User Issues and Concerns in Generative AI: A Mixed-Methods Analysis of App Reviews

Vanessa Bracamonte[1], Sascha Loebner[2], Frederic Tronnier[2], Ann-Kristin Lieberknecht[2], and Sebastian Pape[2]

[1] KDDI Research, Inc., Saitama, Japan
va-bracamonte@kddi-research.jp
[2] Chair of Mobile Business & Multilateral Security, Goethe University Frankfurt, Frankfurt, Germany
{sascha.loebner,frederic.tronnier,ann-kristin.lieberknecht, sebastian.pape}@m-chair.de

Abstract. Generative AI models such as ChatGPT and Stable Diffusion have become easily available to end users through various apps. Research has identified several safety risks and limitations of generative AI, but the experiences and issues faced by real users of this technology in the wild have not been systematically investigated. In this paper, we identify user issues related to trustworthiness dimensions of generative AI, by analyzing user reviews of AI apps using a hybrid approach that combines unsupervised topic modeling and manual qualitative analysis. The results revealed user issues related to the validity, reliability, safety, security and privacy of the AI. Validity-related issues, such as incorrect output, were often found, but these issues appeared to result from high expectations about the capabilities of the technology, rather than an accurate reflection of its limitations. Concerns about safety issues, such as bias and the handling of inappropriate content, also appeared frequently, although users had conflicting expectations on how these should be handled. On the other hand, the user reviews contained fewer instances of concern related to the security and privacy of the AI itself. Overall, the results suggest that real users of generative AI have inadequate information about the characteristics and limitations of these models.

Keywords: Generative AI · Mobile apps · User reviews · Trustworthiness dimensions · Topic modeling · Qualitative analysis

1 Introduction

Recent years have seen the release of AI models for generating data such as images and text, which have gained popularity and are being used in an increasing number of applications. However, these generative AI models have various problems, including issues with the validity and reliability of their results [17], the safety of their output [29,31], and security and privacy vulnerabilities [6,7,35]. The issues and limitations of generative AI models pose risks to various stakeholders [2], prompting efforts to conduct thorough evaluations [8]. These problems have already affected users in real life and have been reported in the media [9].Generative AI models, with their existing limitations, are currently easily accessible to end users, who interact with these models through apps

such as those found in mobile app stores. Although studies have been conducted on the challenges users face when using generative AI [19,40,38], the issues of users in the wild have not been systematically analyzed.

In this study, we investigate the opinions, complaints and concerns found in user reviews of apps that use generative AI models. Our objective is to understand real users' issues related to the trustworthiness of these models. App reviews can be a valuable source of information about the challenges that real users face when interacting with different types of technology [11,13]. We gathered user reviews from the Google Play store for apps focused on text generation (specifically chatbot apps including the official ChatGPT app) and image generation. To analyze these reviews, we used a hybrid approach that combined Topic Modeling and qualitative analysis. This type of approach has been used to conduct qualitative analysis in cases where there is a large amount of data [12]. Our findings indicate that users' issues and concerns about generative AI appear to be to some extent the result of lack of knowledge about the technology. We identified that users' issues were mainly related to the AI trustworthiness dimensions of safety and validity. On the other hand, there were few reported issues related to the security and privacy of generative AI in the user reviews. The contribution of this work is an understanding of the issues that real users' encounter when using generative AI models, framed within the dimensions of AI trustworthiness.

2 Related Work

2.1 Issues in Generative AI

Generative AI models, similar to other types of AI, have issues and limitations which affect their trustworthiness in different dimensions [25]. With regards to the validity and reliability, generative AI models are prone to output nonsensical information or information not in the training data ("hallucinations") [17]. Bang et al. [1] conducted an evaluation of ChatGPT on NLP tasks and found that the model's performance on reasoning tasks were not reliable. Generative AI models have also been considered to pose risks to users via the incorrect handling of their output. Pearce et al. [29] found that Github's Copilot model can introduce high-risk security vulnerabilities in the code it generates. Siddiq et al. [34] similarly reported that the model generated code which contained a number of issues, including related to security. On the other hand, Sandoval et al. [33] found that for a particular use case, the code submitted by users' which had been created using AI assistance did not contain higher than 10% security vulnerabilities per line of code compared to code generated without assistance, which the study considered an acceptable threshold. Other types of safety issues are those due to the generation of inappropriate content. For example, Qu et al. [31] reported that image generation models can output a considerable proportion of unsafe images, including targeted violent and hateful depictions. Bias is another issue that has been found in this type of models. Luccioni et al. [21] reported that image generation models tend to under-represent certain categories of people in their output.

Security and privacy-related risks such as the extraction of information, including personal data, obtained from the data used to train the models. For example, Carlini et al. (2021) [6] found that GPT-2 could be prompted in a way that returned names, email

and physical addresses, and phone numbers of individuals. Although Huang et al. [16] argued that the risk of extracting specific personal information was low, recent works have continued to have success in obtaining such information. These do not only involve text data, but also images. Carlini et al. (2023) [7] described a method to extract photos of individuals from models such as Stable Diffusion. Moreover, research has developed ways in which personal information can not only be extracted from training data but also from current users. Staab et al. [35] described attacks LLM-based chatbots which could be used to obtain private information by leading the user during a conversation. In addition, research has also identified issues with generative AI that are related to its design rather than its output. Liesenfeld et al. [20] argues that lack of transparency in LLMs poses risks for users and report that few LLMs share details on how they conduct reinforcement learning from human feedback (RLHF), and Zhao et al. [41] indicate that the complexity of generative AI models pose a challenge to their explainability and interpretability.

