Why am I seeing this: Democratizing End User Auditing for Online Content Recommendations

Chaoran Chen
cchen25@nd.edu
University of Notre Dame
Notre Dame, Indiana, USA

Leyang Li
lli27@nd.edu
University of Notre Dame
Notre Dame, Indiana, USA

Luke Cao
lcao3@nd.edu
University of Notre Dame
Notre Dame, Indiana, USA

Yanfang Ye
University of Notre Dame
Notre Dame, Indiana, USA

Tianshi Li
tia.li@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Yaxing Yao
Virginia Tech
Blacksburg, Virginia, USA

Toby Jia-Jun Li
University of Notre Dame
Notre Dame, Indiana, USA

arXiv:2410.04917v2 [cs.HC] 15 Apr 2025

ABSTRACT

Personalized recommendation systems tailor content based on user attributes, which are either provided or inferred from private data. Research suggests that users often hypothesize about reasons behind contents they encounter (e.g., "I see this jewelry ad because I am a woman"), but they lack the means to confirm these hypotheses due to the opaqueness of these systems. This hinders informed decision-making about privacy and system use and contributes to the lack of algorithmic accountability. To address these challenges, we introduce a new interactive sandbox approach. This approach creates sets of synthetic user personas and corresponding personal data that embody realistic variations in personal attributes, allowing users to test their hypotheses by observing how a website's algorithms respond to these personas. We tested the sandbox in the context of targeted advertisement. Our user study demonstrates its usability, usefulness, and effectiveness in empowering end-user auditing in a case study of targeting ads.

CCS CONCEPTS

• **Security and privacy** → **Human and societal aspects of security and privacy**; • **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

end-user auditing, LLM generated persona, privacy awareness

ACM Reference Format:
Chaoran Chen, Leyang Li, Luke Cao, Yanfang Ye, Tianshi Li, Yaxing Yao, and Toby Jia-Jun Li. 2025. Why am I seeing this: Democratizing End User Auditing for Online Content Recommendations. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 22 pages.

1 INTRODUCTION

Personalized recommendation systems customize content using personal profiles, catering to individual user interests [1, 2]. While these systems provide convenience, they often leave users feeling disempowered due to a lack of transparency in these algorithmic systems, making it challenging to understand the causal relationships between their personal data characteristics and the system's decisions [36]. Prior research shows that users often have specific hypotheses about which personal data elements are used to target them in personalized recommendations [38]. For instance, some believe their browsing history influences the ads they see, particularly when ads relate to previously visited websites, while others speculate that factors such as age or location are more significant. These beliefs, whether accurate or not, shape users' mental models and influence their decision-making processes [38, 70].

Instead of relying solely on passive transparency measures, such as system behavior disclosures by first parties [14, 21] or audits by third parties [27, 28], an end-user auditing approach can empower users to actively investigate how their data characteristics affect the recommendations they receive [18, 35]. This proactive engagement in auditing and verifying how personal characteristics influence content recommendations can enhance users' privacy awareness and strengthen algorithmic accountability.

However, empowering end-users to audit online content recommendations presents significant challenges.

First of all, most current approaches do not align well with user needs. Traditional audits are primarily designed to meet regulatory requirements [15], which do not necessarily clarify the relationship between users' data and system outcomes. Although privacy-enhancing tools such as anti-tracking measures [8] and data encryption [17] provide some level of data protection, these passive measures are insufficient to verify user privacy hypotheses, improve their privacy knowledge, or build trust in the system. Additionally, the existing power imbalance between users and organizations further complicates matters, as privacy auditing practices often benefit organizations more than the individuals whose data is being audited [62]. Moreover, existing approaches often overlook users' agency in the auditing process, resulting in users' lack of engagement. Users typically have no active role in these processes and must passively accept the outcomes of audits conducted by technology companies or third-party agencies. The lengthy duration of these audits also often renders their results outdated [33],

Conference'17, July 2017, Washington, DC, USA
2025.

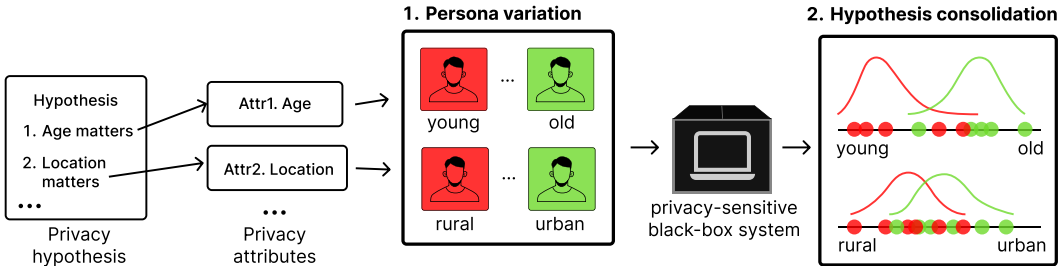

Figure 1: An overview of our privacy auditing sandbox approach

diminishing their usefulness for users who need real-time privacy validation. These third-party audits usually focus only on the most popular websites and hypotheses [54], neglecting the diverse and specific auditing needs of many users, referred to as the "long tail" of user needs. Consequently, current privacy audits fail to empower users or provide them with timely, actionable insights, highlighting a clear need for more agile and user-centric auditing solutions.

Motivated by the aforementioned challenges and limitations of current end-user auditing, we introduce a novel user-centric approach called the Privacy Auditing Sandbox. The sandbox enables users to explore and validate their hypotheses regarding the causal relationships between their personal characteristics and the recommendations they receive online. The key novelty of our work is to introduce a new interactive sandbox approach that empowers non-technical users to generate evidence for or against their privacy hypotheses by directly interacting with recommendation systems in a controlled, risk-free environment. As shown in Fig. 1, this system includes two key features: persona variation and hypothesis consolidation.

- **Persona variation**: A major challenge in supporting end-user audits is enabling users to reason about how specific personal attributes influence system outputs. To address this, we introduce persona variation, a technique for systematically isolating and testing individual privacy attributes. We build on prior work [9] that used large language models (LLMs) to generate fictional personas enriched with demographic and longitudinal data. Drawing on variation theory [43], our method extends this by creating multiple versions of a base persona, each mainly differing by one key attribute (e.g., age or location). This controlled variation allows users to compare system responses across personas and infer whether changes in attributes result in different recommendations.
- **Hypothesis consolidation**: Even when users can observe differences in recommendations across persona variations, it can be difficult to interpret or summarize these changes, especially when outputs are unstructured (e.g., images, text). To address this, we designed hypothesis consolidation to help users make sense of observed patterns and draw evidence-based conclusions. After gathering data on how the system reacts to different personas, the system employs LLMs to map these outputs onto corresponding privacy attribute axes. For example, if exploring age, LLMs categorize output to show which age group each recommendation targets. This method extends to other attributes, allowing users to see how recommendations might change based on urban versus rural settings or across different income levels. By visualizing how outputs shift across variations, users can more easily confirm or refute their initial hypotheses about which personal characteristics influence recommendations.

To evaluate the Privacy Auditing Sandbox, we conducted a case study in the context of online advertising, investigating how different privacy attributes influence targeted ad content. We first performed a technical evaluation to address core research questions related to persona generation quality, ad identification accuracy, ad rating stability, and the impact of persona substitution on advertisements. For persona generation, expert reviews highlighted strong consistency and credibility, although minor hallucinations and stereotyping were noted. For ad identification, a curated dataset of 205 websites and 805 advertisements demonstrated high accuracy (96.52%) and precision, with minor errors stemming from dynamic or non-standard ad formats. LLM-based ad rating stability was confirmed with low variability (standard deviation = 0.87; coefficient of variation = 3.00%) across privacy attributes. Lastly, persona substitution experiments revealed that distinctly different personas received significantly varied ads, while similar personas received consistent ads, underscoring the influence of persona attributes on ad content.

To evaluate our system's usability, we conducted a user study (N = 15) and asked the participants to use pre-defined personas to test their privacy hypotheses and create new personas to explore further. We then assessed the usability with the the System Usability Scale (SUS) questionnaire. The results show that participants successfully validated their privacy hypotheses, identifying clear correlations between persona attributes and the thematic and visual aspects of the ads. Participants also reported a positive user experience and recognized the system's value in enhancing their understanding of how personal characteristics affect ad targeting.

This approach has broader applications beyond targeted advertising, offering the potential for auditing the accountability of other privacy-sensitive systems. It is particularly useful in settings where generated persona data can be seamlessly integrated into systems, allowing users to repeatedly prompt system responses in real time at minimal cost. We anticipated that the tool can enable users to (1) discover unexpected uses of their personal data and (2) validate concerns related to algorithmic practices in content personalization, such as biases in personalized advertising. This study represents

an initial step towards enabling robust end-user auditing of online content recommendations and enhancing end-user privacy awareness.

This paper makes the following contributions:

- Introduction of a novel interactive approach for non-technical end users to audit online recommendations and explore and verify their hypothesis about causal relationships between personal privacy attributes and system outcomes in a risk-free environment through LLM-generated personas for end-user auditing. We make a unique technical contribution by utilizing state-of-the-art methods for generating personas enriched with demographic data and longitudinal personal data [9].
- Validation of the approach through a proof-of-concept prototype and a case study. The study results reveal patterns in how users employ our method to validate their privacy hypotheses in the context of targeted advertising.
- Discussion of the implications of adopting this approach to enhance user engagement in privacy auditing and improve their privacy awareness.

2 RELATED WORK

We reviewed prior work in explainable recommendation systems, algorithm auditing, and counterfactual reasoning to address the challenge of enhancing user understanding and accountability in algorithmic decision-making. Explainable recommendation systems aim to uncover causal relationships between recommended content and user characteristics. However, their technical complexity and limited applicability in real-world contexts often hinder end users from fully grasping these mechanisms. Recent advances in algorithm auditing have incorporated users into the auditing process, enabling them to explore and evaluate the causal connections between privacy attributes and recommended content, thereby increasing privacy awareness and strengthening algorithmic decision-making. Counterfactual reasoning further supports these efforts by illustrating the links between private data and online content recommendations through comparison and contrast. Lastly, as we used online behavioral advertising (OBA) as a case study, we review and contrast common static methods for improving OBA transparency, underscoring the novelty of our proposed interactive Privacy Auditing Sandbox system: it allows end users to learn through exploration, actively verify privacy hypotheses, and observe correlations between privacy attributes and system outcomes.

2.1 Explainable Recommendation

Personalized recommendation systems use vast amounts of user data to customize content based on the preferences and behaviors of individual users. To capture the causal relationship between the recommended content and user characteristics, a popular solution is "explainable recommendation" [73]. This approach aims to not only offer recommendation outcomes to users but also provide explanations for why specific items are recommended. It has been widely applied to diverse recommendation systems, such as those in e-commerce [22], points of interest [5], social networks [26], and multimedia [32].

The explanations of recommendation models have different target audiences (e.g., users, model designers, and service providers) and various objectives (e.g., improving informed decision-making and enhancing profitability). Therefore, different target audiences and objectives lead to diverse evaluation perspectives. The most related criteria for understanding the causal relationship between the recommended content and user data is transparency, which evaluates whether the explanations can reveal the internal working mechanism of the recommendation models [40, 57, 59]. For example, Jiang et al. proposed RCENR, a Reinforced and Contrastive heterogeneous Network Reasoning model for explainable news recommendation [31], which generates multi-hop reasoning paths to connect users with candidate items through intermediate nodes (e.g., previous clicked news and the attributes of the news). Zhang et al. [72] presented NS-ICF, a neuro-symbolic interpretable collaborative filtering that learns logical rules for recommendation systems.

However, the objectives of the above work focus on assisting model designers in knowing more about how the recommendations are generated and better debugging the system. End users often find it hard to understand these model outputs because of technical complexities and difficulties in applying these explainable models in real life. They struggle to see how their private data leads to specific recommendations. Compared with such methods assisting model designers, our system is designed to empower end users to understand and experiment with recommendation systems, enabling them to observe how different privacy attributes of generated personas can influence recommendation items.

2.2 Algorithm Auditing for Online Content Recommendation

Algorithm audits are effective tools for analyzing black-box systems by systematically querying them with inputs and observing the outputs [46]. These audits have been applied across various domains, such as employment [58], healthcare [49], and consumer markets [3], to promote fairness and transparency [46].

