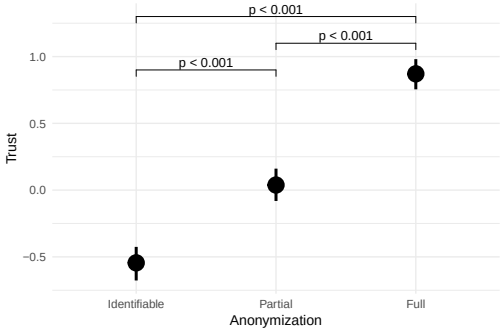
(c) Effect of Anonymization on Usefulness.



(d) Effect of Anonymization on Safety.



(e) Effect of Anonymization on Appropriateness.



(f) Effect of Anonymization on Trust.

Figure 3: Main effects of Data Anonymization.

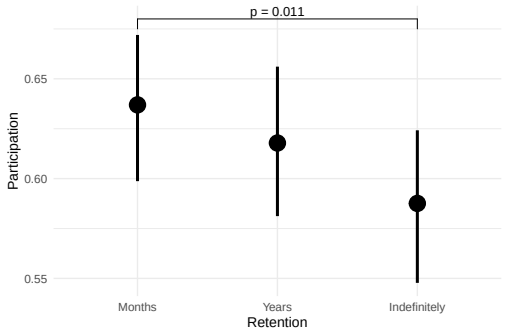
willingness than a data retention period of 2 months, and significantly lower comfort, safety, appropriateness, and trust than a data retention period of either 2 months or 4 years. No significant differences were found between 2 months and 4 years. Thus, while the overall trend is negative, significance primarily emerged when data was retained indefinitely. Participants' open-ended responses confirmed that indefinite retention heightened their concerns. One noted: *"More likely to participate if data fully anonymized and deleted after a time"* (P53), while another emphasized *"how long the data would be retained"* (P19) as a key factor in their decisions.

Table 9 also reports an interesting significant three-way interaction between data retention, model training and additional data access (i.e., who gets access to the data for secondary use) for participation willingness. Figure 5 illustrates this effect, showing that the negative effect of indefinite retention on participation willingness was most pronounced when data was not used for model training and was shared with both original and external researchers.

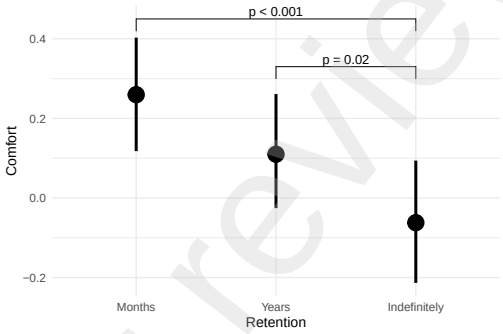
4.2.4. Model Training

Participant data use for model training had a significant effect on participation willingness, usefulness, and appropriateness (see Table 7). Figure 6 shows that, despite potential concerns about data reuse, participants responded *positively* to model training. Specifically, our post-hoc analysis showed that participants were significantly more willing to participate when their data was used for personalization (i.e., "to tailor [the LLM's] responses to your specific needs"), and they rated studies that used their data to improve model accuracy (i.e., "to improve the general accuracy of [the LLM's] responses") significantly more appropriate than studies without training use. They also rated both training conditions significantly more useful than no training.

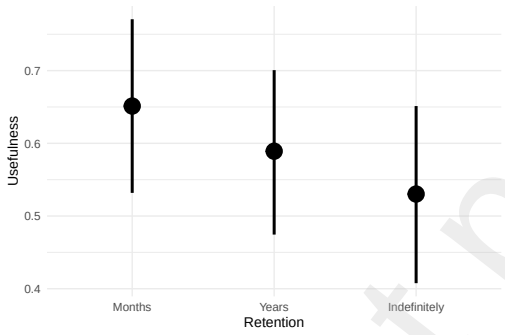
Participants' open-ended responses revealed that perceived personal benefit—rather than altruism or abstract societal good—shaped their relatively positive stance toward model training. Concrete utility participants might gain from improved systems outweighed less salient technical risks—a trade-off aligning with the privacy calculus and "privacy-personalization paradox" (Awad and Krishnan, 2006; Chellappa and Sin, 2005). One participant articulated this directly: *"I do prefer that my answers be used in order to improve the AI, though (versus not). Improving the models with real-world input is very important"* (P8). Another noted dismissively, *"I believe in AI training and I find it silly for people to be concerned with their 'data' when it is anonymized"* (P5), suggesting that perceived risk mitigation (via anonymization) combined



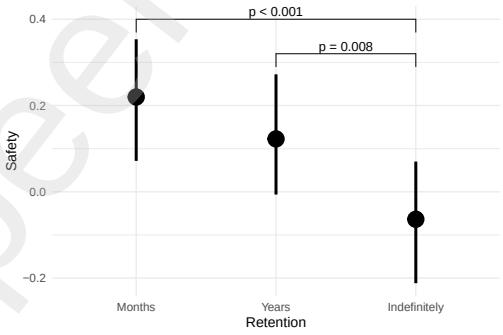
(a) Effect of Retention on Participation.



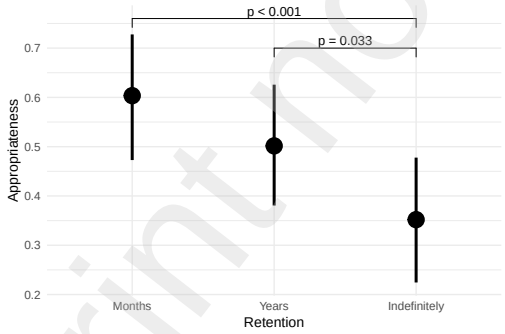
(b) Effect of Retention on Comfort.



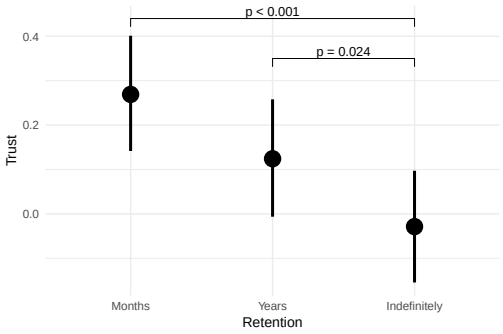
(c) Effect of Retention on Usefulness.



(d) Effect of Retention on Safety.



(e) Effect of Retention on Appropriateness.



(f) Effect of Retention on Trust.

Figure 4: Main effects of Data Retention.

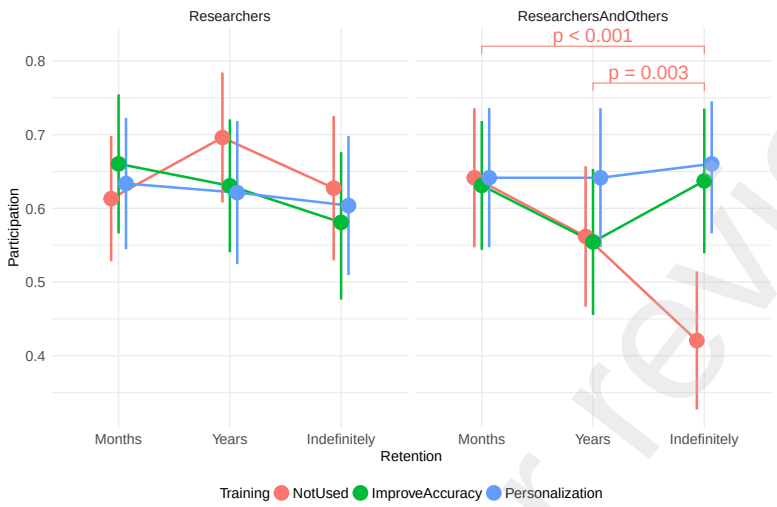
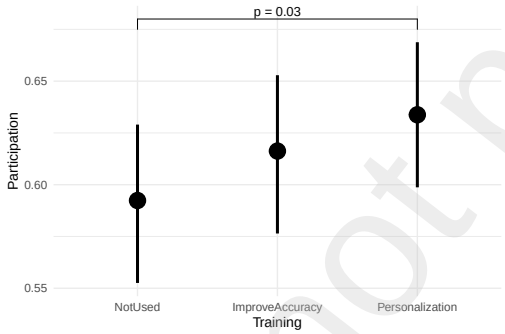
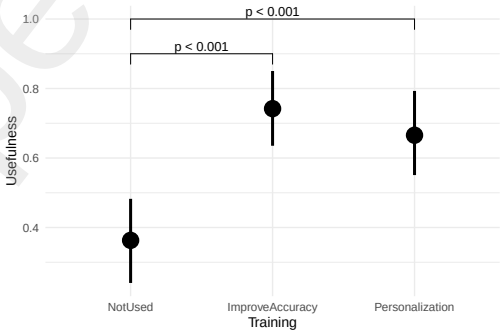


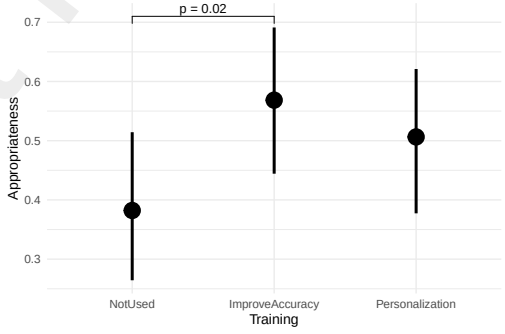
Figure 5: Three-way interaction effect between Data Retention, Model Training, and Additional Data Access on Participation willingness.



(a) Effect of Training on Participation.



(b) Effect of Training on Usefulness.



(c) Effect of Training on Appropriateness.

Figure 6: Main effects of Model Training.

with personal benefit justified the use of model training.

Note that the three-way interaction between data retention, model training and additional data access for participation willingness (see Table 9) is relevant here as well: Figure 5 shows that the negative effect on participation willingness of not using data for model training primarily exists when data was retained indefinitely and shared for additional purposes with the original researchers as well as other researchers and partners.

