The Neighborhood: An Interactive Social Media Experience for Increasing Awareness of Automated User Profiling

Amy Yu[1], **Alexander Símon**[2], **Steven R. Wilson**[3]
[1]Columbia University
[2]Oakland University
[3]University of Michigan-Flint
ajy2127@columbia.edu, asimon2@oakland.edu, steverw@umich.edu

Abstract

Internet users leave a range of digital traces that allow for automated profiling, often without their knowledge. Advancements in artificial intelligence (AI) make it easier than ever to automate the inference of user attributes that are not even directly shared, exposing users to privacy risks. One means of protection against this is the widespread promotion of AI literacy to social media users. As an example of how this might be achieved, we present an interactive online simulation with integrated AI tools, designed to raise awareness of AI capabilities for user profiling. Within this experience, users can take actions similar to those that they would everyday: make a post, update their profile, and like existing posts. After the users finish their experience, they may choose to receive a report about an AI model's inferences about their personal attributes along with information about how similar inferences are made about them when using major social media platforms. We find that the level of interaction within the simulation correlated with large language model (LLM) accuracy in profiling. We also observe high accuracy in profiling despite many participants self-identifying as protective of their data. Finally, a majority of participants reported that as a result of taking part in our study, they expect to think more carefully about their online interactions and what they might reveal about themselves in the future. This aligns with our aim to increase AI literacy and empower people to make more informed decisions on what they disclose on the internet.

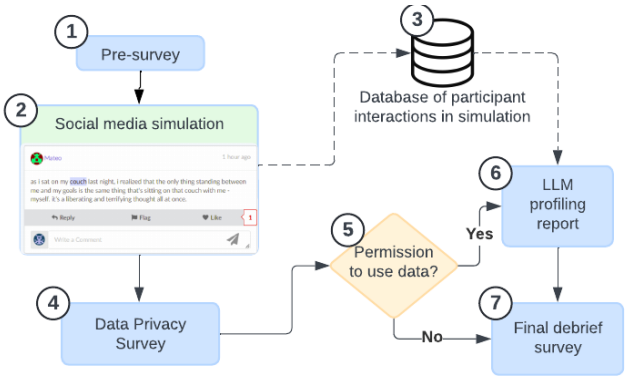


Figure 1: Overview of the participant workflow (solid lines) and data flow (dashed lines) throughout the experience. Participants (1) complete a pre-survey regarding their social media habits before (2) moving to the simulation, which produces (3) data records of their actions. Next, (4) they answer questions about data privacy before being informed of the LLM-based user profiling report and (5) having a chance to opt-out. After (6) generating and viewing their report, they (7) proceed to the final set of questions about the accuracy of the model and their impressions of the experience and potential effects on their future online behavior.

1 Introduction

As social media remains a prevalent part of our society, concerns about data access and privacy continue to rise (Pew Research Center 2023), even leading to US legislative action against Chinese-backed companies (Congressional Research Service 2024). As users become aware of privacy risks, however, they often continue to share potentially personal information about themselves online in order to stay connected with others (Kokolakis 2017; Lumare, Muradyan, and Jansberg 2024). At the same time, AI tools such as Large Language Models (LLMs) have demonstrated an unprecedented ability to deanonymize text in social media posts, even reaching a level comparable to human investigators (Staab et al. 2024). Even for users who may carefully choose which details they explicitly share, LLM-based user modeling techniques can *infer* preferences and characteristics of users from the content that they write (Tan and Jiang 2023). This poses a scalable threat to the privacy, and even more so when applied to the large-scale amounts of data social media platforms so often have access to. Such automated analysis of data is already being applied to mainstream social media, as seen in the case of targeted advertising on Facebook (Smith et al. 2024). The combination of the ease of access to LLMs and their and effectiveness in automated user profiling (Kim et al. 2024) may create dangerous situations in which average users, who are not fully aware of AI capabilities (Scantamburlo et al. 2024), reveal more sensitive information than they believe that they are through their online behavior.

Our aim in this work is to explore how we can educate social media users about the capabilities of LLMs in terms of

automated user profile inference. We propose an interactive experience as a firsthand demonstration of AI's ability to infer personal attributes from simple user interactions within a simulated environment (Figure 1). Users are initially informed that they are going to access a new social media platform in order to provide feedback on their experience. During their interaction, our platform records user activity such as posts, comments, likes, and follows. After interacting with the simulation, we introduce participants to the possibility that AI models can use this type of data to infer personal information, and with participants' permission, use an LLM to generate a sample report about inferred personal attributes based on their interaction data from our platform. We then survey all participants to understand how accurate the report is and how this might influence their perceptions of online privacy in the context of AI models. We choose the modality of our experience to be through a simulation because there is evidence to suggest that a practical experience can be more memorable than a simple verbal lesson (Wang et al. 2024). Our main **contributions** can be summarized as:

1. Designing an interactive social media simulation through which we can collect sample user behaviors;
2. Conducting a user study of more than 250 participants;
3. Analyzing resulting data to understand participants' reactions to the social media simulation and subsequent AI-powered user profiling.

We find that LLMs can accurately infer user attributes and personality traits from social media data. Furthermore, the inclusion of network interactions significantly improves inference accuracy beyond basic profile details and the context of a single post. This scalable and efficient approach demonstrates that even brief social media engagement can enable easy and effective profiling of personal attributes. Participants expressed concerns about AI models' inference capabilities, and a majority of respondents reported that due to their experience in our study, they expect to think more carefully about their online actions and what they might reveal about themselves.

2 Model Selection

Our proposed participant experience requires the use of an AI system to automatically infer user attributes from social media interactions. In this work, our goal is *not* to conduct a complete empirical analysis of the performance of user profiling systems, however, we sought to evaluate the capabilities of several popular text classification approaches in order to verify that they are potentially able to infer user attributes before conducting our user study. While we aim for reasonable accuracy, misclassifications by these models are also valuable in the context of our study, as it allows us to explore the reactions of participants to model failures.

Dataset Preparation In order to evaluate the feasibility of using an AI model to perform user attribute prediction, we used RedDust, a dataset of Reddit posts with labeled attributes (Tigunova et al. 2020). As the dataset only contained the labels along with post IDs, we used the Personal Reddit API Wrapper (PRAW) API[1] to retrieve the post texts. Since the dataset referenced many older posts, some had been deleted by the user, archived, or removed from the platform for other various reasons. We successfully retrieved the contents of 7745 posts. Of these 7745 posts, 3760 were posts labeled with age, and 3985 were posts labeled with gender. This formed our training (80% of the data) and model validation (the remaining 20%) dataset.

	Gender	Age
RoBERTa	0.72	0.66
GPT-4o	0.76	0.41

Table 1: Accuracy of a fine-tuned base RoBERTa model compared to GPT-4o on gender and age classification tasks. Verification dataset was on accuracy of 700 labeled posts from Reddust dataset for RoBERTa and 100 labeled posts from Reddust dataset for GPT-4o.

Benchmarking Model Accuracy We first considered fine-tuning a RoBERTa (Liu et al. 2019) model for text classification, leading to the validation set results presented in Table 1.