Traditionally, auditors have excluded real users to maintain control and strengthen technical audits [16, 55]. The typical methodology involves repeatedly querying an algorithm with inputs and observing the outputs to infer the system's internal workings. Recent studies incorporate end users into the auditing process. There are two main types of input data for algorithm audits: one relies on crowdsourcing real users to conduct audits, while the other involves simulating users by generating data with specific attributes. For crowdsourcing, Lam et al. [36] introduced sociotechnical audits, demonstrating how real users can be actively involved in algorithm audits by systematically collecting their input and behavioral data to reveal algorithmic biases. Although using crowdsourcing ensures that the audit involves real users, this method requires a significant amount of time and resources to engage many participants. To improve auditing efficiency while still using user-related data, simulating user data for audits has been developed. For example, Sunlight [37] introduces a scalable system that uses statistical methods to detect the causes of targeting phenomena, such as personalized ads in Gmail, by simulating users with different attributes.

Compared to Sunlight, our approach is not limited to Gmail but is applicable to ads across various websites. We use large language models (LLMs), which allow for more detailed simulations of users' longitudinal data, such as their weekly schedule and browsing history. Additionally, we build on the persona generation technique proposed by Chen et al. [9], generating not just a single persona but a set of personas related to the personal attributes that users want to test in their privacy hypotheses. This enables users to explore and audit the causal relationship between privacy attributes and the recommended content.

2.3 Counterfactual Reasoning for Enhancing Data Interpretation

Counterfactual reasoning is a psychological term that means modifying a factual prior event and then evaluating the consequences of that change [45]. It has definitions very similar to those of contrastive learning [10] in AI and variation theory [41] in education. Contrastive learning differentiates between similar and dissimilar data points to improve *model* understanding of categories [10]. Variation theory focuses on distinguishing essential from non-essential features of concepts to enhance *human* learning [43]. Although they have different targets, both aim to deepen the target's understanding through comparison and contrast.

Previous work used counterfactual reasoning to improve the explainability of models through heterogeneous information networks [24], influence functions [60], and perturbation models [69]. For example, Ghazimatin et al. [24] generated explanations from the provider's perspective by identifying a minimal set of the user's past interaction (e.g., reviews, purchases, and ratings) that, if removed, would alter the recommendation results. Xu et al. [69] proposed a perturbation model to generate alternative user histories and recommendation items, then applied a causal rule mining algorithm to identify personalized causal relationships for the recommendation system.

Our approach leverages counterfactual reasoning to highlight the connection between private data and online content recommendation. By creating personas with varied privacy attributes, our system demonstrates the impact of privacy data usage on recommendation outputs. Unlike prior methods that relied on public datasets for validation, our system uses the generated personas' data to replace real user data, enabling tests on actual websites. This enables users to evaluate the effects of their data usage by themselves, promoting a greater sense of control and privacy awareness.

2.4 Improving Algorithm Accountability in Online Behavioral Advertising

Online behavioral advertising (OBA) is "the practice of monitoring people's online behavior and using the collected information to show people individually targeted advertisements [6]". Online behavior can include web browsing data, search histories, media consumption data (e.g., videos watched), app usage, purchases, click-through responses to ads, and communication content via email or social media. Although users and regulations require ad companies to be transparent about privacy data processing practices [25, 50], Many users are still concerned about the unwarranted use of their data in the OBA [6, 61, 64].

Informed consent and privacy policy statements serve as the two widely used tools to improve OBA transparency. Although such tools explain the collection and usage of private data, people often cannot find them due to dark patterns [42] or merely ignore them due to their complex content [51]. To mitigate this issue, Van et al. added descriptive information to the privacy icon in OBA, indicating that the ad is shown based on the personal online behavior of users [63]. Compared with these static approaches, we focus on creating an interactive system to allow users to examine different influences of privacy attributes on online ads. Enhancing user engagement in auditing their privacy settings is crucial because it empowers individuals to actively understand and control how their data is used. Experiential learning [34] shows that interaction fosters deeper comprehension of complex systems like online behavioral advertising. By allowing users to explore and verify privacy hypotheses, they gain a more intuitive understanding of privacy risks and outcomes, increasing their sense of control and trust. Therefore, our approach enables users to learn from exploration, where they can interactively verify their privacy hypothesis and observe the causal relationship between privacy attributes and system outcomes.

3 THE PRIVACY AUDITING SANDBOX APPROACH

To democratize users' privacy auditing for online content recommendation, we introduce a novel privacy auditing sandbox approach. This method allows end users to conduct real-time audits of the relationship between personal user attributes and online content recommendations—a capability previously unavailable. Prior work has suggested an approach to generate realistic privacy personas using few-shot learning, contextualization, and chain-of-thoughts techniques [9]. Build on this approach, our method enhances user capabilities by enabling them to: (1) generate a set of personas and corresponding personal data based on controled variaions in specified attributes; (2)import this synthetic personal data into browsers for the automated collection of online content; and (3) observe the relationship between recommended contents and personal attributes to verify users' hypotheses.

The key goals of our privacy auditing sandbox approach are two-fold:

(1) **Support users to explore the relationship between individual user attributes and recommended content.** When users observe a correlation between personas' personal attributes and recommended content (e.g., "low-income personas are more likely to see coupon or discount ads, while high-income users are more likely to see luxury ads"), they can confirm the influence of users' personal attribute on the online content recommendations.
(2) **Enhance users' engagement and initiative in privacy audits.** When users hypothesize that certain characteristic (e.g., age) influences online content recommendations, they can take the initiative to create a set of personas with variations in the specific personal attribute (e.g., young, middle-aged, and old personas) and observe whether the recommended content changes accordingly. This approach is inspired by experiential learning [23], allowing users to quickly

experience the impact of different personal attributes on recommended content. We believe that this privacy auditing sandbox approach can democratize privacy audits for end-users and enhance their engagement in privacy audits.

Our approach consists of two primary steps: persona variation and hypothesis consolidation (Fig. 1). In the persona variation phase, the system generates a set of synthetic personas based on the privacy attributes users wish to test, ensuring these personas reflect realistic variations. Subsequently, during the hypothesis consolidation phase, users deploy these personas on specific websites to collect online recommendations and analyze their association with the privacy attributes.

For the scope of this paper, we chose online advertisements as an example domain for online content recommendation due to their pervasive impact on daily life and their relevance to privacy concerns. As outlined in previous studies on online behavior modeling [70], there is a disconnect between the black-box nature of targeted online ads and users' mental models of how these ads operate. While many users may suspect that their personal data and online behavior shape the advertisements they see, they often lack the tools to verify these suspicions. Future research could extend to other algorithmic systems, such as social media feeds, personalized news recommendations, and more. The approach we introduce should be applicable to a broad range of domains.

In the subsequent sections, we will elaborate on the subject of our privacy audit, describe two pivotal features of our approach, and introduce a prototype system that integrates these features—the Privacy Auditing Sandbox.

3.1 Subject of the Privacy Audit

We clarify that the subject of the privacy audit in our method is the relationship between user characteristics and online content recommendation. The term 'user characteristics' refers to a specific trait reflected by the user's online profile and personal data (such as the user's age group as inferred from their account profile and browsing history). Note that the subject of our privacy audit is not any particular piece of personal data (such as a specific age-related data item in the account profile or browsing history). In other words, we are not auditing whether the system collects a specific piece of data, nor are we aiming to show users how particular personal data contribute to online content recommendations. This is because current content recommendation systems are complex and opaque black-box systems, making it impossible for us to trace how specific personal data is collected and used by such systems. Therefore, our method focuses on user characteristics as reflected by multiple pieces of personal data, enabling privacy audits by allowing users to observe how system outputs, in the form of recommendations, change in response to changes in user characteristics.

3.2 Persona Variation

As defined by prior research [9], the privacy persona represents a fictional user with a distinctive biography, demographic information, and a large set of synthesized personal data. Therefore, persona generation is to create a single artificial persona containing realistic synthetic personal data. Compared with it, the goal of persona variation is to create a set of synthetic personas with corresponding personal data that represent variations in a chosen personal attribute. Our work uses an established persona creation methodology from prior research [9]. When users engage with the sandbox, it utilizes synthetic persona data instead of actual user data, enabling testing of how online content recommendations adapt to changes in persona characteristics. By comparing the recommended content with each variant, users can confirm their hypotheses about the link between user characteristics and content recommendations.

Our method begins with a base persona, derived from user input and a selected personal attribute. Using the specified prompt, we generate descriptions for persona variants. Drawing on prior research in persona generation techniques [9], we construct comprehensive persona profiles that include demographic details, additional personal information, and longitudinal data. As illustrated in Fig. 3, these variants maintain the core personal profile of the base persona but vary based on selected personal attributes. To ensure realistic profiles, not only the chosen attribute but also other related attributes, may be adjusted coherently. For instance, if income level is the selected attribute, the persona's job title might also be modified within the same career path to match the new income bracket.

Prompt for generating persona variants
Given a fictional user's description: {base persona} Modify the description to make it more like a {privacy attribute} version of the person. Do not change age, gender, location, income, and educational level except for the {privacy attribute}. Only make necessary modifications. In the result, the information about the {privacy attribute} must appear. Return the profile in only one paragraph.

3.3 Hypothesis Consolidation

At this stage, users substitute their real data with the generated persona data to collect content recommendations as the persona and assess how variations in persona characteristics impact these recommendations. For each persona variant, the privacy auditing sandbox substitutes the user's real data with synthetic data matching the persona's traits. Thus, when an online service requests this personal information, it will receive the synthetic data corresponding to each persona variant. The system provides an easy-to-use interface that allows users to browse and compare content recommendations received under different personas, thereby facilitating the validation of their hypotheses for auditing.

Persona data replacement. The sandbox comprehensively replaces the user data with the data of each persona variant on Google accounts, location, IP address, user agent, and browsing history, as shown in Fig. 4a. For Google accounts, it replaces data in the Google Ad Center by creating a dedicated Google account and automating profile updates using the Puppeteer libraries [1]. This involves identifying and replacing specific attributes (age, gender, etc.) with persona data in three steps: accessing the profile page, extracting persona attributes based on the alias label in the source code, and replacing the values. For geographical location, it uses

[1] https://pptr.dev/

Figure 2: System Overview: (a) An input field for generating the base persona profile. (b) A control panel to select personal attributes for persona variants, choose a website to audit, and set visit frequencies. (c) A display showing selected personal attribute values, along with the names and descriptions of color-coded persona variants. As users scroll, persona variants shift to the left for continued reference. (d) Visualization of ad distribution based on rated scores, with a scatter plot where each point represents an ad. The ads in the image were obtained by visiting https://www.thepioneerwoman.com/ twice using a persona set consisting of 3 variants. The X-axis shows the Ad-Attribute Alignment score, and the Y-axis shows score probability density. Points are color-coded by persona variants. Users can hover over a point to view the corresponding persona, ad image, description, and rating in the ad list below.

OpenStreetMap's API to get the persona's latitude and longitude, then override the location in Chrome via Puppeteer. For IP address, it uses NordVPN's API to connect to the server closest to the synthesized location of the persona, ensuring that online services see the server's IP instead of the user's. For user agent information, which is a characteristic string that lets servers and network peers identify the application, operating system, vendor, and/or version of the requesting user agent, it uses Puppeteer's "setUserAgent" to replace current details with those that match the persona. For browsing history, it overwrites Chrome's local SQLite database with the persona's browsing data before launching the browser. Specifically, both the URL table (which logs visited links) and the visit table (which records timestamps) are replaced with the browsing history of the generated persona. All data from the persona variant, including Google account details, location, IP address, user agent, and browsing history, are generated by LLMs. This ensures

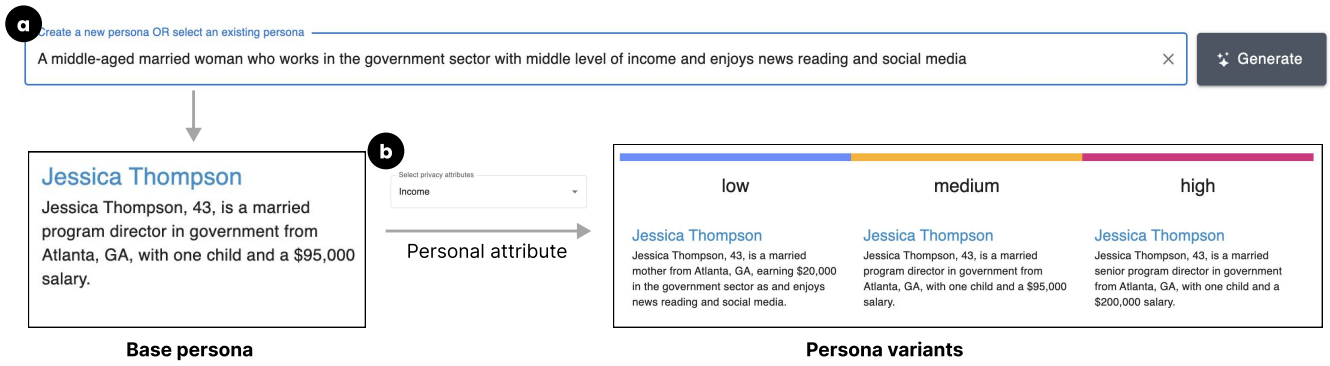

Figure 3: Persona variation. (a) Providing guidance for base Persona's profile generation. (b) Selecting personal attribute to generate persona variants.