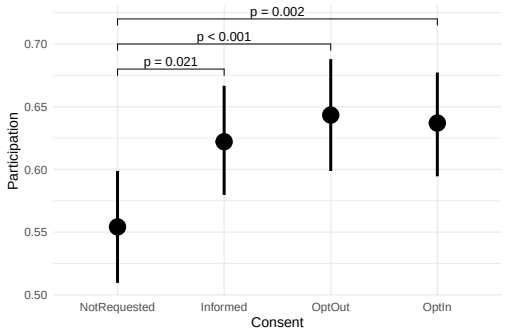
4.2.5. Additional Consent

The way the described study handled additional consent for secondary data use had a significant effect on participation willingness and all attitudinal evaluations except usefulness (see Table 7). This parameter varied how participants were asked for consent if their data were to be reused for secondary purposes: In the *Not Requested* condition, participants were told that no further consent would be sought beyond their initial agreement to the study. The *Informed* condition involved notifying participants if their data were to be used for additional purposes beyond the original study. In the *Opt-Out* condition, participants were told they would be notified of potential future uses of their data and given the opportunity to decline them. Finally, in the *Opt-In* condition, participants were notified of possible future uses of their data and asked to explicitly approve them.

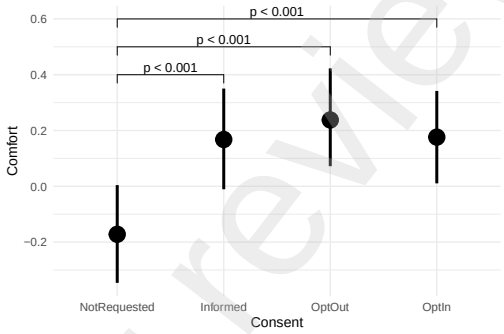
Figure 7 presents the differences between these four additional consent mechanisms. While the post-hoc tests showed significantly lower levels of participation willingness, comfort, safety, and trust when no additional consent was requested compared to the other additional consent mechanisms, there are no significant differences among the informed, opt-out, and opt-in consent mechanisms. These findings suggest that while users deem it important to be informed about additional uses of their data, they do not require the opportunity to control such additional uses of their data.

Importantly, differences between the consent conditions depended on the values of several other parameters. These interaction effects are listed in Table 9 and described below.

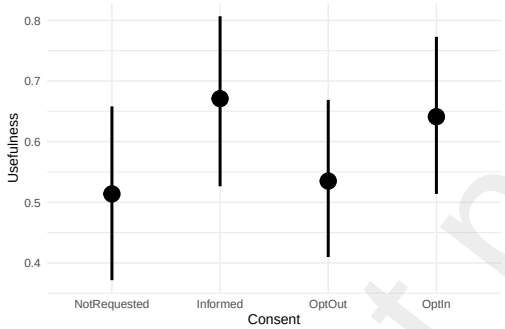
Consent × Retention.. Table 9 shows significant omnibus interaction effects between additional consent and data retention for participation willingness, usefulness, and comfort. Figure 8 presents these interaction effects. In general, there are larger differences between the consent conditions when the data is retained for a longer time period (i.e., the post-hoc tests for participation



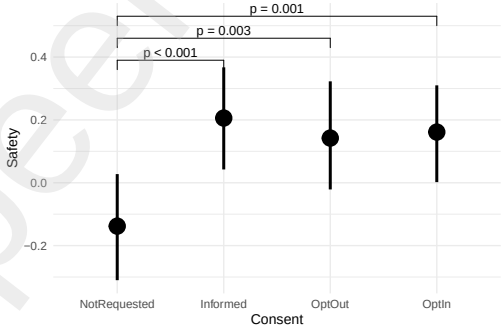
(a) Effect of Consent on Participation.



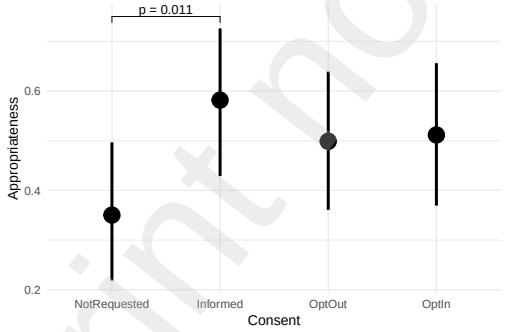
(b) Effect of Consent on Comfort.



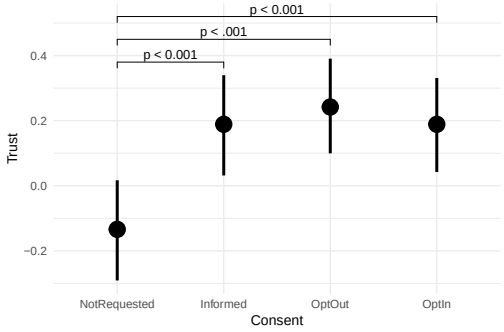
(c) Effect of Consent on Usefulness.



(d) Effect of Consent on Safety.



(e) Effect of Consent on Appropriateness.

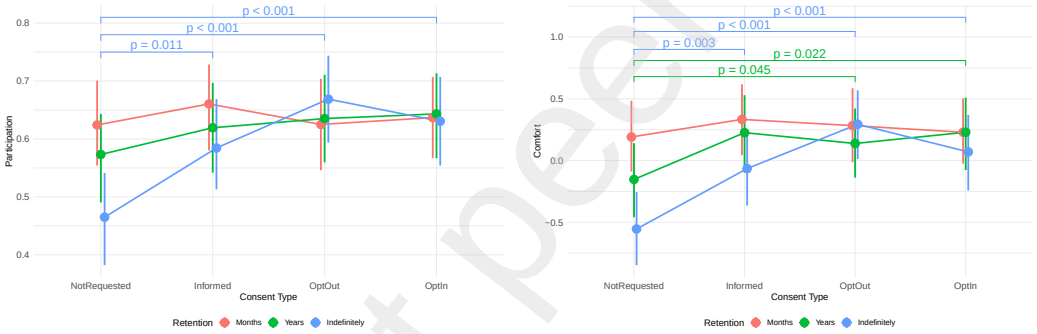


(f) Effect of Consent on Trust.

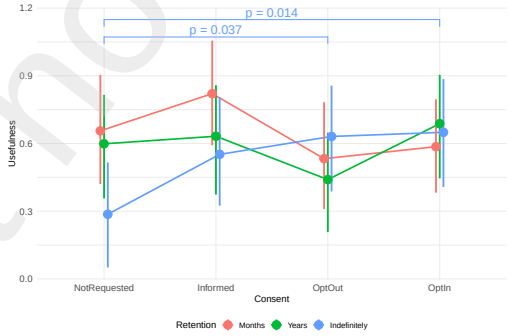
Figure 7: Main effects of Additional Consent.

willingness and usefulness of the consent conditions were only significant when retention was indefinite; for comfort they were only significant when retention was either 4 years or indefinite).

These results suggest that the importance of additional consent grows with the duration of retention: Short-term data retention appears acceptable without additional consent mechanisms, but when data is retained indefinitely participants want to at least be informed about additional uses of their data. One participant's reasoning exemplified this logic: *"It depended on how long they were keeping my data, if my information was public and for how long they were keeping my data for. Some of the ones that made my information public I didn't mind because there was an option to opt out while others made me feel uncomfortable"* (P38).



(a) Effect of Consent and Retention on Participation. (b) Effect of Consent and Retention on Comfort.

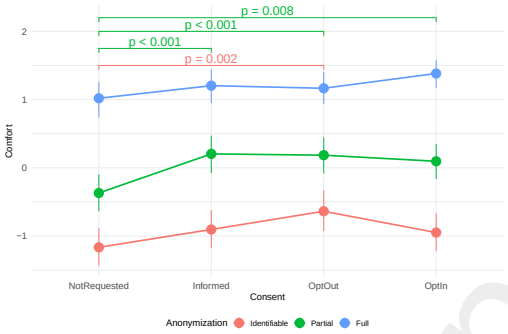


(c) Effect of Consent and Retention on Usefulness.

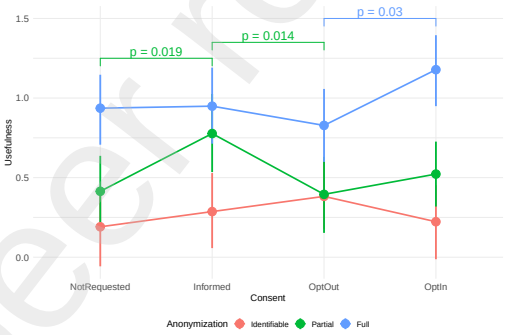
Figure 8: Interaction effects of Additional Consent and Data Retention.

Consent × Anonymization (× Training).. Table 9 shows significant interaction effects between additional consent and anonymization for usefulness

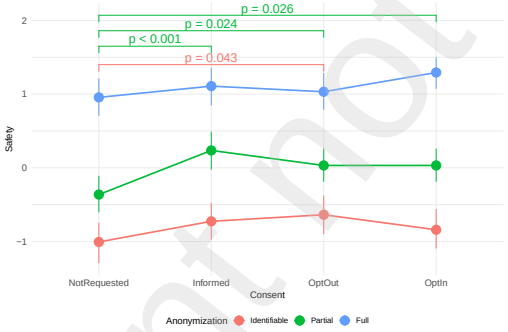
and appropriateness, with marginally significant (but consistent) interaction effects for comfort and safety. This table also shows significant three-way interaction effects between consent, anonymization and model training in terms of comfort, safety, trust, usefulness and participation willingness (with the latter two being marginally significant). Figure 9 shows the two-way interactions, while Figure A.15 in Appendix Appendix A shows the three-way interactions. Below we describe the two-way interactions, highlighting notable deviations depending on model training (i.e., the three-way interactions) where appropriate.



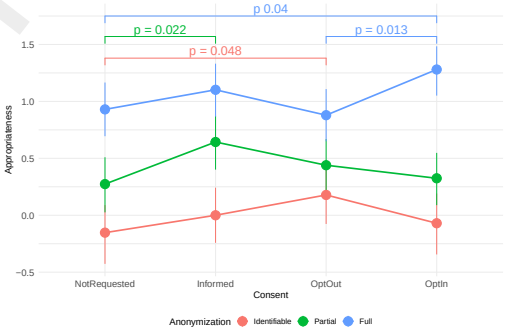
(a) Effect of Consent and Anonymization on Comfort.