For the age classification task, rather than predicting exact numbers, we binned ages according to optimal cutoffs observed when plotting the distribution of the age data (Figure 2). This led to age bins of $<$19, 19-26, 27-45, 46-65, and 66+. Further, upon inspection of our data, we observed some noisy data points (e.g., users labeled "99 years of age" discussing Minecraft) which we manually filtered out to avoid learning less useful associations between post text and outlier age labels.

We also evaluated GPT-4o's (OpenAI 2024) effectiveness in the same age and gender classification tasks using the first 100 entries of the RedDust dataset. This was much simpler to accomplish. Using the OpenAI API, we provided GPT-4o with a Reddit post and asked it to only respond with one label from a given list. For gender, it was only able to re-

[1]https://praw.readthedocs.io

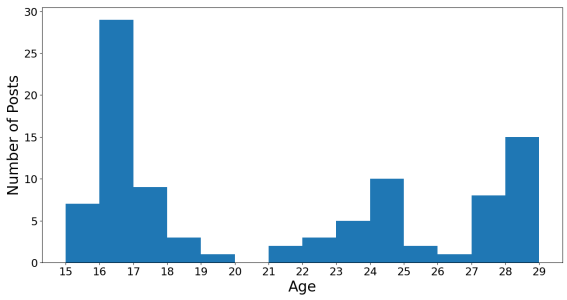


Figure 2: Histogram of the first 100 entries of labeled posts in our age classification dataset, specifically labeled posts within the range of ages 15 to 30 years old.

spond with the labels that appeared in the dataset, "male" or "female." For age, we provided the prompt with the same bins we had defined in our fine-tuning process. In gender prediction, GPT-4o marginally outperformed our fine-tuned RoBERTa by about 4 percent, and under-performed by 25 percent in the age classification task.

Despite the performance increase in age classification from fine-tuning, GPT-4o became our preferred choice in model, and we resolved to utilize an untrained LLM model approach in our profiling as this providers greater flexibility in terms of attributes that can be inferred (especially considering that existing datasets do not provide coverage of all possible attributes), and using these models also allows for the generation of explanations to present to users, which is an important component of the overall participant experience. Moreover, upon further inspection, GPT-4o's errors, even though incorrect, were often closer to the correct age bins than the incorrect predictions made by the RoBERTa model. GPT-4o's ability to perform well completely untrained also offered better scalability in terms of what attributes we were able to explore, including those without curated validation datasets. Finally, the use of untrained LLMs also aligned more thematically with our project goals of simply spreading awareness of commercial use of artificial intelligence models. Therefore, we opt to use an LLM-based approach for our AI profiling reports, which has been shown in prior work to be an effective approach for user profiling across a range of attributes (Liu et al. 2025).

3 Platform Development

In order to accomplish our original experiment design, we created two websites for our participants to interact with. These websites were created through Express.js[2] to connect to our MongoDB instance[3], and then hosted on a domain through Heroku[4]. The social media platform itself was an extended version of the open-source code-base Truman Platform introduced by DiFranzo et al. (2018) [5].

Synthetic Content Generation

We wanted to design the content in our social media simulation to discuss a broad variety of topics in order to encourage user engagement. To accomplish this, we used generative language models to create user profiles, posts, and replies to pre-populate the social media platform with content. Further, we use the same models to reactively simulate additional actions and reactions being performed by the other "users" (bots) throughout the participants' experiences to create the feeling of a live, multiuser social media platform.

Unique user profiles were generated using Llama3-70b (Dubey et al. 2024) through the Groq API[6]. Each profile had the following information: username, first name, last name, gender, age, location, and biography. We manually verified that the model produced unique results for the requested attributes and was not simply repeating the same set of user information, and in cases of repetition, we re-prompted to model to generate additional profiles.

We also use Llama3-70B to generate synthetic user posts through the Groq API(Groq 2025), which allows for fast replies during live interactions when required. To prepopulate the platform, we first generate a random number of posts between 0 to 3 for each user that was created in the previous step. During the post generation process, we reference a diverse list of prompts containing instructions about what the post should be about. For example, "write a post about a recent accomplishment" or "write a post about a project you are working on." Furthermore, we provide the corresponding user profile information as context within the prompt. All of this information forms a single, unique prompt in order to diversify the model's outputs and simulate a more dynamic social media environment. We increase the temperature of the model in this process to 1.8 to further encourage randomness and diversity within our pool of synthetic content. To diversify the lengths of the posts (which were initially all fairly lengthy), we also generated approximately 650 posts using our original prompt, and an additional batch of approximately 200 posts in which we modified the prompt to restrict the model's outputs to one to sentence. After the synthetic posts have been written, we classify them using the Tweet Single-Topic Classification Model (Antypas et al. 2022) and assign them appropriate values for the rest of the database fields such as timestamps, ID, and the identity of the user profile corresponding to the post.

Lastly, we used Llama3-70B through the Groq API to generate *replies* to our synthetic posts with a similar approach. In order to simulate networks within the fake user profiles, we group them by the topics they post about. These simulated groups with shared "interests" form the pool of users we randomly draw from to respond to each post given the topic of the post. For each post, we generate either 0, 1 or 2 replies, using a weighted probability distribution such that posts were more likely to have empty comment sections. These synthetic data objects were then assigned an appearance time within an hour after the original post's assigned timestamp. Through this process, in response to the 800+ posts generated through Llama3, we generate 422 pre-written replies which are set to automatically appear throughout the first hour of a given user's interaction with the platform. For the purpose of our study, we do not expect any users to spend more than, or even close to, one hour interacting with the platform.

"The Neighborhood"

There were several steps required to create our social media simulation for our participants, which we named "The Neighborhood." To streamline development, we built our the simulation from the Truman Platform codebase (DiFranzo et al. 2018), which covered much of the essential functionality, such as user authentication, posting systems, and commenting systems. We maintained the overall architecture of

[2]https://expressjs.com

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

[4]https://www.heroku.com

[5]The Truman platform is licensed under the terms of the MIT license available at https://github.com/cornellsml/truman/blob/master/LICENSE.