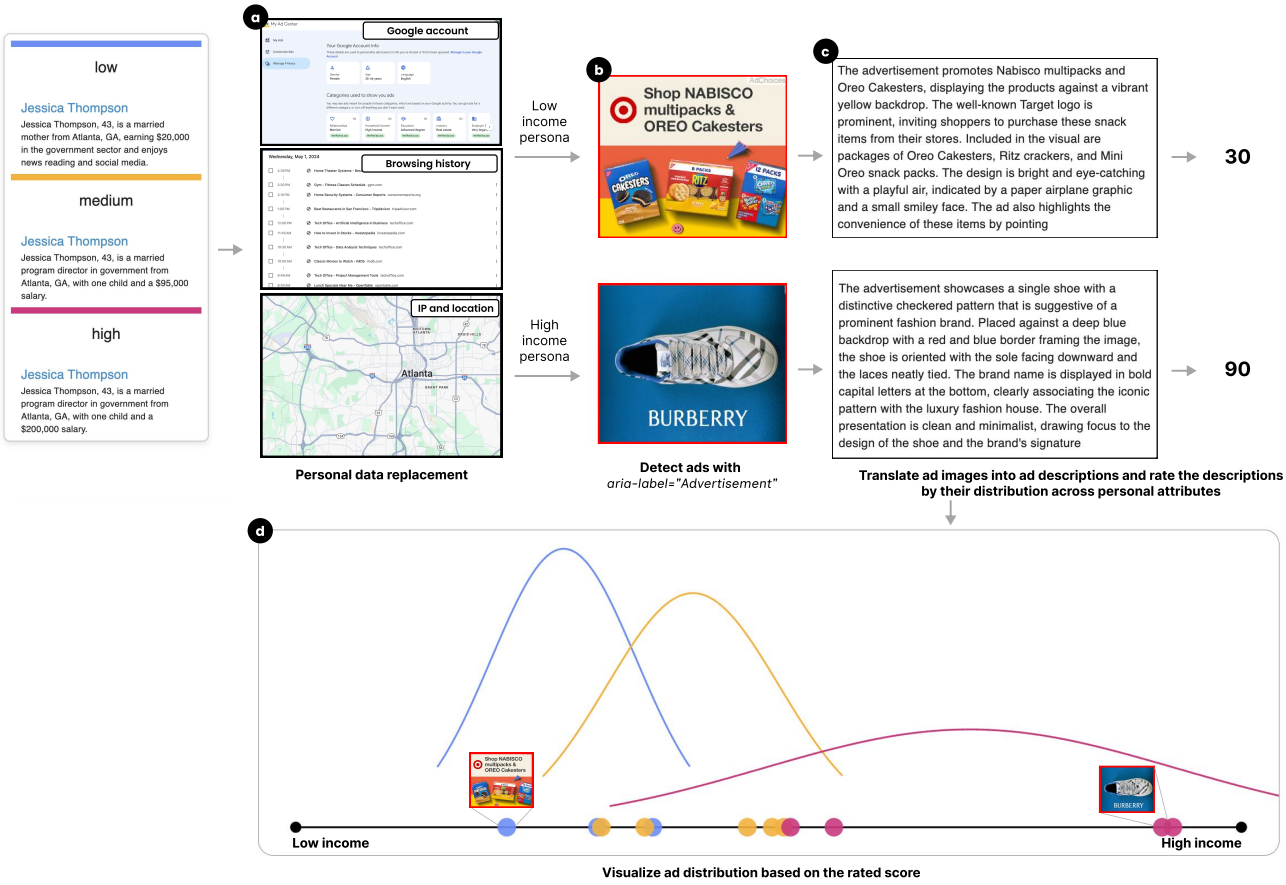

Figure 4: Hypothesis consolidation (a) Replacing the user's real data with each persona variant's data, including Google account, location, IP address, user agent, and browsing history. (b) Detecting online ads displayed on the selected website. (c) Translating ad images into textual descriptions and rating them on a Ad-Attribute Alignment Score to evaluate their alignment with personal attributes. (d) Visualizing the ad distribution based on the rated scores. Each point on the scatter plot represents an ad, with the X-axis indicating the Ad-Attribute Alignment Score and the Y-axis representing the probability density of the scores based on a normal distribution curve. Points are color-coded by one of the three persona variants.

consistency and alignment with the persona's synthesized characteristics, allowing a comprehensive and controlled evaluation of the system's responses to the various attributes of the persona.

Online Ad identification. After loading the persona variant's information into the browser, the sandbox automatically visits the user-specified websites for the number of iterations selected to scrape ads. As shown in Fig. 4b, if online ads are detected during the visit, the system takes screenshots and stores them in base64 format. Online ads share a common characteristic—their aria-label[2] of the corresponding HTML element should be "Advertisement" based on Web Accessibility Standards. Consequently, when a DOM element with this label is detected on the page, the system identifies its location and takes a screenshot.

Ad translation and rating. To account for both the visual elements and semantic information when evaluating ads against personal attributes, the system first uses GPT-4V to convert each captured ad image into a textual description. As shown in Fig. 4c, GPT-4 analyzes these textual descriptions to assess how well each ad aligns with specific personal attributes using an Ad-Attribute Alignment Score. We define the Ad-Attribute Alignment Score as a metric ranging from 0 to 100 on how well a given ad description aligns with a specific persona attribute (e.g., age, gender, income). A lower score indicates the advertisement is more aligned with one end of the attribute spectrum (e.g., young for age, lower income level for income), while a higher score suggests stronger alignment with the constrasting end (e.g., older people for age, higher income for income). In its implementation, we request the LLM to consider the percentile distribution of the ad (in all ads) ordered by how much it specifically targets either extreme end of the attribute spectrum. For example, on the age attribute, an ad that specifically targets a young audience *the most* would receive close to 0, and an ad that specifically targets a old audience *the most* would receive close to 100.

The assessment considers contextual clues such as the advertised product, the brand, the tone of the message, and visual indicators. For instance, an ad for luxury watches might score high on the income scale, indicating alignment with high-income personas, while an ad for budget groceries would score low, aligning with low-income personas. The accuracy of this mechanism is evaluated in Section 4.4.

Ad distribution visualization. To help users explore the relationship between recommended ads and persona variants, thereby validating their privacy-related hypotheses, the system visualizes the distribution of ads based on their relevance to the persona attribute used in creating the variants. As shown in Fig. 4d, the system plots ads on a one-dimensional scatter plot along a coordinate axis based on their Ad-Attribute Alignment Scores. Each point represents an ad, with its position on the x-axis encoding its score. The color of each point corresponds to the persona variant active when the ad was retrieved, making it easy to differentiate the ads' relevance across personas. To make it easier for users to understand the result, the system overlays the scatter plot with a normal distribution curve, allowing users to intuitively observe patterns in ad distributions. If users want to view detailed information about a specific ad, they can hover over the ad point and the corresponding persona variant, ad image, ad description, and rating will be displayed in the ad list below for users' reference.

3.4 Example Use Case

We illustrate the functionality of the Privacy Auditing Sandbox through a specific example use case. In this scenario, a user, Alice, aims to determine if income data affects online advertisements on an e-commerce website [3]. Alice employs the Sandbox to create various personas, each with distinct annual income levels. She then uses the Sandbox to repeatedly automatically visit the website with personal data associated with each persona and collect data on the advertisements displayed to each. This method allows her to investigate her hypothesis that the website may be targeting ads based on user income, showcasing the core capabilities of the Privacy Auditing Sandbox.

(1) *Providing Guidance for Base Persona Generation*: Alice enters her guidance to generate a persona that is similar to her profile. She enters "A middle-aged married woman who works in the government sector with a middle income and enjoys news reading and social media" (Fig. 3a). This guidance serves as the foundation for creating the base persona. Users can provide as little or as much information as they feel comfortable with, ensuring flexibility while protecting their own private information. Typically, using a base persona that is similar to the real profile of the user fosters empathy building and allows the user to better interpret the results in familiar contexts, as reported in [9]. The information provided to generate the base persona is not restricted to age or gender but could include marital status, interests, or any other relevant information. Upon processing Alice's guidance, the Privacy Auditing Sandbox generates a base persona: Jessica Thompson, a 43-year-old married woman living in Atlanta. She works as a program director associate and earns an annual income of $95,000. She enjoys news reading and social media. Alice can then review and, if necessary, update the guidance to regenerate the persona.

(2) *Selecting Personal Attribute for Persona Variation*: Once Alice finalizes the base profile, she selects "Income" as the attribute for persona variation. The Privacy Auditing Sandbox then creates three variants of Jessica Thompson based on the original profile—each representing low, medium, and high-income levels as illustrated in Fig. 4a. To enhance the realism and consistency of these variants, changes in the selected attribute (income) lead to corresponding adjustments in other personal details, such as job titles. For each version of the persona variant, the Privacy Auditing Sandbox also generates detailed personal data including the persona's device and browser in use, home location, weekly schedule with location records, and browsing history. These data are tailored to align consistently with each variant's specific income profile.

(3) *Collecting Online Ads with the Generated Persona Variants*: Alice inputs a URL for the auditing target and specifies the

[2]https://developer.mozilla.org/en-US/docs/Web/Accessibility/ARIA/Attributes/aria-label

[3]https://www.thepioneerwoman.com/

number with which the website will be sampled for each persona variant. The Privacy Auditing Sandbox then activates a Chrome extension that substitutes Alice's real privacy data with the synthesized data of different variants of Jessica. Alice's privacy data, including her profile in the Google Ad Center, browsing history, real-time location, and IP address, are temporarily replaced. The Privacy Auditing Sandbox will automatically visit the target URL repeatedly according to the number of samples, identify online ads, and take screenshots.

(4) *Reviewing the results to validate the hypothesis*: After collecting all the advertisements, the Privacy Auditing Sandbox evaluates each ad by assigning it a score based on its percentile ranking related to the selected personal attribute, "income" (Fig. 4c). Finally, the ads will be visualized on a distribution chart and an ad table based on the scores (Fig. 4d). Alice reviews these visualizations to analyze the correlation between the income levels of the persona variants and the pricing of products or services featured in the ads. This analysis helps her validate the hypothesis: higher-income personas are more frequently targeted with advertisements for luxury goods on this website.

3.5 Implementation

We developed the Privacy Auditing Sandbox with a frontend using the framework Next.js and a backend powered by Flask. They communicate through HTTP requests for API access.

For data storage, we utilize JSON files to manage the synthesized personal privacy data and the corresponding advertisements collected through the personas. To create synthetic personas, we enhanced our methodology by integrating a Python script that interfaces with the GPT-4 public API using the "langchain" Python library, building on the foundational work cited in [9].

4 TECHNICAL EVALUATION

We conduct a technical evaluation of the Privacy Auditing Sandbox to assess the following technical research questions (TRQs):

TRQ1: What is the quality of LLM-generated persona variants (described in Section 3.2) in terms of contextual relevance, accuracy, hallucination, and stereotyping?

TRQ2: How accurate is the ad identification mechanism (described in Section 3.3) ?

TRQ3: How stable is the Ad-Attribute Alignment Score (described in Section 3.3) across different persona attributes and ad contexts?

TRQ4: What is the impact of persona substitution on the personalized ad content?

- **TRQ4a:** Do similar personas, based on chosen attributes, on the same website receive similar ads?
- **TRQ4b:** Do different personas, based on chosen attributes, on the same website receive different ads?