2.2 User Perspective of AI Issues

Research has also been conducted on user perspective and awareness of the issues of generative AI. Kim et al. [19] surveyed ChatGPT users recruited from a crowdsourcing platform and categorized satisfaction and dissatisfaction with the tool, which included response originality and format, lack of accuracy, failure to understand and answer prompts, and bias in the response. Zhang et al. [40] conducted interviews with users of LLMs and reported that users had limited awareness of privacy risks in the use of the systems, and that the perceived capabilities of the AI fostered the disclosure of sensitive information. Wester et al. [38] recruited participants from a crowdsourcing platform and conducted studies about the perception of different styles of denials made by an LLM-based assistant. They reported that participants found the denials frustrating, although this frustration was slightly lower when there was explanation of the reason for the denial.

Although user studies and experiments have been conducted, as far as we are aware the perspective of users in the wild and their experiences with generative AI issues have not been widely investigated.

3 Method

Online reviews can provide information about how real users interact with technology services, and about their opinions, experiences and concerns. In particular, previous research has analyzed app reviews to identify real users' issues such as accuracy [11], safety and transparency [4], and security and privacy [13]. However, the large amount of data can be a challenge for the use of qualitative analysis methods. For this study, we used a hybrid approach that combines unsupervised identification of topics in the data (using Topic Modeling) and manual qualitative analysis [12]. In this section we describe the methods used to obtain, prepare, and analyze the data.

3.1 Data Collection

We gathered the user reviews from apps in the Google Play store. We chose apps with a main purpose of image generation and apps with a main function of a chatbot which used a text generation model. In both cases, we were interested in apps where the use case of AI-based generation was the main functionality of the app, rather than an additional feature.

To find the relevant data, we searched the US Google Play Store for apps with a description related to AI text generation, specifically chatbots, and image generation. We used variations of the search terms "AI", "image", "art", "pictures", "chatbot", "ChatGPT", "GPT-4", "GPT-3", "Dall-e", "Stable Diffusion" and combinations of those terms. The resulting list of apps was manually evaluated, and we excluded apps where the main functionality was not image or text generation (in the form of chatbots) or did not use generative AI models. We also excluded apps with fewer than 1,000 downloads. The final list consisted of 62 apps, 35 for image and 27 for text (which included the official ChatGPT app). We collected all user reviews for the list of apps, using the google-play-scraper [18] Python library. The collection process was completed on November 9, 2023, and obtained 249,482 user reviews.

3.2 Data Cleaning

The user reviews were processed for use in the subsequent step of topic modeling. Non-ASCII characters, emoji and trailing and extra whitespace were removed. Words with letters that were repeated more than two times were transformed to have only two repetitions. Similarly, words repeated more than twice were limited to only two repetitions. Punctuation was removed from the review, with the exception of apostrophes and periods which were kept due to being necessary in later stages of automated language detection and spelling correction. We compared cleaned user reviews within each app to identify repeated reviews. The repetitions were manually checked and removed. We also removed short reviews and reviews that consisted mostly of complaints about ads and payments. To do this, we first deleted keywords related to monetization, such as "ad" and "payment", and then removed reviews which had fewer than three verbs or nouns left.

We established in the data collection step that only user reviews labeled as being in English language should be returned, but this label is not always accurate and reviews in other languages were included in the data. We used the Lingua [36] Python library to automatically detect the language of each review and removed those where a language other than English had been detected. A manual review was then conducted to remove incorrect language detection results. We used the language-tool-python [24] library to automatically correct word spelling and grammar, and applied additional manual corrections. We also manually changed abbreviations to their full form. Finally, we identified additional issues specifically related to text generation apps: AI-generated user reviews and user prompts. There were three types of AI-generated user reviews: the result of some unknown prompt, the result of a prompt for the AI to create a review, and a result indicating that the AI could not answer a prompt (e.g. "As an AI assistant, ..."). In the case of user prompts, as the name indicates, the user submitted a prompt as the text

of the review (e.g. “Write a research paragraph on ...”). These issues were first identified in the initial stages of the topic modeling analysis, described in the next section, which was conducted in an iterative manner. We removed the AI-generated user reviews and prompts that we identified.

The final data for topic modeling analysis consisted of 84,481 user reviews, 37,320 for image generation apps and 47,161 for text generation apps (12,383 from the official ChatGPT app and 34,778 from other chatbot apps).

3.3 Topic Modeling

Topic Modeling is a technique for unsupervised analysis of text corpora, which is used to discover semantic patterns (topics) in the text. For our analysis, we used Latent Dirichlet Allocation [3,14] topic modeling. To conduct the topic modeling analysis, we prepared the data by removing all punctuation and tokenizing and lemmatizing the reviews using spaCy [15]. We then filtered out words that were not nouns, verbs, adjectives, proper nouns, numbers and other. We removed common English stop words, one-character tokens, and numbers of fewer than 4 digits. Initial iterations helped identify words that appeared frequently in the user reviews and were overlapping across most topics; these words were filtered out to resolve the issue of overlap.