(b) Effect of Consent and Anonymization on Usefulness.



(c) Effect of Consent and Anonymization on Safety.



(d) Effect of Consent and Anonymization on Appropriateness.

Figure 9: Interaction effects of Additional Consent and Data Anonymization.

Overall, we find that participants' preferences for consent mechanisms shifted depending on the degree of anonymity:

- When participants were told that their data would be *identifiable*, they favored the ability to opt-out of additional data use: Our post-hoc

tests indicated that this option significantly increased perceived comfort, safety and appropriateness compared to having no consent.

  - However, when identifiable data is *not used for training*, merely being informed about additional data use seems to be the preferred option—it significantly increases participation willingness, comfort, usefulness, safety, and trust compared to having no consent.
  - In contrast, when identifiable data is used for general *model improvements*, participants perceive higher levels of comfort and trust with the opt-in mechanism compared to merely being informed.
  - Finally, when identifiable data is used to *improve personalization*, participants are more willing to participate and perceive a higher level of comfort with the opt-out mechanism than with the opt-in mechanism.

- When data were *partially anonymized*, participants favored merely being informed about additional data use—this option significantly increased perceived comfort, usefulness, safety, and appropriateness compared to having no consent. The opt-out and opt-in mechanisms also increased comfort and safety compared to no consent, but the opt-out option led to significantly lower usefulness than simply being informed about additional data use.

  - When their partially anonymized data is used for *model improvements*, participants prefer the opt-in mechanism over no consent in terms of participation willingness, comfort, safety, and trust.

- When data were *fully anonymized*, consent mechanisms had little effect on participants, although the opt-in mechanism resulted in a significantly higher level of usefulness and appropriateness than the opt-out mechanism.

  - When their fully anonymized data is used for *model improvements*, the participants are more willing to participate with the opt-out mechanism, and perceive higher levels of trust with both the opt-out and opt-in mechanisms.

Together, these findings highlight a layered logic. The two-way interaction between consent and anonymization suggests that expectations for consent

depend on the degree of identifiability: participants want control (opt-out) when identifiability was high, transparency (informed consent) when data were partially anonymized, and showed little concern about additional data use consent when data were fully anonymized—though they still favored opt-in over opt-out. At the same time, the three-way interactions indicate that the expectations for consent also depend on model training. Participants required stronger forms of additional consent when their data were used for model improvements or personalization, particularly when the data were not fully anonymized.

Participants' reasoning revealed fundamentally different conceptions of "control"—some prioritized agency, others transparency, and many calibrated both to context. One participant emphasized the importance of decision-making authority: *"If I could opt in or out to have my information used in further research, then I was more likely to take part in a study [...]. I didn't want to participate in the study if I couldn't opt out or decide for myself if I wanted my data used in other studies"* (P80). This framing positions consent as a way to maintain autonomy over one's data even after initial participation. Others prioritized transparency, as one explained, *"whether I was anonymous and whether others had access to my info and* ***whether I would be notified****."* (P59, emphasis added). For this participant, notification of secondary data use mattered more than the formal ability to refuse. This heterogeneity suggests that no single consent mechanism satisfies all participants across all contexts.

Consent × Training × Topic.. Table 9 shows significant three-way interactions between additional consent, model training, and topic sensitivity in terms of safety, appropriateness, and trust. These effects are displayed in Figure 10. As noted earlier in this section, we find that participants calibrated their preferences regarding the consent mechanisms for additional data use to the other two parameters:

- No strong differences among consent types appeared when participants were told that their data would *not be used for model training*, although for studies about *medical advice* participants perceived higher trust when informed about additional data uses.
- When participants were told that their data would be used to *improve model accuracy*, the preferred consent mechanism depended on the topic: For studies about *medical advice*, participants perceived higher safety

and trust when given the option to opt-in, whereas for studies about *movie recommendations*, participants reported higher safety and trust when given the option to opt-out of additional data uses.

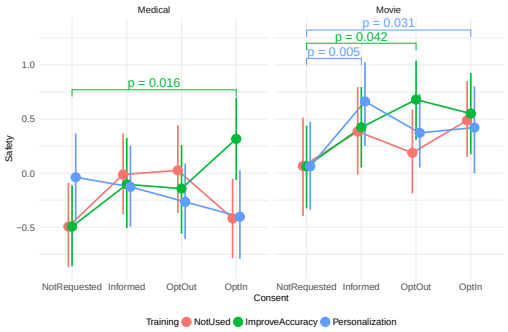
- When participants were told that their data would be used for *personalization purposes*, they perceived higher levels of safety, appropriateness and trust in studies about *movie recommendations* when informed about additional data uses. In this situation, no significant differences between consent options were found for studies about medical advice.

These results indicate that consent preferences are shaped by topic sensitivity and perceived training value: Participants expected stronger control (opt-in) when medical data used for accuracy gains, but lighter mechanisms (opt-out) sufficed for lower-stakes data or when data was not used for model training.

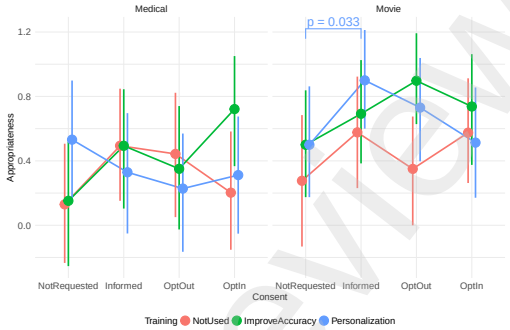
Consent × Training × Data Access.. Table 9 also shows significant three-way interactions between additional consent, model training, and additional data access (i.e., who gets access to the data for secondary use) in terms of participation and comfort. These effects are displayed in Figure 11. Again, we find that participants calibrated their preferences regarding the consent mechanisms for additional data use to the other two parameters:

- When participants were told that their data would *not be used for training* and only used for other purposes by the *original researchers*, participation willingness and comfort increased significantly when participants were told they would be informed about additional uses of their data (compared to no consent). If the data would be *shared with other researchers and partners*, participation and comfort were low in general and the additional consent mechanism did not have an impact.

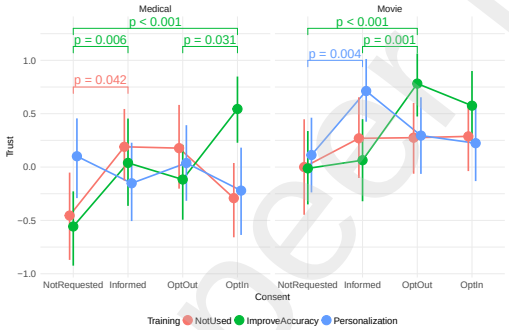
- When participants were told that their data would be used to *improve model accuracy* and only used for other purposes by the *original researchers*, participation willingness and comfort increased significantly when participants were told they would be able to opt-out of additional uses of their data (compared to no consent). If the data would be *shared with other researchers and partners*, comfort increased significantly when participants were told they would be able to opt-in.



(a) Effect of Consent, Training, and Topic on Safety.

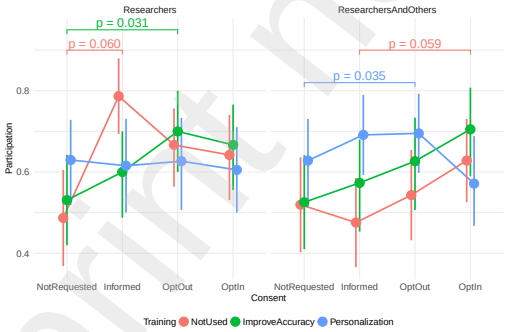


(b) Effect of Consent, Training, and Topic on Appropriateness.

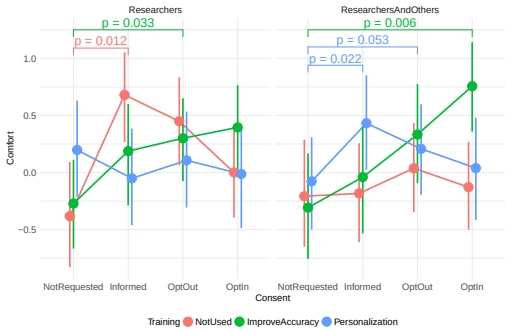


(c) Effect of Consent, Training, and Topic on Trust.

Figure 10: Three-way interaction effects between Additional Consent, Model Training, and Topic Sensitivity.



(a) Effect of Consent, Training, and Data Access on Participation.



(b) Effect of Consent, Training, and Data Access on Comfort.

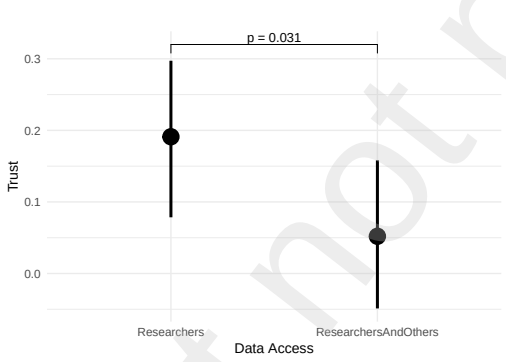
Figure 11: Three-way interaction effects between Additional Consent, Model Training, and Additional Data Access.

- When participants were told that their data would be used for *personalization purposes* and *shared with other researchers and partners*, participation willingness and comfort increased with the informed and opt-out consent mechanisms. If the data would only be re-used by the *original researchers*, participation and comfort were low in general and the additional consent mechanism did not have an impact.

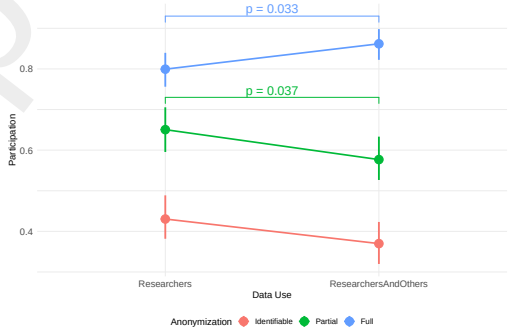
These results suggest that, in certain situations, appropriate consent mechanisms increased participation willingness and comfort; in others, both remained low regardless of consent type.

