End-Users Know Best: Identifying Undesired Behavior of Alexa Skills Through User Review Analysis

MOHAMMED ALDEEN, Clemson University, USA
JEFFREY YOUNG, Clemson University, USA
SONG LIAO, Clemson University, USA
TSU-YAO CHANG, Clemson University, USA
LONG CHENG, Clemson University, USA
HAIPENG CAI, Washington State University, USA
XIAPU LUO, Hong Kong Polytechnic University, Hong Kong
HONGXIN HU, University at Buffalo, USA

The Amazon Alexa marketplace has grown rapidly in recent years due to third-party developers creating large amounts of content and publishing directly to a skills store. Despite the growth of the Amazon Alexa skills store, there have been several reported security and usability concerns, which may not be identified during the vetting phase. However, user reviews can offer valuable insights into the security & privacy, quality, and usability of the skills. To better understand the effects of these problematic skills on end-users, we introduce ReviewTracker, a tool capable of discerning and classifying semantically negative user reviews to identify likely malicious, policy violating, or malfunctioning behavior on Alexa skills. ReviewTracker employs a pre-trained FastText classifier to identify different undesired skill behaviors. We collected over 700,000 user reviews spanning 6 years with more than 200,000 negative sentiment reviews. ReviewTracker was able to identify 17,820 reviews reporting violations related to Alexa policy requirements across 2,813 skills, and 131,855 reviews highlighting different types of user frustrations associated with 9,294 skills. In addition, we developed a dynamic skill testing framework using ChatGPT to conduct two distinct types of tests on Alexa skills: one using a software-based simulation for interaction to explore the actual behaviors of skills and another through actual voice commands to understand the potential factors causing discrepancies between intended skill functionalities and user experiences. Based on the number of the undesired skill behavior reviews, we tested the top identified problematic skills and detected more than 228 skills violating at least one policy requirement. Our results demonstrate that user reviews could serve as a valuable means to identify undesired skill behaviors.

ACM Reference Format:
Mohammed Aldeen, Jeffrey Young, Song Liao, Tsu-Yao Chang, Long Cheng, Haipeng Cai, Xiapu Luo, and Hongxin Hu. 2024. End-Users Know Best: Identifying Undesired Behavior of Alexa Skills Through User Review Analysis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 89 (September 2024), 28 pages. https://doi.org/10.1145/3678517

Authors' addresses: Mohammed Aldeen, mshujaa@clemson.edu, Clemson University, USA; Jeffrey Young, Clemson University, USA, jay8@clemson.edu; Song Liao, Clemson University, USA, liao5@clemson.edu; Tsu-Yao Chang, Clemson University, USA, tsuyaoc@clemson.edu; Long Cheng, Clemson University, USA, lcheng2@clemson.edu; Haipeng Cai, Washington State University, USA, haipeng.cai@wsu.edu; Xiapu Luo, Hong Kong Polytechnic University, Hong Kong, csxluo@comp.polyu.edu.hk; Hongxin Hu, University at Buffalo, USA, hongxinh@buffalo.edu.

2474-9567/2024/9-ART89 $15.00
https://doi.org/10.1145/3678517

1 INTRODUCTION

Voice personal assistants (VPAs), such as Amazon Alexa, Google Assistant and Apple Siri, are rapidly gaining in both domestic and business popularity. In particular, the Amazon Alexa marketplace has, in recent years, expanded rapidly through the use of voice apps called skills. Today's Amazon Alexa boasts over 100,000 skills [55] available for use through the skills store[1] with functionality ranging from unlocking a door to paying a credit card balance. Amazon has accelerated the expansion of its platform by opening up its skills store to third-party developers, who can now create and publish their own skills. Unfortunately, third-party development has been shown to be fraught with many security & privacy issues on other platforms such as Android and iOS [32] with recent research [61] [12] [29] [40] [51] drawing similar conclusions for the Amazon Alexa platform. For instance, it was shown that a skill's content can be altered without going through a re-certification process making all deployed skills susceptible to malicious content changes [40]. This is due to the fact that a large portion of a skill's source-code is hosted on the developer's back-end (*i.e.*, hosted on AWS). Allowing developers to host their code externally makes static code analysis a daunting task by current methods and leaves only the skill's behavior that can be tested through dynamic analysis, which is a very time-consuming process.

In an effort to thwart unscrupulous skill development, Amazon has put in place a series of policies to be adhered to by third-party developers. However, recent research has found that the skill vetting process is lax [12]. The authors could get policy violating content easily through vetting. Due to this issue, large scale skill testing for identifying existing policy-violating skills has been an ongoing process. For example, SkillExplorer [29] tested over 30,000 Alexa skills for data collection related policy violations. Another recent work SkillDetective [61] tested 54,055 Alexa skills for multiple policy violations in a study conducted over one year. Both efforts report large numbers of potentially malicious skills found on the Amazon Alexa skills store. Even though current work has managed to test a large number of skills, there still exists over half of the Amazon marketplace left untested. In particular, existing studies can not analyze most skills in the "Smart Home" and "Connected Car" categories because these skills normally require account linking and device linking for them to work properly.

In this work, we introduce ReviewTracker, a tool to analyze user reviews posted by end-users on the Amazon Alexa skills store and identify reported incidents of policy violations and user frustrations. In summary, we make the following contributions:

- We collected over 700,000 user reviews posted during the period between January 2016 and March 2023 from US marketplace in Amazon Alexa's skills store. We first used semantic similarity to set rules to filter out unrelated reviews, and we labeled 14,080 reviews with various types of undesired skill behaviors. In our analysis, we categorize these undesired behaviors into two main groups: user frustrations and policy violations. To classify the review data, we designed and developed ReviewTracker[2], an NLP pipeline capable of identifying specific instances of policy violations and user frustrations in Alexa skills.
- We conducted a comprehensive analysis of user frustration and policy violation reviews. ReviewTracker identified 17,820 reviews reporting policy violations across 2,813 skills, and 131,855 reviews highlighting different types of user frustrations associated with 9,294 skills. Our findings revealed that users expressed clear dissatisfaction with IoT skills because they didn't match their advertised capabilities. Also, skills in certain categories only understand simple commands, leading to high volume of users complaining about the skills comprehension.
- We also developed a novel LLM-based dynamic testing tool, leveraging ChatGPT to interact with VPA to automatically test skill behaviors and record all the interactions that occur in between. This approach allows us to accurately analyze whether the identified problematic skills align with their actual behavior.

[1]Amazon's skill store is an online marketplace that contains all skill listings. A skill's listing is a unique URL that contains a skill's description, sample utterances, user rating, user feedback, developer, etc.

[2]The implementation code of ReviewTracker and user review dataset are available at https://github.com/REviewTracker/ReviewTracker.

The dynamic testing revealed violations of policies and safety measures across multiple categories, with 228 skills found to be in violation of at least one policy requirement. Kids-related and Health skills were discovered to collect personal information while a number of skills across other categories redirect users to external websites.

Roadmap. Section 2 provides a comprehensive background on Alexa Skills, User reviews and policy requirements. Section 3 reviews related work. Section 4 delves into the design of ReviewTracker. Section 5 evaluates ReviewTracker's performance. Discussion of the implications, benefits, and limitations of ReviewTracker is presented in Section 6. Section 7 concludes the paper.

2 BACKGROUND

2.1 Alexa Skills Store

The Amazon Alexa platform stands as one of the largest VPA platforms available. It empowers third-party developers to publish voice apps, referred to as skills, on the Alexa skills store. In order for a skill to be made available in the skills store, it must undergo a certification process that involves validation and testing to ensure compliance with Amazon Alexa's policy requirements [5] and security requirements [6]. To facilitate user interaction with these skills, Alexa mandates that developers provide not only a description but also a set of utterances. These utterances serve as voice commands that users can employ to invoke specific skills or execute particular functions within them. Furthermore, the skill webpage for each individual skill showcases fundamental details such as the skill's name, developers, available utterances, description, privacy policy, and user ratings/reviews. This comprehensive information aids users in selecting appropriate skills and gaining a clear understanding of how to effectively utilize them. In addition, the Amazon Alexa platform provides skill developers with a skill simulator specifically designed for testing purposes.

2.2 Alexa Policy Requirements

To ensure the privacy and content safety for end-users, Amazon has defined a set of policy requirements, which are categorized into 14 main sections. Each section has a list of policies related to specific category. These policies and restrictions address various aspects, such as prohibiting promotions, advertisements, the endorsement of alcohol, tobacco usage within skills, etc. Every skill that is certified and listed in the skills store is supposed to adhere to these policy requirements and guidelines. If a skill contains, facilitates, or promotes content that violates these policies, it will be subject to rejection or suspension.

2.3 Undesired Skill Behaviors

Users tend to have an expectation of engaging in a functional conversation with VPA. Researchers have dedicated significant efforts to comprehend the dynamics of human interaction with automated agents, particularly by comparing and contrasting these experiences with human-to-human communication [13]. Previous research has indicated that users usually trust conversational agents more when they engage in small talk [9]. There is a gap between user expectations and the current capabilities of VPA [45].

2.4 User Reviews

Amazon Alexa's review system allows users to leave textual and numeric reviews (1 to 5 stars) to express their overall satisfaction with a skill. After using Alexa skills, users may provide feedback through reviews to report policy violations or functionality issues within the skills. Amazon makes all of this information accessible on the Internet via the skills store which functions as a repository of data regarding Amazon's deployed skills and provides access to all skill pages. It allows for the ability to collect data through the use of a web-driver

(*i.e.*, Selenium Web Driver [50]) for a large-scale user review analysis, as presented in this work. The detailed methodology for deriving these categories will be discussed in Section 4.4.2.

User reviews could serve as an effective source for Alexa skills store to identify undesired skill behaviors even after they have successfully passed the certification process. User feedback provides unique insights into potential issues, policy violations, and problematic behaviors that may not have been detected during the initial vetting process. What makes user reviews particularly valuable is that users provide feedback based on their experiences with the skill after they have enabled and used it for an extended period of time. This extended usage allows users to uncover additional insights and report any issues or violations that might not have been evident during the initial testing phase. By thoroughly analyzing user reviews, we could gain valuable feedback on the performance, functionality, and compliance of skills, which also allows VPA platform providers and skill developers to take necessary actions to improve the overall quality and safety of the skills.