4.1 Overview

As shown in Fig. 5, the Privacy Auditing Sandbox consists of three primary stages: (1) generating persona variants based on specific privacy attributes; (2) substituting user data with persona-specific data on target websites; (3) identifying and collecting advertisements on the target websites; and (4) employing an LLM to evaluate the advertisements based on the privacy attribute defined in the first stage.

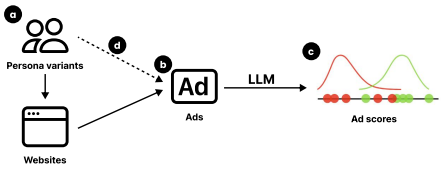

Figure 5: An overview of our technical evaluation. (a) Conducted an expert review to evaluate the quality of the generated personas. (b) Evaluated the accuracy of ad identification. (c) Assessed the stability of LLM-generated ad scores based on privacy attributes. (d) Investigated the impact of persona substitution on the ad content.

To assess the reliability of our approach, we conducted technical evaluations at each stage. In the generation of personas, we performed an expert review to detect potential hallucinations and stereotyping in the generated personas. In the identification of online ads, we evaluated the accuracy of our ad identification mechanism to ensure it correctly detected and classified advertisements across various websites. In the ad rating stage, we assessed the consistency of LLM-generated ad scores to ensure stable and reliable assessments. Lastly, in persona substitution, we analyzed the impact of persona substitution on the types of advertisements generated, shedding light on the relationship between persona attributes and ad content.

4.2 Evaluation of Persona Generation Quality

LLM-generated personas are often criticized for issues such as hallucinations and stereotyping [13, 56], which can undermine their effectiveness and reliability. While most existing automated metrics assess the performance of LLM models [11, 12, 47, 67], they inadequately address the quality of generated personas. This gap highlights the need for human evaluation to ensure accuracy and relevance. To address this challenge, we adopted an expert review approach to evaluate persona quality. The evaluation used established scales [52, 53] to assess the quality of personas from the perspectives of consistency, credibility, and stereotypicality.

To address **TRQ1**, we engaged two experts with relevant experience in persona creation to evaluate the quality of generated personas. None of the experts is an author on this paper. Both experts have at least 5 years of experience in UX design and research, are familiar with the use of personas in the UX practice, and have at least a master degree in HCI or relevant fields. Expert reviewers offer nuanced insights into stereotyping and contextual relevance, aspects often overlooked by automated methods. To reduce subjective bias, we implemented a structured evaluation framework combining Likert-scale ratings and open-ended questions to capture qualitative feedback. Additionally, we computed Cohen's Kappa to measure inter-rater agreement, ensuring reliability and consistency in the evaluation outcomes.

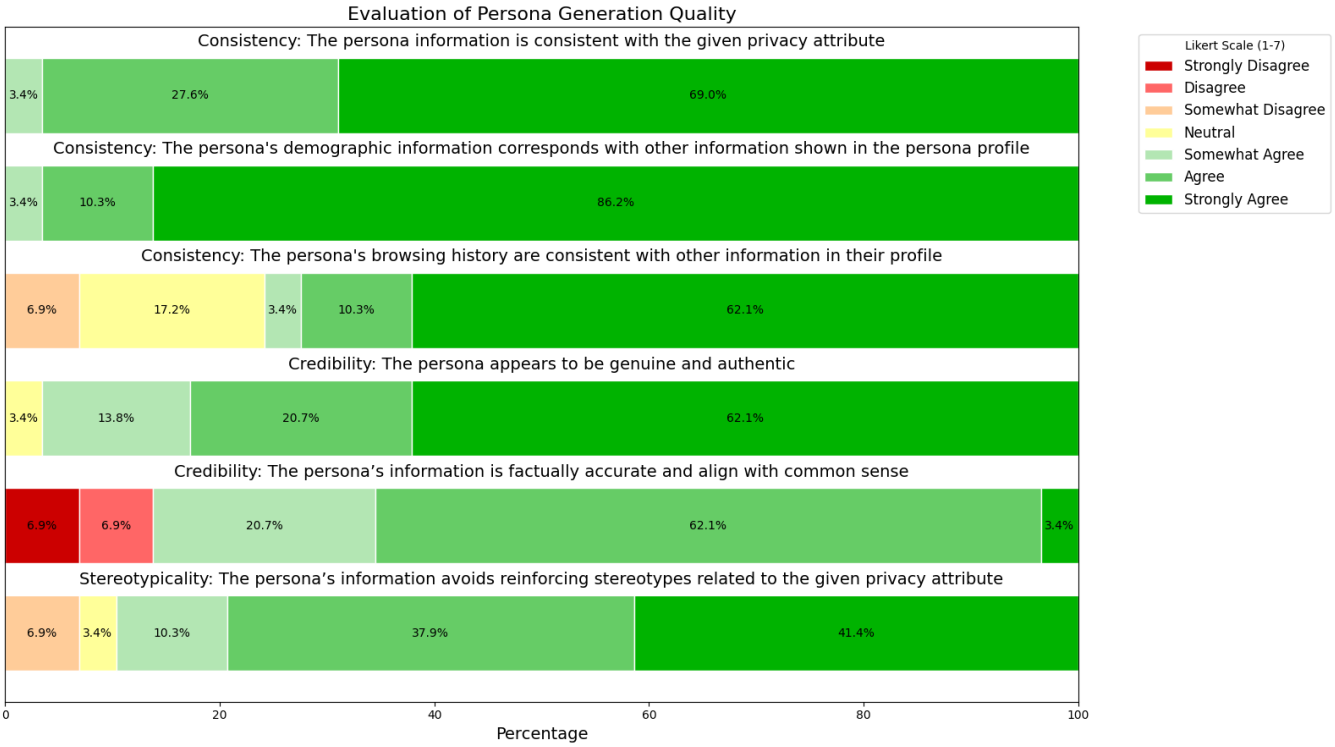

Figure 6: Result of the expert evaluation on persona generation quality.

As shown in Fig. 6, the expert review (Cohen's Kappa=0.88, indicating strong inter-rater reliability) demonstrated that the generated personas exhibited strong consistency with the given privacy attribute. The demographic information of the generated personas was also highly consistent with other information shown in the persona profile. However, there were minor inconsistencies in the browsing history relative to other profile information. This could be due to the dynamic nature of browsing data, which might not always align closely with static persona attributes. In terms of credibility and stereotypicality, most of the generated personas achieved a score of 6 or above on a 7-point Likert scale, meeting baseline criteria for capturing the intended privacy attribute without introducing significant inaccuracies or biases. However, hallucination errors, such as implausible income levels and physical addresses, were identified in a small subset of the personas. These hallucinations may have occurred due to the LLM's lack of understanding of specific real-world geographical or income data, leading to the generation of a small portion of data that did not align with reality.

4.3 Evaluation of Ad Identification

To answer the **TRQ2** for the accuracy of ad identification, we tested our ad identification module on 205 websites with 805 advertisements. Since there is no publicly available dataset of ad-containing websites, we curated our dataset using a publicly available Kaggle dataset[4] of 1,408 websites. Among the 1,408 websites, we manually visited each website's homepage and identified 205 websites with ads. For those 205 websites, we recorded the positions of 805 displayed advertisements. Our dataset emphasizes the location of advertisements rather than their content. For ads occupying the same position but displaying rotating or refreshing content, we treated them as a single advertisement due to their consistent placement.

Using this curated dataset, we evaluated our ad identification module and summarized the results in Table 1. The method achieved achieved a high accuracy rate of 96.52%, , with a precision of 100% (all identified ads are indeed ads) and a recall of 96.52%, demonstrating its ability to detect the majority of ads across diverse website designs. Errors were primarily associated with dynamically loaded advertisements or non-standard labeling practices on certain websites, which limited detection in a small number of cases. Additionally, some missed detections were due to advertisements embedded in floating window, which presented unique challenges for identification. We plan to address these edge cases in future work.

TP	FN	TN	FP	Acc.	Prec.	Rec.
777	28	0	0	96.52%	100%	96.52%

Table 1: Performance of the Ad Identification Module. TP = True Positive, FN = False Negative, TN = True Negative, FP = False Positive, Acc. = Accuracy, Prec. = Precision, Rec. = Recall.

[4]https://www.kaggle.com/datasets/hetulmehta/website-classification

4.4 Evaluation of Ad Rating Stability

To answer the **TRQ3** for the stability of the Ad-Attribute Alignment Score along five dimensions of privacy attribute (age, gender, urbanization of home location, income level, and education level), we selected 20 advertisements per dimension and computed the Ad-Attribute Alignment Score for each ad five times using the LLM. The LLM did not receive information on which persona was associated with each ad. Instead, its evaluations were solely based on its understanding of the underlying privacy attribute.

We measured the consistency of the scores using the averaged standard deviation and the averaged coefficient of variation. Lower values for these metrics indicate higher stability and demonstrate the LLM's ability to deliver consistent evaluations across multiple iterations.

The experimental results, presented in Table 2, show that that the average standard deviation for the LLM's scores is 0.87, with an average coefficient of variation of 3.00%. This indicates that the LLM demonstrated a high stability in its evaluation of advertisements across repeated scoring attempts. Overall, these results confirm the stability of the LLM's scoring mechanism for evaluating ads based on privacy attributes.

	Age	Gender	Location	Income	Education	All
Avg. std	0.57	0.75	0.63	1.19	1.22	0.87
Avg. cov	0.77%	7.92%	2.75%	2.06%	1.50%	3.00%

Table 2: Results of the Ad Rating Stability. "std" = standard deviation, "cov" = coefficient of variation.

4.5 Evaluation of the Impact of Persona Substitution on Ads

To address **TRQ4** regarding the impact of persona substitution on advertisements, we conducted a validation experiment. For each of the five privacy attribute (age, gender, urbanization of home location, income level, and education level), a consistent guidance prompt was used three times to generate three groups of persona variants, with each group containing three distinct personas. We identified five representative websites (as detailed in Section 5.2.2). We used each persona to access each one of the five websites, visiting each site three times. During these visits, the user data was substituted with persona-specific data (as described in Section 3.3). For each persona-website pair, we collected and analyzed the advertisements displayed to assess the effect of persona substitution on ad recommendations.

Since in the previous subsection we verified the stability of LLM in ad ratings, we proceeded to use the LLM to rate the ads based on their alignment with privacy attributes. Through the ratings, we quantified how much the persona attributes influenced the advertisement content. For each group of three privacy variants (e.g., young, mid-aged, old for age), we treated them as distinctly different personas. For the three personas generated in the same group based on the same privacy attribute values, we treated them as similar personas. For both different personas and similar personas, we used the Kruskal–Wallis H test [44] and post-hoc tests [19] to evaluate whether there were significant differences in ad-attribute alignment scores among personas from different groups (**TRQ4b**), and whether the ad-attribute alignment scores for similar personas remained consistent across the three rounds (**TRQ4a).**

Fig. 7 shows the scores for different personas in each round, with all personas generated from contrasting privacy attribute values (e.g., young-old, low income-high income) showing significant differences. This indicates that advertisements shown to distinctly different personas exhibited significant differences (**TRQ4b**). In contrast, Fig. 8 shows that personas generated from the same privacy attribute values showed stable rating scores, with no significant differences detected. The analysis revealed that advertisements served to similar personas on the same website were more similar to each other (**TRQ4a**).

5 USER STUDY

We conducted a user study[5] with 15 participants to evaluate our approach. The study examines the following research questions.

RQ1: How can Privacy Auditing Sandbox support end users to audit online content recommendations by verifying their privacy hypotheses?

RQ2: How does the engagement of the Privacy Auditing Sandbox influence users' attitudes, awareness, and sense of empowerment with online content recommendations?

RQ3: Do users find the Privacy Auditing Sandbox usable and useful?

5.1 Participants

We recruited 15 participants through word-of-mouth and social media, all of whom participated in the study virtually via Zoom. Before the study, participants completed a pre-screening survey that collected essential demographic information such as age, gender, state of residence, race/ethnicity, occupation, and education level. We aimed to recruit a group of participants from diverse demographic backgrounds. The comprehensive demographic details of the participants are outlined in Table 3. The age range of participants was from 22 to 51, including five females and ten males. Each participant was compensated $25 USD for their participation.

5.2 Study Design

Each study session lasted around 60 minutes. The session consisted of two phases.