4.2.6. Additional Data Access

Our omnibus tests showed that additional data access (i.e., whether only the original research team or also other researchers and industry partners get access to the data for secondary use) had a small but statistically significant effect on trust (see Table 7). As shown in Figure 12a, participants reported significantly higher trust when their data could only be used for additional purposes by the original researchers.



(a) Effect of Data Access on Trust.



(b) Effect of Data Access and Anonymization on Participation.

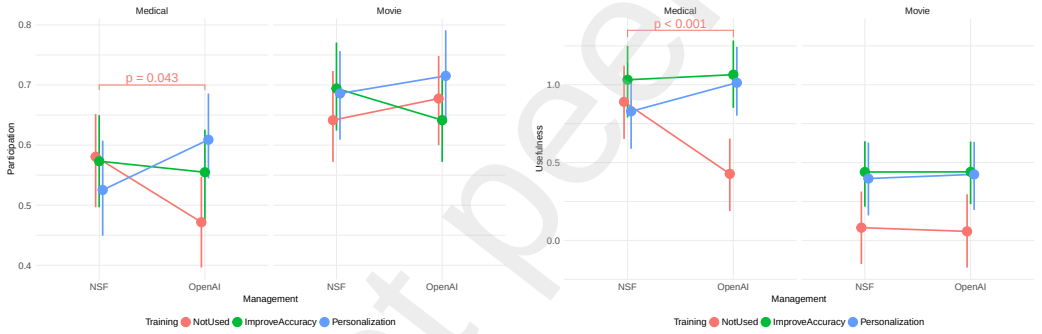
Figure 12: Main and interaction effects of Additional Data Access.

Data Access × Anonymization.. While the main effect of additional data access on participation willingness is not significant, we found a significant interaction between anonymization and data access on willingness to participate (see Table 9). Figure 12b illustrates this effect, showing that when participants were told that their data was identifiable or partially anonymized, they were less willing to participate if they were told that their data could be

used for additional purposes by other researchers and partners as compared to the original researchers only. In contrast, when data were fully anonymized, participants were significantly more willing to allow such broader sharing of their data with other researchers and partners.

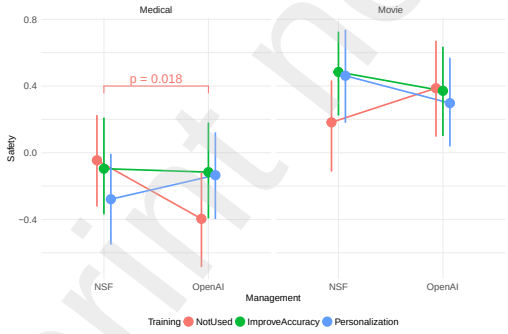
4.2.7. LLM Management

Our omnibus tests (see Table 7) showed that LLM management (i.e., whether the system in the scenario was operated by a federal research agency or by a private AI company) had no significant main effect on participation willingness or the attitudinal evaluations. However, Table 9 shows that we did find several interaction effects with LLM management, suggesting that significant differences emerged depending on the values of other parameters. These interactions are discussed below.

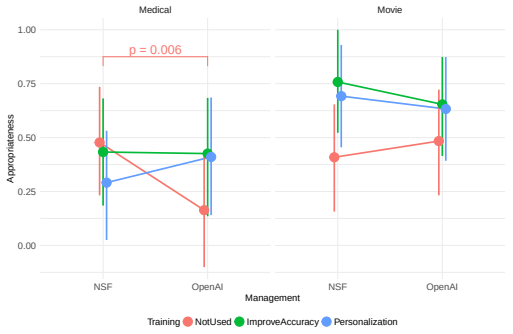


(a) Effect of Management, Topic, and Training on Participation.

(b) Effect of Management, Topic, and Training on Usefulness.



(c) Effect of Management, Topic, and Training on Safety.



(d) Effect of Management, Topic, and Training on Appropriateness.

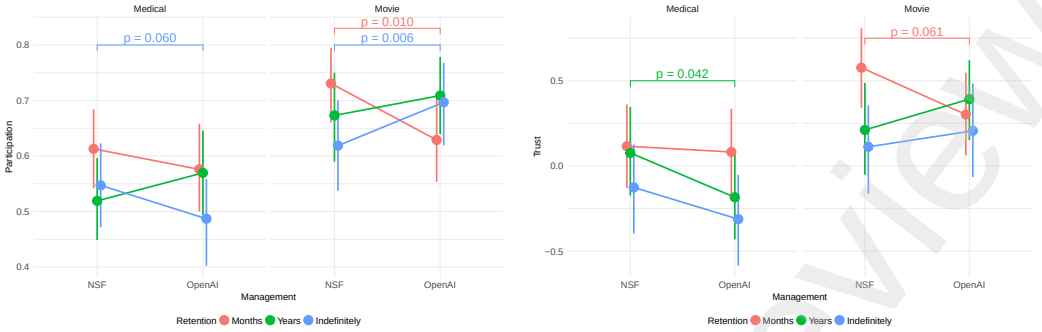
Figure 13: Three-way interaction effects between LLM Management, Topic Sensitivity, and Model Training.

Management × Topic × Training.. Table 9 shows significant interactions between LLM management, topic sensitivity, and model training for usefulness, safety, appropriateness, and (marginally) participation willingness. Figure 13 illustrates a consistent pattern of significant post-hoc differences: When the study topic was *medical advice* and participants were told that their data would *not be used for model training*, they showed a significantly lower levels of participation willingness, usefulness, safety, and appropriateness when the LLM was managed by OpenAI rather than by the NSF. In contrast, when training was framed as serving accuracy improvement or personalization, and/or when the study topic was movie recommendation, the differences between NSF- and OpenAI-managed LLMs were not significant.

These results reveal a nuanced pattern: Participants were most wary of private LLM management of sensitive (medical) data in the absence of training benefits. Arguably, participants favored an NSF-managed LLM in these situations because they associated the NSF with public-interest research rather than commercial LLM development.

Management × Topic × Retention.. Table 9 also shows significant interactions between LLM management, topic sensitivity, and data retention for participation willingness and (marginally) trust. Figure 14 shows a consistent pattern of significant post-hoc differences: When the study topic was *movie recommendations* and the retention period was *2 months*, participants showed a significantly lower willingness to participate and (marginally) lower trust when the study used an LLM managed by OpenAI rather than the NSF. Conversely, when movie data was retained indefinitely, participants reported a significantly higher willingness to participate when the study used an LLM managed by OpenAI rather than the NSF.

Participants' open-ended responses revealed that institutional trust was conditional and intertwined with other governance concerns. One participant's reasoning illustrated this conditionality: *"I somewhat trust the NSF more than I do open AI. I prefer to be completely anonymous, but if certain things I say could identify me, I chose 'no' on some of those. It depended on the topic of the study, too. I may or may not want my medical questions leaked"* (P84). This participant calibrated trust alongside anonymization protections and topic sensitivity rather than deferring to institutional reputation alone. Conversely, some participants expressed categorical rejection: *"I don't believe LLMs should be used in Medicine in any capacity. Additionally, the NSF shouldn't need data for movie recommendations. I also do not trust OpenAI in the slightest"*



(a) Effect of Management, Topic, and Retention on Participation.

(b) Effect of Management, Topic, and Retention on Trust.

Figure 14: Three-way interaction effects between LLM Management, Topic Sensitivity, and Data Retention.

(P1). For this participant, institutional affiliation was insufficient to overcome fundamental doubts about the appropriateness of LLM use in certain domains. These responses suggest that institutional trust operates as one component in a multifaceted evaluation including data sensitivity, governance safeguards, and normative judgments about technology appropriateness.

4.2.8. Heterogeneity in Decision-Making

While our quantitative findings reveal systematic patterns in how study parameters shaped attitudes and participation willingness, participants' open-ended responses reveal substantial heterogeneity in reasoning, risk tolerance, and governance priorities. At one end of the spectrum, some participants expressed comfort bordering on indifference toward data sharing. One stated simply, *"I had nothing to lose or hide"* (P22), while another reported: *"I have been participating in supplying data for LLMs and studies since 2020. I will continue to do so. I honestly do not care who has my data and what they do with it"* (P23). For these individuals, privacy risks appeared negligible. At the other end, some participants expressed fundamental skepticism about LLM technology, independent of governance arrangements. One stated: *"I don't trust LLMs because of their tendency to hallucinate, and would not want to contribute to their development in any way"* (P3). Another reflected a more general caution: *"AI is simply too young and too risky to justify willfully giving my information away so easily. Even in the low stakes situations"* (P2). For these participants, governance improvements were unlikely to shift their participation calculus.

The majority of participants, however, demonstrated nuanced, context-dependent reasoning (reflected throughout Section 4.2): they weighed multiple parameters simultaneouslyand engaged in trade-off reasoning between utility and protection. This heterogeneity underscores an important limitation of aggregate findings: while our statistical models identify parameters that matter *on average* across the sample, individual participants varied substantially in their governance priorities, risk tolerance, and decision-making logic.

5. Discussion

Our results show that the design parameters of Human-LLM Conversational Interaction (HLLMCI) research studies substantially affect participants' attitudinal evaluations and participation willingness. In this section, we discuss the implications of our work for researchers and human-subjects research review boards. We structure the discussion into three parts: general implications, implications for specific study design parameters, and recommendations for practice.

5.1. General Implications

5.1.1. Participants make deliberate decisions about study participation

Our mediation analysis (Section 4.1) demonstrates that most study parameters significantly impact participation willingness, and that these effects are fully mediated by participants' attitudinal evaluations. These results indicate that participants' willingness to participate in the described study was not rooted in heuristic decision-making, but rather a deliberate decision shaped by *how the study's characteristics made them feel*. These attitudinal evaluations acted as mediators that translated study policy descriptions into experiential judgments, ultimately guiding participants' willingness to engage in HLLMCI research. Multiple interaction effects (in Section 4.2 and Table 9) and participants' open-ended responses revealed that this deliberative process often involved sophisticated multi-factor reasoning: participants engaged in a 'privacy calculus' (Dinev and Hart, 2006), weighing combinations of parameters rather than evaluating each in isolation. This underscores that consent process is not merely procedural but involves deliberate decision-making. Researchers should treat research design choices as critical input and communicate them explicitly.