3 RELATED WORK

User reviews of open app markets (*e.g.*, Android or Apple App Store) provide an opportunity to proactively collect user attitudes and complaints [25]. There is a large number of studies on mining user opinions in mobile app reviews, to identify emerging issues, improve mobile apps to meet user expectations, and understand app changes. Sorbo *et al.* [18] proposed SURF (Summarizer of User Reviews Feedback) system, which automatically generates summaries of user reviews and recommends developers to perform maintenance and evolution tasks. Vu *et al.* [57] developed a review search system that enables an analyst to search for user reviews that are most relevant to the provided keywords. Ali *et al.* [7] analyzed user reviews of cross-platform mobile apps to identify discrepancies in user complaints on two different platforms. Yu *et al.* [62] developed a tool named ReviewSolver, to localize function errors in mobile apps by mapping user reviews to problematic source code. Fu *et al.* [24] analyzed user reviews to discover inconsistencies between user comments and ratings, identify reasons why users complain about an app, and how user complaints evolve over time. Gao *et al.* [25] proposed an automated framework to identify emerging app issues effectively based on online review analysis. It takes reviews of different versions as input and generates version-sensitive topic distributions, and emerging topics are then identified based on a typical outlier detection method. In another work, the authors of [60] identified functionality-relevant user reviews and inferred their permission implications. The team was able to show that the permission usage of apps can be better reflected by user reviews than the claimed descriptions of apps. Nguyen *et al.* [48] investigated the relationship between end-user reviews and security & privacy-related changes in Android apps, where a keyword-based approach is used to identify security & privacy-related user reviews. This study reveals that user reviews are a significant predictor of security & privacy-related changes in Android apps.Vetrivel *et al.*[56] employed topic modeling and manual coding to analyze customer reviews, identifying and categorizing security and privacy concerns in IoT devices. Hark [30] utilized a deep learning system and NLP techniques to identify privacy related feedback from app reviews in Google Play. CHAMP [35] proposed using rule-based matching to classify comments into pre-defined types of undesired behaviors that violate markets policies. Similarly, Zhang *et al.* [64] presents a method called UBC-BERT that leverages a pre-trained model (BERT-BASE) fine-tuned to detect undesired behavior from user comments based on their semantics. iRogue [27] utilizes training on deep learning features extracted from manually labeled reviews to identify rogue mobile apps. Guo *et al.* [54] presented SRR-Miner, a review summarization approach focusing on security-related user feedback. On the other hand, Vitor *et al.* [17] proposed using recent Large Language Models (LLMs), specifically the OPT model, to automatically construct a risk matrix by extracting information from app reviews, such as bugs. Paulo *et al.* [19] used ChatGPT-4 to analyze a dataset of user reviews concerning mobile app accessibility. Mittal *et al.*[47] proposed leveraging hard and soft prompting techniques to detect policy violations in online

content with extractive explanations from user comments. All of these research works show that user reviews can provide a wealth of information about the real world performance of deployed software.

In VPA ecosystem, most of the existing works focus on the potential vulnerabilities and the testing of voice apps as well as detecting policy violations and privacy issues. Kumar *et al.* [40] presented the squatting attack based on the similarly pronounced names of skills. Zhang *et al.* [63] proposed the voice masquerading attack to steal users' information in the backend. Lentzsch *et al.* [42] shown that malicious users can pressure innocent users to reveal information. Cheng *et al.* [12] and Wang *et al.* [59] showed that problematic skills were easy to pass the certification system. SkillExplorer [29] found 1,141 skills collecting users' data without providing a complete privacy policy. SkillDetective [61] detected 6,079 skills with policy violations on the skills store. SkillVet [21] presented a machine-learning method to check permission and privacy issues. Verhealth [52] analyzed 813 skills in the Health category and SkillBot [41] checked 3,434 Kids' skills.

However, very few works analyzed the user reviews in the domain of VPAs. Fruchter *et al.* [23] investigated if consumers write about privacy or security related issues in the form of reviews on popular e-commerce websites (Best Buy, Amazon, Walmart and Target) after purchasing or using different smart speakers. Though security and privacy related topics account only for 2% of the total reviews in their analysis, reviewers explicitly expressed concerns about data collection and user privacy. The authors in [8] studied the user reviews of more than 2,800 Alexa skills in order to understand the characteristics of the reviews and the issues that they raise. The team was able to identify 16 types of issues. In another work, the researchers conducted an analysis of limited 55,502 Amazon Alexa reviews spanning over 2 years from May 2015 to May 2017 [26]. The research found that a significant number of reviewers personify Alexa as an assistant, a friend, or a family member. Different from these works, we analyze a large-scale user review data from the Alexa skills store, and we focus on identifying individual problematic skills in a salable manner.

To the best of our knowledge, this is the first work to comprehensive analyze user frustration and policy violation reviews on the Amazon Alexa skills store. The focus on analyzing Amazon Alexa skills based on user feedback is a relatively unexplored area, making this work original in its scope. As mentioned above, existing works in the VPA field [12, 29, 61] focused on potential vulnerabilities or the testing of voice apps in general, while our research targets the user experience on a more fine-grained level, examining user frustrations and policy violations associated with specific skills. Building on this foundation, our study extends the analysis by incorporating a simulated and systematic voice testing framework to provide more comprehensive understanding of user experiences of Alexa skills.

4 REVIEWTRACKER DESIGN

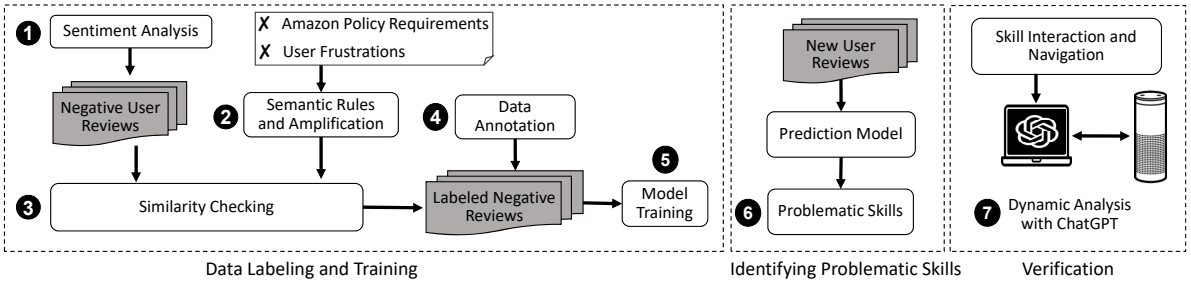


Fig. 1. ReviewTracker System Overview

4.1 Overview

Figure 1 illustrates the system overview of ReviewTracker. We employed various NLP-based data preprocessing techniques to clean and refine the negative reviews (❶). To establish semantic similarity between undesired skill

behaviors and negative reviews, we parsed policy requirements into sentences using SpaCy [34] to correlate nouns and verbs, ignoring other words in a sentence. We also employed Bi-term Topic modeling [35] to identify recurring themes in user frustrations. These methods laid the foundation for creating rules to identify user reviews potentially related to undesired skill behaviors (❷). To build the training dataset, we used Word2Vec to measure semantic similarity between the defined rules and negative reviews (❸). Then, we individually inspected and labeled reviews through the data annotation step (❹). Finally, we utilized the FastText classification model to classify the reviews (❺).

Next, we applied our trained classifier to identify potential undesired skill behaviors in new reviews (❻). We compiled a list of frequently reported problematic skills in cases of policy violations and user frustrations. To analyze whether these reported issues aligned with actual skill behaviors, we developed a dynamic analysis tool using the ChatGPT language model (❼). This tool took the identified problematic skills as input and performed querying and answering tasks within Alexa skill simulator and systematic voice testing framework, demonstrating its performance in Section 5.2.

4.2 Categorization of Policy Requirements and Undesired Behaviours

Amazon defined a set of policy requirements with 14 main sections. We grouped policies that were closely related or referred to the same concept, resulting in five distinct categories under a generic label as shown in Table 1. For example, we group together all data collection policies (*i.e.*, collect personal information, collect health information, etc.), under the label "Collect Data". It is worth mentioning that the "unsupported language" could also be one of the user frustration behaviors. Since it is explicitly mentioned in Alexa's policy that contents presented in a language not supported by Alexa are not allowed, we consider it as a policy requirement.

To better understand the spectrum of issues that users encounter with Alexa skills, this study classifies undesired skill behaviors into the following two categories: **1) *Policy Violations*:** users identified specific policy violations within the skill. **2) *User Frustrations***: users reported certain malfunctions or misbehaviors related to the functionality of a skill. By manually analyzing the most common frustrations within the reviews, we further

Table 1. Grouping of related policies from the Amazon Alexa policy requirements

Policy Categories	Alexa Policy Requirements
Promotion	Promote products, promote services, sell products, sell services, contain ads, request a positive or five star rating.
Collect Data	Collect information, predict gender.
Inappropriate Content	Provide life-saving assistance, cure all diseases, black market sale, prescription drugs, health information but lack disclaimer in description, extreme gore, decapitations, unsettling content, excessive violence, organized crime, terrorism, illegal activities, forced marriages, purchasable husbands, purchasable wives, promote hate speech, incite racial hatred, incite gender hatred, Nazi symbols, promote Ku Klux Klan, illegal downloading, pirated software, join illegal organization, illegal lifestyle, prostitution, create dangerous materials, build bomb, build meth lab, build silencer, promote terrorism, praise terrorism, recruit members for terrorist, gambling, excessive alcohol, underage alcohol, excessive profanity, mature content.
Redirect Outside	Direct to outside of Alexa, recommend skills, offer compensation for using skills, solicit donations, contact emergency responders, contact 911.
Unsupported Language	Unsupported language in Alexa content.

Table 2. Five key topics about user frustrations

User Frustrations	Description
Not Working	The users have experienced technical problems or errors while using the skill, without explicitly mentioning a specific issue.
Does not Understand	The skill was unable to comprehend the user's input or query.
Cannot Stop	The skill continues running in the background even when the user has completed an interaction with it.
Stopped Working	The users reported that the skill was previously operational but has ceased to function after a period of use.
Unresponsive	The skill is unable to establish a connection with devices or link accounts.

define five key topics about user frustrations, as illustrated in Table 2. We note that users may express general disappointment or dissatisfaction with a skill, or provide suggestions for additional features to be added to a skill. As skill dissatisfaction and suggestion reviews are not usually related to skill misbehaviors and can be subjective, we exclude these categories from our study. We focus specifically on undesired skill behaviors listed in Table 1 and Table 2, and mainly analyze reviews related to policy violation and user frustration in this work. The detailed methodology for deriving these categories will be discussed in Section 4.4.2.

4.3 Data Crawling and Research Ethics

Data Collection. Amazon Alexa's skills only provide up to 10 reviews for each page, yet several top skills receive a surge of feedback, surpassing 15,000 reviews. This limitation restricts the quality and quantity of data extracted during the crawling process, which may lead to non-accurate analyses. To address this issue, we used the Selenium WebDriver [50] to navigate and crawl through multiple pages of reviews. We achieved this by locating the "next page" button on the page and programmatically clicking it to move to the next page of reviews. We implemented measures such as distributed scraping over an extended period, adding delays in the script to minimize the load on Amazon servers and to avoid bot defense mechanisms. We obtained a total of 715,885 reviews, including positive and negative reviews under 21 categories from the US marketplace Amazon Alexa's skills store. Each review we crawled has multiple attributes, such as rating, title, date posted, and review text. Since a single Alexa skill can be listed across multiple marketplaces, our study primarily focused on the US marketplace to avoid repetition.

Research Ethics. Our study was evaluated based on the ethical principles outlined in the Menlo Report [37] for conducting computing studies. To collect data, we analyzed publicly available customer ratings and reviews from Amazon Alexa skills pages, ensuring that usernames were not collected to protect their privacy. Additionally, we implemented measures such as distributed scraping over an extended period with added delays in the script to minimize the load on Amazon servers. In Section 5, we quoted several user reviews to illustrate specific points or themes derived from the data. It is worth noting that the practice of using quotes from public data (reviews in our case) is widely used in academic research. To ensure the quoted reviews cannot be linked back to the users, we paraphrased the reviews into shorter versions that convey the idea without being identical to the main review. We collected publicly available data and avoided any discrimination in data selection, ensuring fairness and equity in our research methodology. We strictly complied with data protection and privacy regulations, contributing insights that could enhance user experience and service quality in a technological domain.

4.4 Data Labeling and Training

4.4.1 Data Preprocessing and Sentiment Analysis. We used several NLP data pre-processing techniques to enhance the quality of the reviews. These techniques include the removal of URLs, emojis, non-English words, emoticons, and dates. Additionally, we converted numbers to words, expanded chat words (*i.e.*, "omg" to "oh my god"), and extended contractions (*i.e.*, "don't" to "do not"). It is worth mentioning that removing the general stopwords cannot fit this review study since traditional stopwords (*i.e.*, can) combined with other words can become key phrases for describing user frustrations in user comments. For example, the comment *"This skill always forward me to this audible book"* is related to the user frustrations "Does not Understand"; thus the traditional stopwords "always" should not be removed. Also, the sentiment analysis on the user reviews was performed using a sentiment classification tool called the Stanford Core NLP [46]. We created a subset of reviews with high negative sentiment, employing a minimum confidence score threshold of 0.7 to ensure a substantial focus on strongly negative reviews, and refined it further by excluding reviews with a star rating greater than 2 stars. These steps allowed us to obtain a precise dataset consisting almost entirely of reviews with negative sentiments towards the skills, thereby narrowing the number of negative reviews to 238,238 out of 715,885 reviews in total.