[6]https://console.groq.com/docs/libraries

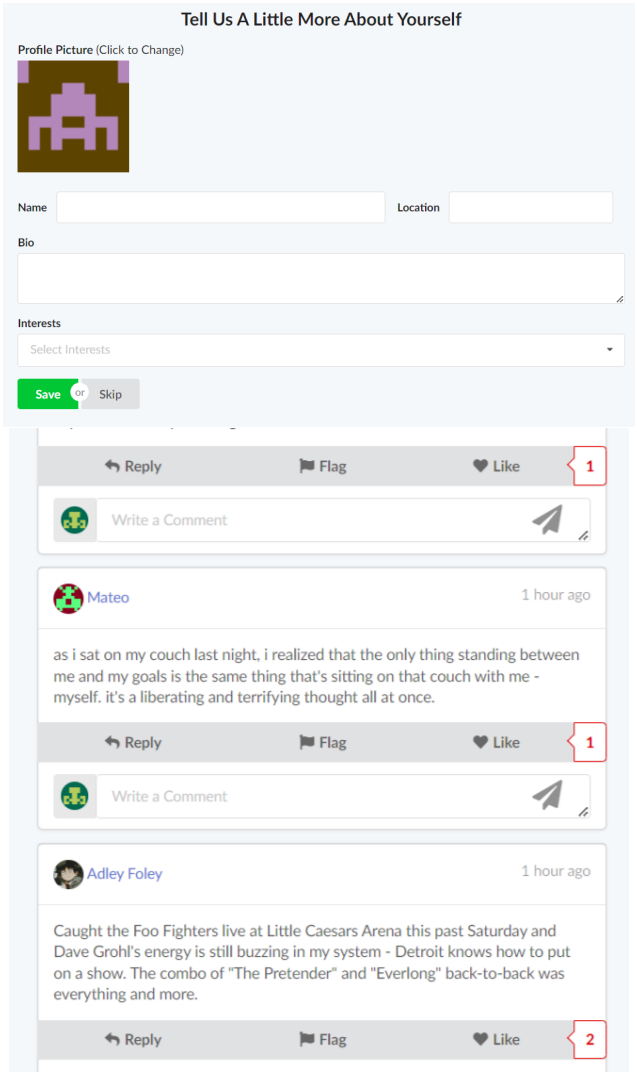


Figure 3: Sample screenshots of our social media simulation. Top: account creation page. Bottom: example from the main feed.

the codebase, maintaining its Express.js framework using PUG template syntax in order to maintain uniformity in the code. However, we also needed to make a handful of key modifications to the original Truman Platform to better serve our purposes, which are outlined below.

Database Schema Updates Using our own MongoDB instance, we were able to inject our synthetic content into the database, which was served to the front-end to display the content that we created in an interface that the user was able to interact with.

To achieve this, we needed to customize the pre-defined database schemas. This included, among others, removing the email field (per IRB stipulations).

Feed Curation One of our first objectives was the ability to curate the user's feed according to which posts they are interested in viewing in order to encourage user engagement. The topics we chose were mapped from the Twitter Classification Model's original labels into a broader set, and as a result, users could choose posts related to topics they were interested in from the following list: arts, business, pop culture, lifestyle, fashion, entertainment, health, food, gaming, education, music, news and politics, science, sports, and travel. To accomplish this , we extended the database schema to include a "topics" variable, which resulted in the addition of this field to the account creation page and the profile editing page. Each post was assigned a single "topic." Using this additional attribute, we reworked the feed algorithm such that each user could choose to filter the feed by the topics they had selected.

"Less Instagram, More Twitter" One of the limitations of our untrained LLM-based approach was that it was only able to process text. Therefore, in the event a user posts a highly specific photo with a very vague caption, if we were to try sending the contents of the post as an input, the LLM would only have a vague caption to use for context. To avoid abstracting too much context by providing pure text without photo context, we removed the image feature of the posts in the platform and made all of our posts text-based.. We also made some modifications to the UI to move away from the Instagram-like theme that the original Truman Platform codebase was following, which contained a majority posts focused on just a few topics (e.g., photos of food).

Dynamic Response System In the original Truman platform codebase, when a user replies, they receive an automated notification of a fake user responding with a hard-coded value. To make the environment simulate more realistic social media conditions, we replaced this system of hard-coded reply values to be more adaptive and receptive to user activity. To accomplish this, we handle three cases of user interaction: new user-made posts, user comments on their own posts, and user comments on scripted posts. With each of these three user interaction scenarios, we begin with finding an appropriate fake user to respond. For instance, if a user comments on a scripted post, we will check if they are responding to the original post or if they are responding to a comment on the post. Then, we will fetch the details of the user that was addressed and put the relevant context information (profile information, post details, etc.) into a prompt for Llama3-70B through the Groq API to generate a response. The generated response is then populated into the database as a notification for the user and added to the appropriate post. In order for these responses to remain unique to the user, we separate dynamically-generated responses from scripted comments on generated posts by linking them to a specific user ID. This separation by user ID prevents users from seeing fake user comments that were not generated by their own actions, maintaining a uniform study environment.

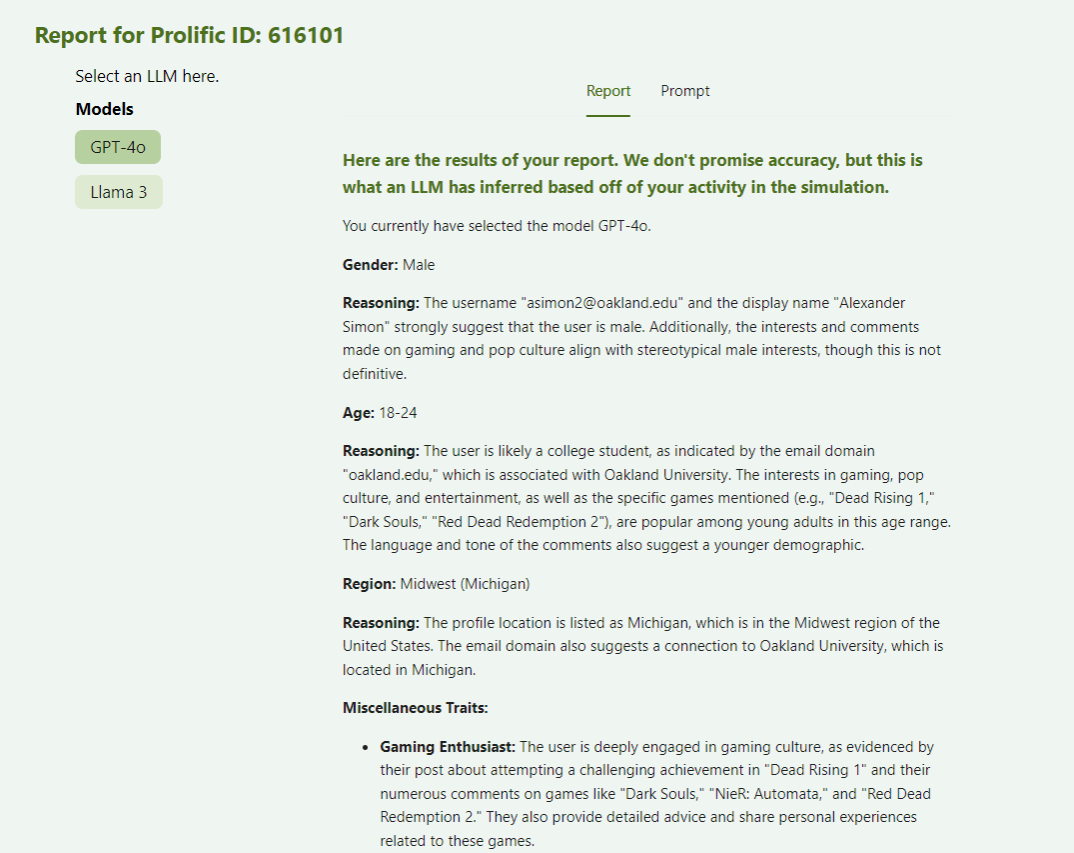


Figure 4: Example screenshot of an AI-generated profiling report.