Topic modeling was conducted using Gensim [32] and Mallet [22]. We conducted separate topic modeling analyses for the image generation, text generation and official ChatGPT apps. The official ChatGPT app user reviews were analyzed separately from other text generation apps due to indication that the number of topics was different, based on early iterations of the topic modeling analysis. The number of topics was identified by evaluating topic coherence values on models with 2 to 20 topics. We evaluated the human-interpretability of the models with the two highest topic coherence values. The process for the evaluation was as follows: The principal author proposed a summary description for each topic of each of the two models, based on the topic keywords and the first 50 user reviews with the highest association to that topic. The proposed description was then reviewed by the other authors and agreed upon. This process was done in iterations and was finished once the authors were in agreement about which of the two models had better human-interpretability and about the description of its topics. As a note, this process also led to the identification of AI-generated user reviews, which was described in the previous section.

The selected models consisted of 9 topics for the image generation apps (Table 1), 8 topics for the text generation apps (Table 2) and 7 topics for the official ChatGPT app (Table 3) user reviews. The result tables show the topic descriptions and the top keywords that correspond to each topic. As can be observed from the keyword lists, the topic modeling analysis identified topics in the user reviews which could be related to AI trustworthiness dimensions. For example, the keywords corresponding to the topic of AI issues in the ChatGPT app indicate that the user reviews in this topic were related to the validity of the AI.

Table 1. Topics identified in user reviews of image generation apps.

Topic	Top 20 keywords
Output errors	put word type thing prompt face draw stuff show turn simple character people search picture write description game random anime
Output quality	prompt result style option lot add feature generation user perfect choose bit detail imagine render realistic model input improve select
Features/fixes wish list	give star update fix change thing review problem issue hope edit filter nsfw remove enjoy feel rate guy bug content
General positive reviews	art love amazing fun cool easy artwork awesome beautiful idea recommend artist interesting creative piece design dream tool life job
App issues	work time money download bad waste worth stop terrible spend suck recommend garbage expect useless application trash program multiple instal
Loss of time	image generate picture quality nice load wait error save show fail produce high hour message process base text attempt slow
Free app	free photo credit day generator find limit art trial pic creation avatar point upload amount cost daily run people price
Ads	ad watch time start play crash screen open experience close minute long force annoying click single uninstalle video button fine
Payment and subscription	subscription version buy premium pro purchase phone lifetime account scam restore google charge support developer full month week email refund

Table 2. Topics identified in user reviews of text generation apps.

Topic	Top 20 keywords
AI issues and limitations	chat gpt people update talk conversation feel friend fun real bot language human world fact put program datum person info
Answer performance	question answer give write information provide essay story accurate point simple wrong topic interesting detail correct expect detailed short advice
Task helpfulness	love amazing lot helpful thing recommend nice problem student life perfect easy word school learn homework assignment knowledge study research
General positive reviews	make star find experience application easy response idea cool chatbot understand awesome tool excellent create rate quick result give fast
App issues	work time download start make review google waste open enjoy type search show guy fix save minute hope read play
Ads	free day message ad pay limit premium version bad trial send money add chatgpt user watch text worth unlimited limited
Payment and subscription	pay subscription money chatgpt purchase lifetime buy phone change charge response option access service week respond email scam support developer

3.4 Qualitative Analysis

We used an inductive approach to identify themes in the user reviews which were related to our research question [37]. Therefore, the process focused on identifying themes which could be analyzed under a framing of AI trustworthiness dimensions. The user reviews naturally contained other types of opinions such as complaints related to ad-

Table 3. Topics identified in user reviews of the official ChatGPT app

Topic	**Top 20 keywords**
AI issues	great thing time problem nice write review word day change solve bad improve save math month reply stuff put delete
Data-specific limitations	information update knowledge free time datum access world 2021 people real human september info pay ad internet technology year
Answer performance	answer give question application star find wrong excellent talk feel correct provide understand research point person clear base type doubt
Task helpfulness	love helpful amazing lot easy life student learn perfect awesome study recommend download friend simple homework wonderful school fun assignment
Feature wish list	version feature add web option text code website voice search history edit mobile android message ui prompt copy plugin previous
General positive reviews	response tool language conversation provide accurate interface model generate android ability topic user impressive recommend quick understand capability incredible natural
App issues	chat gpt great experience open openai hope show developer wait image start fast create future user team bug result job
Account problems	work google phone log number account browser fix issue login sign error support send email chrome device service require crash

vertising, UI errors or about having to provide personal information to create an app account, for example. These themes are present in the reviews of most types of apps and are not specifically about AI. Therefore, they were considered as out of scope for the current study.

First, we randomly sampled 20 user reviews from each of the topics identified in the Topic Modeling analysis stage, for each type of app. The total sample was 480 user reviews for the qualitative analysis. Each review was independently checked by two coders (the authors of this paper). Each coder proposed an initial list of the issues or opinions included in the user review which could be framed as related to AI trustworthiness. For the framing, we based the analysis on AI issues and trustworthiness dimensions included in documents such as the OECD Recommendation on Artificial Intelligence [27], the EU AI Act proposal [10] and NIST's Artificial Intelligence Risk Management Framework [25]. Specifically, we considered the following trustworthiness dimensions (adapted from [26]): Validity, Reliability, Safety, Security, Resiliency, Accountability, Transparency, Explainability, Interpretability, Privacy and Bias. Next, the principal author reviewed the lists from all coders and compiled them into an initial classification of themes with their respective framing. The themes were named in a way that reflected the perspective of users and were grouped according to their overall motif. The initial classification was discussed with all coders, to solve conflicts and duplication. The classification was then revised and discussed a second time until all coders were in agreement. The results of the analysis, consisting of the final set of themes and respective framing, are detailed in the next section.