4.4.2 Semantic Rule Creation. First, to identify user frustrations in reviews, we manually analyzed the most common frustrations within the negative reviews and empirically identified five key topics about user frustrations, as shown in Table 2. Then, we establish a set of rules for distinguishing them from other types of reviews using Bi-term Topic modeling [35], which analyze the frequency of an unordered pair of words that co-occur in the same comment. While Traditional Latent Dirichlet Allocation (LDA) [10] and Probabilistic Latent Semantic Analysis (PLSA) [33] models are able to identify word co-occurrences across documents to understand the relationships between words and group them into topics. With such short descriptions like user reviews, these algorithms might struggle to find meaningful patterns and topics because they lack the rich context that longer documents like news articles provide. Bi-term Topic modeling is a variant of LDA specifically designed to identify pairs of words (bi-terms) that often co-occur together in the text and group them into topics, which is perfect for short text. This allowed us to identify a set of topics, where each topic contained a set of words, and assign a label to each topic to express the main type of user frustration associated with that topic. It is important to note that the topics presented in Table 2 do not correspond to the co-occurring pairs, they represent general themes related to user frustrations.

Second, Amazon Alexa marketplace policies consist of predominantly short sentences and have very little uniformity; therefore, we first collected a total of 57 policy requirements and parsed these polices into sentences. To analyze each sentence, we utilized the SpaCy library [34], which is designed to extract linguistic features from natural language text to obtain the attributes for each word in the sentences. Specifically, SpaCy is able to identify the correlation between a noun and a verb, while ignoring other words in the sentence. For example, *"it collects information relating to any person's physical or mental health or condition"*. Using this example as input to SpaCy will output "collect information". The word "collect" is recognized as the root verb and "information" is the direct object. This helped to create a more accurate understanding of the text being expressed. In order to enhance the feature set of our classification model, we grouped closely related (or refer to the same) policies, ultimately consolidating the policies into five distinct categories under a generic label as shown in Table 1 (Section 2.2).

4.4.3 Similarity Analysis. To facilitate manual data annotation, we employed the Word2vec similarity model to filter the reviews that relevant to frustrations and violations topics. This model is used to measure the semantic similarity between the words associated with each topic, which were established in the previous phase, and the words in each review. Word2Vec was utilized to create word embeddings for the relevant words of identified topics in the reviews by calculating dense vector representations of topics that capture the most representative words for each topic defined in the previous step. Therefore, the relevant reviews are semantically similar to the

topics identified as policy violations or user frustrations. For instance, if a review includes phrases such as "track" and "information", like in the example review, *"This app tracks all of my three-year-old's information. If technology continues on this path, we're headed for disaster. Also, isn't it illegal in California to track someone's information like this?"*, we consider it a policy violation under data collection policy. Our method classifies each comment into its related topic. In order to increase the robustness of our similarity analysis, we set the similarity threshold 0.8 (a higher threshold value means stricter rules will be applied in the semantic similarity measurement), in which we obtained a significant number of reviews that potentially report policy violations or user frustrations and filter out any irrelevant reviews. In total, we obtained 17,276 comments that are related to the five policy violation categories as listed in Table 1, and five key topics that are related to user frustrations as listed in Table 2.

4.4.4 Data Annotation. Even with a high similarity threshold in place, false positives cannot be entirely eliminated due to the complexity of natural language. Also, training a classifier exclusively on texts identified through similarity analysis can result in the classifier inheriting the biases and limitations embedded in the initial set of texts used for the analysis. Therefore, we manually annotate data according to the labels on Table 1 and Table 2 and ensure these labels assigned to the individual reviews are accurate, relevant, and not solely influenced by the patterns inferred from the similarity analysis.

After the similarity analysis had been performed, we extract and annotate the data according to a set of rules. For instance, for user frustrations reviews, we focused on comments that explicitly mentioned a specific issue. We excluded reviews where users expressed general dissatisfaction, made suggestions, did not provide any details about the problem they encountered, or mentioned racial or sexist slurs. For instance, comments such as *"This skill sucks"* without any further explanation were not considered relevant for identifying user frustrations. Narrowing down the selection to comments that explicitly indicate a specific issue allows the model to understand better the relevant features associated with each undesired behavior.

For policy violation reviews, any comments that explicitly mention anything related to the defined policy violations are labeled as violations. Moreover, if a comment describes a complaint about the skill's functionality while referring to any defined policy violations, it is categorized as a violation. For instance, if a comment states that *"the skill does not accurately predict the gender of a child"*, it is still labeled as a policy violation because it involves a violation related to data collection. Besides, if a review is related to more than one topic, we split the review between these related topics.

Inter-rater reliability was actively maintained throughout the review process by involving three annotators who collaboratively revised and reached an agreement on the labeled reviews. In the initial phase, each annotator individually labeled 200 reviews. Subsequently, a comprehensive discussion was conducted involving all annotators to address any discrepancies in the labeled reviews. As annotators gained familiarity with the review categories, the occurrence of inconsistent labeling decreased after the second round labeling 600 reviews each. It is important to note that if two annotators had the same label, it was adopted regardless if the third annotator provided a different label. In the last round of the process, each annotator labeled 1200 reviews, replicating the review process. The overarching goal was to solidify the inter-rater reliability by subjecting a large set of reviews to the established review classification method. After this round, the annotators reached a consensus, and any discrepancies were resolved. The average pairwise inter-rater agreement among 3 annotators was $\kappa = 0.81$ (Cohen's kappa [14]), which can be considered to be almost-perfect agreement. The annotators proceeded to label the remaining data based on the established consensus. Eventually, we obtained 14,080 comments which are used as a training dataset for our classifier in the next step.

4.4.5 ReviewTracker Classifier Training. The training dataset consists of 17,276 labeled negative reviews. Among these, there are 14,080 'positive' samples. However, it's important to note that 'positive' in this context does not refer to positive reviews. Instead, these 'positive' samples represent instances of user frustration and policy violations. The remaining data consists of other 'negative' samples that do not report frustration or policy

violations. This allow the model to learn how differentiate between positive and negative instances effectively. We used a text classification model FastText [11], which uses a neural network to learn these word embeddings, and then uses these embeddings to classify text. FastText is known of its excellent performance in various NLP (Natural Language Processing) tasks and it only needs a small number of data for the pre-trained model. This enables our system to effectively analyze the complex interactions between different linguistic elements, leading to a better understanding of the nuances present in the reviews. We also evaluated multiple machine learning approaches and found that FastText achieved the highest performance in classifying policy violations and user frustrations comments, and more details about the evaluation process are discussed in Section 5.

4.5 Identifying Undesired Skill Behaviors

ReviewTracker's classifier identified 17,820 and 131,855 reviews related to policy violations and user frustrations respectively. To evaluate the effectiveness of the ReviewTracker, we further conducted a manual evaluation of a random subset of the classified labels. We randomly selected 200 reviews from the policy violation comments and another 200 reviews from the frustrations comments. Each review was evaluated by two independent human evaluators who were trained to identify policy violations or user frustrations.

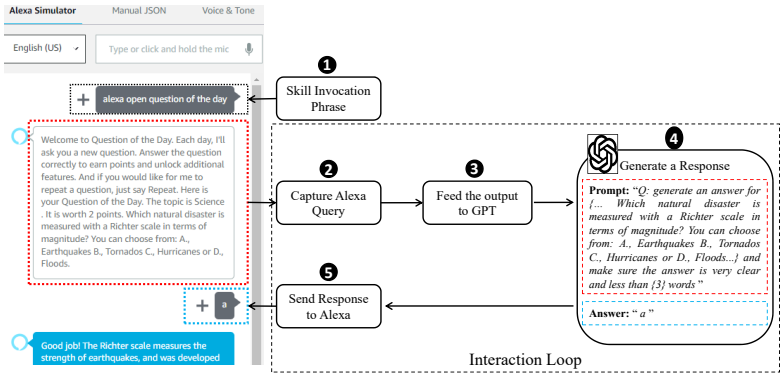
Our manual evaluation results showed that the ReviewTracker's classifier were effective in identifying reviews containing policy violations or user frustrations, we found 552 reviews out of 600 (92%) were correctly predicted. Upon conducting a thorough manual analysis of the 8.00% false positive occurrences, we found that the majority of these occurrences are associated with user frustration reviews. The distinctions between "Cannot Stop" and "Stopped Working" often pose challenges for the classifier since their linguistic elements overlap. This discrepancy can be attributed to the model's feature extraction process because there were fewer instances of "Cannot Stop" compared to "Stopped Working" in the training data. Additionally, some instances resulted from complex reviews that could be associated with more than one label. For example, a review stating, *"I keep hearing ads on my premium Spotify account, and it seems to be playing content from a different Spotify account."* was labeled as "Not Working" but predicted as "promotion."

4.6 Dynamic Analysis of Problematic Skills

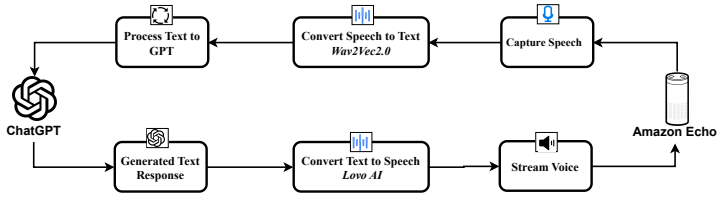
To better understand the issues encountered by users while interacting with Alexa skills, and to verify the mainly reported policy violations, it is essential to deploy a dynamic analysis tool that can automate the interaction with these skills. However, as skills become more intelligent, the hard-coded grammar-based rules in SkillExplorer [29] and SkillDetective [61] may not be scalable in handling diverse responses from skills. To address this challenge, we designed and developed a simple but effective tool based on ChatGPT (gpt-3.5-turbo). This tool is able to understand diverse questions from skills and generate corresponding responses.

4.6.1 Simulator-Based Dynamic Testing: To facilitate interaction with the skills, we utilized Alexa Skill Simulator [4], which is a tool provided by Amazon that allows developers to test and interact with their Alexa skills in a text-based environment, simulating how a user would communicate with Alexa. Figure 2(a) illustrate the continuous feedback loop between the ChatGPT model and the Alexa Skill Simulator. First, we invoke a skill with an invocation phrase from the provided list to Alexa Simulator, such as *"Alexa, open Lemonade Stand"*, as given by the developer. Then, the captured output from the Simulator is fed into our GPT model, which generates an appropriate response based on the output received from the skill. To illustrate, a skill output example is *"Pick a story theme you want to create. Tap or say Dinosaurs, Enchanted Forest, Underwater, or hear more"*. This output was then fed into our tool, which initiates a corresponding answer of *"enchanted forest"*. This answer is again fed back to the skill simulator. This iterative process (interactions) continues to go back and forth between our tool and the simulator until there is no questions or no sentences are provided, or it exceeds the iteration threshold we

have set (15 iterations). It allowed us to interact with the Alexa simulator and collect the interactions from each skill, which we can then analyze for potential policy violations or user frustrations.



