Analyzing Ad Prevalence, Characteristics, and Compliance in Alexa Skills

Aafaq Sabir, Abhinaya S.B., Dilawer Ahmed, Anupam Das
North Carolina State University
{*asabir2, asrivid, dahmed2, anupam.das*}*@ncsu.edu*

***Abstract*—With the rapid adoption of smart voice assistants like Amazon Alexa and the potential for more growth with large language model-powered assistants, as well as the introduction of "advertising ID" within Alexa, it is inevitable that advertisements (ads) will become prevalent on such platforms if not already. Although Alexa permits third-party developers to include ads within voice apps (known as "skills") and enables targeted advertisement through ad identifiers, Alexa also lists an ad policy that restricts ads within skill responses, notifications, or reminders except in defined cases. However, it remains unclear whether all developers comply with these policies or attempt to bypass vetting processes to publish non-compliant ads. This paper presents the first large-scale analysis of advertising on the Alexa platform, examining ad prevalence, characteristics, and adherence to platform policies. We introduce an automated ad detection method using a fine-tuned large language model (LLM) with 88.92% accuracy and, using chain-of-thought (CoT) prompting, achieve 94.52% accuracy in identifying potential policy-violating ads. Analyzing 45,477 Alexa skills, we find that 13.58% include ads or promotional content, with themes such as travel and entertainment. Notably, some ads come from skills by Amazon-promoted agencies like "Vixen Labs" while others are generated by agencies solely focused on voice assistant platforms, such as "Skilled Creative." Our model identifies approximately 29.18% of ads as possible policy violations. We reported our findings to Amazon, resulting in a bug bounty reward. The proposed system aims to enhance Alexa's vetting by automatically flagging potential ad violations and demonstrates how fine-tuned LLMs can support policy enforcement on voice platforms.**

1. Introduction

Voice assistants like Amazon Alexa, Google Assistant, and Apple Siri have seen significant growth in recent years [1]. Estimates suggest that by the end of 2024, the number of voice assistants will reach 8.4 billion, a number higher than global population [2]. These assistants engage in pseudo-natural conversations, learning user behavior over time to provide personalized services [3], [4]. Users utilize voice commands for various tasks, including web searches, music streaming, gaming, online shopping, and smart home control [5], [6], [7], [8], [9]. To improve the quality of conversations Alexa recently announced it is testing Alexa powered by LLMs [10], [11]. To further extend the functionality of voice assistants and support the integration with other devices, voice platforms have introduced voice applications. Known as *skills* in the Alexa ecosystem, these voice applications operate on top of Alexa-enabled devices. These apps enable users to interact with various online services and smart home devices through carrier devices such as Amazon Echo. Amazon Alexa, the largest ecosystem, currently boasts over 100,000 skills worldwide, and mainly comprises of third-party skills [12]. Third-party developers can create voice apps but must adhere to Alexa's policies for publication on the official skill store [13]. Consequently, skills undergo a certification process before becoming public [14].

Due to the widespread adoption of voice assistants, they are being utilized to deliver audio-based ads. Despite lacking a visual interface, voice assistants offer a platform for ads more interactive than those in traditional web-based ads, given users cannot easily skip ads within voice apps [15], [16], [17], [18]. Consequently, the voice-based advertising market is projected to reach $11.78 billion in 2024 [19]. Research has demonstrated that interactions with Alexa impact not only advertisements on the Alexa platform but also the broader online advertisement market. Alexa interactions can drive ad auction bids up to 30 times higher, influencing the overall online advertising landscape and numerous online advertisers [20]. Notably, Alexa now supports interest-based advertising by providing a unique *customer advertising ID* to skill developers, enabling user tracking across multiple skills, interest profile creation, and the delivery of interest-based targeted ads [21], [22], [23]. However, Alexa mandates that all skills comply with its ad policy [24], [25], which allows ads only for skills that stream content, promote specific products or services, or offer promotional deals in response to explicit customer request.

While Alexa has ad policy requirements in place, the extent to which Alexa properly vets and regulates ads in skills is unknown. Prior studies on Alexa's skill certification process have revealed cases where skills, despite passing the vetting process, still included policy-violating content [26], [27], [28], collected private information violating platform guidelines [29], or made back-end code changes after certification that coaxed sensitive information from users [30]. In this work, as a proof-of-concept, we demonstrate that even now, it is possible for developers to publish policy-violating ads in skills after the certification process. We use this as a motivation to develop a system that automatically detects ads within skills. We further characterize such ads to better identify policy-violating ads.