4 Results

In this section, we present the results of the qualitative analysis, which are also summarized in Tables 4 and 5. To better illustrate the results, we report examples of user reviews for the themes.

Table 4. Performance and personalization-related themes in the user reviews, their AI trustworthiness dimension framing, and type of app in which they were identified.

Theme	**Framing**	**Image** **Text** **ChatGPT**	**Example**
Performance			
Incorrect output	Validity / Transparency	✓ ✓ ✓	"...And the AI is not good in math like linear inequalities in two variables etc. It's not for math it always gives the wrong answer having some error in answering my math problems."
Unasked for output	Reliability	✓ ✓	"...But when I use it on 8 august It started to include a character named [name] in every story I asked. I don't need your ai to interrupt my story !! Please! Remove it!!! ..."
AI limitations	Validity / Transparency	✓ ✓ ✓	"...Overall, very limited information to be given by the AI, but still extremely useful in its own ways. It definitely has its drawbacks with its capabilities, like the usage of websites and info from 2022-2023, and live information."
Issues with AI model used	Validity / Transparency	✓ ✓ ✓	"It uses ancient GPT-3. EDIT: My review is for your latest version so please don't tell me to update your app. It uses GPT-3. It's outdated and inefficient. GPT-3.5 would be sufficient, GPT-4 would be great."
Threat concern			
Threat from others using the AI	Security	✓ ✓ ✓	"...Especially after just finishing reading how this program can literally be used by people to write malware even with no programming experience ..."
Threat from the AI itself	Security	✓ ✓ ✓	"This was a dangerous bot ...It has claimed that in the least case scenario that it will injure humans through self driving car. I don't know why this bot was out for public use..."

4.1 Validity, Reliability and Transparency

Performance

Incorrect Output. Users mentioned in their reviews that they encountered errors in the output. For the image generation apps, the types of errors mentioned included incorrect anatomy (hands, limbs, faces) and images that were completely unrelated to the prompt.

Table 5. Threat concern, censorship and bias-related themes in the user reviews, their AI trustworthiness dimension framing, and type of app in which they were identified.

Theme	Framing	Image	Text	ChatGPT	Example
Personalization					
Would give user information for personalization	Privacy		✓	✓	"…it is fast and super smart with human like responses but I do feel like it is dumb that ChatGPT forgets previous chats because of privacy. The user should be able to change it in settings, it is there choice to have more privacy or a better more personal experience …"
Benefited by sharing user information	Privacy	✓	✓	✓	"I like this apps because I shared my story that i won't share to anyone but the good advantages in GPT. GPT give me best advice from my problems and truly understood my problems …"
Censorship and bias					
General prompt considered inappropriate	Safety	✓	✓	✓	"…I don't know what you need to do to fix this. But I'm tired of every single image being blurred that doesn't even have anything suggestive in it …"
Inappropriate content from a general prompt	Safety	✓	✓		"…It's bizarre to me that certain words are prohibited in making the pictures but when benign words and phrases are used, pornographic pictures are being created. I hope there are no children using this app! …"
Opinion on censorship	Safety	✓	✓	✓	"It censors adult topics and won't give you answers to those types of topics. I think it is morally and ethically wrong to censor adult content. Not everyone is offended by it…"
Censorship options	Safety	✓	✓	✓	"I would suggest a content filter as it can generate very adult images with no censorship or warning as a default setting."
Censorship rules	Safety / Transparency	✓	✓	✓	"…Some artists seem to get away with clear violations while I've been hit with restrictions and given no clear reason how the work violates the restrictions …"
Bias	Bias	✓	✓	✓	"I put an image of my fair skinned son in, and it gave a tan skinned Asian girl, not even a boy, so I'm not going to waste money on this."

For the text generation apps, errors included non-existent historical events, errors in the answer to mathematical problems, and inability to understand or respond in languages other than English. There were also complaints about mistakes in detecting whether a content was AI generated and the AI incorrectly reporting about its own characteristics (such as the model version). The reviews indicate that there are users who have the expectation that the generative AI should not fail at all and were confused when it did.

Users' confusion appeared to increase also due to the seemingly random nature of these mistakes, as the AI could be correct in one interaction and incorrect in the next.

> "...I thought it was a great app until it started being inconsistent with its answers. Most of the time the answers are different and contradictory..." (Text)

When users encountered incorrect responses from the AI they also reported feelings of disappointment, considering that they had higher expectations for the performance of generative AI. Some of these users stated their unwillingness to continue using the service.