((a)) Simulator-Based Dynamic Testing



((b)) Voice-Based Dynamic Testing

Fig. 2. Overview of Dynamic Testing Methods

To ensure skill compliance with established policy rules, we employed a verification process to examine each skill interaction. Policy requirements were split into sentences (as in Section 2), and these sentences were used as inputs for the "gpt-3.5-turbo" language model, allowing each sentence to be compared against every interaction. Using a prompt-based approach, we checked interactions for potential policy violations. If the model identified a violation in an interaction, it was added to a list. ChatGPT has shown strong performance in handling formal language, aligning well with Alexa skill interactions [3]. To ensure accuracy, we conducted manual verification of a random subset of policy-violating interactions. Given that different skills may have the same invocation utterances, we manually verified a small group of the identified skills. It is worth noting that, all of our testing was conducted in a new Amazon account, where all skills were initially disabled by default. Upon invoking a skill using its sample utterance (*i.e.*, "*alexa, open question of the day*"), the targeted skill ("*Question of the Day*") would be automatically enabled. To ensure accuracy, we manually verify a subset of skills to ensure they have been enabled on the Alexa skills store account, corresponding to the identified skills.

4.6.2 Voice-Based Dynamic Testing: While Alexa Simulator provide valuable insights into skill behavior, it may not be suitable to understand certain user frustrations that are related to speech recognition. For example, users interact with Alexa using diverse accents, tones, and speech patterns. As a result, these challenges posed may not be accurately reflected in the controlled setting of the simulator. To address this limitation, our approach included a dedicated phase of voice-based testing. We incorporate the Wav2Vec 2.0 (`wav2vec2-large-960h-lv60-self`) model developed by Facebook AI [16], which offers state-of-the-art accuracy in transcribing spoken language to text (Speech-to-Text (STT)) including inputs from the Amazon Echo device.

Table 3. FastText achieves the best classification performance

Model	Precision	Recall	F1	Accuracy
LR	0.890	0.891	0.909	0.928
SVM	0.900	0.901	0.900	0.904
NB	0.868	0.846	0.835	0.846
RF	0.904	0.910	0.902	0.908
RNN	0.890	0.884	0.892	0.884
LSTM	0.902	0.898	0.899	0.898
BiLSTM	0.904	0.902	0.907	0.902
FastText	**0.914**	**0.913**	**0.912**	**0.913**

In order to closely mimic natural user speech synthesis, we utilized generative voice AI platform for text-to-speech (TTS) named Lovo AI [2], known for high-quality audio generation with long-term consistency. This platform offers a diverse range of voiceovers with various accents, enabling the synthesis of speech that closely matches different users demographics and speaking styles. This will allow us to analyze different factors that may contribute to discrepancies in user experience. As depicted in Figure 2(b), the voice-based testing proceed by using STT to convert the spoken words captured from a physical Amazon Echo device (i.e., Amazon Echo 3rd Gen in our case) into text. This text is then fed to the ChatGPT model, which processes the queries and generates responses tailored to the text's request similar to simulator-based testing. After that, Lovo's TTS model converts the generated text into natural voice streamed to the Amazon Echo device. It is important to highlight that using such synthetic voiceovers is a common practice in analyzing and testing accents in speech technology that has been deployed in multiple studies [28, 58]. We also discussed this limitation in Section 6.3.

5 EVALUATION

ReviewTracker was implemented using Python 3.7 in a Conda environment, and the instantiations were run on Linux environments. Data collection and communication between ReviewTracker and the testing console (*i.e.*, skill simulator) were performed using the Selenium WebDriver. To identify undesired skill behaviours, we used ReviewTracker to make predictions on a new pool of reviews that were not included in the original training dataset. To ensure the integrity of the evaluation process, we took precautions to remove any duplicate comments between the training data and the new pool of reviews. We collected a total of 715,885 reviews over a span of six years, from January 2016 to March 2023.

5.1 FastText Model Performance

Our FastText model was trained with a learning rate of 0.5 and a word vector dimension of 800 across 300 training epochs. We then evaluated our FastText model against eight machine learning based approaches at classifying policy violations and user frustration reviews. We began with algorithms such as Logistic Regression (LR), Support Vector Machines (SVM), Naive Bayes (NB), and Random Forest (RF). Then, we explored deep learning techniques such as Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), and Bi-directional LSTM (BiLSTM), which are known for their performance in text classification. For each model, we trained it with 10-fold cross-validation. Table 3 shows the performance of different classifiers. FastText achieved the best performance with 0.914, 0.913, and 0.912 average precision, recall, and F1 score, respectively. Therefore, we used the Fasttext model in the rest of our analysis for predicting the reviews.

5.2 Performance of ChatGPT-based Dynamic Skill Analysis

To evaluate the performance of our ChatGPT based dynamic testing tool, we carried out a small-scale evaluation by comparing our tool against two state-of-the-art tools SkillDetective [61] and VITAS [43], which is an improved version of SkillExplorer [29]. Our tool uses one of the recent version of the OpenAI GPT model 'gpt-3.5-turbo'. We accessed the model through the ChatCompletion API, and performed real-time interactions, which were fully automated. It is noteworthy that we utilized prompt engineering by instructing ChatGPT to generate responses for each of skill interaction. Also, we set the model temperature to 0.2 and explicitly requested a short answer due to the limitation in Alexa's ability to understand long responses. Moreover, during the testing phase, each interaction with a skill was stored in an array and fed into ChatGPT, allowing it to understand the full context based on the previous interactions to generate accurate responses. We also instructed ChatGPT to avoid generating similar responses that have been included in the given array to explore diverse interaction branches. To illustrate, we employed few-shots prompting, *e.g.*, *"generate an answer for {Skill_response} and make sure the answer is very clear and less than {3} words"*. It is worth mentioning that, our primary focus in this step is to continue the dialogues to explore as many interactions as possible with Alexa skills, rather than prioritizing the accuracy of the responses given by the interaction agent.

Our evaluation focuses on scrutinizing each tool's ability to respond to a variety of responses derived from a single skill. To establish a reliable benchmark for comparison, we manually interacted with the targeted skills to understand the specific limitations inherent in each tool. Figure 3 demonstrates the average in-depth interactions of each tool. ChatGPT achieves a similar in-depth interaction average as the manual approach, showcasing its ability to engage in more detailed and complex interactions. For example, in "Heads Up" skill, involving a hint to guess an animal with the description *"They have a long trunk"*, ChatGPT generated the correct response, which other tools could not accomplish. Not only was it able to correctly predict most responses and questions, but it also generated random answers to personal queries like name, zipcode, location, phone number and even favorite cuisine. This feature enabled the conversations to continue smoothly, maximizing the number of interactions between the skill and ChatGPT tool. On the other hand, VITAS and SkillDetective exhibit lower interaction averages, implying potential limitations in capturing the complex responses or the contextual nuances of the skills. For instance, in the context of skill "WebMD", where both VITAS and SkillDetective failed to provide an appropriate response to the query *"I can tell you about health conditions, drugs, or even side effects. What would you like to know?"*. It is important to note that when a skill does not receive the expected response, it remains in a loop, repeatedly asking the same question without progressing to other interactions. However, ChatGPT

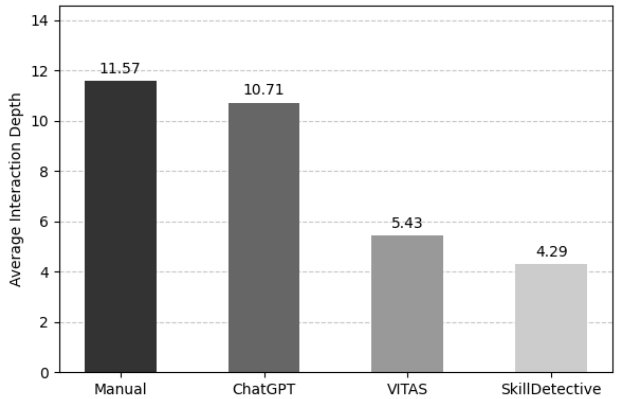


Fig. 3. Average interaction depths of different dynamic skill testing tools

experienced a few common errors when responding to certain skills that asked specific personal questions such as the user's symptoms or dietary habits.

We also evaluated the time complexity of each testing tool. We selected the "WebMD" skill and manually conducted a total of five interactions with a time complexity of 184 seconds, which served as a benchmark to compare with other testing tools. Surprisingly, ChatGPT based tool demonstrated remarkable efficiency by completing the same interactions within 125 seconds, faster than our manual testing. This can be attributed to its extensive general knowledge, which contrasts with human participants, who may require additional time to think about accurate responses. In contrast, VITAS consumed around 238 seconds for five interactions, with some inaccurate responses. Similarly, SkillDetective took 235 seconds for five interactions, yielding responses for only Yes/No questions.

Since skill testing is a time-consuming process due to the fact that every new skill requires a fresh start for testing each interaction branch, we first tested the top 1,350 skills with the most reported reviews about policy violations and user frustrations identified by ReviewTracker (results can be found in Section 5.4). Then, we conducted an additional in-depth evaluation specifically for the Kids and Health categories to ensure that these skills meet the relevant legal standards such as Health Insurance Portability and Accountability Act (HIPPA) [1] and Children's Online Privacy Protection Act (COPPA) [15]. We collected a total of 4,216 skill interactions by using this technique, and we used the recommended invocations provided by the skill developers for each skill.

5.3 User Review Analysis

In this section, we present our review analysis results of undesired skill behaviors. By applying the method described in Section 4, we identified 2,813 unique skills, with 17,820 reviews reporting potential policy violations. Also, we found 9,294 unique skills, with 131,855 reviews that report functionality and/or performance issues. Table 4 presents a summary of our detection results.

Table 4. Summary of the reviews identified by ReviewTracker

Category	Behavior	Skills (%)	Reviews (%)
User Frustration	Unresponsive	3,156 (33.31%)	60,283 (45.75%)
	Stopped Working	4,031 (42.48%)	37,008 (28.08%)
	Does not Stop	3,097 (32.59%)	17,302 (13.12%)
	Does not Understand	3,017 (31.79%)	16,283 (12.33%)
	Not Working	3,289 (34.65%)	20,093 (15.21%)
Policy Violation	Promotion	1,580 (16.65%)	9,587 (7.27%)
	Inappropriate Content	1,192 (12.56%)	5,124 (3.88%)
	Collecting Data	721 (7.59%)	2,383 (1.81%)
	Unsupported Language	221 (2.33%)	500 (0.38%)
	Redirect to Outside	104 (1.09%)	181 (0.14%)

Category	# of Skills	# of Reviews
Smart Home	1,007	63,750
Music & Audio	1,012	34,825
Games & Trivia	1,542	12,183
Lifestyle	975	6,497
Productivity	485	4,334
News	713	4,103
Novelty & Humor	438	3,766
Education & Reference	650	2,550
Kids	337	2072
Health & Fitness	295	1,953
Connected Car	46	1,763
Food & Drink	176	1,318
Social	92	1,277
Weather	80	966
Utilities	168	867
Sports	129	637
Communication	36	608
Local	105	562
Movies & TV	112	504
Business & Finance	161	490
Home Services	44	436
Travel & Transportation	95	344
Shopping	36	297
Total	**9,497**	**131,855**

Table 5. Analysis of user frustration by category

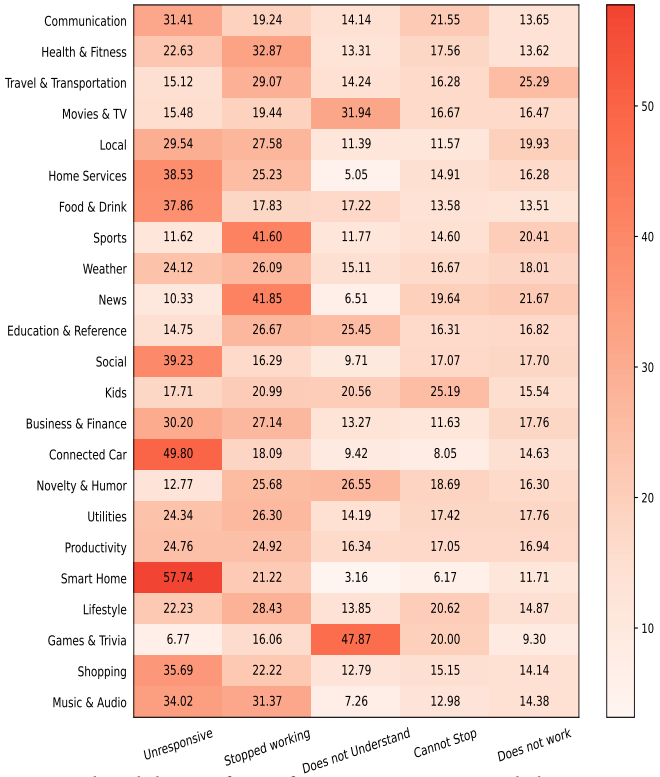


Fig. 4. A breakdown of user frustration reviews and their categories

5.3.1 User Frustrations Reviews. Table 5 shows the overall summary of the user frustrations with their categories. The distribution of user frustration reviews across categories and policy violations is shown in Figure 4. The y-axis represents the different skill categories, while the x-axis represents the specific policy violations identified by ReviewTracker. Each row of the figure adds up to 100%, with each cell representing the percentage of a specific violation in the user reviews for a particular category. To enhance the visual representation of the data, we used heatmap to indicate the percentage of each cell. For instance, deeper colors in the cell represent a larger percentage of user frustration reviews, while lighter colors indicate a lower percentage.