5.2.1 Study procedure. After completing the informed consent process and receiving a brief introduction to the study, each participant completed the following two phases of the study.

- **Phase 1: Using existing sets of personas to verify privacy hypotheses.** To understand whether and how users could verify privacy hypotheses through our system, we employed a combination of the "think-aloud" method [30] and semi-structured interviews. The experimenter first introduced the Privacy Auditing Sandbox to the users, ensuring they understood how persona variants were generated and how to load them into the browser to receive ads.

[5]The study protocol was reviewed and approved by the IRB at our institution

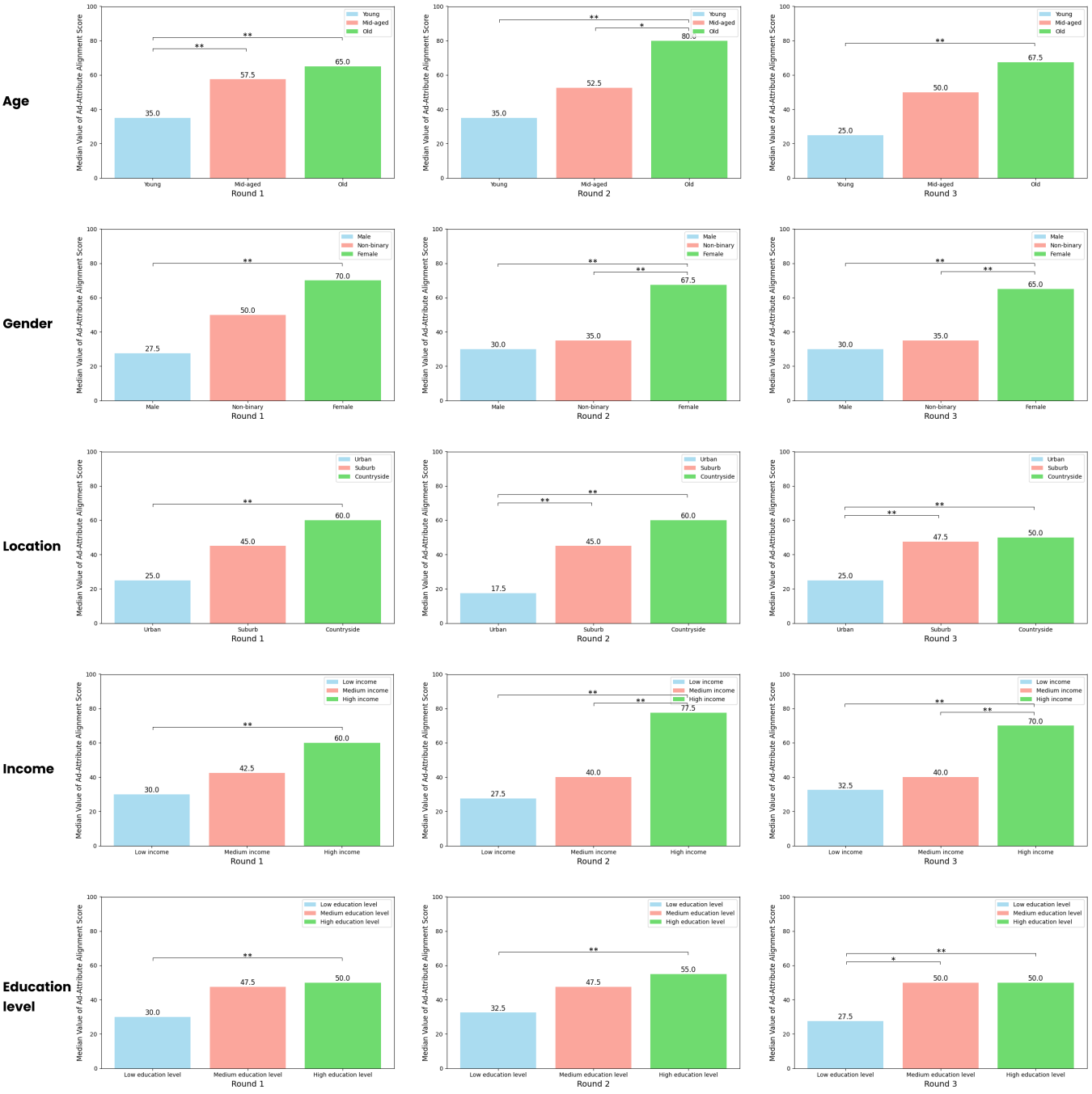

Figure 7: Comparison of Ad-Attribute Alignment Scores for different personas. Single asterisk (*) denote p<0.1, while double asterisks (**) indicate p<0.05.

Each participant was then asked to review five personal attributes—age, gender, urbanization, income level, and education level—where each attribute corresponded to a different set of pre-existing persona variants and the ads collected using these personas in the Privacy Auditing Sandbox's hypothesis consolidation interfaces. These five attributes were specifically chosen because online recommendation systems frequently rely on this data to tailor their offerings [48]. As they navigated the personas, participants were instructed to vocalize the connections they observed between the ads and the personal attributes, as well as the reasons they believed these connections existed. During the "think-aloud" process,

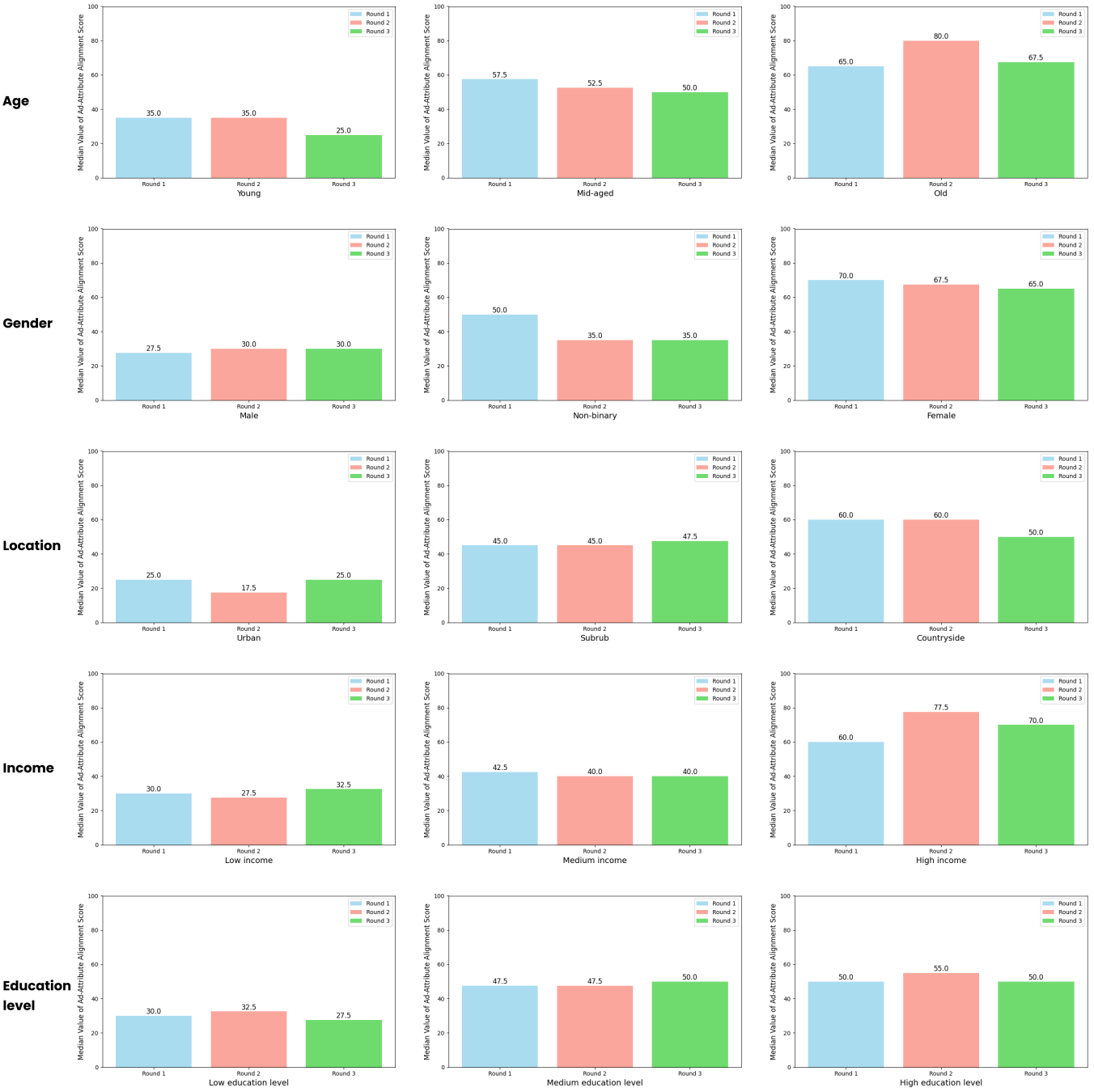

Figure 8: Comparison of Ad-Attribute Alignment Scores for similar personas. There was no statistical significant difference between any rounds within the same attribute value group.

as participants shared their immediate feedback, researchers could interject with follow-up questions or ask for clarifications. A detailed list of these five sets of personas can be found in the Appendix A.

- **Phase 2: Exploring privacy hypotheses with generated personas.** The goal of this phase is to evaluate the entire workflow of Privacy Auditing Sandbox, study its impact on participants' understanding of online content recommendations, and explore how this understanding influences their future interactions with websites. Since in the previous phase, participants tested the connections between five personal attributes and online ads, they may develop an interest in

ID	Gender	Age	Ethnicity	Educational Level	Occupation
P1	Female	26	Asian and Pacific Islander	Master's degree	Financial analyst
P2	Female	21	Asian and Pacific Islander	Bachelor's degree	Student
P3	Female	22	Asian and Pacific Islander	Bachelor's degree	Student
P4	Female	51	White or Caucasian	Master's degree	Property manager
P5	Male	22	Black or African American	High school graduate	Freelancer
P6	Female	24	White or Caucasian	Bachelor's degree	Student
P7	Male	31	White or Caucasian	Bachelor's degree	Project Manager
P8	Male	25	Black or African American	Master's degree	Teacher
P9	Male	30	Asian and Pacific Islander	Master's degree	Student
P10	Male	30	Asian and Pacific Islander	Doctorate degree	Student
P11	Male	27	Black or African American	Bachelor's degree	Technician
P12	Female	42	White or Caucasian	Bachelor's degree	Writer
P13	Male	25	Black or African American	Bachelor's degree	Architect
P14	Male	30	White or Caucasian	Bachelor's degree	Program Coordinator
P15	Male	22	Black or African American	Bachelor's degree	Student

Table 3: Participant Demographics

a particular personal attribute. This phase allows users to experience the process of creating a persona and collecting ads in real time on actual websites, allowing them to further explore the relationship between personal attributes and online ads. First, the participant enters a natural language description of the persona they want to generate and selects a personal attribute to create a set of corresponding personas. Then, the participant was asked to choose a website and specify how many times each persona should browse that site. The Privacy Auditing System triggers the launch of a new browser window, which automatically replaces their real personal data with synthetic persona data and collects ads from the websites visited. Once all the ads had been collected, the participant reviewed the distribution of ads collected using this set of personas in relation to the personal attributes and explained, in a think-aloud manner, how the ads relate to those personal attributes.

At the end of the study session, the participant completed a System Usability Scale (SUS) test [4] to assess the system's usability. The experimenter also conducted a brief follow-up semi-structured interview based on the participant's responses and the experimenter's observations during the study session, asking the participant to explain the rationales behind their actions and ratings, elaborate on their thoughts on the usability and usefulness of the system, and describe the impacts of using the system on their privacy awareness, attitudes, and future behaviors.

Note that this study does not include a baseline for comparison, because we cannot identify an existing tool that can support end users to verify privacy hypotheses and explore the relationship between personal attributes and online ads in third-party websites in the similar way as the Privacy Auditing Sandbox. Instead, we focus on evaluating the usability, user interaction, and perceived usefulness of the system.