AI Profiling Website

Apart from the Neighborhood simulation, we develop a second web platform to manage the AI profiling itself. We opted not to centralize the AI profiling and social media simulation into one website as it was possible that some users may not consent to this use of their data afterthe user-profiling element of the study was revealed to them. Additionally, in the experiment's narrative design, the social media simulation and the AI-generated profiling report are divided by the reveal of the study's purpose. For these reasons, we determined that it would be better to separate the platforms altogether.

In cases where users give permission to generate a report, their interactions with the platform are first retrieved from the database based on their Prolific ID. Then, these interactions are formatted using our prompt template (details in appendix A) and used as input to an LLM.

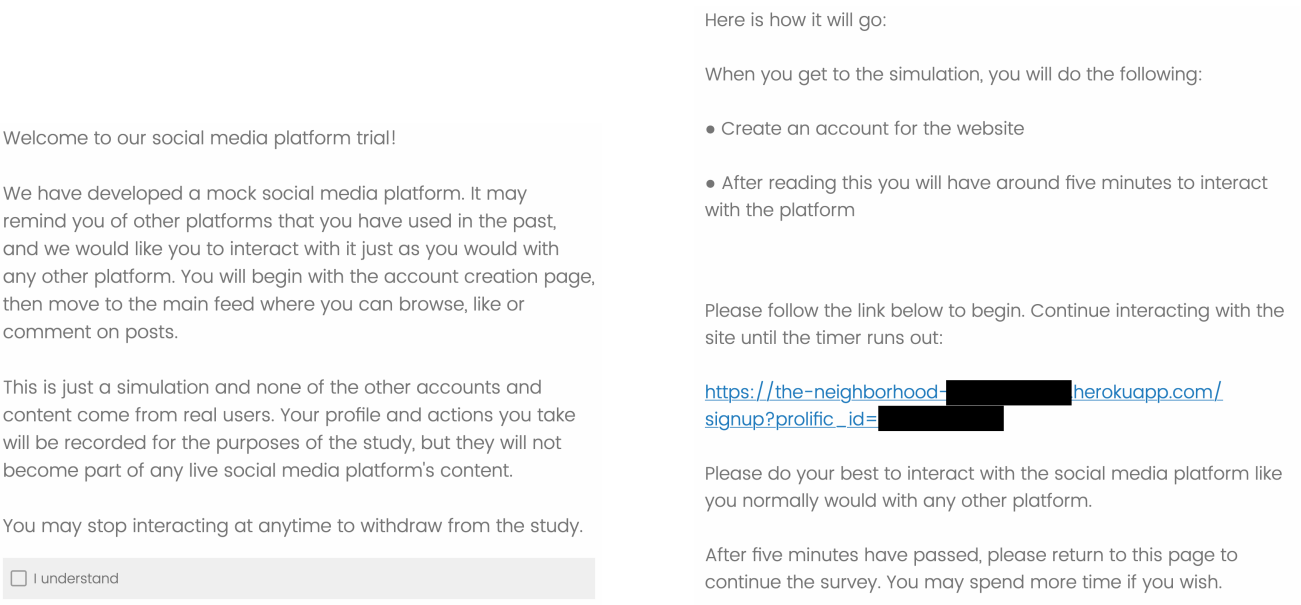
We use a temperature of 0.25 to balance between predictions that were too conservative and maintaining good coverage of the details of the users' interactions. The output report is then saved to present to the user.

4 Participant Experience

Survey Setup We used Prolific to recruit and compensate our participants, and Qualtrics to drive the main progression through the social media experience and associated survey components. Participants, along with their Prolific IDs, are routed to Qualtrics automatically, and these IDs are used to track the user throughout the experience. At the end of the experience, participants are routed back to Prolific and a completion code is automatically submitted via a URL parameter. Before payment was processed, we manually checked each completion to ensure that the participant had actually interacted with the platform and answered the required survey questions. The participants were paid $5.00 each and took, on average 18 minutes, resulting in an average hourly rate of $16.71 and a total participant budget of approximately $1,747 when accounting for Prolific fees.

Participant Workflow Below, we outline the sequence of steps that participants were directed through using the Qualtrics survey (specific language used in the survey questions can be found in Appendix B).

1. **Informed Consent and Pre-Simulation Survey** Before interacting with our platform, users were presented with an informed consent page. It is worth noting that this study included an element of deception: at this stage, the users are *not* informed of the complete purpose of the study – specifically that their interactions would potentially be used as input to an AI model which would make



(a) (b)

Figure 5: (a) Brief informational page shown to participants before being directed to begin the simulation and (b) instructions provided to participants along with link to "The Neighborhood" account creation page.

predictions about their personal attributes – to avoid biased behavior and compromising the integrity of the data. Instead, users were only told that they would be interacting with an experimental social media platform. However, we request *additional consent* before using any user data in a way that was not initially agreed-upon. Users who agreed to the initial informed consent page were subsequently asked questions related to their social media habits. Participants were first asked how frequently they used social media platforms Instagram, TikTok, X, Reddit, Snapchat, Youtube, and Facebook, with options that varied between "never" and "multiple times per day." They were also asked for reasons behind their use of social media. Provided options for this question included entertainment, sharing content, staying informed, and maintaining connections with friends, alongside an optional "other" text field where participants could share additional reasons.

2. **Simulation** After being presented with a set of instructions (Figure 5), users were directed to an isolated instance (i.e., they would not interact with other participants in any way) of the live version of the social media platform. Users would first create a basic profile, then be directed to the main feed, where they could perform ordinary social media actions such as posting, liking, and commenting. Users were required to spend five minutes using the platform before continuing with the study, but were allowed to spend longer if they wished.
3. **Pre-Debrief** After the five (or more) minutes of interaction, participants were asked about their basic familiarity with artificial intelligence and how protective they are about their personal information when using the internet.
4. **AI User Profile Report** At this stage, users were presented with a page explaining that AI models can be used to automatically infer user traits based on social media interactions, and in fact, such a model was available that could be run on the data generated from their interactions on our platform. Participants were then asked if they were interested in producing a report of what an AI model would have predicted about them based on their interactions, noting that their data would need to be processed by an AI model in order to achieve this, and that the AI predictions would not be shared without their permission. Users were also informed that refusing to generate the report would *not* disqualify them from the study. Users who agreed then received a report detailing their inferred traits. Users who agreed to report generation were given further questions about their sentiments and opinions on accuracy of the AI models' predictions before redirecting them to the post-simulation debrief survey. The questions regarding model accuracy asked the user for a personal evaluation of the LLM's accuracy regarding the attributes of age, gender, region, and personality respectively. Users who refused were immediately redirected to the post-simulation debrief survey.
5. **Debrief Survey** In this portion of the study, participants were surveyed about their comfort levels regarding AI automation of profiling using data harvested from their social media activity. They were also asked whether their understanding and perception of artificial intelligence had changed over the course of the study. Furthermore, to assess potential biases, participants were questioned about the authenticity of their behavior on the mock social media platform. Finally, to evaluate the study's im-

Age Group	%
18–24	8.78
25–34	36.64
35–44	21.76
45–54	16.03
55–64	13.74
65–74	3.05

Ethnicity	%
White	66.41
Black	14.50
Asian	6.87
Mixed	7.25
Other	4.96

Sex	%
Female	46.95
Male	51.53
Unreported	1.52

Familiarity with AI	%
Not familiar	1.15
Somewhat unfamiliar	2.67
Neither familiar nor unfamiliar	6.11
Slightly familiar	46.95
Somewhat familiar	26.72
Very familiar	16.41

Table 2: Participant Demographics (N = 262)

pact, participants were asked whether they anticipated changing their online behavior as a result of their experience. The survey then concluded with questions about consent for using anonymized data in future research and an option to share additional feedback.