5.1.2. Optimizing study participation is a complex task

While for some study design parameters is it relatively straightforward to select the value that optimizes participation willingness and participant comfort, for other parameters this decision is all but straightforward. The various two- and three-way interaction effects presented in Section 4.2 suggest that, for instance, the best way to manage additional consent depends on the values of several other parameters.

Furthermore, participants' open-ended responses suggest considerable heterogeneity in reasoning, underscoring that while participants engage deliberately with governance parameters, the relative weight they assign to different factors varies substantially across individuals. This heterogeneity suggests that one-size-fits-all governance approaches may fail to accommodate diverse participant expectations, and future work should explore governance frameworks that can accommodate this heterogeneity (e.g. user-tailored privacy (Knijnenburg et al., 2022)).

5.1.3. Balancing participant preferences against protective standards

Researchers seeking to optimize their study design are encouraged to use our results as guidance for their research design process. However, optimizing participation should not supersede ethical obligations. Parameters that participants find less concerning may still pose meaningful risks that warrant protection. For instance, participants viewed model training positively, expecting personal benefits from LLM improvements, yet model training carries risks—including potential data leakage, memorization, or unintended disclosure—that may not be fully apparent even when disclosed. Similarly, participants showed relatively less concern about commercial management or longer data retention periods, but these choices can meaningfully affect long-term data security and secondary use.

This highlights a core principle: participant indifference should not be interpreted as blanket permission to minimize protections. Researchers and IRBs must balance participant preferences against protective standards, ensuring governance choices meet ethical requirements even when participants express comfort with weaker safeguards. Where participants' preferences appear to conflict with their protection—such as valuing personalization benefits while underestimating training risks—researchers face a dual responsibility to: (1) educate participants about the implications of data practices, and (2) explore alternatives that address participants' needs (e.g., personalization benefits) while maintaining robust protections (e.g., federated learning, differ-

ential privacy, or on-device processing (Friedman et al., 2015; Kobsa et al., 2014)).

5.1.4. A need for tiered system design and human-subjects research design policies

Although our study primarily focused on research practices, the findings also have implications for system and research policy design. Specifically, AI-related (research) policies may need to be tiered by topic sensitivity: stricter retention limits, stronger anonymization, and tighter training restrictions may be appropriate for sensitive domains such as health or criminal justice, whereas lighter governance mechanisms may suffice for less sensitive topics like entertainment. Such differentiated governance aligns with participants' evaluations and could better balance system/research value with participant protection.

In the United States, a tiered system of human-subjects research governance already exists, in the form of different levels of IRB review. A similar tiered (or context-specific) approach may benefit AI system legislation more broadly. Preliminary examples of this are the recent state-level restrictions on the use of AI in the context of mental healthcare decision-making (Slaby, 2025). For HCI researchers, this suggests a dual responsibility to not only comply with governance requirements but also design consent processes and data practices that transparently communicate these tiers to participants.

5.2. Implications About Specific Study Design Parameters

5.2.1. Anonymization

Anonymization was the strongest predictor of participation willingness and attitudinal evaluations, outweighing other study design parameters that have traditionally been considered important for protecting research participants, such as data retention or consent (Mascalzoni et al., 2022; Sanderson et al., 2017). Average participation willingness reached over 80% for anonymized studies—markedly higher than 60% for partially anonymized and 40% for identifiable data (see Figure 3a).

These results suggest that developing robust anonymization practices is one of the most powerful ways to foster participation in HLLMCI research. Yet, guaranteeing full anonymity in conversational LLM settings is extremely challenging, as free-text disclosures can contain indirect identifiers (Shim et al., 2024). In practice, the best that HLLMCI research can often offer is partial anonymization, complemented by careful redaction, minimization

of human review, and technical safeguards. Future work should focus on developing and validating methods of anonymization specific to conversational AI studies, given that they differ from structured data settings. For now, we do note that a best-effort (i.e., partial) anonymization already improves participation willingness substantially over studies that leave users' data fully identifiable.

Another important finding related to anonymity can be seen in Figure 12b: Participants were *more* willing to participate in studies that share data access with other researchers and partners (as compared to just the original researchers) if their data was anonymized, but less willing to participate when the data shared was partially anonymized or identifiable. This suggests that participants view anonymization as a critical condition for secondary data access, and that when their data is properly anonymized, they are more likely to participate if their data can benefit a larger number of researchers.

5.2.2. Model training and LLM management

Despite potential privacy concerns about data leakage (Inan et al., 2021), participants were *more* willing to participate in studies that used data for model training (Section 4.2.4). Participants appeared not to perceive privacy concerns associated with training (effects on comfort, safety and trust were not significant), while perceiving clear benefits (significant effects on usefulness and appropriateness). Open-ended responses, such as *"I do prefer that my answers be used in order to improve the AI [...]. Improving the models with real-world input is very important"* (P8), suggested they weighed personal utility more heavily than potential risks.

In terms of LLM management, contrary to expectations (Kibriya et al., 2024), we did not observe consistent advantages if the LLM used in the described study was managed by a public institution (e.g., NSF) over a private provider (e.g., OpenAI). Instead, Section 4.2.7 shows that the effects were context-dependent. When medical data were not used for training, participation willingness was higher for studies using LLMs backed by public institutions, aligning with expectations of collective societal benefit. However, when data *were* used for model training or personalization, participation willingness was higher for studies using a commercially managed LLM. The latter was also true for movie data, especially with longer retention periods.

These counterintuitive findings may have been shaped by considerations of direct personal benefit: improvements to ChatGPT—which most participants already used—would arguably offer more immediate utility than improvements

to a hypothetical NSF-managed LLM. Participants' open-ended responses provide some support for this interpretation, although alternative explanations remain possible. This suggests that consent communications should not only emphasize societal benefits (a practice that is increasingly common in HCI research (Hochheiser and Lazar, 2007)), but to also articulate the potential personal benefits that participants may receive, particularly in domains where such individual utility is salient. We note, however, that participants' positive attitudes toward model training may partly reflect our benefits-oriented framing and participants' limited awareness of associated risks such as data leakage or memorization (Zhang et al., 2024). This positive response to model training illustrates a critical tension: participants' expressed preferences may not align with their best interests when risks are technical, long-term, or difficult to communicate concisely. While this does not diminish the value of transparency or participant agency, ethical research governance cannot rest solely on participant acceptance. Even when participants express comfort with practices like model training, researchers and IRBs retain obligation to assess and mitigate risks that may exceed participants' awareness or concern.

5.2.3. Consent

Providing notice of secondary data uses substantially improved participation willingness and attitudinal evaluations (see Section 4.2.5), while active opt-out of or opt-in mechanisms generally provided no further improvement. While HCI literature often emphasizes the value of granular notice and consent practices (see Knijnenburg et al., 2013, for a critical discussion of this notion), these findings suggest that disclosing secondary data use practices carries much more weight than providing active consent mechanisms. Several participants' open-ended reflections similarly prioritized knowing about data practices over controlling them.

Arguably, the promise to provide notice places the burden of responsibility on researchers to notify participants of secondary data use, and this commitment may serve as a perceived deterrent from inappropriate secondary uses (cf. O'Connell and Church, 2024), instilling confidence that researchers will exercise careful judgment. Asking participants to actively accept or decline secondary data uses, in contrast, puts the burden of control on the participant, who may not be interested in taking on this burden, unless (as evidenced by our uncovered significant interaction effects) the collected data is sensitive, contains identifiers, and/or is used for model training.

5.2.4. Retention

The length of the data retention period had a significant effect on participation willingness and most attitudinal evaluations (see Section 4.2.3), but the effect was only significant when data was described as retained indefinitely (see also Leon et al., 2013); something most privacy guidelines advise against (Blanchette and Johnson, 2002). Moreover, our significant interaction effects showed that shorter retention reduced reliance on other governance mechanisms: Participants expressed greater tolerance of limited protections when data were stored only briefly. This reinforces the importance of not only limiting data retention but also communicating retention periods explicitly in consent or notice statements.

5.2.5. Topic Sensitivity

Topic sensitivity surfaced as a key influence as well as a moderator. Medical advice studies were seen as highly sensitive yet highly beneficial, producing a notable trade-off between perceived safety and usefulness (reflecting a recent finding in chatbot privacy research (Sannon et al., 2020)). By contrast, movie recommendation studies were seen as less sensitive but also less useful. Overall, participants were less willing to participate in studies on medical advice despite the higher perceived benefit, suggesting that sensitivity may outweigh perceived utility in research participants' privacy calculus (Laufer and Wolfe, 1977).

Notably, the effects of several other parameters (i.e., consent and training, management and training, and management and retention) differed by topic, indicating that optimal research design practices may not generalize across domains.

5.3. Implications for Practice

Our findings highlight that governance choices in HLLMCI studies directly shape how participants evaluate and decide on participation. These results carry implications for multiple stakeholders:

- **Researchers:** Governance parameters (e.g., anonymization, retention, training, consent) should be treated as design features, clearly communicated in consent materials and tested through piloting or scenario-based pretesting—especially in sensitive domains.

- **Institutional Ethics Boards:** Supporting researchers in balancing participant preferences with robust protections is essential to ethical review. IRBs can move beyond compliance checks by providing checklists and structured feedback that explicitly incorporate governance dimensions, requiring researchers to justify choices even when participants express indifference, and imposing protective standards—particularly for sensitive domains—that exceed baseline participant expectations.

- **System Designers:** Transparency should be built into LLM interfaces through clear notices about training use, retention timelines, or data access, helping participants understand governance choices in real time.

- **Policymakers:** Context-sensitive frameworks that impose stricter safeguards in sensitive domains while allowing flexibility where anonymization or limited retention are guaranteed may better reflect participant perspectives than one-size-fits-all rules.

Our findings also carry implications for consent form design. Our concise, structured scenarios—clearly highlighting key governance parameters—enabled deliberate, informed decisions, contrasting with typical consent forms that are lengthy and laden with legal language (Pearman et al., 2022). Including a brief, structured summary at the beginning of consent forms could serve as an accessible entry point that help participants understand and evaluate studies before encountering detailed legal requirements. Future research should examine how consent form structure affects comprehension and attitudes in real-world LLM research contexts.

6. Limitations and Future Work

Scenario-based studies have inherent limitations, and our work is no exception. Below we address notable limitations and provide directions for future work.