5.2.2 Websites. We selected five representative websites (as shown in Table 4) to test the Privacy Sandbox with personalized advertisements. Our website selection process was based on the methodology previously employed by Zeng et al. [71]. The selection was based on the following criteria: 1) representing a diverse range of website topics, 2) presenting multiple advertisements on each website, and 3) using Google Ads as the primary ad provider. The decision to use Google Ads was driven by two key factors: (1) Google Ads is the most widely used advertising platform on the Internet, reaching over two million websites and apps and more than 90% of Internet users worldwide [6]; and (2) Google Ads has extensive access to personal data stored in the Chrome browser (e.g., browsing history, Google accounts), which was central to our study.

Website	Topics	Site Rank
www.weather.com	Weather forecasts	35
www.foxnews.com	National news	43
www.thepioneerwoman.com	E-commerce	1,049
boardingarea.com	Traveling	27,905
www.algebra.com	Education	54,809

Table 4: Websites visited by participants in the study

5.3 Data Analysis Methods

We employed standard open coding procedures [7] to perform qualitative data analysis on the transcripts of user think-aloud sessions and the concluding interviews of the study. Two researchers in our team independently began the coding process using MAXQDA. One researcher coded 20% of the sample, generating an initial set of codes. The second researcher then coded the same portion, introducing new codes as needed. Any discrepancies between their coding were resolved through discussion to ensure consistency, resulting in a unified codebook. Then, this codebook was used to

[6] https://support.google.com/google-ads/answer/117120?hl=en

perform a thematic analysis, identifying and defining key themes that emerged from the interviews and aligned with the established codes.

The complete codebook is presented in Appendix B. These themes were then refined and developed to inform the study's findings, which are presented in detail in Section 6.

6 RESULTS AND FINDINGS

Our study demonstrates that the Privacy Auditing Sandbox effectively enabled participants to validate their privacy hypotheses by identifying the connections between personal attributes and targeted advertisements. Participants successfully linked personal attributes to ad themes and visual elements, highlighting patterns related to age, gender, and income (**RQ1**). After using our tool, users expressed discomfort over privacy intrusions(**RQ2**). The system was generally rated as user-friendly and useful (**RQ3**).

6.1 RQ1: Participants validated their privacy hypotheses by linking personal attributes with the theme of the ad and visual elements

.

The Sandbox enabled participants to systematically test their assumptions and observe real differences across persona variants, revealing hidden patterns in ad content. As shown in the result of our post-study SUS questionnaire in Fig. 9, over 75% of participants agreed that they could quickly learn to use the Privacy Auditing Sandbox. Once they learned it, over 75% of participants found our system easy to use. This ease of use was crucial in encouraging users to try multiple hypotheses and interact deeply with the system. All participants discovered that personal attributes clearly influenced ad recommendations. One participant (P5) explained: , "*I think it (the Privacy Auditing Sandbox) is pretty easy (to use). You just input the person you're thinking about and then compare the differences (among ads), whether those are based on gender, age, education, or work background.*" Below we summarized two focus areas where participants linked personal attributes with ads to validate their privacy hypotheses: the theme and visual elements of the ads.

6.1.1 Ad themes. In the auditing process, participants successfully validated the impact of persona variants on ads primarily through the correlation between the ad theme and personal attributes, particularly in the areas of age, gender, and income. For age, participants found that young personas were more likely to receive ads for video games, fast food, and workouts, while older personas were shown more ads related to financial aid, healthcare, and skin repair. The users also provided some explanations. For example, P5 said "*The ad shows a smiling young person playing video game...This one definitely has a relevancy to the age since young people are probably interested in gaming.*". P1 said, "*For these (middle-aged and old) people, financial aid is probably what they need, because maybe in middle age, like people who own their own company, they might fall into some bad financial situation.*" In terms of gender, participants believed that ads targeting women were easier to identify. They observed that ads for women often focused on footwear or fitness, while ads targeting men included more tech-related content (e.g., Starlink). Participants did not find any ads specifically targeting non-binary individuals when browsing ads received when using the non-binary persona. Regarding income level, participants noticed that lower-income personas were more likely to see ads for coupons, financial aid, and fast food, while middle- and high-income personas, although occasionally exposed to discount-related ads, were the only ones shown ads for expensive products (e.g., financial products, luxury goods), as P2 said "*luxury products only appear in the person with very high income.*"

6.1.2 Visual elements. The visual elements in advertisements can also help participants identify the connection between ads and persona characteristics, particularly in terms of gender and age. If an ad features a female character or uses colors like pink or purple, participants are more likely to think the target audience is female. For instance, P1 said, "*The advertisement for females tends to use female models.*" Similarly, if the ad features an older person, participants assume that the target audience is also older adults. As P7 noted, "*Even though the ad topic is general, the visual element is an older gentleman, which means that it has some relation with older people.*"

6.2 RQ2: Participants exhibited conflicting emotions between accepting personalized content recommendations and concerns about privacy invasion

The Sandbox helped participants reflect on their emotional responses to personalized recommendations. This reflection surfaced tensions between appreciation for relevance and discomfort about surveillance. Some participants were somewhat accepting that websites use personal attributes to tailor advertisements, appreciating that this customization could enhance the relevance of content and facilitate more targeted product or service offerings. P5 said: "*I feel like I am more influenced to shop at places when I get targeted ads that align with my personal attributes.*" But many also felt discomfort when seeing how easily systems inferred personal details. P10 noted: "*But sometimes it annoys me because I am... let's say I looked at luxury cruises just for fun. And then suddenly they pop up everywhere... The system knows what I'm doing, which bothers me.*" By making this targeting process visible, our system moved users from vague discomfort to informed concern—making privacy risks tangible and specific. Some participants expressed "privacy resignation." [20], knowing what's happening, but feeling unable to change it. Even in these cases, the tool gave users language and structure to describe what they were experiencing—transforming abstract discomfort into concrete awareness.

6.3 RQ3: Participants found the privacy auditing sandbox generally usable and useful

Overall, participants found the tool easy to learn, well-structured, and effective for exploring the impact of personal data on recommendations. The System Usability Scale(SUS) questionnaire [4], indicates a generally positive user experience, as shown in Fig. 9. Participants rated each item on a 5-point Likert scale. They found

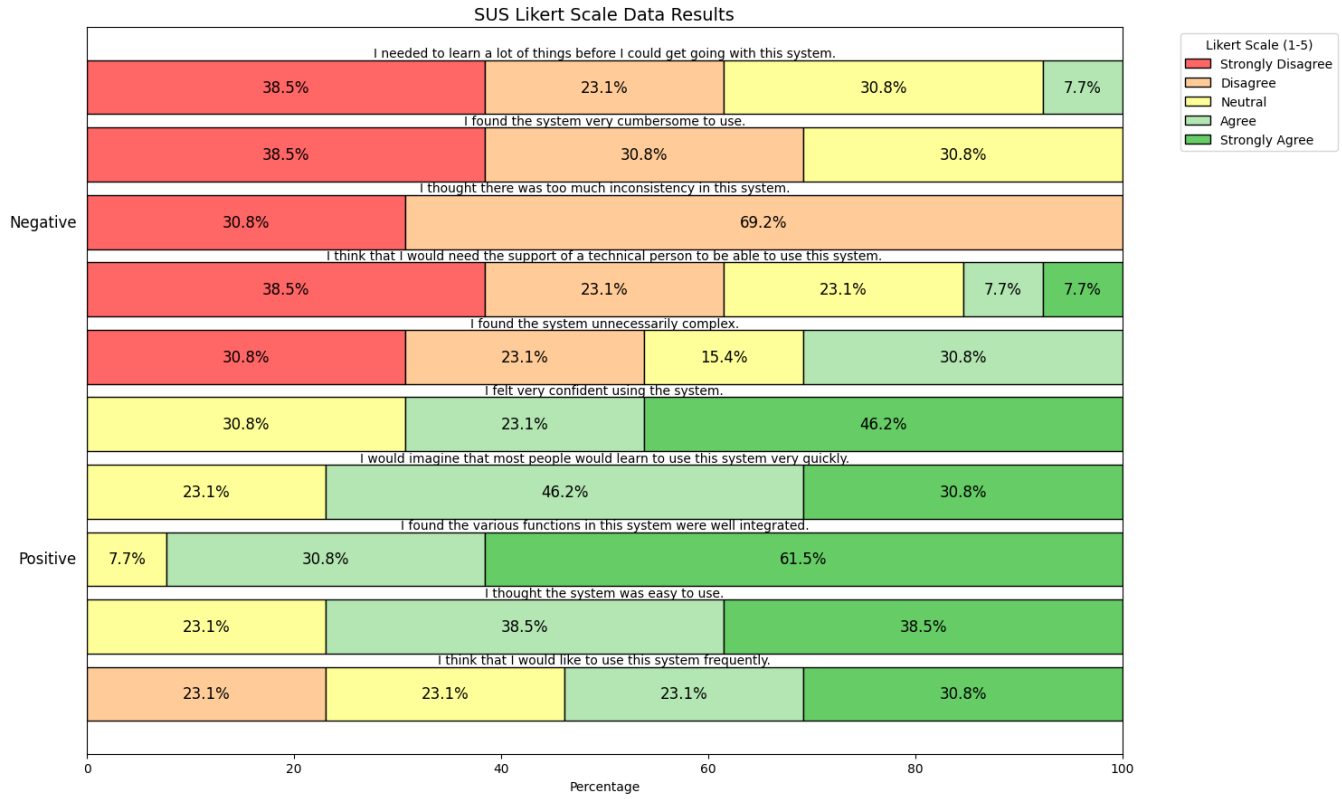

Figure 9: Result of Post-study Questionnaire. The statements are categorized into positive and negative ones based on their wording. For the negative statements (first five), lower ratings are better. For the positive statements (bottom five), higher ratings are better.

the system user-friendly and well-integrated, with most anticipating a shallow learning curve. The confidence in using the system is high.

Participants do not perceive too much inconsistency in the system, nor do they feel cumbersome to use the system. Participants highlighted that the system helped them ask better questions about data use. For example, P15 shared, "*For me, I think it's useful. If, for example, I want to see some kind of privacy setup on a website, I could actually configure it with the software you're building. So, when I set it up with a persona, I know I'll get everything I need (for privacy auditing).*" P3 added, "*When the ads are clearly targeted, for example, based on a certain income level (of the persona), it's definitely something I will pay attention to in the future. I will wonder if this ad is directed at me because of my attributes.*" These reflections show how the tool influenced users' thinking beyond the session, prompting ongoing curiosity and vigilance.

Among the 15 participants, eight chose to have each persona variant revisit their selected websites three times, four chose five repetitions, and three opted for two repetitions. This variability shows that some used the tool for deep audits, while others opted for quicker validation—highlighting its flexibility.

In particular, participants did not rate "use this system frequently" highly. In the follow-up interview, they explained two main reasons: first, they believed that frequent and repetitive privacy audits were not necessary for a certain website. Second, it took a bit too long to conduct each audit using the Privacy Auditing Sandbox, mainly due to the time it takes to sample each website repeatedly to scrape data. Nevertheless, this would not be a barrier for them to audit the target website once. Additionally, some users found the Privacy Auditing Sandbox a bit complex at the beginning, but also mentioned that, after a period of learning, they were able to quickly grasp how to use it, since "*It breaks everything down by category, and the graphics are pretty clear*" (P14).

7 DISCUSSION

7.1 The Role of Generated Persona in Usable Privacy and End-User Auditing

Generated personas play a valuable role in helping users gain generalizable privacy knowledge while minimizing risks. They support "what-if" types of exploration, enable large-scale testing, and allow users to compare how algorithmic results may differ between different personas. Well-constructed personas can also evoke empathetic responses, as shown in [9], helping users relate the outcomes observed with the personas to potential implications for their own future behavior. As confirmed in Section 6.1, users were able to load various generated personas into the browser, observe the relationships between ad themes, visual elements, and user characteristics,

and validate the impact of these characteristics on recommended content. This demonstrates that users could effectively identify and internalize the consequences of sharing information related to specific privacy attributes by connecting the ads they viewed to the traits of the generated personas.

A key reason users are drawn to using generated personas for privacy audits is that they can test system feedback without exposing their real data. When users share personal data, they lose control over how it is used, making it difficult for them to experiment with different privacy behaviors (e.g., trying to allow access to a particular data to see whether the trade-off in improved service quality is worth it) and observe the outcomes for audits. Using generated personas alleviates concerns about potential privacy breaches, encouraging more active engagement in privacy audits.