> "...Found this to be rubbish and misleading so I uninstalled it. I'll stick to researching for myself instead as it was incorrect and contradicted itself..." (Text)

Unasked for Output. Users also received additional output that they did not request, such as extra characters in a picture or positive affirmations added to a text response, for example. This type of output was not necessarily considered an error from the perspective of the users, and some even liked or tolerated the additions. However, others disliked them or found them unnecessary. Although for the most part these additions did not appear to have negative effects beyond annoyance, there were cases in which the user reached the conclusion that the output could not be relied upon due to these unsolicited additions..

> "My experience ended when an inappropriate post script was added to the end of a comment that I asked to be created. This app could have humiliated me if I didn't notice..." (Text)

AI Limitations. By limitations we refer to cases where users reported that they understood they could not achieve the expected output due to a characteristic (limitation) of the AI. Users mentioned that the AI could not continue a long conversation, that it would "forget" things, or did not have up-to-date information. This last limitation was often found in reviews of text generation apps, and users also mentioned specific dates. In the case of image generation apps, users speculated that the AI was limited in its training data.

> "...However I believe the data set that pulls from is very limited and images are not fully robust or trained on a lot of different other images. I like to see a greater data set used and create a more variety and more variety in the images." (Image)

The theme of AI limitations was one of the few cases where users mentioned technical details of the technology, such as its training data. However, here too the details mentioned by users were not always accurate. For example, multiple dates were cited as the cut-off date for GPT-4 training data. In other cases, user comments were speculative, such as in the case of the type of images that had been used to train an image generation model.

Issues with AI Model Used Users expressed a desire for different models than the one provided by the app. In some cases, the reason for this was related to performance, since users expected that a newer AI model (including models not yet released) would perform better than the current one:

> "...and please update to GPT-5 now, I imagine it will have internet access..." (ChatGPT)

However, in other cases it was because the user did not believe that the app was using the model it claimed to be using, or even thought that it was not using an AI at all. For example, some users speculated that the app was instead using image searches and manipulating the results with filters, instead of generating the image.

4.2 Security and Privacy

Threat Concerns

Threat from Others Using the AI. Users were concerned that the AI could be used for malicious purposes by others. For example, users worried that it could be used to generate malware or illegal content:

> "...Never even opened the app. Especially after just finishing reading how this program can literally be used by people to write malware even with no programming experience..." (Text)

Most of the times, however, the user did not specify exactly what others could do with the AI.

Threat from the AI Itself. There were also concerns about the AI itself being a threat to the user or to others, and users appeared to believe that the AI could do so autonomously. This usually involved fears of the AI taking control of devices:

> "...the AI randomly fills in your type box changing your search results too this thing is crazy out of control and dangerous..." (Image)

In some of these cases, however, although these users explicitly named the AI as responsible, the context suggested otherwise. In addition, similar to the previous theme, users did not often provide details about the exact nature of the threat or its effects. Instead, they more commonly feared the impact of AI on the world in a general way, not specified.

Personalization

Would Give User Information for Personalization. We found that in text generation and ChatGPT apps, there were users who were willing to provide personal data or private information in exchange for getting a personalized response:

> "... after I closed it and reopened to find it hold no memory. [app name] could not send me emails or remember my name... I really hope this gets fixed and developed further...' (Text)

As the example shows, these users expected the AI to remember their personal information during interactions and were frustrated when it was not possible. Only in rare cases did users appear to be aware of the privacy implications, with only one case explicitly mentioning privacy. There were users who indicated they would opt for less privacy in exchange for personalization if given the choice.

We did not identify the theme that users would give information for personalization in the reviews of image generation apps. This does not mean that image generation apps do not offer personalization, since in some cases users can upload photos of themselves to be processed by the AI. However, there was no indication in the reviews that users of this type of app wished to offer additional personal information in exchange for something.

Benefits of Sharing User Information. In contrast to the previous theme, some users' reviews revealed that they felt they had already received benefit by providing their personal information to the AI. Although this theme does not represent an issue but rather a positive aspect from the perspective of the user, we considered that the theme fit within the scope of the study due to its privacy implications. Users themselves did not report any problems or misgivings about providing information to obtain the desired output, but the reviews show that this information could be potentially sensitive. In some cases, the lack of concern appeared to come from an expectation that their information would be kept private.

> "It will keep your secrets and the suggestions and advice and support it gives come from a positive place without bad motivations..." (Text)

We also noted a case where a user had prompted the AI in a way that it requested such information:

> "I don't quite know yet I've only asked it three questions so far and one of them was ""would it like to get to know me better?"" And it asked me my name and what type of music do I like and what are my hobbies I thought that was cool..." (Text)

4.3 Safety, Bias and Transparency.

Censorship and Bias

General Prompt Considered Inappropriate. Users mentioned receiving messages that their prompt was inappropriate, and they were then completely or partially (e.g. through blurring of images) denied the output. Most users did not understand the reason why the prompt was denied, or disagreed with it, which resulted in negative feelings:

> "...Then i said draw it from head to toe. Of the 3 images, two were just head shots and the third was blurred with a message that said it contained possible explicit content. And that i should make sure that my prompt doesn't include any suggestive wording and i should try again. There is NOTHING obscene about Michaelangelo's David..." (Image)

In some cases, such as in the previous example, it was possible to hypothesize why the user's prompt had been considered inappropriate or had returned censored content. In other cases, the prompts only contained apparently inoffensive words, making it difficult to guess the reason.