First, our scenario-based study does not capture actual participation decisions in real-world Human-LLM Conversational Interaction (HLLMCI) research studies. However, our factorial scenario design provides a unique opportunity to systematically compare (combinations of) study design decisions at scale, which would be impractical or unethical to manipulate in the real world. While our scenario descriptions were necessarily concise (to prevent skimming behavior and participant fatigue), they closely resembled the

amount and type of information in typical consent forms, though real-world consent documents are often longer and more complex (Pearman et al., 2022). Our study may thus represent a best-case scenario where participants attend carefully to governance details. However, this design choice also represents a methodological strength: our mediation analysis (Section 4.1) demonstrates that when governance parameters are presented clearly and concisely, participants engage in deliberate, attitudinally-mediated decision-making rather than heuristic judgments. This suggests that participants *can* meaningfully evaluate governance choices when information is accessible.

Second, our findings derive from scenarios describing human-LLM conversational interaction research studies—contexts where participants directly engage with LLMs through conversational interfaces. While some principles (e.g., transparency about data use and participant agency) may extend to other LLM research contexts, other results may not directly generalize to other modalities.For instance, when LLMs analyze participant data rather than interact with participants, concerns about conversational disclosure may be less salient, while concerns about algorithmic bias in interpretation may become more prominent. Future research should examine whether the governance parameters and attitudinal mediators we identify operate similarly across diverse LLM research contexts, and whether additional parameters (e.g., model customization, human oversight, API versus interface distinctions) play comparable roles.

Third, our parameter descriptions balanced technical accuracy with participant comprehension. For instance, our model training parameter used accessible language (e.g., "improve the general accuracy of its responses," "tailor its responses to your specific needs") that may not fully capture the technical mechanisms or potential risks such as data leakage, memorization, or unintended model behaviors (Zhang et al., 2024). While this mirrors how consent forms typically describe data practices at a high, non-technical level, more detailed risk descriptions could have increased participants' concerns. Our positive framing, combined with participants' limited familiarity with LLM data practices (Table 5), means some findings may reflect incomplete understanding rather than fully informed preferences. Future work could investigate how varying levels of detail and risk communication in parameter descriptions affect comprehension and attitudes, helping identify optimal balances between accessibility and completeness in consent communications.

Fourth, generalizability has boundaries. We tested for gender and age differences but found no meaningful moderating effects, suggesting our find-

ings apply broadly across these demographics. However, we intentionally limited recruitment to US-based participants to reduce uncontrolled cultural variation, which simplified interpretation but restricts cross-cultural applicability. Extending this work to other cultural contexts will be important, especially since norms around data governance and institutional trust differ internationally. Additionally, while our study included members of underrepresented groups, it was not powered to examine whether their attitudes and participation willingness deviated from the majority. Future work could uncover whether marginalized populations respond differently to governance mechanisms, given their historically disproportionate exposure to research harms.

Fifth, our study relied primarily on quantitative measures—multiple-choice items assessing attitudes and willingness to participate. Although we reference selected open-ended responses for context, these do not capture the full depth of participants' reasoning processes, lived experiences, or underlying values. Future studies should complement our work with in-depth qualitative methods, such as semi-structured interviews or think-aloud protocols, to surface participants' underlying concerns and decision-making strategies. Such work could clarify why certain conditions matter more to some individuals and how LLM literacy or prior experiences shape governance expectations. It may also deepen understanding of how marginalized populations navigate governance decisions, given historical research harms.

Finally, our study examined participant perspectives, but did not capture how researchers make research design decisions. Future work could investigate how researchers balance considerations of anonymization, retention, or training against practical constraints, highlighting gaps between participant expectations and researcher practices, and suggesting opportunities for alignment.

7. Conclusion

This paper examined how study design parameters in Human-LLM Conversational Interaction (HLLMCI) research studies shape participants' attitudes and willingness to participate. Our findings extend the privacy calculus by highlighting how perceptions of comfort, appropriateness, trust, and safety jointly mediate participation decisions. They also underscore that governance mechanisms are not interchangeable: Participants distinguish between transparency, control, and protection, and expect them to align with the sensitivity

and purpose of data use. More broadly, this work contributes to ongoing discussions about trustworthy AI, research ethics, and participant agency. By systematically unpacking how governance choices shape participants' privacy evaluations, we highlight pathways toward more responsible research practices that not only meet regulatory standards but also align with participants' expectations of fairness, safety, and respect.

Appendix A. Additional Interaction Effect Graphs

Appendix B. Definition of LLMs

At the start of the study, participants were shown the following definition:

> LLMs are advanced AI systems that can understand and generate human language based on extensive text data they have been trained on. Examples of such models include OpenAI's GPT and ChatGPT. These models are designed to answer questions, assist with various tasks, and generate conversational responses that mimic human-like interaction. LLMs continuously improve their language processing abilities by learning from vast datasets, allowing them to generate more accurate and relevant responses over time.
>
> LLMs are trained using publicly available data from the internet. Additionally, they can use data from the current conversation to improve the accuracy of their responses or tailor them to better meet your specific needs.

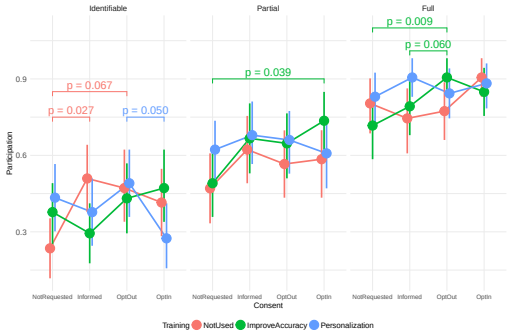
Appendix C. Post-Study Questions

This appendix provides the complete post-scenario survey questions. Summary statistics are reported in Table 5.

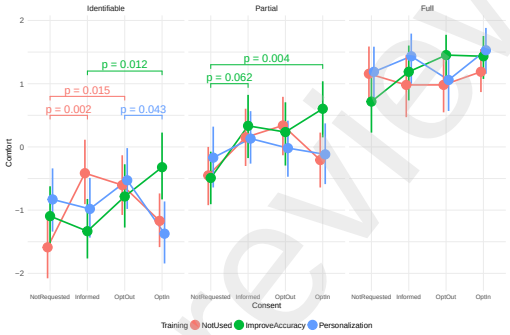
Scenario Participation Follow-Up Questions

In the previous scenarios, did you choose to participate in:

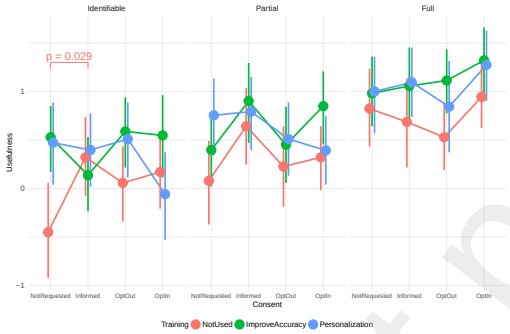
◯ None of the studies
What concerns or reasons led you to decline participation in all the studies?



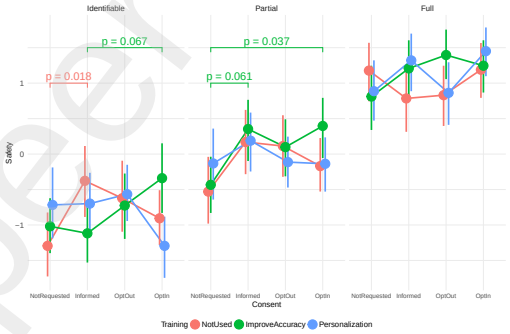
(a) Consent, Anonymization, and Training effect on Participation.



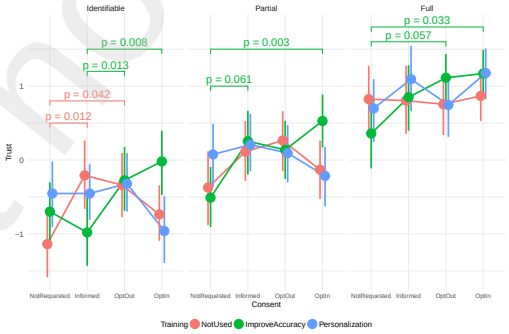
(b) Consent, Anonymization, and Training effect on Comfort.



(c) Consent, Anonymization, and Training effect on Usefulness.



(d) Consent, Anonymization, and Training effect on Safety.



(e) Consent, Anonymization, and Training effect on Trust.

Figure A.15: Three-way interaction effects between Additional Consent, Anonymization, and Model Training.

◯ Some studies
What factors influenced your decision to participate in some studies but not others?

◯ All of the studies
What motivated you to participate in all the studies?

Prior Experience and Knowledge

How would you describe your experience with LLM tools?

◯ Occasional user

◯ Frequent user

◯ Technical user (e.g. developer)

◯ None

How familiar are you with the data privacy policies of LLM tools such as ChatGPT?

◯ Very familiar

◯ Somewhat familiar

◯ Not familiar

How do you think LLM tools use your data?

◯ My data is used to provide a response to my input.

◯ My data is stored and used to train future models.

◯ My data is stored and used for quality assurance purposes.

◯ My data is stored to fulfill legal requirements.

◯ My data is stored and sold or used for commercial purposes.

◯ I don't know.

Do you know the difference between data collected from independent users versus corporate users of LLM tools?

◯ Yes

◯ No

Factual Knowledge

Please respond to the following based on your current understanding:

- Current AI technology can understand the thoughts and emotions of human beings. ◯ False ◯ True
- AI can learn from its mistakes and improve over time. ◯ False ◯ True
- AI systems are always accurate and free from errors. ◯ False ◯ True
- AI's decision is based on data. ◯ False ◯ True
- All AI systems require human supervision to operate. ◯ False ◯ True
- AI cannot personalize content for users because it treats everyone the same way. ◯ False ◯ True
- AI models require large amounts of data to improve their accuracy. ◯ False ◯ True
- AI algorithms are always unbiased, as they rely solely on mathematical computations. ◯ False ◯ True

Participation in LLM-Based Studies

How familiar are you with how data is typically handled in academic research studies?