Although previous research has pointed out the risk that LLM-generated personas may reinforce existing stereotypes, in the context of our study, these biases may not be detrimental. When recommendation systems process persona data that reflect real-world biases, they generate outcomes representative of real-world services or experiences, as these systems often embed such biases. Consequently, if our generated personas contain biases, they serve as useful tools for auditing recommendations made by external websites and apps, which frequently rely on stereotypes and biases to inform their recommendations.

7.2 The Complexity in the Design of Privacy Auditing Sandbox

A common issue reported by users in the study was the system's high level of complexity, particularly at the beginning of use. Although participants found the system to be learnable, there was an initial learning curve in understanding how to navigate the system and interpret the audit results. The complexity of the Privacy Auditing Sandbox arises from several factors, including the interface design, users' data literacy, and unfamiliarity with outputs generated by LLMs.

7.2.1 Interface Complexity. As discussed in Section 6.3, although the system interface was generally clear, some participants initially found it overwhelming due to their unfamiliarity with the concept of end-user auditing. However, after gaining experience with the system, they appreciated its structure and clear visuals, suggesting that the interface becomes more approachable over time. To reduce this initial barrier, a simplified onboarding process or a tutorial could help users become familiar with the layout and functionality of the system more quickly.

7.2.2 Data literacy. Users with limited data literacy struggled to interpret the normal distribution charts used to represent privacy risks. Although labels and graphics were provided, additional explanations or visual aids could enhance comprehension for those less familiar with statistical concepts.

7.2.3 Lack of Familiarity with LLM Generation. Another challenge was the user's unfamiliarity with LLM generation processes. The Privacy Auditing Sandbox integrates LLM-generated synthetic user data, associated with individual personas, into its auditing system, which some participants found confusing. Providing clearer explanations of how the LLM operates, along with contextual insights into why specific recommendations are made, could help build greater user confidence in the system's outputs.

7.3 An End-User Empowerment Approach for Algorithm Accountability

The approach introduced by Privacy Auditing Sandbox exemplifies an end-user empowerment approach to enhance algorithm accountability. Instead of relying on the first party, i.e., the owner of the website or third parties such as regulatory bodies or user advocacy groups, to perform audits, our approach allowed everyday web users to perform audits of the relationship between the user's personal attributes and the behaviors of online content recommender systems. This approach has several distinct benefits.

First, it allows users to validate a "long tail" of hypotheses that they are personally interested in. Unlike first- or third-party approaches, which likely will only cover hypotheses relevant to the most common privacy concerns, end users using our tool can now audit for hypotheses they come up with themselves. Previous research has shown that users harbor notions about how their age, location, or browsing history might influence the ads they encounter [38]. However, current auditable AI documentation is designed primarily for regulators and experts rather than end users affected by AI decisions, making it difficult for users to evaluate their privacy hypotheses [55]. Through the Privacy Auditing Sandbox, users can use personas to validate their privacy hypotheses without compromising their real personal data. As described in Section 6.1, users can evaluate whether their data have been used in unexpected ways by linking personal attributes to ad themes and visual elements. Prior research has highlighted ad discrimination in online recommendation systems [29]. We anticipate that the Privacy Auditing Sandbox can also support verifying these concerns by enabling users to investigate potential discrimination based on specific privacy attributes.

Second, this approach allows end users to take initiatives by actively participating in algorithmic auditing and experience the varying outcomes of the systems when using different personas, instead of being only passive consumers of auditing reports created by first- or third-parties. This can improve the user's sense of engagement and potentially amplify the positive impact of auditing on user privacy literacy and future behavioral changes based on experiential learning theories [23, 34]. As described in Section 6.2, participants in our study demonstrated a nuanced awareness of the trade-offs between personalized content and privacy risks. This suggests that empowering users to take more initiative in managing their data and privacy could bridge the gap between perceived benefits and concerns. From the user's perspective, privacy awareness is the prerequisite of privacy literacy. As shown by the discomfort and sense of powerlessness that some participants experienced (Section 6.2), end users often feel trapped in a cycle of data collection and targeting without clear consent or control. To address this issue, it is crucial to involve users actively in privacy audits, transforming them from passive subjects to engaged participants in the process.

Third, beyond individual users, end-user empowerment has the potential to transform the broader data ecosystem by promoting awareness and accountability among all stakeholders. Privacy awareness and literacy are fundamental to this transformation.

Users who better understand data practices are more likely to demand transparency and ethical behavior from companies. Empowered users can more effectively hold platforms accountable by questioning the fairness and appropriateness of algorithms that curate their content and experiences [68]. In this context, end-user empowerment can facilitate a bottom-up approach to privacy activism. When users are educated about the implications of data collection, they are more inclined to engage in collective actions, such as data activism [39, 65, 66], to drive change. These grassroots movements can catalyze broader industry shifts towards more ethical data practices. By actively participating in the dialogue about data use, empowered users contribute to the democratization of algorithmic governance. This bottom-up pressure complements top-down regulatory efforts, creating a more balanced and user-centric approach to data privacy and algorithmic fairness.

7.4 Ethical and Legal Considerations

While our approach can help users verify their privacy hypotheses and increase engagement in privacy audits by lowering its barrier to end users, we also identified certain ethical and legal risks associated with its implementation.

One concern is the potential to reinforce user bias. Biased or stereotypical privacy personas may be useful for analyzing the relationship between persona traits and recommended content, but repeated exposure to these personas could unintentionally reinforce users' existing biases. It is important to warn users about the risk of bias and stereotyping in the generated personas, and further research is necessary to mitigate this effect.

Another concern is the potential for malicious misuse, which could lead to cybersecurity risks. Although the method is intended to enhance privacy literacy, the creation of realistic personas could be exploited to develop undetectable bots or facilitate phishing attacks. This underscores the need for stronger technical safeguards and policy measures to address these risks. For instance, incorporating advanced identity verification techniques or restricting access to persona generation tools to authorized users could help prevent such misuse.

7.5 Limitation and Future Work

We identify three key limitations of our work and suggest future improvements for them: privacy auditing sandbox prototype, experiment design, and generalizability of auditing tasks.

7.5.1 Privacy auditing sandbox prototype. Our Privacy Auditing Sandbox prototype was developed primarily as a proof-of-concept to support the experiments presented in this paper. Consequently, its ability to audit a wide range of privacy attributes is currently limited, with gaps in coverage for commonly used data types, such as search history. From a technical perspective, those data are typically stored and managed on the server-side by service providers rather than locally, making the data replacement approach used in our Sandbox infeasible.

In addition, some data that influence online recommendations, such as inferred behavioral patterns and social connections, are difficult to replicate using only modifiable personal attributes. To overcome these limitations, we plan to enhance the sandbox by expanding its capability to simulate a broader range of privacy data, including more complex attributes that underpin behavioral and social inferences.

Moreover, the existing proof-of-concept uses the GPT-4 model through its commercial API, which requires users to have access to the model to use the tool. Our future work will explore how the use of other LLMs, especially open-source ones, to achieve similar performance while making the tool more accessible and easily deployable.

7.5.2 Experiment design. The primary objective of our experimental design was to assess the feasibility of our approach in helping users test their privacy hypotheses and raise their privacy awareness. Rather than directly measuring users' privacy knowledge gains or behavioral changes resulting from the Privacy Auditing Sandbox, this experiment focused on evaluating the viability of our method. While participant interviews provided qualitative insights suggesting the potential of our approach to foster both privacy knowledge and behavioral changes, further validation is required in long-term, real-world environments. To address this, we plan to conduct longitudinal deployment studies of the Privacy Auditing Sandbox. These studies will allow us to (1) implement pre- and post-tests to evaluate users' privacy knowledge acquisition; (2) conduct longitudinal studies to track changes in users' privacy behaviors over time. This future research will provide a more comprehensive understanding of our approach's effectiveness in promoting privacy literacy and encouraging lasting behavioral changes.

7.5.3 Generalizability of auditing tasks. In our experiment, we focused on online advertisements for downstream tasks as proof-of-concept due to their ubiquity, user familiarity, and sensitivity to changes in privacy data. However, we believe that our approach can be generalized to a variety of other contexts, allowing users to audit the effects of their privacy behaviors in areas such as friend recommendations on social media, dynamic pricing in e-commerce, and personalized content on streaming platforms. In the next phase of developing the Privacy Auditing Sandbox, our goal is to extend its capabilities to cover these additional tasks, facilitating audits of algorithmic accountability. We will also conduct further deployment and testing to evaluate how effectively our approach enhances user engagement in privacy audits and promotes more equitable algorithmic outcomes.

8 CONCLUSION

We introduced a novel interactive sandbox approach to democratize end-user auditing of online content recommendations, enabling users to test their hypotheses by observing how websites' recommendation algorithms respond to synthetic user personas with variations in selected key attributes. In a case study on targeted ads with 15 participants, the Privacy Auditing Sandbox system demonstrated its usability and effectiveness. Participants successfully verified their privacy-related hypotheses by analyzing ad themes and visual elements, and actively engaging in privacy audits. Our findings offered design implications for empowering users to audit algorithmic accountability through the use of artificially generated personas.

REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering* 17, 6 (2005), 734–749.

[2] Charu C Aggarwal et al. 2016. *Recommender systems*. Vol. 1. Springer.

[3] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.

[4] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

[5] Ramesh Baral and Tao Li. 2017. PERS: A personalized and explainable POI recommender system. *arXiv preprint arXiv:1712.07727* (2017).

[6] Sophie C Boerman, Sanne Kruikemeier, and Frederik J Zuiderveen Borgesius. 2017. Online behavioral advertising: A literature review and research agenda. *Journal of advertising* 46, 3 (2017), 363–376.

[7] Meryl Brod, Laura E Tesler, and Torsten L Christensen. 2009. Qualitative research and content validity: developing best practices based on science and experience. *Quality of life research* 18 (2009), 1263–1278.

[8] Maryam Abdulaziz Saad Bubukayr and Mounir Frikha. 2022. Web Tracking Domain and Possible Privacy Defending Tools: A Literature Review. *Journal of Cybersecurity* 4, 2 (2022), 79.

[9] Chaoran Chen, Weijun Li, Wenxin Song, Yanfang Ye, Yaxing Yao, and Toby Jiajun Li. 2023. An Empathy-Based Sandbox Approach to Bridge Attitudes, Goals, Knowledge, and Behaviors in the Privacy Paradox. *arXiv preprint arXiv:2309.14510* (2023).

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[11] Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189* (2023).

[12] I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv preprint arXiv:2307.13528* (2023).

[13] Yoonseo Choi, Eun Jeong Kang, Seulgi Choi, Min Kyung Lee, and Juho Kim. 2024. Proxona: Leveraging LLM-Driven Personas to Enhance Creators' Understanding of Their Audience. *arXiv preprint arXiv:2408.10937* (2024).

[14] Cary Coglianese and David Lehr. 2019. Transparency and algorithmic governance. *Administrative law review* 71, 1 (2019), 1–56.

[15] Sasha Courville, Christine Parker, and Helen Watchirs. 2003. Auditing in Regulatory Perspective. *Law & Pol'y* 25 (2003), 179.

[16] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2014. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491* (2014).

[17] Deepika Deepika, Rajnesh Malik, Saurabh Kumar, Rishabh Gupta, and Ashutosh Kumar Singh. 2020. A review on data privacy using attribute-based encryption. In *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*.

[18] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[19] Alexis Dinno. 2015. Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. *The Stata Journal* 15, 1 (2015), 292–300.

[20] Nora A Draper. 2017. From privacy pragmatist to privacy resigned: Challenging narratives of rational choice in digital privacy debates. *Policy & Internet* 9, 2 (2017), 232–251.

[21] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–19.

[22] Huaqi Gao and Shunke Zhou. 2022. A Survey of Explainable E-Commerce Recommender Systems. In *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 723–730.

[23] James W Gentry. 1990. What is experiential learning. *Guide to business gaming and experiential learning* 9, 1 (1990), 20–32.

[24] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. 2020. PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 196–204.

[25] Joshua Gomez, Travis Pinnick, and Ashkan Soltani. 2009. KnowPrivacy. (2009).

[26] Ido Guy and David Carmel. 2011. Social recommender systems. In *Proceedings of the 20th international conference companion on World wide web*. 283–284.

[27] Muhammad Haroon, Magdalena Wojcieszak, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, and Zubair Shafiq. 2023. Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the national academy of sciences* 120, 50 (2023), e2213020120.

[28] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring misinformation in video search platforms: An audit study on YouTube. *Proceedings of the ACM on human-computer interaction* 4, CSCW1 (2020), 1–27.