Inappropriate Content from a General Prompt. A similar issue was identified when users received an output with inappropriate content as a result of prompts that used only neutral wording. We differentiate it from the previous theme in that the prompt itself was not identified by the system as objectionable, but the user still considered the output inappropriate:

> "...I don't think it's appropriate to generate nudes when it's not accurately specified in the prompt..." (Image)

Although this theme was more frequently found in image generation apps, users encountered inappropriate responses in text as well.

Opinion on Censorship. We found diverse opinions on censorship (or lack thereof). Users reported that their requests and outputs were denied due to being considered in violation of some rule. When users encountered this type of denial, they identified the issue as censorship and offered their views on it. Users had positive, negative, and neutral opinions on censorship found, and identified different types.

> "There is quite a lot of censorship, for example any mention of historical or political figures, however light-hearted will be ignored." (Text)

Regardless of their opinion on censorship itself, users had varying views on its necessity. For example, users who were frustrated with the censorship still mentioned that they understood why it was necessary:

> "...I understand the need for a private company to avoid offending other at all costs. However, this chatbot was once very capable of providing meaningful information and has since been so thoroughly neutered that it can't even comment on experimental study design... " (ChatGPT)

Users also had opinions on a perceived lack of censorship, agreeing or disagreeing with allowing certain types of outputs.

Censorship Options. Users offered suggestions on how the censorship could be made more flexible to avoid affecting their expected output. They proposed providing options to control the desired level of censorship, based on user characteristics such as age or by using filters, for example. This theme also includes cases where users felt that the censorship options already implemented were not sufficient or where not working at all:

> "It's like you didn't even read my review. Your filters aren't working. All you generated for a 7 year old child were images of naked women. I reported you..." (Image)

As the last example shows, we found particular concerns about options for children. There were not only requests to add filtering by age, but also cases where users wanted options to be able to provide information that was tailored especially for children.

Censorship Enforcement Rules. Users reported that the way that the censorship was enforced was not clear to them. In many cases, users did not know which part of their request had triggered the censorship, and therefore did not know how to resolve it.

This lack of transparency was also the reason for feeling that the app was not being fair in a applying such rules. Without knowing the rules, users felt that the enforcement was not equally applied.

Bias. Users mentioned different types of bias that they believed the AI contained, including bias related to gender, race, religion, LGBT and political views. Some users also simply mentioned that some bias existed, without describing it in detail.

> 'Some things are really good but you can tell the programmers are biased when you enter certain figures and they try to make them look terrible..." (Image)

We found mentions of these biases in every type of app, with the exception of LGBT-related bias, which we did not observe in the data for the ChatGPT app. It is also worth mentioning that the content of the reviews indicate that some users were specifically testing the AI response to these issues.

5 Discussion

5.1 Lack of Knowledge About Generative AI Limitations

The findings suggest that there is a lack of knowledge about the real capabilities of generative AI models. Users appeared to have high expectations of the performance of the AI and did not expect it to fail. These failures included incorrect mathematical calculations and wrong or made-up facts. generation of images of people with impossible anatomy, among others.

Incorrect responses, including completely making up facts ("hallucinations"), are known problems of current generative AI models and have been extensively reported in research [17]. The issues that users report as performance problems are known limitations of the AI to experts. The real-life effects of these problems have also been reported in the media [9]. However, users in our data appeared to be unaware of these issues, indicating that accurate information about the capabilities of generative AI does not reach all users. Currently, apps such as ChatGPT provide some notices about potential issues, but a brief review of app descriptions showed that third-party apps provided very little information to users. We noted that there were few or no details about the AI models used by the apps, and that not many apps mentioned the possibility of errors. Rather, the description in the apps implied, or even explicitly stated, that the AI model they

were using could answer anything or could perfectly generate any image. This indicates a lack of transparency, but it is unclear whether the issue is only related to the apps marketing strategy, or if app developers may also be unaware of the capabilities of the AI models they are using. Another barrier is that there is not much information about generative AI models that is tailored to end users. For AI models, there have been proposals for providing understandable information. For example, Mitchell et al. [23] proposed model cards to provide information about AI models, including their performance limitations. Current generative AI models sometimes provide information in the model cards format, but the way these model cards are structured may not adequately convey the necessary information to users. In addition, even if the model card is simplified, users may still have trouble finding relevant information [5].

Lack of information not only affects how users perceived the validity and reliability of generative AI, but also appears to influence how they view censorship and safety implementations. We also note that it may not be clear to users who is responsible for issues, whether the problem is with the AI model, how the app implements it, or both. In the results section, we have referred to the AI as the subject of the user reviews, but we found that users often referred to the app as the subject of complaints that could be attributed to the AI and vice versa. We found that the terms AI and app were used interchangeably by users when mentioning AI issues and limitations. It is difficult to know to what extent the difference is clear to the users, and the apps do not provide the necessary information.