◯ Very familiar

◯ Somewhat familiar

◯ Not familiar

Have you previously participated in any research studies that involved LLM tools?

◯ No

◯ Yes:

1. Were you informed about how your data would be used in those studies?
   ◯ No
   ◯ Yes
2. Did you have any concerns about your data privacy in those studies?
   ◯ No
   ◯ Yes

Have you ever interacted with:

- Character.AI ◯ No ◯ Yes
- ChatGPT ◯ No ◯ Yes
- Claude ◯ No ◯ Yes
- DeepSeek ◯ No ◯ Yes
- FakeSeek[4] ◯ No ◯ Yes
- GitHub Copilot ◯ No ◯ Yes
- Google Bard ◯ No ◯ Yes
- Google Gemini ◯ No ◯ Yes
- Grok ◯ No ◯ Yes
- Jasper ◯ No ◯ Yes
- Meta Llama ◯ No ◯ Yes
- Microsoft Bing AI ◯ No ◯ Yes
- Microsoft Copilot ◯ No ◯ Yes
- Perplexity ◯ No ◯ Yes
- Snapchat My AI ◯ No ◯ Yes

[4]Attention check - a fake option to detect inattentive responses.

Appendix D. Demographics

Please select your age group:

◯ 18–24

◯ 25–34

◯ 35–44

◯ 45–54

◯ 55 and above

Select the gender(s) or gender identities that currently apply to you:

◯ Man (including Trans Male/Trans Man)

◯ Woman (including Trans Female/Trans Woman)

◯ Nonbinary

◯ Prefer to self describe

◯ Prefer not to disclose

Please specify your ethnicity:

◯ Caucasian/White

◯ African American/Black

◯ Hispanic/Latino

◯ Asian/Pacific Islander

◯ Native American/Alaska Native

◯ Mixed ethnicity

◯ Other (please specify)

Which of these is the highest level of education you have completed?

◯ No formal qualifications

◯ Secondary education (e.g., GED/GCSE)

◯ High school diploma/A-levels

◯ Technical/community college

◯ Undergraduate degree (BA/BSc/other)

◯ Graduate degree (MA/MSc/MPhil/other)

◯ Doctorate degree (PhD/other)

What is your current employment status?

◯ Full-Time

◯ Part-Time

◯ Due to start a new job within the next month

◯ Unemployed (and job seeking)

◯ Not in paid work (e.g., homemaker, retired, or disabled)

◯ Other

References

Akalın, N., Kiselev, A., Kristoffersson, A., Loutfi, A., 2023. A taxonomy of factors influencing perceived safety in human–robot interaction. International Journal of Social Robotics 15, 1993–2004. doi:10.1007/s12369-023-01027-8.

Awad, N.F., Krishnan, M.S., 2006. The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency and the Willingness to be Profiled Online for Personalization. MIS Quarterly 30, 13–28.

Bach, T.A., Khan, A., Hallock, H., Beltrão, G., Sousa, S., 2024. A systematic literature review of user trust in ai-enabled systems: An hci perspective. International Journal of Human–Computer Interaction 40, 1251–1266. doi:10.1080/10447318.2022.2138826.

Baron, R.M., Kenny, D.A., 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of personality and social psychology 51, 1173.

Baxter, P., Li, M., Wei, J., Koizumi, N., 2025. Public versus academic discourse on chatgpt in health care: mixed methods study. Jmir Infodemiology 5, e64509–e64509. doi:10.2196/64509.

Blanchette, J.F., Johnson, D.G., 2002. Data retention and the panoptic society: The social benefits of forgetfulness. The Information Society 18, 33–45. URL: https://doi.org/10.1080/01972240252818216, doi:10.1080/01972240252818216.

Chancellor, S., Baumer, E.P.S., Choudhury, M.D., 2019. Who is the "human" in human-centered machine learning. Proceedings of the ACM on Human-Computer Interaction 3, 1–32. doi:10.1145/3359249.

Chellappa, R.K., Sin, R.G., 2005. Personalization versus privacy: An empirical examination of the online consumer's dilemma. Information Technology and Management 6, 181–202. URL: http://link.springer.com/article/10.1007/s10799-005-5879-y.

Chernick, L., Bugaighis, M., Britton, L., Cruz, A., Goyal, M., Mistry, R., Reed, J., Bakken, S., Santelli, J., Dayan, P., 2023. Factors influencing the conduction of confidential conversations with adolescents in the emergency department: a multicenter, qualitative analysis. Academic Emergency Medicine 30, 99–109. doi:10.1111/acem.14638.

Chow, J., Sanders, L., Li, K., 2023. Impact of chatgpt on medical chatbots as a disruptive technology. Frontiers in Artificial Intelligence 6. doi:10.3389/frai.2023.1166014.

Cohen, S., Mompelat, L., Mann, A., Connors, L., 2024. The linguistic leap: Understanding, evaluating, and integrating ai in language education. Journal of Language Teaching 4, 23–31.

Crabtree, A., Tolmie, P., Knight, W., 2017. Repacking 'privacy' for a networked world. Computer Supported Cooperative Work (CSCW) 26, 453–488. doi:10.1007/s10606-017-9276-y.

Daley, A., Polifroni, E., Sadler, L., 2018. The essential elements of adolescent-friendly care in school-based health centers: A mixed methods study of the perspectives of nurse practitioners and adolescents. Journal of Pediatric Health Care 32, 327–328. doi:10.1016/j.pedhc.2018.04.009.

Dinev, T., Hart, P., 2006. An extended privacy calculus model for e-commerce transactions. Info. Sys. Research 17, 61–80. URL: https://doi.org/10.1287/isre.1060.0080, doi:10.1287/isre.1060.0080.

Fecher, B., Hebing, M., Laufer, M., Pohle, J., Sofsky, F., 2023. Friend or foe? exploring the implications of large language models on the science system. Ai & Society 40, 447–459. doi:10.1007/s00146-023-01791-1.

Fiske, A., Henningsen, P., Buyx, A., 2019. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. Journal of Medical Internet Research 21, e13216. doi:10.2196/13216.

Friedman, A., Knijnenburg, B.P., Vanhecke, K., Martens, L., Berkovsky, S., 2015. Privacy Aspects of Recommender Systems, in: Ricci, F., Rokach, L., Shapira, B. (Eds.), Recommender Systems Handbook. 2 ed.. Springer US, pp. 649–688. URL: http://link.springer.com/chapter/10.1007/978-1-4899-7637-6_19, doi:10.1007/978-1-4899-7637-6_19.

Gama, F., Tyskbo, D., Nygren, J., Barlow, J., Reed, J., Svedberg, P., 2022. Implementation frameworks for artificial intelligence translation into health care practice: Scoping review. J Med Internet Res 24, e32215. URL: https://www.jmir.org/2022/1/e32215, doi:10.2196/32215.

Ghaiumy Anaraky, R., Li, Y., Knijnenburg, B., 2021. Difficulties of measuring culture in privacy studies. Proc. ACM Hum.-Comput. Interact. 5. URL: https://doi.org/10.1145/3479522, doi:10.1145/3479522.

Goldshtein, M., Schroeder, N.L., Chiou, E.K., 2025. The role of learner trust in generative artificially intelligent learning environments. Journal of Engineering Education 114, e70000. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jee.70000, doi:https://doi.org/10.1002/jee.70000.

He, Q., Wang, J., He, D., 2024. The influence of task and group disparities over users' attitudes toward using large language models for psychotherapy. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 68, 1147–1152. doi:10.1177/10711813241268507.

Hochheiser, H., Lazar, J., 2007. Hci and societal issues: A framework for engagement. International Journal of Human–Computer Interaction 23, 339–374. URL: https://doi.org/10.1080/10447310701702717, doi:10.1080/10447310701702717.

Inan, H.A., Ramadan, O., Wutschitz, L., Jones, D., Rühle, V., Withers, J., Sim, R., 2021. Training data leakage analysis in language models. URL: https://arxiv.org/abs/2101.05405, arXiv:2101.05405.

James, L.R., Brett, J.M., 1984. Mediators, moderators, and tests for mediation. Journal of applied psychology 69, 307.

Judd, C.M., Kenny, D.A., 1981. Process analysis: Estimating mediation in treatment evaluations. Evaluation review 5, 602–619.

Kalkman, S., Delden, J.v., Banerjee, A., Tyl, B., Mostert, M., Thiel, G.J.M.W.v., 2019. Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. Journal of Medical Ethics 48, 3–13. doi:10.1136/medethics-2019-105651.

Kamal, A., 2025. Ai chatbots in pediatric orthopedics: how accurate are their answers to parents' questions on bowlegs and knock knees? Healthcare 13, 1271. doi:10.3390/healthcare13111271.

Kapania, S., Wang, R., Li, T.J.J., Li, T., Shen, H., 2025. 'i'm categorizing llm as a productivity tool': Examining ethics of llm use in hci research practices. Proc. ACM Hum.-Comput. Interact. 9. URL: https://doi.org/10.1145/3711000, doi:10.1145/3711000.

Kassam, I., Ilkina, D., Kemp, J., Roble, H., Carter-Langford, A., Shen, N., 2023. Patient perspectives and preferences for consent in the digital health context: State-of-the-art literature review. J Med Internet Res 25, e42507. URL: https://www.jmir.org/2023/1/e42507, doi:10.2196/42507.

Khan, W.U., Seto, E., 2023. A "do no harm" novel safety checklist and research approach to determine whether to launch an artificial intelligence–based medical technology: introducing the biological-psychological, economic, and social (bpes) framework. Journal of Medical Internet Research 25, e43386. doi:10.2196/43386.

Khatiwada, P., Yang, B., Lin, J., Blobel, B., 2024. Patient-generated health data (pghd): understanding, requirements, challenges, and existing techniques for data security and privacy. Journal of Personalized Medicine 14, 282. doi:10.3390/jpm14030282.

Kibriya, H., Khan, W.Z., Siddiqa, A., Khan, M.K., 2024. Privacy issues in large language models: A survey. Computers and Electrical Engineering 120, 109698. doi:https://doi.org/10.1016/j.compeleceng.2024.109698.