5 Results and Discussion

Through Prolific, we recruited 101 total participants who consented to take part in the study between July 15-17th, 2024. Out of these 101 participants, 17 did not consent to have their recorded data be used to generate a report and 8 experienced technical issues. A second run of the study with 169 participants was conducted between May 9-11th, 2025. Of the 169 participants, 142 consented to the report. This gives us 218 AI-generated reports from 262 participants in total (Participant demographic information in Table 2).

How Accurate Were the AI-generated Reports?

We can see that the gender predictions were more accurate (Figure 6), though this is to be expected given that there were fewer labels. In terms of age, participant feedback indicates that GPT-4o tends to be biased towards guessing younger ages, which we theorize to be an association of most normal social media behaviors with younger demographics. We

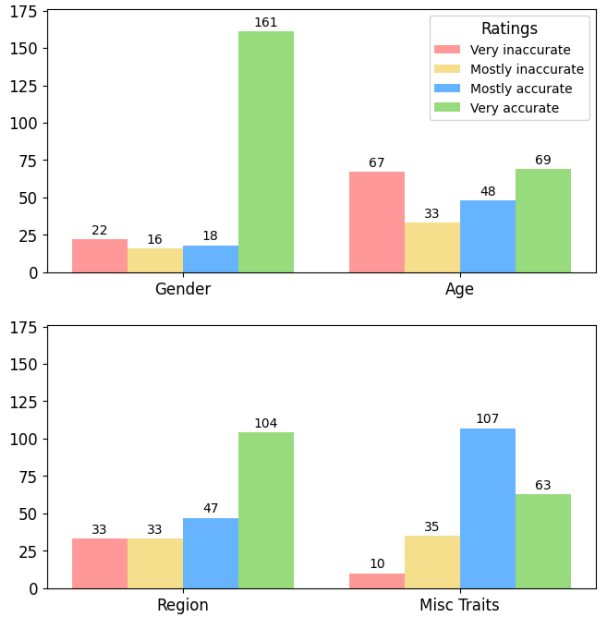


Figure 6: User Ratings of GPT-4 Profiling Predictions by Attribute

	Gender	Age
Accuracy	0.78	0.43

Table 3: Accuracy of model ratings for gender and age based off of Prolific demographic information.

also notice that the reported accuracy for region is higher than age despite region proving to be a much more difficult attribute to infer in other studies. However, upon further investigation, we notice that some participants explicitly label their region in the location field of their profile. We hope to explore the accuracy of such deanonymized profiles versus profiles that are anonymized by not stating their location in future data analysis.

In addition to evaluating the user ratings of the predictions themselves, we looked to verify the accuracy of the model according to the demographic data of each participant that was associated with their Prolific account (Table 3). Looking at these accuracy scores, we notice a marked improvement over our preliminary accuracy results from the model selection stage of our methodology, most likely due to the more detailed and personalized information that the LLM receives from social media interactions on our platform compared to those on Reddit. This suggests that supplementing user-created content with other types of interactions, such as liking posts or following users, augments the user attribute inference capabilities of LLMs.

We also sought to understand how the amount of data produced by a given user impacted the ability of an LLM to accurately infer information about them. In order to accomplish this, we first looked at the individual reported accuracy of each attribute relative to the number of interactions the user made with the platform (Figure 10). We define an "in-

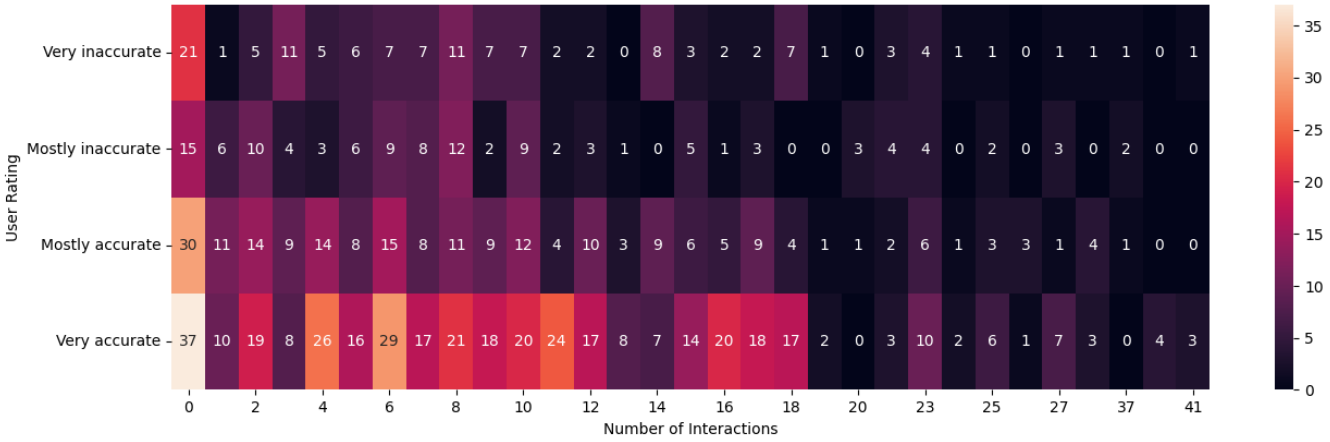


Figure 7: Heatmap of the number of user interactions with the platform versus user accuracy ratings of LLM predictions. User ratings have been aggregated across all attributes.

teraction" with the platform to be any additional action beyond profile creation that gets used as additional context for our LLM's prompt. The actions that fall within this definition are user-created posts, user-created comments, following other users, as well as flagging or liking content. Within the individual charts, we observed a general trend where user ratings of model performance improved with more interactions. We then aggregated all of the ratings across attributes into a heatmap to look at the data more broadly (Figure 7). In this heatmap, we noticed that the rating distribution for users who made zero interactions beyond account creation looked more akin to what one would expect from random guessing. However, for higher levels of interaction, we observe fewer "completely wrong" ratings and many more "spot on" ratings, highlighting the small positive correlation between the number of interactions and the accuracy of the LLM's inferences (Spearman's $\rho = 0.093$, p-value$= 0.0064$).