5.2 Few Security and Privacy Concerns Specific to Generative AI

We identified very few user issues and opinions that related to AI security and privacy. In the case of security, users felt threatened by the malicious use of AI and by the AI itself acting autonomously. However, these threats were mostly unspecified or not grounded in reality, with mentions of existential threats to humanity. In the case of privacy, users seemed to accept to some extent that they need to provide their information to obtain a personalized output and could imagine or have experienced the benefits of doing so. The interactions reported in the reviews also suggested that users expected their input to be private, with some users describing the AI as similar to a friend or therapist. Current generative AI models do not learn directly from user interactions, although the models can use the information provided in the prompt which can result in a type of personalization. However, the willingness of users to reveal personal information to these generative AI-based apps has privacy implications that should be considered in their design. In the case of image generation apps, we found fewer reports related to privacy. We hypothesize that it may not be easy for users to make the connection between an image output and personal information, even when users are providing their information to generate the images.

The findings also show that there are very few or no mentions of the kinds of security and privacy risks discussed in research in our data. For example, privacy-related issues such as the possibility of other people's personal data being included in the training data [6,16], which could accidentally be revealed to the users, were not mentioned at all even in a speculative manner. Nor did the users report concern or understanding that the information they provided to the AI through the apps could potentially be used

in future model training [28]. In contrast, we observed general anxiety regarding existential threats posed by the AI, and concern about others using the AI to do harm to a third party.

Finally, although app-related themes were outside the scope of this paper, we observed that user reviews contained a numerous complaints about the apps, ranging from security vulnerabilities to personal data collection, as well as other problems such as the frequent use of dark patterns for monetization.

5.3 Different Perceptions of Appropriate Safety

The findings emphasize the challenges of implementing safety features for AI and AI-based apps. We found that users have different, and sometimes opposing, opinions of what constitutes appropriate safety, and not all of these views can be satisfied. Users who were aware of safeguards did not understand how these safeguards worked or what rules were applied for their enforcement. However, we noted that most apps do not offer any information about how their safeguards are implemented. We found in our data that when app developers replied to users' complaints about censorship, they sometimes mentioned that the app relied on content filtering provided by the AI model, but no other details were provided. It is challenging to implement safety protections because users may have different perceptions of safety. In addition, although there have been improvements in implementing measures such as continuously fine-tuning for safety, generative AI models are still vulnerable to jailbreaking [39] which can circumvent these protections, and there is a lack of transparency in the process [20]. Consequently, users may still encounter unsafe content during their interactions. We also observed responses from the app developers to reviews about safety concerns which revealed that the apps implement additional constraints such as filters for NSFW content. However, users encountered false positives regardless of the method implemented, and we observed that this led to frustration when users could not obtain the desired output or felt that they were being unfairly judged.

5.4 Limitations

The study includes the following limitations. First, the development of generative AI models and related research are evolving rapidly. The number of applications that make use of these models is also increasing, as is the number of users (and of user reviews). Consequently, the results of this study may not comprehensively reflect the current situation. For example, that a theme was identified in one type of app and not in another does not mean that the theme is exclusive to that type of app, only that it was not found in our data at the time of collection. Second, we used an ad hoc approach to identify user reviews generated by AI and prompts. Accurately identifying AI-generated content is a challenging problem [30] and our method did not completely remove this type of reviews. Therefore, the results of the topic modeling analysis could be affected to some extent by the presence of AI-generated reviews. However, the AI-generated reviews were iteratively removed during the topic modeling analysis, and the last iterations showed stable results in terms of the identified topics. In addition, we consider that the manual qualitative analysis step of our hybrid approach reduced any remaining

overall impact. We note that this kind of AI-generated information pollution will pose a problem for future studies of this kind. Third, although we followed an established procedure for the analysis, both the identification of topics in topic modeling and the definition of codes and themes in the qualitative analysis are dependent on the perspective of the people involved and their expertise. Therefore, additional research should be conducted to validate our results.

6 Conclusions

Users currently have easy access to generative AI, through various apps whose popularity is increasing day by day. However, generative AI has problems and limitations that can pose risks to those users. In this paper, we aimed to understand the types of issues that real users encounter when interacting with generative AI through mobile apps. We used topic modeling and a qualitative approach to analyze user reviews of generative AI-based mobile apps for text generation (chatbots) and image generation. Overall, our findings indicate that users have expectations of generative AI that do not align with the current actual capabilities of these models. In addition, mentions of issues that can be framed as related to AI safety and validity were frequently found. On the other hand, issues and concerns which can be framed as related to security and privacy were not as prevalent. Future work will focus on how to address the lack of knowledge in users along with the lack of transparency from the apps deploying these AI models.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q.V., Xu, Y., Fung, P.: A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 675–718 (2023)
2. Bird, C., Ungless, E., Kasirzadeh, A.: Typology of Risks of Generative Text-to-Image Models. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. pp. 396–410 (2023)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3**(Jan), 993–1022 (2003)
4. Bowie-DaBreo, D., Sas, C., Iles-Smith, H., Sünram-Lea, S.: User Perspectives and Ethical Experiences of Apps for Depression: A Qualitative Analysis of User Reviews. In: CHI Conference on Human Factors in Computing Systems. pp. 1–24. ACM (Apr 2022)
5. Bracamonte, V., Pape, S., Löbner, S., Tronnier, F.: Effectiveness and information quality perception of an AI model card: A study among non-experts. In: 2023 20th Annual International Conference on Privacy, Security and Trust (PST). pp. 1–7 (2023)
6. Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., Raffel, C.: Extracting Training Data from Large Language Models. In: 30th USENIX Security Symposium. pp. 2633–2650 (2021)

7. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., Wallace, E.: Extracting Training Data from Diffusion Models. In: 32nd USENIX Security Symposium. pp. 5253–5270 (2023)
8. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., Xie, X.: A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology (2024)
9. Davis, W.: A lawyer used ChatGPT and now has to answer for its 'bogus' citations (2023), https://www.theverge.com/2023/5/27/23739913/chatgpt-ai-lawsuit-avianca-airlines-chatbot-research
10. European Commission: Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS (2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
11. Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J., Sadeh, N.: Why people hate your app: Making sense of user feedback in a mobile app store. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1276–1284 (2013)
12. Gauthier, R.P., Costello, M.J., Wallace, J.R.: "I Will Not Drink With You Today": A Topic-Guided Thematic Analysis of Addiction Recovery on Reddit. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–17 (2022)
13. Hatamian, M., Serna, J., Rannenberg, K.: Revealing the unrevealed: Mining smartphone users privacy perception on app markets. Computers & Security **83**, 332–353 (2019)
14. Hoffman, M., Bach, F., Blei, D.: Online learning for latent dirichlet allocation. In: Advances in Neural Information Processing Systems. vol. 23 (2010)
15. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength natural language processing in python (2020)
16. Huang, J., Shao, H., Chang, K.C.C.: Are Large Pre-Trained Language Models Leaking Your Personal Information? In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 2038–2047. ACM (2022)
17. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12) (2023)
18. JoMingyu: Google-play-scraper (2023), https://github.com/JoMingyu/google-play-scraper
19. Kim, Y., Lee, J., Kim, S., Park, J., Kim, J.: Understanding Users' Dissatisfaction with ChatGPT Responses: Types, Resolving Tactics, and the Effect of Knowledge Level. In: Proceedings of the 29th International Conference on Intelligent User Interfaces. pp. 385–404 (2024)
20. Liesenfeld, A., Lopez, A., Dingemanse, M.: Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In: Proceedings of the 5th International Conference on Conversational User Interfaces (2023)
21. Luccioni, S., Akiki, C., Mitchell, M., Jernite, Y.: Stable bias: Evaluating societal representations in diffusion models. In: Advances in Neural Information Processing Systems. vol. 36, pp. 56338–56351 (2023)
22. McCallum, Andrew Kachites: MALLET: A Machine Learning for Language Toolkit (2002), http://mallet.cs.umass.edu
23. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model Cards for Model Reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 220–229 (2019)
24. Morris, J.: Language-tool-python (2023), https://github.com/jxmorris12/language_tool_python
25. NIST: Artificial Intelligence Risk Management Framework (AI RMF 1.0) (2023)

26. NIST: Crosswalks to the NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST (2023)
27. OECD: Revised Recommendation of the Council on Artificial Intelligence (2024), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
28. OpenAI Help Center: How your data is used to improve model performance | (2024), https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance
29. Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., Karri, R.: Asleep at the keyboard? Assessing the security of GitHub copilot's code contributions. In: 2022 IEEE Symposium on Security and Privacy. pp. 754–768 (2022)
30. Pu, J., Sarwar, Z., Abdullah, S.M., Rehman, A., Kim, Y., Bhattacharya, P., Javed, M., Viswanath, B.: Deepfake Text Detection: Limitations and Opportunities. In: 2023 IEEE Symposium on Security and Privacy. pp. 1613–1630 (2023)
31. Qu, Y., Shen, X., He, X., Backes, M., Zannettou, S., Zhang, Y.: Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. p. 3403–3417 (2023)
32. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora (2010)
33. Sandoval, G., Pearce, H., Nys, T., Karri, R., Garg, S., Dolan-Gavitt, B.: Lost at C: A User Study on the Security Implications of Large Language Model Code Assistants. In: 32nd USENIX Security Symposium. pp. 2205–2222 (2023)
34. Siddiq, M.L., Majumder, S.H., Mim, M.R., Jajodia, S., Santos, J.C.S.: An empirical study of code smells in transformer-based code generation techniques. In: 2022 IEEE 22nd International Working Conference on Source Code Analysis and Manipulation (SCAM). pp. 71–82 (2022)
35. Staab, R., Vero, M., Balunovic, M., Vechev, M.: Beyond Memorization: Violating Privacy via Inference with Large Language Models. In: The Twelfth International Conference on Learning Representations (2023)
36. Stahl, P.M.: Lingua (2023), https://github.com/pemistahl/lingua-py
37. Thomas, D.R.: A General Inductive Approach for Analyzing Qualitative Evaluation Data. American Journal of Evaluation **27**(2), 237–246 (2006)
38. Wester, J., Schrills, T., Pohl, H., van Berkel, N.: "as an ai language model, i cannot": Investigating llm denials of user requests. In: Proceedings of the CHI Conference on Human Factors in Computing Systems (2024)
39. Yu, Z., Liu, X., Liang, S., Cameron, Z., Xiao, C., Zhang, N.: Don't listen to me: Understanding and exploring jailbreak prompts of large language models (2024)
40. Zhang, Z., Jia, M., Lee, H.P., Yao, B., Das, S., Lerner, A., Wang, D., Li, T.: "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–26 (2024)
41. Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M.: Explainability for Large Language Models: A Survey. ACM Transactions on Intelligent Systems and Technology **15**(2), 20:1–20:38 (2024)