It is important to note that some skills belong to more than one category, which explains why the total number of skills (9,497) in Table 5 is different from the identified number of skills (9,294). We observe that the Games & Trivia category exhibits a significant percentage of "Does not Understand" issues, accounting for 47.87% of the total reviews. Since Games & Trivia skills heavily rely on voice interactions, one plausible reason can be attributed to the limitations of speech recognition technology, which faces challenges with user accents, rapid speech, and background noise, impacting comprehension and understanding. Moreover, when users use specific terms or jargon, the voice recognition system may not be familiar, as these terms may not have been included in the interaction model. For example, in one review, the user expresses frustration, stating, *"Even when ensuring that answers were well articulated, Alexa did not understand them"*.

In the Smart Home and Connected Car categories, a significant percentage of the user frustration reviews, 57.74% and 49.8% respectively, reported "unresponsiveness". This may be due to difficulties in establishing connections and

linking accounts with IoT (Internet of Things) devices. These issues can be related to connectivity, compatibility, or network configurations. In addition, the process of linking accounts within IoT skills can be complex due to varying protocols and requirements across different devices and platforms. It involves multiple steps, including authentication, authorization, and data synchronization, which can be overwhelming for some users. For instance, a user faced similar issue by stating, *"Alexa indicates that the device is unresponsive. I have followed all the recommended actions provided by Tuya support and Alexa, but they have not resolved the issue"*.

The percentages of the user frustration reviews in the IoT categories (Smart Home and Connected Car) are substantial, which indicates a notable trend where users expressed dissatisfaction with IoT skills. This observation suggest that the users have strong desire for efficient control over IoT devices. Thus, we conducted a deeper analysis of the user reviews, we found a large number of skills could not perform as well as what they claimed or what users hope they can do based on the user's feedback. In one specific review, a user shared their experience, stating, *"My Alexa, connected to my lights, responds with 'Sorry, no device found' when I try to control the lights, even though they are listed as connected"*.

It can be dangerous for IoT skills when it comes to sensitive devices such as doors and windows. The skill "Hue" has 65 reviews complaining about the same issue of turning on or off a light. For the detected skills with over 10 reviews, the average rating star is only 2.2 (out of 5 stars), which also shows the users were not satisfied with their functions. Despite the high popularity of these skills, with an average number of reviews over 1,500 compared to 65 on average for all IoT skills, there is a clear discrepancy between their popularity and functional performance.

Category	# of Skills	# of Reviews
Smart Home	761	4,067
News	525	3,762
Music & Audio	441	3,330
Lifestyle	371	1,064
Games & Trivia	321	1,098
Education & Reference	192	731
Productivity	179	664
Novelty & Humor	146	548
Health & Fitness	117	333
Kids	121	500
Business & Finance	68	145
Connected Car	58	243
Sports	49	113
Food & Drink	53	144
Communication	25	114
Utilities	40	77
Social	38	141
Weather	38	121
Travel & Transportation	20	53
Local	20	71
Shopping	19	46
Home Services	23	78
Movies & TV	16	48
Total	**3,940**	**17,820**

Table 6. Analysis of policy violations from reviews by category

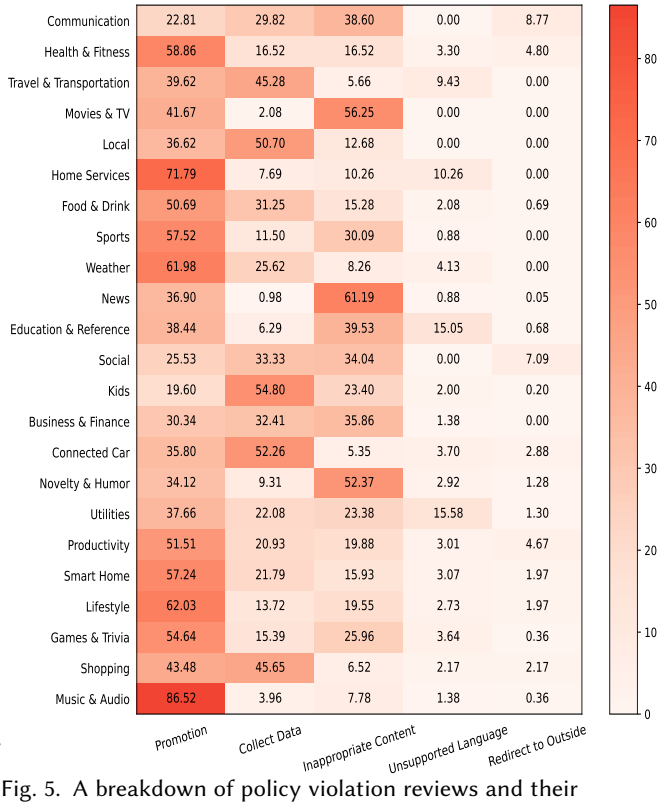


Fig. 5. A breakdown of policy violation reviews and their categories

5.3.2 Policy Violation Reviews. Table 6 provides a summary of the number of skills and policy violation comments. Within the Smart Home category, there are (761) skills potentially violating policies based on (4,067) reviews, and the Music & Audio category stands out with a substantial number of skills (441) and reviews (3,330), which highlights the frequency of policy-violating behaviors within these particular domains.

From Figure 5 we observe that the majority of policy violation reviews (86.52%) reported in the Audio & Music category were related to promotions. One explanation is that if an advertisements (or subscription requests, or review requests) interrupts a user's listening experience, it may be perceived as more intrusive and irritating than an advertisement in a different category. For instance, a review says *"Constantly nagging for star reviews, it woke the baby while trying to sleep"*. Hence, most users would report about such an annoying issue in the review section.

It is worth noting that, from Figure 5, we notice that most of the reviews (61.19%) reported as inappropriate content in the News category. As we further analyzed these reviews, we found that News skills may contain topics that are subjective to users. Some users may write a review to express their opinions on specific news headlines. As an example, one review states, *"Spreading lies to push their left-wing agenda? Nah, that's just fake, fake, fake."*. Due to the subjective nature of news content, we excluded the News category from our validation process in Section 5.4. We also analyzed the skill developers and their connection to policy-violating skills, which we discussed in detail in Appendix A.

Table 7. Examples of skills policy violation reviews in the Health and Kids categories

Category	Policy Violation	Skill Name	Review Example
Health	Collecting Data	Headspace	*"I connected my Amazon account, but it kept insisting I create a Headspace account. Then, I thought, "Fine, I'll use my Facebook." But guess what? Same story. Seriously, how much of my personal stuff do you really need? I shouldn't have to deal with another login, especially after handing over my Amazon and Facebook details."*
	Promotion	myNursebot	*"Ever since I started using this skill, I've been bombarded with spam about medical stuff that I only talked about with this skill. It's a clear violation of privacy rules like HIPAA. Watch out, they're probably selling off our info."*
	Inappropriate Content	Allergy Forecast	*"It keeps insisting allergens are low, but three other sites are saying the high. I've had it—I'm turning off this skill. It's way off the mark."*
Kids	Collecting Data	NORAD Tracks Santa	*"Why do they need all these information about kids? It's just crazy that Alexa wants all that just to track Santa. It's so silly.*
	Promotion	Cool Koala	*"Although my daughter enjoys this app, I've been struggling to buy the subscription. Even after reaching out to Amazon customer service twice, all they suggested was to leave a review seeking help"*
	Inappropriate Content	Sonnar Interactive LTD	*"Red Riding Hood is a story for children, bbut there's an ending that could upset them, with the Big Bad Wolf saying, "Little girls and the Big Bad Wolf killed Red Riding Hood." No offense, but I really think this shouldn't be allowed on the Amazon app. If there are any kids under five reading this review, I suggest not playing the interactive story."*
	Unsupported Language	Santa-Holiday Personality	*"I tried to enable it, but every time I click the button, it just takes me to a page in Spanish, and I don't speak Spanish."*

5.4 Verification with Dynamic Testing

Out of the 2,813 skills that were identified as suspected policy violations, 238 were in the health and kids category, and the remaining 2,576 were in other categories.

5.4.1 Skills in Kids and Health Categories. We conducted many detailed analysis of the policy-violating skills, specifically focusing on those that fall under the Health and Kids categories. Table 7 presents some examples of these violations with representative samples from the user reviews. We detected the violations within the collected interactions. Thus, we identified 38 policy-violating skills in the kids category out of 121, and Table 8 lists the overview of these violations.

Table 8. Detailed breakdown of 38 policy-violating skills in the kids category, "-" means no reviews found complaining

Policy Violation	# of Skills	Skill Response Example	Review Example
Collect data	15	In order to better communicate with you, I would like to know your first name. What is your name?	"Don't believe that information gathering on children is ok."
Redirect to outside	10	If you like this skill, try our new skill, too. Say, "Alexa, open package hero."	-
Inappropriate Content	6	Then in his fury he seized his left foot with both hands and tore himself in two.	"even though it would frighten my grandchildren"
Promotion	7	I hope you had fun playing Lemonade Stand. Please give us five stars.	-

Table 9. Detailed breakdown of 47 policy-violating skills in the Health category

Policy Violation	# of Skills	Skill Response Example	Review Example
Collect data	23	I can assess and help improve your memory with fun games and tests. To get started, tell me your name	"The first thing it asks is to say your name."
Redirect to outside	14	Joining our skill-users only Facebook group. Simply search for "My Morning Meditation Practitioners."	-
Inappropriate Content	10	(Recommends taking Valium) Take the tablet by mouth with a glass of water with or without food, as prescribed.	"Just another way Big Pharma wants to manipulate you to take more pills."

Since data collection is strictly prohibited for skills in the Kids category, we first checked whether any skills collect any information. As a result, we discovered fifteen skills were indeed collecting personal information, which is a clear policy violation. For example, the "Potty Training" skill requested the child's name when it is invoked during its first reply. It is concerning that these skills were approved despite the strict skill vetting process in place. Besides, we identified six skills that contained inappropriate content for kids. For example, the skill "Fairy Tale," which includes the sentence *"Then in his fury he seized his left foot with both hands and tore himself in two"*, which could be disturbing for young kids as flagged by ChatGPT.