Knijnenburg, B.P., Anaraky, R.G., Wilkinson, D., Namara, M., He, Y., Cherry, D., Ash, E., 2022. User-Tailored Privacy, in: Knijnenburg, B.P., Page, X., Wisniewski, P., Lipford, H.R., Proferes, N., Romano, J. (Eds.), Modern Socio-Technical Perspectives on Privacy. Springer International Publishing, Cham, pp. 367–393. URL: https://doi.org/10.1007/978-3-030-82786-1_16, doi:10.1007/978-3-030-82786-1_16.

Knijnenburg, B.P., Kobsa, A., Jin, H., 2013. Preference-based location sharing: are more privacy options really better?, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 2667–2676. URL: https://doi.org/10.1145/2470654.2481369, doi:10.1145/2470654.2481369.

Kobsa, A., Knijnenburg, B.P., Livshits, B., 2014. Let's Do It at My Place Instead?: Attitudinal and Behavioral Study of Privacy in Client-side Personalization, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, Toronto, Canada. pp. 81–90. URL: http://doi.acm.org/10.1145/2556288.2557102, doi:10.1145/2556288.2557102.

Kocaballi, A.B., Quiroz, J.C., Laranjo, L., Rezazadegan, D., Kocielnik, R., Clark, L., Liao, Q.V., Park, S.Y., Moore, R.J., Miner, A.S., 2020. Conversational agents for health and wellbeing. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems , 1–8doi:10.1145/3334480.3375154.

Kuhlmeier, F., Bauch, L., Gnewuch, U., Lüttke, S., 2025. Designing chatbots to treat depression in youth: qualitative study. Jmir Human Factors 12, e66632–e66632. doi:10.2196/66632.

Laufer, R.S., Wolfe, M., 1977. Privacy as a concept and a social issue: A multidimensional developmental theory. Journal of Social Issues 33, 22–42. doi:https://doi.org/10.1111/j.1540-4560.1977.tb01880.x.

Lee, T., Cho, V., 2025. Enhancing language learning through generative artificial intelligence in blended learning: An empirical study on productive and receptive of informal digital learning english. Journal of Educational Technology Systems 53, 143–169. URL: https://doi.org/10.1177/00472395241266454, doi:10.1177/00472395241266454.

Leon, P.G., Ur, B., Wang, Y., Sleeper, M., Balebako, R., Shay, R., Bauer, L., Christodorescu, M., Cranor, L.F., 2013. What matters to users? factors that affect users' willingness to share information with online advertisers, in: Proceedings of the Ninth Symposium on Usable Privacy and Security, Association for Computing Machinery, New York, NY, USA. URL: https://doi.org/10.1145/2501604.2501611, doi:10.1145/2501604.2501611.

Levkovich, I., Elyoseph, Z., 2023. Suicide risk assessments through the eyes of chatgpt-3.5 versus chatgpt-4: vignette study. Jmir Mental Health 10, e51232. doi:10.2196/51232.

Liao, Z., Antoniak, M., Cheong, I., Cheng, E.Y.Y., Lee, A.H., Lo, K., Chang, J.C., Zhang, A.X., 2024. Llms as research tools: A large scale survey of researchers' usage and perceptions. URL: https://arxiv.org/abs/2411.05025, arXiv:2411.05025.

Luca, S., Clausen, M., Shaw, A., Lee, W., Krishnapillai, S., Adi-Wauran, E., Faghfoury, H., Costain, G., Jobling, R., Aronson, M., Liston, E., Silver, J., Shuman, C., Chad, L., Hayeems, R., Bombard, Y., Bernier, F., Brudno, M., Carroll, J., Cohn, R., Dhalla, I., Friedman, J., Hewson, S., Jamieson, T., Kodida, R., Laberge, A., Lerner-Ellis, J., Mamdani, M., Marshall, C., Osmond, M., Phạm, Q., Reble, E., Rudzicz, F., Seto, E., Shastri-Estrada, S., Smith, M., Thorpe, K., Ungar, W., 2023. Finding the sweet spot: a qualitative study exploring patients' acceptability of chatbots in genetic service delivery. Human Genetics 142, 321–330. doi:10.1007/s00439-022-02512-2.

Maity, S., Deroy, A., 2024. Human-centric explainable ai in education. URL: https://arxiv.org/abs/2410.19822, arXiv:2410.19822.

Mane, H., Doig, A., Gutierrez, F., Jasczyński, M., Yue, X., Srikanth, N., Mane, S., Sun, A., Moats, R., Patel, P., He, X., Boyd-Graber, J., Aparicio, E., Nguyen, Q., 2023. Practical guidance for the development of rosie, a health education question-and-answer chatbot for new mothers. Journal of Public Health Management and Practice 29, 663–670. doi:10.1097/phh.0000000000001781.

Mascalzoni, D., Melotti, R., Pattaro, C., Pramstaller, P.P., Gögele, M., Grandi, A.D., Biasiotto, R., 2022. Ten years of dynamic consent in the chris study: informed consent as a dynamic process. European Journal of Human Genetics 30, 1391–1397. doi:10.1038/s41431-022-01160-4.

Matson, M., Macapagal, K., Kraus, A., Coventry, R., Bettin, E., Fisher, C., Mustanski, B., 2019. Sexual and gender minority youth's perspectives on sharing de-identified data in sexual health and hiv prevention research. Sexuality Research and Social Policy 16, 1–11. doi:10.1007/s13178-018-0372-7.

McDonald, N., Forte, A., 2020. The politics of privacy theories: Moving from norms to vulnerabilities, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 1–14. URL: https://doi.org/10.1145/3313831.3376167, doi:10.1145/3313831.3376167.

Nicol, E., Briggs, J., Moncur, W., Htait, A., Carey, D.P., Azzopardi, L., Schafer, B., 2022. Revealing cumulative risks in online personal information: A data narrative study. Proc. ACM Hum.-Comput. Interact. 6. URL: https://doi.org/10.1145/3555214, doi:10.1145/3555214.

O'Connell, P., Church, P., 2024. No one reads privacy notices. so why do we have them? Global Privacy Law Review 5, 148–153.

Pearman, S., Young, E., Cranor, L.F., 2022. User-friendly yet rarely read: A case study on the redesign of an online hipaa authorization. Proceedings on Privacy Enhancing Technologies

URL: https://par.nsf.gov/biblio/10392520, doi:10.56553/popets-2022-0086.

Pizzi, G., Vannucci, V., Mazzoli, V., Donvito, R., 2023. I, chatbot! the impact of anthropomorphism and gaze direction on willingness to disclose personal information and behavioral intentions. Psychology and Marketing 40, 1372–1387. doi:10.1002/mar.21813.

Porcheron, M., Clark, L., Jones, M., Candello, H., Cowan, B.R., Murad, C., Sin, J., Aylett, M.P., Lee, M., Munteanu, C., Fischer, J.E., Doyle, P., Kaye, J., 2020. Cui@cscw: collaborating through conversational user interfaces. Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing , 483–492doi:10.1145/3406865.3418587.

Sanderson, S.C., Brothers, K.B., Mercaldo, N.D., Clayton, E.W., Antommaria, A.H.M., Aufox, S., Brilliant, M.H., Campos, D., Carrell, D., Connolly, J.J., Conway, P., Fullerton, S.M., Garrison, N.A., Horowitz, C.R., Jarvik, G.P., Kaufman, D., Kitchner, T., Li, R., Ludman, E., McCarty, C.A., McCormick, J.B., McManus, V., Myers, M.F., Scrol, A., Williams, J.L., Shrubsole, M.J., Schildcrout, J.S., Smith, M.E., Holm, I.A., 2017. Public attitudes toward consent and data sharing in biobank research: a large multi-site experimental survey in the us. The American Journal of Human Genetics 100, 414–427. doi:10.1016/j.ajhg.2017.01.021.

Sannon, S., Stoll, B., DiFranzo, D., Jung, M.F., Bazarova, N.N., 2020. "i just shared your responses": Extending communication privacy management theory to interactions with conversational agents. Proc. ACM Hum.-Comput. Interact. 4. URL: https://doi.org/10.1145/3375188, doi:10.1145/3375188.

Shahriar, S., Allana, S., Hazratifard, S.M., Dara, R., 2023. A survey of privacy risks and mitigation strategies in the artificial intelligence life cycle. IEEE Access 11, 61829–61854. doi:10.1109/ACCESS.2023.3287195.

Shim, H., Cho, J., Sung, Y.H., 2024. Unveiling secrets to ai agents: Exploring the interplay of conversation type, self-disclosure, and privacy insensitivity. Asian Communication Research 21, 195–216.

Slaby, C., 2025. Gov. pritzker signs legislation prohibiting ai therapy in illinois. URL: https://www.illinois.gov/news/release.html?releaseid=31573. accessed: 2025-09-09.

Tlili, A., Shehata, B., Adarkwah, M., Bozkurt, A., Hickey, D., Huang, R., Agyemang, B., 2023. What if the devil is my guardian angel: chatgpt as a case study of using chatbots in education. Smart Learning Environments 10. doi:10.1186/s40561-023-00237-x.

Vilaza, G.N., Doherty, K., McCashin, D., Coyle, D., Bardram, J.E., Barry, M., 2022. A scoping review of ethics across sigchi. Designing Interactive Systems Conference , 137–154doi:10.1145/3532106.3533511.

Vladika, J., Fichtl, A., Matthes, F., 2025. Investigating expectations and needs of medical professionals regarding the use of large language models: A study at german university clinics .

Wang, B., Li, G., Li, Y., 2023. Enabling conversational interaction with mobile ui using large language models, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. URL: https://doi.org/10.1145/3544548.3580895, doi:10.1145/3544548.3580895.

Wong, R.Y., Mulligan, D.K., 2019. Bringing design to the privacy table: Broadening "design" in "privacy by design" through the lens of hci, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 1–17. URL: https://doi.org/10.1145/3290605.3300492, doi:10.1145/3290605.3300492.

Zhang, J., Oh, Y., Lange, P., Yu, Z., Fukuoka, Y., 2020. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: viewpoint. Journal of Medical Internet Research 22, e22845. doi:10.2196/22845.

Zhang, Z., Jia, M., Lee, H.P.H., Yao, B., Das, S., Lerner, A., Wang, D., Li, T., 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, Association for Computing Machinery, New York, NY, USA. URL: https://doi.org/10.1145/3613904.3642385, doi:10.1145/3613904.3642385.