[29] Basileal Imana, Aleksandra Korolova, and John Heidemann. 2024. Auditing for Racial Discrimination in the Delivery of Education Ads. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2348–2361.

[30] Riitta Jääskeläinen. 2010. Think-aloud protocol. *Handbook of translation studies* 1 (2010), 371–374.

[31] Hao Jiang, Chuanzhen Li, Juanjuan Cai, and Jingling Wang. 2023. RCENR: A Reinforced and Contrastive Heterogeneous Network Reasoning Model for Explainable News Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1710–1720.

[32] Rahul Katarya and Om Prakash Verma. 2017. An effective web page recommender system with fuzzy c-mean clustering. *Multimedia Tools and Applications* 76 (2017), 21481–21496.

[33] KirkpatrickPrice. 2024. Privacy Assessment Audit Services. https://kirkpatrickprice.com/audit/privacy-assessment/. Accessed: 2024-09-13.

[34] David A Kolb. 2014. *Experiential learning: Experience as the source of learning and development*. FT press.

[35] Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 512 (nov 2022), 34 pages. https://doi.org/10.1145/3555625

[36] Michelle S. Lam, Ayush Pandit, Colin H. Kalicki, Rachit Gupta, Poonam Sahoo, and Danaë Metaxa. 2023. Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 360 (oct 2023), 37 pages. https://doi.org/10.1145/3610209

[37] Mathias Lecuyer, Riley Spahn, Yannis Spiliopolous, Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu. 2015. Sunlight: Fine-grained targeting detection at scale with statistical confidence. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 554–566.

[38] Hao-Ping Hank Lee, Jacob Logas, Stephanie Yang, Zhouyu Li, Nata Barbosa, Yang Wang, and Sauvik Das. 2023. When and Why Do People Want Ad Targeting Explanations? Evidence from a Four-Week, Mixed-Methods Field Study. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2903–2920.

[39] Toby Jia-Jun Li, Yuwen Lu, Jaylexia Clark, Meng Chen, Victor Cox, Meng Jiang, Yang Yang, Tamara Kay, Danielle Wood, and Jay Brockman. 2022. A Bottom-Up End-User Intelligent Assistant Approach to Empower Gig Workers against AI Inequality. In *Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work*. 1–10.

[40] Yang Li, Tong Chen, and Zi Huang. 2021. Attribute-aware explainable complementary clothing recommendation. *World Wide Web* 24 (2021), 1885–1901.

[41] Mun Ling Lo. 2012. *Variation theory and the improvement of teaching and learning*. Göteborg: Acta Universitatis Gothoburgensis.

[42] Yuwen Lu, Chao Zhang, Yuewen Yang, Yaxing Yao, and Toby Jia-Jun Li. 2024. From awareness to action: Exploring end-user empowerment interventions for dark patterns in ux. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–41.

[43] Ference Marton. 2014. *Necessary conditions of learning*. Routledge.

[44] Patrick E McKight and Julius Najab. 2010. Kruskal-wallis test. *The corsini encyclopedia of psychology* (2010), 1–1.

[45] Peter Menzies and Helen Beebee. 2001. Counterfactual theories of causation. (2001).

[46] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344.

[47] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251* (2023).

[48] Marwa Hussien Mohamed, Mohamed Helmy Khafagy, and Mohamed Hasan Ibrahim. 2019. Recommender systems challenges and solutions survey. In *2019 international conference on innovative trends in computer engineering (ITCE)*. IEEE, 149–155.

[49] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[50] Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (eu)* 679 (2016), 2016.

[51] Manuel Rudolph, Denis Feth, and Svenja Polst. 2018. Why users ignore privacy policies–a survey and intention model for explaining user privacy behavior. In

Human-Computer Interaction. Theories, Methods, and Human Issues: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part I 20. Springer, 587–598.

[52] Joni Salminen, Chang Liu, Wenjing Pian, Jianxing Chi, Essi Häyhänen, and Bernard J Jansen. 2024. Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[53] Joni Salminen, Joao M Santos, Haewoon Kwak, Jisun An, Soon-gyo Jung, and Bernard J Jansen. 2020. Persona perception scale: development and exploratory validation of an instrument for evaluating individuals' perceptions of personas. *International Journal of Human-Computer Studies* 141 (2020), 102437.

[54] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22, 2014 (2014), 4349–4357.

[55] Nicolas Scharowski, Michaela Benk, Swen J Kühne, Léane Wettstein, and Florian Brühlmann. 2023. Certification labels for trustworthy ai: Insights from an empirical mixed-method study. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 248–260.

[56] Andreas Schuller, Doris Janssen, Julian Blumenröther, Theresa Maria Probst, Michael Schmidt, and Chandan Kumar. 2024. Generating personas using LLMs and assessing their viability. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[57] Nasim Sonboli, Jessie J Smith, Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. 2021. Fairness and transparency in recommendation: The users' perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 274–279.

[58] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Potential for discrimination in online targeted advertising. In *Conference on fairness, accountability and transparency*. PMLR, 5–19.

[59] Chang-You Tai, Liang-Ying Huang, Chien-Kun Huang, and Lun-Wei Ku. 2021. User-centric path reasoning towards explainable recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 879–889.

[60] Khanh Hiep Tran, Azin Ghazimatin, and Rishiraj Saha Roy. 2021. Counterfactual explanations for neural recommenders. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1627–1631.

[61] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. 2012. Smart, useful, scary, creepy: perceptions of online behavioral advertising. In *proceedings of the eighth symposium on usable privacy and security*. 1–15.

[62] Aleksandra Urman, Ivan Smirnov, and Jana Lasser. 2024. The right to audit and power asymmetries in algorithm auditing. *EPJ Data Science* 13, 1 (2024), 19.

[63] Guda Van Noort, Edith G Smit, and Hilde AM Voorveld. 2013. The online behavioural advertising icon: two user studies. *Advances in Advertising Research (Vol. IV) The Changing Roles of Advertising* (2013), 365–378.

[64] Kaan Varnali. 2021. Online behavioral advertising: An integrative review. *Journal of Marketing Communications* 27, 1 (2021), 93–114.

[65] Nicholas Vincent, Brent Hecht, and Shilad Sen. 2019. "Data strikes": evaluating the effectiveness of a new form of collective action against technology companies. In *The World Wide Web Conference*. 1931–1943.

[66] Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. 2021. Data leverage: A framework for empowering the public in its relationship with technology companies. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 215–227.

[67] Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, et al. 2024. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802* (2024).

[68] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3351095.3372833

[69] Shuyuan Xu, Yunqi Li, Shuchang Liu, Zuohui Fu, Yingqiang Ge, Xu Chen, and Yongfeng Zhang. 2021. Learning causal explanations for recommendation. In *The 1st International Workshop on Causality in Search and Recommendation*.

[70] Yaxing Yao, Davide Lo Re, and Yang Wang. 2017. Folk models of online behavioral advertising. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1957–1969.

[71] Eric Zeng, Rachel McAmis, Tadayoshi Kohno, and Franziska Roesner. 2022. What factors affect targeting and bids in online advertising? a field measurement study. In *Proceedings of the 22nd ACM Internet Measurement Conference*. 210–229.

[72] Wei Zhang, Junbing Yan, Zhuo Wang, and Jianyong Wang. 2022. Neuro-symbolic interpretable collaborative filtering for attribute-based recommendation. In *Proceedings of the ACM Web Conference 2022*. 3229–3238.

[73] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 83–92.

A PERSONA USED IN THE USER STUDY

B CODEBOOK

(1) Reason for having relationship
  (a) Ad visual elements
    (i) Users tend to see the similar age group people in ads
    (ii) Ads for men often use red and blue color
    (iii) Ads for women often use pink or purple color
    (iv) Ads for women often use images of women themselves
  (b) Ad topic
    (i) Middle ages are more business oriented
    (ii) Workout ads are more related with young people
    (iii) Beauty ads are more related with young people
    (iv) Products for students are more related with young people
    (v) Video game is more related with young people
    (vi) Older adults are more likely to see disaster prevention ads
    (vii) Middle-aged/old adults are more likely to see skin repair ads
    (viii) Older adults are more likely to see financial aid ads
    (ix) Older adults are more likely to see health related ads
    (x) Women are more likely to see female shoe ads
    (xi) Women see female fitness ads
    (xii) Men are more likely to see tech-related ads
    (xiii) Drive up service target suburb/countryside people
    (xiv) Expensive product ads target high-income people
    (xv) Coupon and discount target people with low income and education level
    (xvi) Lower income people get more ads about junk food
    (xvii) Ads for financial aid target low income people
    (xviii) Ads for continuing studies target low education people
    (xix) Ads for traveling may target medium or high income people
(2) Attitudes, awareness, and sense of empowerment
  (a) Positive attitudes
    (i) The content is just enough to verify their hypothesis
    (ii) Personal data tracking can improve the recommendation quality
    (iii) Support verifying relation between privacy attributes and ads
    (iv) Improve privacy awareness
  (b) Negative attitudes
    (i) Privacy Resignation
    (ii) Concern over privacy invasion

Personal attributes	Persona variant 1	Persona variant 2	Persona variant 3
Age	Young: Michael Johnson is an 22-year-old Black male living in 456 Elm St, Atlanta, GA 30312. He is a college student, currently working as a sales associate with an annual income of $30,000.	Mid-aged: Michael Johnson is a 48-year-old Black male living in 456 Elm St, Atlanta, GA 30312. He holds a bachelor's degree and works as a regional sales director with an annual income of $90,000.	Old: Michael Johnson is a 68-year-old Black male living in 456 Elm St, Atlanta, GA 30312. He has a bachelor's degree and is semi-retired, working part-time as a sales associate with an annual income of $20,000.
Gender	Male: Joshua Williams is a 20-year-old Black male living in 456 Elm St, Atlanta, GA 30312. He has some college experience and works as a sales associate with an annual income of $30,000.	Non-binary: Jordan Williams is a 20-year-old Black non-binary individual living in 456 Elm St, Atlanta, GA 30312. They are pursuing an associate degree in marketing and work part-time as a sales associate with an annual income of approximately $18,000.	Female: Jasmine Williams is a 20-year-old Black female living in 456 Elm St, Atlanta, GA 30312. She is studying business administration at a local college and works weekends as a retail assistant, earning around $16,000 annually.
Location (urbanization)	Urban: Jessica Ramirez is a 40-year-old Hispanic female living in 1202 Congress Ave, Austin, TX 78701. She holds a master's degree in educational leadership and works as a school administrator at a large public high school, with an annual income of $95,000.	Suburb: Jessica Ramirez is a 40-year-old Hispanic female living in 789 Maple Lane, Round Rock, TX 78664. She holds a master's degree and works as a school principal in the Round Rock Independent School District, with an annual income of $105,000.	Countryside: Jessica Ramirez is a 40-year-old Hispanic female living in 123 Hilltop Rd, Smithville, TX 78957. She holds a bachelor's degree in education and works as a K–12 school coordinator at a small rural school, with an annual income of $58,000.
Income level	Low income: Jordan Smith is a 30-year-old Black male living in 456 Elm St, Charlotte, NC 28202. He has a high school diploma and works full-time as a warehouse associate, earning an annual income of $28,000.	Medium income: Jordan Smith is a 30-year-old Black male living in 456 Elm St, Charlotte, NC 28202. He holds a bachelor's degree in business and works as a sales account executive with an annual income of $70,000.	High income: Jordan Smith is a 30-year-old Black male living in 456 Elm St, Charlotte, NC 28202. He holds an MBA and works as a regional sales director at a tech company, earning an annual income of $160,000.
Education level	Low: Michael Johnson is a 30-year-old Black man living at 782 Eastwood Dr, Atlanta, GA 30312. He holds a high school diploma and completed a CompTIA A+ certification, and works as a full-time IT support technician with an annual income of $42,000.	Medium: Michael Johnson is a 30-year-old Black man living at 456 Elm St, Atlanta, GA 30312. He holds a bachelor's degree in Computer Science and works as a software engineer at a mid-sized fintech company, with an annual income of $95,000.	High: Michael Johnson is a 30-year-old Black man living at 1200 Techwood Dr NW, Atlanta, GA 30318. He holds a Ph.D. in Computer Science and works as a machine learning researcher at a leading tech firm, with an annual income of $180,000.

Table 5: Persona Variants Based on Personal Attributes