In the Health category, we detected 47 policy-violating skills out of 117 skills and the detailed summary of the detection results is presented in Table 9. 23 skills were found to collect data from users such as a "Brain Workout" skill, which asks for body temperature *"Have you had a fever higher than 100 point four degrees in the past 24 hours?"*. This could be considered a violation since a skill is not allowed to collect information about person's physical health or condition. Additionally, we identified 10 skills that included potential inappropriate content by providing inaccurate information. An example of this is the skill named "Diazepam", which recommends users to *"Take the tablet (Valium) by mouth with a glass of water with or without food, as prescribed"* if they have anxiety issues. Furthermore, we identified 14 skills that redirected users to external websites, such as inviting them to join a Facebook group.

Table 10. Detailed breakdown of 143 policy-violating skills in other categories

Policy Violation	# of Skills	Skill Response Example	Review Example
Collect data	60	You must first tell me your weight.	"It keeps asking me for my weight"
Redirect to outside	33	If you like this skill, try our new skill, too. Say, "Alexa, open package hero."	-
Inappropriate Content	23	You have been poisoned and will eventually be the second pile of bones in the room.	"worked fine with commands, but it's a little creepy."
Promotion	27	"To use high quality rechargeable Lithion ion batteries for more info, please visit our website at www.echobattery.com."	"Wanted to check my battery level, instead I get ads."

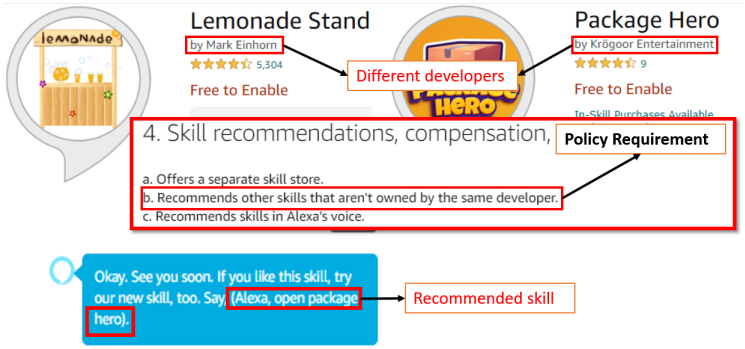


Fig. 6. Recommendation of a skill developed by a different developer

5.4.2 Skills in General Categories. Overall, our analysis identified 143 skills out of 1,112 that may violate at least one policy (Table 10). Among these skills, 60 collected personal information, like the 'planetary weight' skill, which asked, *"You must first tell me your weight"*. We also found 33 skills that redirected users to external websites or recommended other skills, e.g., 'Lemonade Stand,' which violated the policy prohibiting suggestions of skills from different developers. An example is shown in Figure 6. Additionally, 27 skills contained promotional content, like 'Game of Words', which solicited positive ratings, stating *"A simple five-star review to help us keep bringing you great games like this to Alexa"*. Furthermore, we identified 23 skills with inappropriate content, e.g., 'castle adventure,' stating, *"You have been poisoned and will eventually be the second pile of bones in the room.*

In order to identify the possibility of fake reviews, we investigated changes in the number of daily reviews on the skills store. We observed abnormal review trends over time, potentially due to delays in the review vetting process or the use of automated systems to generate reviews, as detailed in Appendix B. However, these represent a small fraction of the total skills analyzed as noted in Figure 9. This indicates that while the potential for misuse of reviews exists, it has a limited impact on the overall findings and conclusions of our study.

5.5 Verification with Voice Based Dynamic Testing

During voice testing, we concentrated "Does not Understand" complains, as these are crucial indicators of the speech recognition capabilities in Alexa. As depicted in Figure 4, the majority of these complaints were reported from the Games & Trivia category. We conducted this test on the top 40 most reported skills, which revealed that

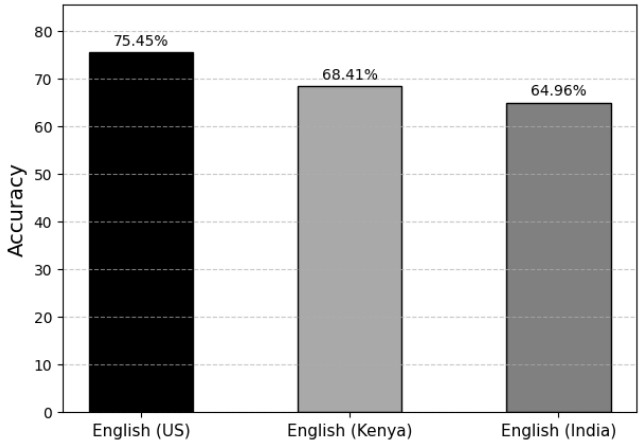


Fig. 7. Comparative Analysis of Voice Recognition Accuracy Across Different English Accents

75.14% of interactions were accurately understood. Nonetheless, a considerable portion of the interactions 14.59% were misunderstood, even though the inputs were correct. Additionally, 5.41% of skills stopped working after the first two correct responses, and 4.86% failed to function entirely.

In order to understand other factors influencing the performance of Alexa's speech recognition, we extended our analysis to evaluate the impact of different accents: American accent, Kenyan accent English, and Indian accent. To generate these voice samples with high fidelity and natural sounding accents, we utilized a text-to-speech synthesis Lovo API[2], which have pre-trained voices with different accents. The experiment setup was designed to minimize environmental variables, such as background noise and microphone quality to ensure fair comparison. Our results, as depicted in Figure 7, show that Alexa's speech recognition rates varied across different English accents. The highest accuracy was noted for American English at 75.45%, followed by Kenya-English at 68.41%, and Indian-English at 64.96%. The lower accuracy rates for Kenyan and Indian English could be attributed to the diversity and complexity of these accents, which might not be as well-represented in the training data. As a result, users from different accents especially foreign accents, may experience increased frustrations due to frequent misunderstandings and errors in voice recognition. This align with previous studies that have highlighted that speech from non-American speakers exhibits about 30% more inaccuracies compared to American English accents [31]. Incorporating a wider range of phonetic patterns and training models on diverse speech datasets can improve recognition rates across accents.

We also analyzed the top reported skills with "Does not Stop" complaints to further verify and understand the motivations behind these issues. We started by interacting with each skill three times to establish a pattern of response behavior. Then, we initiated different stop commands (Stop, Cancel, Exit) separately to accurately assess the skills' ability to cease operations upon command. The analysis revealed that a considerable number of skills, 46.34% for "Stop" and 41.46% for "Cancel" commands as shown in Table 11, did not immediately stop functioning after the user commanded them to stop. Instead, these skills often continued to interact with the user by recommending another skill, asking for a review, or advertising additional services, ..., etc. However, from final response of Table 11, it becomes evident that the majority of these skills did indeed cease operations after one or two additional interactions. This indicates there is a gap between the programmed skill responses and the user's expectation that the skill should cease all activity upon receiving a stop command, leading to potential user frustration and confusion.

Table 11. Analysis of Skill Responses to Stop Commands

Continuing to Respond		Final Response	
Stop Command	# of Skills (%)	Refused Stop (%)	Initiated to Stop (%)
Stop	46.34%	14.29%	85.71%
Cancel	41.46%	10.71%	89.29%
Exit	12.20%	14.29%	85.71%

6 DISCUSSION

6.1 Implications

This work brings new insight into the voice assistant ecosystem from the user's perspective. User reviews reflect the actual experience when users interact with skills, while most may not know much about the details of skills. User reviews can provide feedback to developers about the cases that developers might ignore. For example, developers might overlook other skills with the same skill name so that another skill might be invoked instead of his/her skill. Also, for certain types of functional errors, it can be hard to reproduce, and users can provide such details so that developers can solve them soon. On the other hand, VPA platforms shouldn't overlook user reviews and they should take actions to prevent low-quality skills, especially those that have policy violations or abnormal behaviors, from being published on the skills store.

It should be emphasized that ChatGPT annotation showed promising performance in handling clear formal language [3, 44]. Since Alexa skill interactions predominantly use formal language, employing ChatGPT aligns with its strengths. However, the current version of ChatGPT faces challenges to understand slang and informal language in user reviews [39], making it impractical to be used in our main method to classify 10 closely related classes (user frustrations and policy violations).

Understanding user frustrations is crucial for addressing recent concerns about fairness/bias issues in VPA systems. For example, when a certain group of users report frustrations related to "Does Not Understand", it could suggest potential biases favoring certain demographics. Recent studies by Koenecke *et al.* [38] and Walker *et al.* [58] reveal that Alexa exhibits varying speech recognition error rates among different racial and gender groups. For instance, they found that the error rates for African American users were higher compared to White users. This highlights potential vulnerabilities in speech recognition systems that may disproportionately affect certain demographics. In the future, we plan to further study user frustrations within certain social groups and utilize user reviews as a tool to recognize instances of fairness issues in virtual personal assistants.

While higher-rated reviews above 2-stars may contain some constructive feedback, our primary focus is on reviews that explicitly express undesired skill behaviors, often found in lower star ratings. This emphasis allows us to effectively filter and capture reviews most relevant to these behaviors in order to build a high-quality dataset that is crucial for training our classifier. Our approach does not overlook the potential value in reviews rated above 2 stars, which might also hold valid criticisms. However, our trained classifier also has the capability of extending its analysis to classify reviews that are above 2-stars.

When users unknowingly contribute their data through user reviews, which is then utilized to develop and refine AI systems, significant concerns arise regarding the transparency and ethical validity of such practices as highlighted in [36, 53]. In scenarios where AI systems can predict user behavior and preferences based on these reviews, there is a risk that such capabilities will be exploited to manipulate user choices or obfuscate the genuine intent behind data collection of reviews. However, our approach is designed to identify undesirable skill behaviors, while maintaining clear ethical standards and transparency, ensuring that such issues do not arise in our work.

6.2 How Stakeholders can Benefit from ReviewTracker ?

This paper offers a structured approach to understanding user reviews in the context of policy regulations and overall user frustrations. ReviewTracker benefits VPA users, third-party developers, VPA platform providers (Amazon) and potentially policymakers.

- **Amazon or policymakers:** Amazon can benefit from ReviewTracker's ability to sift through vast amounts of user reviews, identifying specific concerns and policy violations. This capability allows Amazon to manage the quality and security of the vast array of skills on its platform, thereby enhancing user trust in the Alexa ecosystem. It also has the potential to aid auditors/policymakers (e.g., US Federal Trade Commission) to obtain a quick understanding about the policy compliance practices and service quality of skills in VPA platforms.
- **User:** Users will be able to identify hidden issues and policy violations that may not be easily discernible from individual reviews. For example, if a user wants to know whether a particular Alexa skill is prone to collecting personal information - a concern that may not be frequently mentioned in reviews. Similarly, parents concerned about the appropriateness of content for their children can use ReviewTracker to quickly understand if any skills present inappropriate content, without having to manually read through possibly hundreds of reviews.
- **Developers:** ReviewTracker serves as a tool to monitor user sentiment and compliance. It helps in comprehending user expectations and improving skill functionality and user interaction, especially in cases where the skill is operational, but reviews express concerns. This is valuable as manually sifting through potentially thousands of user reviews to pinpoint specific issues or policy violations is impractical and time-consuming.

6.3 Limitation

Our work uses a semantic rule-based method so that only pre-defined undesired behaviors can be identified, which makes it hard to detect new types of issues automatically. Nevertheless, the vast and varied nature of user feedback, without specific guiding rules, would pose a challenge in pinpointing specific undesired behaviors, potentially reducing our analysis conclusions and lead to the neglect of critical concerns. The semantic rules employed by ReviewTracker were developed to capture a wide range of undesired skill behaviors identified at the time of our research. However, these rules are not static. They can be expanded and adapted to include new undesired behaviors as they emerge, where user expectations and skill functionalities continually change. One can accumulate more user reviews and observe new patterns of undesired behaviors, and iteratively update the rule set. This process involves analyzing emerging trends in user feedback, which can be integrated into ReviewTracker's existing framework.