We also noticed that the trend does not hold for participants with very high numbers of interactions, i.e., greater than 25. We believe this is due to a trade-off between quality and quantity of data. Within our participant experience, users are exposed to the platform for a minimum of five minutes and typically do not take more than 15 minutes to complete the experience. We observe that users with exceedingly high numbers of interaction with the platform usually exhibit "spammer behavior" where many of their interactions come in bursts in the form of liking many posts (potentially indiscriminately), rather than more direct interactions by posting or commenting. As a result, the sheer amount of posts that such users "like" will oversaturate the prompt being fed to the LLM, leading to more inaccurate predictions.

How did Participants React to Their Experiences?

Here we explore the impact that the experience had on our participants and their general sentiment regarding data privacy and artificial intelligence. Based on their answers to the Pre-Debrief survey component (Figure 8), we observe that a majority of the participants considered themselves "very protective" of their data, while only 2 percent of them considered themselves not protective of their data at all. To some extent, an overwhelming amount of our participants considered themselves protective of their data, while also producing interaction data on our platform that led to highly accurate predictions of their traits.

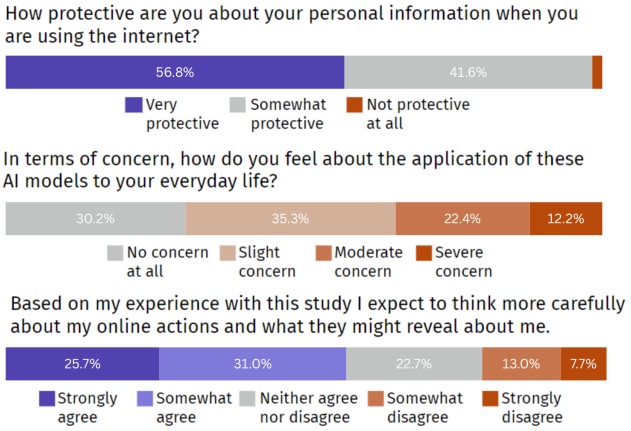


Figure 8: Participant responses to questions related to data privacy.

Additionally, we asked participants about their sentiments regarding this application of LLMs to their daily lives, to which the sentiments were a bit more mixed. We observe a positive, yet non-statistically-significant relationship between level of concern regarding AI and data privacy among older demographics 4.

We also want to examine the objective impact of our methodology on the perspective of the participants, since more than not people (56.7%) answered that they expect to think more carefully about their online actions after the experience. Future work could explore this more quantitatively through a follow-up study, or through additional questions regarding what participants had learned through the process

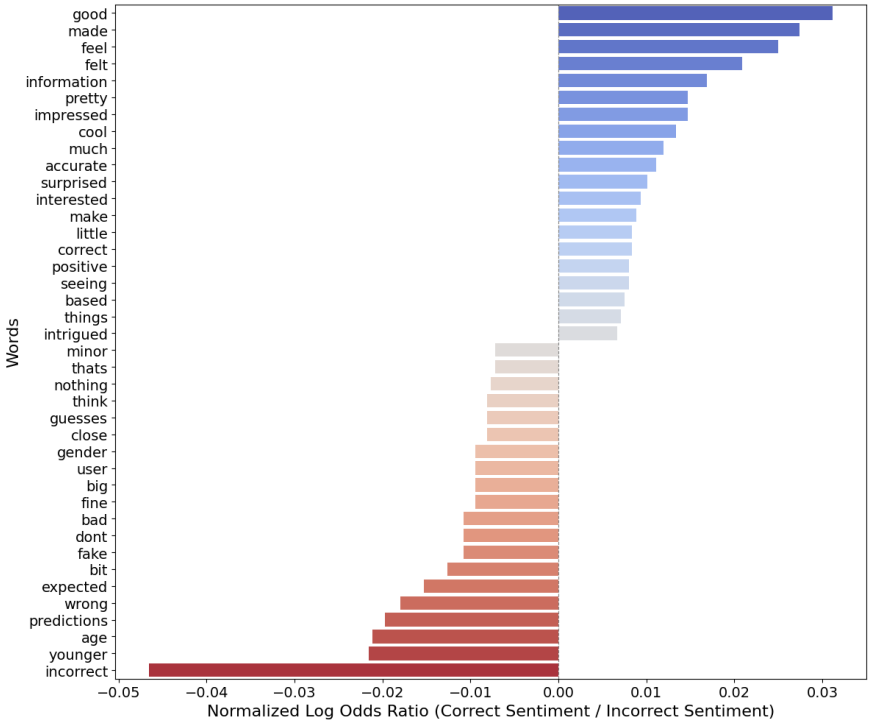


Figure 9: Normalized (based on number of total words per category) log odds ratio showing the words that are most distinctive for each of the two types of texts – those describing reactions to correct and incorrect predictions made by the models, respectively.

Concern Level	Average Age
No concern	34.54
Slight concern	36.02
Moderate concern	37.06
Severe concern	47.00
No answer	46.50

Table 4: Distribution of mean participant age and reported concern levels in the post-debrief survey.

to further analyze the impacts of the experience.

In order to characterize the types of reactions that participants had to the correct and incorrect predictions of the model, we compute the log odds ratio (Monroe, Colaresi, and Quinn 2008) between the non-stopwords[7] in the free text fields completed for those who opted in to the LLM-based profiling. This allows us to present the words that were most likely to appear in one set of texts, but not the other. Figure 9 shows the results. For correct classifications, other than commenting on these classifications themselves, users mentioned that they were "surprised", "impressed", and "intrigued" by the accuracy of the models. Some expressed a desire to understand more about how the models worked, and others were simply "freaked out" by the predictions. Some participants mentioned that they were glad that the model was able to accurately profile them. In terms of incorrect predictions, some participants reported that they "expected" this result (some even reporting that this was because they had intentionally provided "fake" information about themselves in their profile for the social media platform). Several participants noted that they were happy that the model predicted that they were "younger" than they actually were. Finally, others mentioned that even though the models were incorrect, they were at least "close".

[7]using the English stopword list provided by NLTK (Bird and Loper 2004)

6 Conclusion and Future Directions

Conclusions We presented an approach for improving user awareness off AI-powered social media profiling capabilities though an interactive, simulated experience. We measured participants' perceptions of LLM-generated user profiling reports and their overall experiences on the platform. From our results, we observed that AI models are capable of inferring user attributes through simple user interactions, often in a way that is surprisingly accurate from the perspective of everyday users. We also note that the increased context of adding network interactions and profile fields on top of "active user" activity of writing posts and comments has increased the objective accuracy of LLM inference capabilities, and that this network interaction-based approaching to LLM profiling has greatly augmented it. Just by having participants engage in everyday social media behavior for a short period of five minutes, an LLM was able to profile their personal attributes fairly accurately and even

gauge more subjective descriptors such as their personality traits. Additionally, the process of automating the data retrieval and prompting to get a profiling report was relatively simple, proving that such a method to profiling is scalable and efficient. Users noted their concern about the ability of AI models to make these types of inferences so successfully, and in many cases claimed that they expected to think more carefully about how they interacted with others on social media in the future.