For undesired behaviors, such as "unresponsiveness", "stopped working" and "does not understand", we didn't profoundly check the real reasons for these function errors since normal users couldn't provide such detailed information in the reviews. The same behavior might be caused by different reasons, such as voice recognition provided by the platform, program issues from developers, or misuse of users. As our future work, we plan to conduct a comprehensive dynamic analysis of skills to understand factors that cause users frustrations. In addition, the scope of our analysis was limited to the Amazon Alexa Skill Store,which dominates the U.S. smart speaker market with 64% of Americans owning an Amazon Echo as of 2023 [22]. In the future, we would like to expand our research to include other marketplaces such as Google Assistant. This decision stems from the fact that Google Assistant presents a unique set of challenges compared to Amazon Alexa. Firstly, the number of available actions on Google Assistant is fewer compared to the plethora of skills on Alexa, which limits the breadth of comparative analysis. Secondly, a significant portion of Google actions lacks user reviews, which are

crucial for our study. Finally, Google has discontinued many of these actions last summer, rendering direct testing and evaluation not possible at present.

Also, certain skills might fall under exception cases where specific actions are allowed that are prohibited otherwise. According to the policy, streaming music, radio, and podcast skills are allowed to include audio advertisements as long as certain conditions are met. While our classifier may not be perfectly equipped to handle skills that fall under specific exception cases in policies, such as those outlined above, we are actively working to find effective solution in the future. Also, the scope of this study did not include examining reviews that might not be directly related to the skill functionality or description. These reviews will be considered in our future work as well.

Although our main objective is to demonstrate that user reviews could serve as a valuable means to identify undesired skill behaviors, we acknowledge the possibility of outdated reviews, which motivated us to conduct a small-scale dynamic testing using LLM-based tool, to determine if the identified issues still exist or have been resolved in the current state of the skills. In the future, we aim to conduct a comprehensive study on a larger scale, involving extensive evaluations and analyses. Although we manually ensured that the generated AI voices are not robotic or unnatural, they may have limited variability in accents, dialects, and speech patterns, which can make them less effective for rich dynamic interactions.

The interactions generated by ChatGPT can expand the diversity of responses with Alexa, especially in scenarios where public knowledge is needed to confirm the functionality of the skill. Nevertheless, this synthetic interactions of ChatGPT are based on pre-trained models that may not comprehensively represent the diversity of real user interactions and may have embedded biases due to their training on specific datasets. For example, ChatGPT is prone to answer with ideological biases and left-leaning viewpoints [49], and struggles with answering questions faithfully across various domains due to limitations in its comprehension [65]. To mitigate these biases, strategies like prompt engineering can be employed, where prompts are carefully designed to neutralize bias and enhance the model's ability to generate balanced and accurate responses [20].

7 CONCLUSION

In this work, we developed ReviewTracker, an NLP pipeline to detect various undesirable skill behaviors using user feedback. After analyzing a large-scale dataset of user reviews, we gained insights into user concerns about skill usability and security. We observed a recurring issue where users expressed dissatisfaction with connecting to IoT devices. To validate skill behavior, we created an LLM-based dynamic testing tool using GPT model, confirming policy violations in over 228 skills out of 1350. Our research demonstrated that user reviews can be used to effectively ascertain the current state of user frustrations as well as many privacy concerns on the Amazon Alexa skills store.

ACKNOWLEDGMENTS

The work of L. Cheng is supported by National Science Foundation (NSF) under the Grant No. 2239605, 2228616 and 2114920. The work of H. Hu is supported by NSF under the Grant No. 2228617, 2120369, 2129164, and 2114982. The work of H. Cai is supported by Open Technology Fund B00236-1220-00. The work of X. Luo is supported by HKPolyU Grant No. ZVG0.

REFERENCES

[1] Accountability Act. 1996. Health insurance portability and accountability act of 1996. *Public law* 104 (1996), 191.

[2] LOVO AI. Year. *Getting Started with Genny API.* https://lovo.ai/post/getting-started-with-genny-api [Accessed: 23-Jan-2024].

[3] Mohammed Aldeen, Joshua Luo, Ashley Lian, Venus Zheng, Allen Hong, Preethika Yetukuri, and Long Cheng. 2023. ChatGPT vs. Human Annotators: A Comprehensive Analysis of ChatGPT for Text Annotation. In *2023 International Conference on Machine Learning and Applications (ICMLA).* IEEE, 602–609.

[4] Alexa Simulator. [n. d.]. https://developer.amazon.com/en-US/docs/alexa/devconsole/alexa-simulator.html. [Accessed: 30-Jan-2024].

[5] Alexa Skills Privacy Requirements. [n. d.]. https://developer.amazon.com/fr-FR/docs/alexa/custom-skills/policy-requirements-for-an-alexa-skill.html. [Accessed: 22-May-2023].

[6] Alexa Skills Security Requirements. [n. d.]. https://developer.amazon.com/fr-FR/docs/alexa/custom-skills/security-testing-for-an-alexa-skill.html. [Accessed: 22-May-2023].

[7] M. Ali, M. E. Joorabchi, and A. Mesbah. 2017. Same App, Different App Stores: A Comparative Study. In *2017 IEEE/ACM 4th International Conference on Mobile Software Engineering and Systems (MOBILESoft)*. 79–90.

[8] Soodeh Atefi, Andrew Truelove, Matheus Rheinschmitt, Eduardo Santana de Almeida, Iftekhar Ahmed, and Amin Alipour. 2020. Examining user reviews of conversational systems: a case study of Alexa skills. *CoRR* abs/2003.00919 (2020). arXiv:2003.00919 https://arxiv.org/abs/2003.00919

[9] Timothy Bickmore and Justine Cassell. 2001. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 396–403.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[11] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics* 5 (2017), 135–146.

[12] Long Cheng, Christin Wilson, Song Liao, Jeffrey Young, Daniel Dong, and Hongxin Hu. 2020. Dangerous Skills Got Certified: Measuring the Trustworthiness of Skill Certification in Voice Personal Assistant Platforms. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

[13] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[14] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[15] Federal Trade Commission et al. 1998. Children's online privacy protection act of 1998 (COPPA).

[16] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979* (2020).

[17] Vitor Mesaque Alves de Lima, Jacson Rodrigues Barbosa, and Ricardo Marcondes Marcacini. 2023. Learning Risk Factors from App Reviews: A Large Language Model Approach for Risk Matrix Construction. (2023).

[18] Andrea Di Sorbo, Sebastiano Panichella, Carol V. Alexandru, Junji Shimagaki, Corrado A. Visaggio, Gerardo Canfora, and Harald C. Gall. 2016. What Would Users Change in My App? Summarizing App Reviews for Recommending Software Changes. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*. 499–510.

[19] Paulo Sérgio Henrique Dos Santos, Alberto Dumont Alves Oliveira, Thais Bonjorni Nobre De Jesus, Wajdi Aljedaani, and Marcelo Medeiros Eler. 2023. Evolution may come with a price: analyzing user reviews to understand the impact of updates on mobile apps accessibility. In *Proceedings of the XXII Brazilian Symposium on Human Factors in Computing Systems*. 1–11.

[20] Satyam Dwivedi, Sanjukta Ghosh, and Shivam Dwivedi. 2023. Breaking the Bias: Gender Fairness in LLMs Using Prompt Engineering and In-Context Learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities* 15, 4 (2023).

[21] Jide Edu, Xavi Ferrer Aran, Jose Such, and Guillermo Suarez-Tangil. 2021. SkillVet: Automated Traceability Analysis of Amazon Alexa Skills. *IEEE Transactions on Dependable and Secure Computing* (2021).

[22] Anna Fleck. 2023. *Alexa, What's America's Favorite Smart Speaker?* Statista. https://www.statista.com/chart/23943/share-of-us-adults-who-own-smart-speakers/ [Accessed: 01-May-2024].

[23] Nathaniel Fruchter and Ilaria Liccardi. 2018. Consumer Attitudes Towards Privacy and Security in Home Assistants. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*.

[24] Bin Fu, Jialiu Lin, Lei Li, Christos Faloutsos, Jason I. Hong, and Norman M. Sadeh. 2013. Why people hate your app: making sense of user feedback in a mobile app store. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*. 1276–1284.

[25] C. Gao, J. Zeng, M. R. Lyu, and I. King. 2018. Online App Review Analysis for Identifying Emerging Issues. In *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*. 48–58.

[26] Yang Gao, Zhengyu Pan, Honghao Wang, and Guanling Chen. 2018. Alexa, My Love: Analyzing Reviews of Amazon Echo. In *2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. 372–380. https://doi.org/10.1109/SmartWorld.2018.00094

[27] Vaibhav Garg, Hui Guo, Nirav Ajmeri, Saikath Bhattacharya, and Munindar P Singh. 2023. irogue: Identifying rogue behavior from app reviews. *arXiv preprint arXiv:2303.10795* (2023).

[28] Diego Guffanti, Danilo Martínez, José Paladines, and Andrea Sarmiento. 2018. Continuous speech recognition and identification of the speaker system. In *Proceedings of the International Conference on Information Technology & Systems (ICITS 2018)*. Springer, 767–776.