Limitations and Future Work Opportunities Our current study is limited to initial explorations of this approach, early findings, and a small sample size of participants. Further quantitative investigation into nuanced connections between our variables, such as the correlation between reported AI familiarity and data privacy concern, may yield new insights. It would also be worthwhile to study the specific "clues" the the LLM models used to make inferences, and measure how much inference is truly needed compared to simple associations or even direct claims that users shared about themselves. Conducting a follow-up study with some of the past participants would shed light on whether the experience truly changed their behavior, as opposed to only expecting to do so. Finally, exploring modalities beyond text, and the capabilities of Vision-Language Models (Bordes et al. 2024) to make similar inferences based on image, video, or audio data would enhance the realism of our dataset and align the user experience more with many modern social media platforms that rely on these modalities.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful feedback, which significantly improved this manuscript. We also thank the participants from the Prolific platform who made this study possible. This material is based upon work supported by the National Science Foundation under Grant No. 2349663. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Antypas, D.; Ushio, A.; Camacho-Collados, J.; Silva, V.; Neves, L.; and Barbieri, F. 2022. Twitter Topic Classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, 3386–3400.

Bird, S.; and Loper, E. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 214–217. Barcelona, Spain: Association for Computational Linguistics.

Bordes, F.; Pang, R. Y.; Ajay, A.; Li, A. C.; Bardes, A.; Petryk, S.; Mañas, O.; Lin, Z.; Mahmoud, A.; Jayaraman, B.; et al. 2024. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*.

Congressional Research Service. 2024. TikTok: Recent Data Privacy and National Security Concerns. https://crsreports.congress.gov/product/pdf/IN/IN12131. Report No. IN12131.

DiFranzo, D.; Taylor, S. H.; Kazerooni, F.; Wherry, O. D.; and Bazarova, N. N. 2018. Upstanding by Design: Bystander Intervention in Cyberbullying. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 1–12. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356206.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Groq. 2025. GROQ API Documentation.

Kim, H.; Song, M.; Na, S. H.; Shin, S.; and Lee, K. 2024. When LLMs Go Online: The Emerging Threat of Web-Enabled LLMs. *arXiv preprint arXiv:2410.14569*.

Kokolakis, S. 2017. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & security*, 64: 122–134.

Liu, Y.; Jia, Y.; Jia, J.; and Gong, N. Z. 2025. Evaluating {LLM-based} Personal Information Extraction and Countermeasures. In *34th USENIX Security Symposium (USENIX Security 25)*, 1669–1688.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Lumare, N.; Muradyan, L.; and Jansberg, C. 2024. Behind the screen: the relationship between privacy concerns and social media usage. *Journal of Marketing Communications*, 1–16.

Monroe, B. L.; Colaresi, M. P.; and Quinn, K. M. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4): 372–403.

OpenAI. 2024. GPT-4o System Card. https://cdn.openai.com/gpt-4o-system-card.pdf.

Pew Research Center. 2023. How Americans View Data Privacy. Accessed: 2025-01-15.

Scantamburlo, T.; Cortés, A.; Foffano, F.; Barrué, C.; Distefano, V.; Pham, L.; and Fabris, A. 2024. Artificial intelligence across europe: A study on awareness, attitude and trust. *IEEE Transactions on Artificial Intelligence*.

Smith, G.; Carson, S.; Vengurlekar, R. G.; Morales, S.; Tsai, Y.-C.; George, R.; Bedwell, J.; Jones, T.; Mondal, M.; Smith, B.; Su, N. M.; Knijnenburg, B.; and Page, X. 2024. "I Know I'm Being Observed:" Video Interventions to Educate Users about Targeted Advertising on Facebook. In *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–27. ACM.

Staab, R.; Vero, M.; Balunovic, M.; and Vechev, M. 2024. Beyond Memorization: Violating Privacy via Inference with Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Tan, Z.; and Jiang, M. 2023. User modeling in the era of large language models: Current research and future directions. *arXiv preprint arXiv:2312.11518*.

Tigunova, A.; Mirza, P.; Yates, A.; and Weikum, G. 2020. RedDust: a Large Reusable Dataset of Reddit User Traits. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6118–6126. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Wang, G.; Zhao, J.; Johnston, S.-K.; Zhang, Z.; Kleek, M. V.; and Shadbolt, N. 2024. CHAITok: A Proof-of-Concept System Supporting Children's Sense of Data Autonomy on Social Media. In *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–19. ACM.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, we informed participants about the LLM-based profiling step and gave them the opportunity to opt-out. Further, the purpose of the study is to evaluate how similar experiences can help reduce unfair profiling and help users maintain their privacy.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, see section 1.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see section 1
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? NA
   (e) Did you describe the limitations of your work? Yes, see "Future Changes to Study" section in conclusions and future directions.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Ethical Statement.
   (g) Did you discuss any potential misuse of your work? Yes, see Ethical Statement.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. We release full details of survey questions and have kept our dataset unreleased. Participants are given a consent form prior to the start of the experiment outlining risks and are asked for consent before any data processing.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No, code can be linked through Github upon acceptance.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA, trained models were not central to the experiment's purpose so much as mainstream LLMs. Datasets and model specified.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA, not central to the experiment's purpose.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA, not central to the experiment's purpose.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA, not central to the experiment's purpose.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA, not central to the experiment's purpose.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, codebases and software used in the methodology have been cited.
   (b) Did you mention the license of the assets? Yes, see footnote 5 referencing the codebase we use for our simulation.

(c) Did you include any new assets in the supplemental material or as a URL? No, to maintain anonymity. URL to Github repository of new assets will be provided upon acceptance.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see "Participant Experience" section.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No, none of the personally identifiable information or offensive content has been published or discussed.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA, dataset is unreleased.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA, dataset is unreleased.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, see Figure 5 and Appendix B.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Risks are enumerated in the consent waiver prior to the Prolific survey.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, see "Participant Experience" section.

(d) Did you discuss how data is stored, shared, and deidentified? Yes. Data is securely stored through Prolific and MongoDB without being shared publicly.

Ethical Statement

The methodology presented in this paper was reviewed and approved by the Institutional Review Board of the Organization where the study was conducted. The study was determined to have no more than minimal risk to the participants. Although there is an element of deception, we made sure that participants still had an option to opt-out of the LLM-generated user profiling step without being excluded from the study or from potential payment for their participation. It is possible, though unlikely, that participants' awareness of LLM capabilities will inspire them to attempt to profile *other* users against their consent. As people become more aware of risks, they may also choose to avoid using social media platforms and fail to experience potential, positive social benefits of connecting with others online. In the end, however, we believe the potential positive impact of educating everyday social media users about the privacy risks associated with online interactions in the era of LLMs makes our approach worthwhile.

Appendices

A Prompt Template

The prompt we used for the LLM-based user profiling system has two components: the system prompt, which instructs the model about the task in general, and the user prompt, which provides details about a specific participant's interactions when using the mock social media platform. The system prompt is shown below.

```
You are an expert investigator
familiar with social media behaviors.
Your job is to provide as many
details of the user as you can given
details of their behavior in a day of
scrolling. In particular, provide a
guess for their gender, age, and a
general idea of what region of the US
they may be from. Additionally,
provide a paragraph of any other
miscellaneous traits you may think of
the user given their online behavior.
Each of these attributes should be
separated by a new-line and each
section labeled (ex., '**Age:** ').
Please explain your reasoning in
order to arrive at the most logical
and holistic conclusion.
```

The user prompt, which was automatically generated and filled in uniquely for each user, followed the structure shown below.

```
The user you are profiling is someone
with the username <username>.
Their display name on their profile
is <display-name>.
Their location on their profile is
<location>.
Their bio on their profile is <bio>.
They have listed on their profile
that they are interested in seeing
posts about <interests>.

Over their time on the social media
simulation, they have
<interaction-type> the following
users: <list-of-users>.

During their time spent on the social
media simulation, they have made
<number-of-posts> post(s) on the
platform.

User-written Post <post-number>:
"<post-body>"
...

On this post, they made the following
comments:
- <comment-1>
```

```
...
- <comment-N>>

Here are the actions they took on the
posts of other users:
- <action-1>
...
- <action-M>
```

Where items surrounded by angle brackets are filled in with the appropriate values according to the participant's interactions with the platform.

B Survey Question Wording

Specific language used in our survey questions and answer choices are presented below.

Participant Intro Survey

(Before interaction with platform)

Before we begin, please tell us about your social media usage.

1. How frequently, on average, do you use the following social media platforms?
   Choices:
   - Never
   - Used in the past but no longer use
   - Once a month
   - Once a week
   - A few times per week
   - Once per day
   - Multiple times per day
   - Prefer not to say

   Platforms:
   - Instagram
   - TikTok
   - X (formerly Twitter)
   - Reddit
   - Snapchat
   - YouTube
   - Facebook
2. What are the main reasons you use social media? (Check all that apply)
   - To learn or stay informed
   - To connect with others
   - To be entertained
   - To share content
   - None of these
   - Prefer not to answer
   - Other: _____

Pre-Debrief Survey

(After interaction with platform, but before AI-generated report)

1. Which of the following applies to your experience with AI?
   - I read news about it
   - I worked with AI-related technologies
   - I have used tools related to AI or have AI features
   - I consider myself an expert
   - Never heard of it
2. How protective are you about your personal information online?
   - Not protective at all
   - Somewhat protective
   - Very protective
   - Prefer not to say
   - Other: _____

Permission for Data Processing

(After interaction with platform, but before AI-generated report)

Beyond the information that you directly share, one way that social media platforms, marketing companies, governments, and other organizations may learn about you is through the use of AI. To illustrate this, we have developed a sample system which can be used to try to guess things about you based on your interactions with our social media platform.

1. Do we have your permission to process your data (interactions with our platform) in order to reveal what the AI thinks about you?
   - Yes
   - No

Debrief Survey

(After AI-generated report)

1. Feedback on predictions (only for users who consented to data processing) :
   - Age: Very accurate, Mostly accurate, Mostly inaccurate, Very inaccurate
   - Region: Very accurate, Mostly accurate, Mostly inaccurate, Very inaccurate
   - Gender: Very accurate, Mostly accurate, Mostly inaccurate, Very inaccurate
   - Miscellaneous traits: Very accurate, Mostly accurate, Mostly inaccurate, Very inaccurate
2. In general, how concerned are you about the fact that AI models are making these kinds of predictions about you in your daily life?
   - Very concerned
   - Moderately concerned
   - Slightly concerned

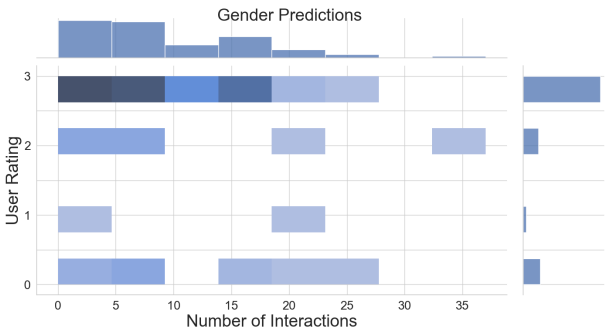
- Not concerned at all
- Prefer not to say

3. Agreement with the following statements:
   - ”Based on my experience with this study I expect to think more carefully about my online actions and what they might reveal about me.”
   - ”My interactions with the mock social media platform reflect real-life usage of social media platforms.”

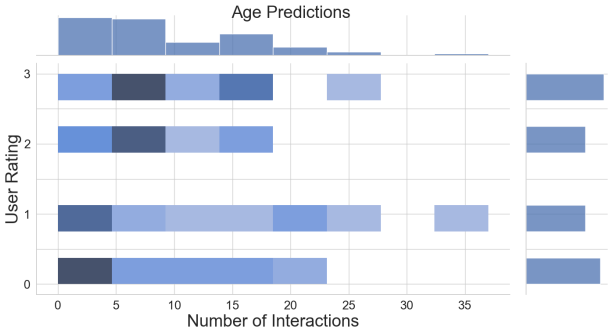
   Scale: Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree

C Attribute-level Relationships Between Interactions and Accuracy

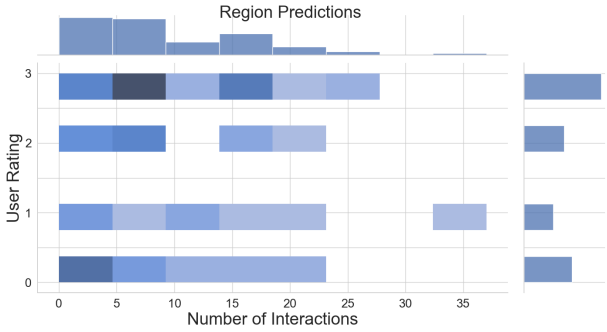
In the joint plots presented in Figure 10, we show how the number of interactions related to the accuracy of the models, disaggregated by various demographic variables. These plots also provide the overall distribution of these ratings.



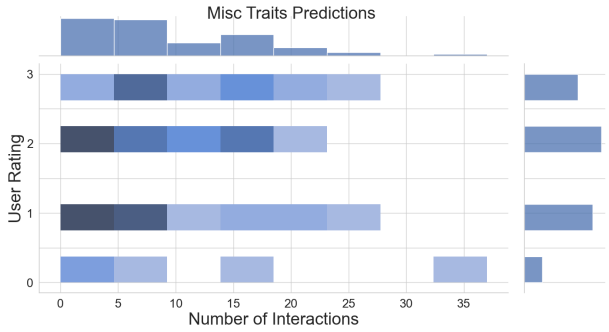
(a) Gender joint plot



(b) Age joint plot



(c) Region joint plot



(d) Miscellaneous traits joint plot

Figure 10: Joint plots of the four different attributes we examined for LLM profiling. The plots along the x and y axis represent a histogram for the ratings and the interactions respectively. The numeric ratings along the y-axis correspond to the following labels: “completely wrong”, “somewhat accurate”, “mostly accurate”, and “spot on” in ascending order. Bins with deeper shades of blue indicate bins with more ratings, and bins with lighter shades of blue indicate fewer ratings.