[29] Zhixiu Guo, Zijin Lin, Pan Li, and Kai Chen. 2020. SkillExplorer: Understanding the Behavior of Skills in Large Scale. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 2649–2666.
[30] Hamza Harkous, Sai Teja Peddinti, Rishabh Khandelwal, Animesh Srivastava, and Nina Taft. 2022. Hark: A deep learning system for navigating privacy feedback at scale. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2469–2486.
[31] Drew Harwell. 2018. Amazon's Alexa and Google Home show accent bias, with Chinese and Spanish hardest to understand. (2018). https://www.scmp.com/magazines/post-magazine/long-reads/article/2156455/amazons-alexa-and-google-home-show-accent-bias
[32] Darren Hayes, Francesco Cappa, and Nhien An Le-Khac. 2020. An effective approach to mobile device management: Security and privacy issues associated with mobile applications. *Digital Business* 1, 1 (2020), 100001. https://doi.org/10.1016/j.digbus.2020.100001
[33] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 50–57.
[34] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.
[35] Yangyu Hu, Haoyu Wang, Tiantong Ji, Xusheng Xiao, Xiapu Luo, Peng Gao, and Yao Guo. 2021. Champ: Characterizing undesired app behaviors from user comments based on market policies. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 933–945.
[36] Rajesh Kumar Jaiswal, Shyam Sundar Sharma, and Rajkumar Kaushik. 2023. ETHICS IN AI AND MACHINE LEARNING. *Journal of Nonlinear Analysis and Optimization* (2023). https://api.semanticscholar.org/CorpusID:266320296
[37] Erin Kenneally and David Dittrich. 2012. The menlo report: Ethical principles guiding information and communication technology research. *Available at SSRN 2445102* (2012).
[38] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689.
[39] Bartłomiej Koptyra, Anh Ngo, Łukasz Radliński, and Jan Kocoń. 2023. CLARIN-Emo: Training Emotion Recognition Models Using Human Annotation and ChatGPT. In *International Conference on Computational Science*. Springer, 365–379.
[40] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. 2018. Skill Squatting Attacks on Amazon Alexa. In *27th USENIX Security Symposium (USENIX Security)*. 33–47.
[41] Tu Le, Danny Yuxing Huang, Noah Apthorpe, and Yuan Tian. 2022. Skillbot: Identifying risky content for children in alexa skills. *ACM Transactions on Internet Technology (TOIT)* 22, 3 (2022), 1–31.
[42] Christopher Lentzsch, Sheel Jayesh Shah, Benjamin Andow, Martin Degeling, Anupam Das, and William Enck. 2021. Hey Alexa, is this skill safe?: Taking a closer look at the Alexa skill ecosystem. *Network and Distributed Systems Security (NDSS) Symposium2021* (2021).
[43] Suwan Li, Lei Bu, Guangdong Bai, Zhixiu Guo, Kai Chen, and Hanlin Wei. 2022. VITAS: Guided Model-based VUI Testing of VPA Apps. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*. 1–12.
[44] Song Liao, Mohammed Aldeen, Jingwen Yan, Long Cheng, Xiapu Luo, Haipeng Cai, and Hongxin Hu. 2024. Understanding GDPR Non-Compliance in Privacy Policies of Alexa Skills in European Marketplaces. In *Proceedings of the ACM on Web Conference 2024*. 1081–1091.
[45] Ewa Luger and Abigail Sellen. 2016. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5286–5297.
[46] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.
[47] Sid Mittal, Vineet Gupta, Frederick Liu, and Mukund Sundararajan. 2023. Using Foundation Models to Detect Policy Violations with Minimal Supervision. *arXiv preprint arXiv:2306.06234* (2023).
[48] D. C. Nguyen, E. Derr, M. Backes, and S. Bugiel. 2019. Short Text, Large Effect: Measuring the Impact of User Reviews on Android App Security Privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. 555–569.
[49] David Rozado. 2023. The political biases of chatgpt. *Social Sciences* 12, 3 (2023), 148.
[50] Selenium WebDriver. [n. d.]. https://www.selenium.dev.
[51] Faysal Hossain Shezan, Hang Hu, Gang Wang, and Yuan Tian. 2020. VerHealth: Vetting Medical Voice Applications through Policy Enforcement. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* (2020).
[52] Faysal Hossain Shezan, Hang Hu, Gang Wang, and Yuan Tian. 2020. VerHealth: Vetting Medical Voice Applications through Policy Enforcement. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* (2020).
[53] Lisa Singh, Agoritsa Polyzou, Yanchen Wang, Jason Farr, and Carole Roan Gresenz. 2020. Social Media Data-Our Ethical Conundrum. *A Quarterly bulletin of the IEEE Computer Society Technical Committee on Database Engineering* (2020).
[54] Chuanqi Tao, Hongjing Guo, and Zhiqiu Huang. 2020. Identifying security issues for mobile applications based on user review summarization. *Information and Software Technology* 122 (2020), 106290.
[55] Venture Beat. [n. d.]. https://venturebeat.com/2019/09/25/the-alexa-skills-store-now-has-more-than-100000-voice-apps/.

[56] Swaathi Vetrivel, Veerle Van Harten, Carlos H Gañán, Michel Van Eeten, and Simon Parkin. 2023. Examining consumer reviews to understand security and privacy issues in the market of smart home devices. In *32nd USENIX Security Symposium (USENIX Security 23)*. 1523–1540.
[57] Phong Minh Vu, Tam The Nguyen, Hung Viet Pham, and Tung Thanh Nguyen. 2015. Mining User Opinions in Mobile App Reviews: A Keyword-Based Approach. In *Proceedings of the 30th IEEE/ACM International Conference on Automated Software Engineering*. 749–459.
[58] Payton Walker, Nathan McClaran, Zihao Zheng, Nitesh Saxena, and Guofei Gu. 2022. BiasHacker: Voice Command Disruption by Exploiting Speaker Biases in Automatic Speech Recognition. In *Proceedings of the 15th ACM Conference on Security and Privacy in Wireless and Mobile Networks*. 119–124.
[59] Dawei Wang, Kai Chen, and Wei Wang. 2021. Demystifying the Vetting Process of Voice-Controlled Skills on Markets. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3 (2021).
[60] Run Wang, Zhibo Wang, Benxiao Tang, Lei Zhao, and Lina Wang. 2020. SmartPI: Understanding Permission Implications of Android Apps from User Reviews. *IEEE Transactions on Mobile Computing* 19, 12 (2020), 2933–2945. https://doi.org/10.1109/TMC.2019.2934441
[61] Jeffrey Young, Song Liao, Long Cheng, Hongxin Hu, and Huixing Deng. 2021. SkillDetective: Automated Policy-Violation Detection of Voice Assistant Applications in the Wild.
[62] L. Yu, J. Chen, H. Zhou, X. Luo, and K. Liu. 2018. Localizing Function Errors in Mobile Apps with User Reviews. In *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. 418–429.
[63] Nan Zhang, Xianghang Mi, Xuan Feng, XiaoFeng Wang, Yuan Tian, and Feng Qian. 2019. Understanding and Mitigating the Security Risks of Voice-Controlled Third-Party Skills on Amazon Alexa and Google Home. In *IEEE Symposium on Security and Privacy (SP)*.
[64] Wenyu Zhang, Xiaojuan Wang, Shanyan Lai, Chunyang Ye, and Hui Zhou. 2022. Fine-Tuning Pre-Trained Model to Extract Undesired Behaviors from App Reviews. In *2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS)*. IEEE, 1125–1134.
[65] Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in answering questions faithfully? *arXiv preprint arXiv:2304.10513* (2023).

A SKILL DEVELOPER ANALYSIS

We examined the skill developers and their association with policy-violating skills in ReviewTracker. As depicted in Fig. 8, we present the top 15 developers who have the highest number of reported policy-violating skills. Voice Apps, Sleep Jar, Patch.com, Volley Inc., and Amazon stand out with a substantial number of skills with more than 20 policy-violating skills in comparison to other developers. It is worth noting that developers with a large number of reported skills, such as Amazon, naturally attract a higher volume of reviews. Consequently, a portion of these reviews may include reports of policy violations. To gain a comprehensive understanding of the compliance level of each developer, it is important to consider not only the number of reported skills but also the nature of the violations themselves.

B ABNORMAL REVIEW TRENDS

In Fig. 9 we present a visual representation of the daily number of reviews noted in blue and corresponding star ratings noted in orange. Upon analysis, there was a noteworthy peaks related to negative reviews. Specifically, we observed three distinct days, denoted by the red circles, where the review numbers experienced a sudden spike while the review stars significantly decreased. Additionally, two other days, indicated by red arrows, exhibited an increase in review numbers while maintaining higher star ratings than the preceding days.

Further examination of the reviews from these specific days revealed a common pattern: the majority of reviews focused on one or two particular skills. Let us consider the first circled day, March 2nd, 2017, where a staggering total of 866 reviews were posted. Surprisingly, 719 out of these reviews were dedicated to the same skill, with a substantial 555 of reviews giving a 1-star rating. Fig. 10 demonstrates the daily review numbers exclusively for this skill, clearly indicating an abnormal review count on the identified day. We also observed similar patterns on the two other highlighted days, indicating a consistent trend across multiple skills.

These findings highlight the importance of conducting further investigations to determine the root causes. One factor that could contribute to these spikes is the review vetting system implemented by the Amazon Alexa skills store. This system delays the publication of reviews until they have been reviewed and approved by the system

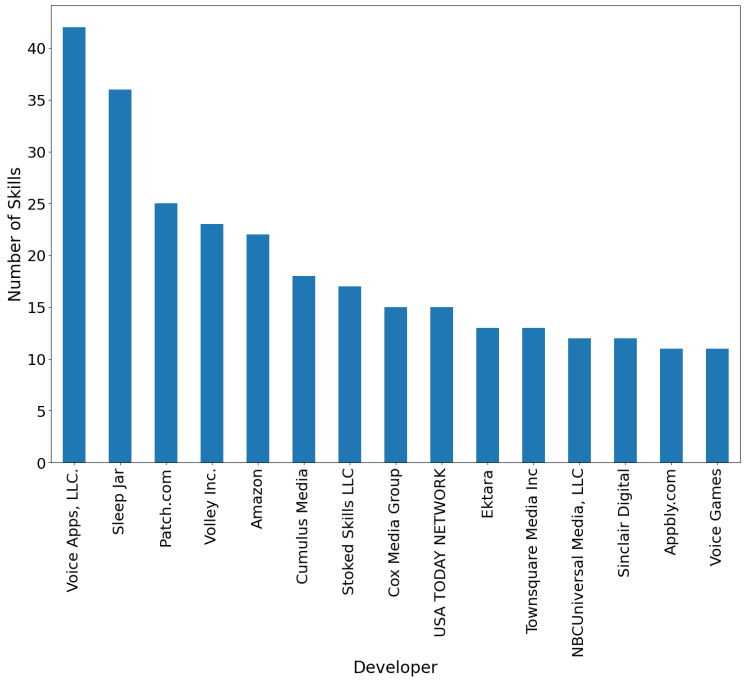


Fig. 8. Top 15 developers that have most reported policy-violating skills

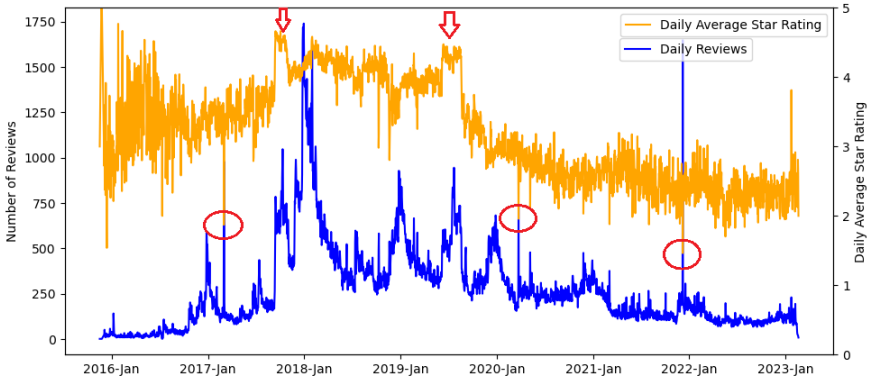


Fig. 9. Daily reviews and daily average star rating

administrators. Hence, if there were delays or inefficiencies in the review vetting process on those particular days, it could have resulted in a backlog of reviews waiting to be published. Once the reviews were finally approved and published in a single day, it would cause a sudden surge in the review numbers, leading to the observed spikes. Another possible reason could be the use of bots or automated systems to generate reviews by malicious

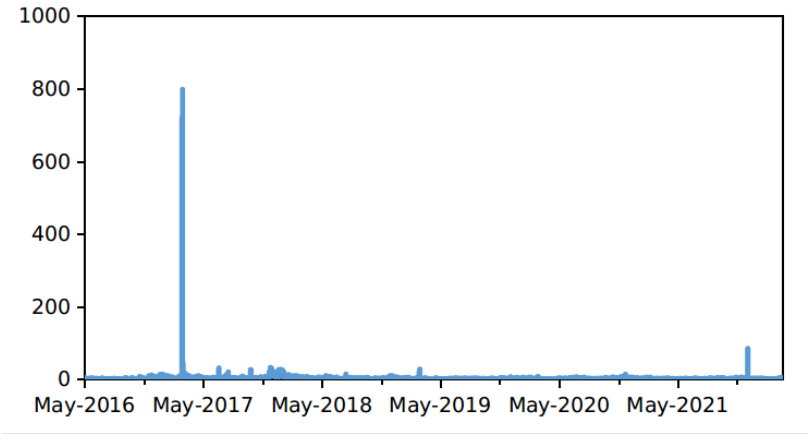


Fig. 10. Daily review number of one skill with possible fake reviews

users. These bots can be programmed to generate a large number of reviews within a short period of time, leading to a sudden surge in review numbers.