Specifically, we conduct the first large-scale empirical analysis on ads in Alexa skills to answer the following research questions: **RQ1:** *How can we effectively detect advertisements within Alexa skill responses?* We develop an LLM-based automated mechanism to detect ads in Alexa skills, which we optimize through prompt engineering and fine-tuning. Unlike online/web ads, audio ads in Alexa skills may not be explicitly declared or marked, making them difficult to detect and regulate. Our detection mechanism addresses this issue by analyzing the content of skill responses to identify advertisements. **RQ2:** *How prevalent are advertisements or promotional content in Alexa?* We perform automated interactions with Alexa skills published in the US Alexa skill store and analyze their runtime responses to determine the prevalence of ads or promotional content within them. We leverage a fine-tuned LLM to automatically detect ads in skill responses. **RQ3:** *What are the characteristics of ads or promotional content in Alexa skills? Are there specific companies developing these ad-supported skills?* We characterize the nature of ads by clustering the ad content and identify advertising agencies and skill-building companies that collaborate with Alexa. **RQ4:** *What is the prevalence of ads that violate Alexa's ads policy?* We extend the automated ad detection system to automatically identify policy-violating advertisement using a *chain-of-thought* (CoT) prompting approach. Our proposed detection system can augment existing vetting procedures by dynamically analyzing skill responses before and after publication (i.e., periodic) to ensure continuous policy adherence, a feature that can complement the existing vetting process by Alexa. In summary, this paper makes the following contributions:

- We develop an automated ad detector using LLMs to identify ads within Alexa skill responses. Unlike visual ads, voice-based ads lack easily distinguishable cues (e.g., visual icons), requiring content analysis to detect ads. We optimized LLM performance through prompt engineering and fine-tuning, achieving an average detection accuracy of 88.92%, significantly outperforming zero-shot learning and transformer models like BERT (§ 4).
- We use the Chain-of-Thought (CoT) prompting technique to guide decisions on the detection of non-compliant ads in Alexa skills. By prompting ad-detection LLMs to explain their decisions, we create a context for the subsequent compliance validation step. This approach achieves 94.52% accuracy in validating compliance of detected ads with Alexa's ad policy (§ 5).[1]
- We perform a large-scale analysis of ads in Alexa skills covering dynamic responses from 45,477 Alexa skills. Using our ad detection model, we find 6,095 (i.e., 13.4%) skills running ads, with 11,335 unique skill responses containing ads. Our findings indicate that running policy-violating ads through skills is not merely a proof-of-concept; rather, we observe that 29.18% of detected ads potentially exhibit non-compliant behavior. To the best of our knowledge, we made the *first* attempt to automatically detect ads within Alexa skill responses and also created the first automated technique to flag potentially policy-violating ads (§ 5).
- We categorize the ads discovered in Alexa skills into various themes and topics (e.g., entertainment, travel services) through agglomerative clustering on the TF-IDF features of the LLM-based summaries. We also identify several ad intents, including promoting specific products or businesses. Moreover, our analysis revealed skill-building agencies that promote ads in Alexa skills, featuring well-known agencies like ”Vixen Labs” [31] as well as agencies solely focusing on voice assistant platforms, such as ”Skilled Creative.” [32] (§ 5).
- We open-source all research artifacts, including our benchmark dataset comprising 1,400 manually labeled skill responses (700 responses containing ads and 700 not containing ads) [2]. Additionally, we have reported our findings through Amazon's bug bounty program, and have recently been awarded a bug bounty [33].

1. This paper presents an automated approach to detect implicit promotional content within Alexa skill responses and assess its compliance with Alexa's content policy [24]. Here, "ad" or "advertisement" refers to any text with a promotional tone, regardless of its policy compliance. Policy compliance is only determined once text is considered promotional.

2. Background

Publishing Voice Apps. Amazon Alexa enables third-party developers to create and publish skills using an online developer console. This console streamlines the development, testing, debugging, and deployment of both the frontend and backend of skills. Within the console's *build* tab, developers can set invocation phrases, define intents, and manage slots for data input through natural language. After defining the frontend, developers create the backend to process frontend requests using Alexa's built-in code editor. Each frontend-defined intent requires a corresponding backend handler, executing the desired action and preparing a user response. The skill dashboard offers the option to host the backend on Amazon, but developers also have the flexibility to host it on an external server.

Alexa's developer console includes a voice assistant emulator for skill testing during development. Once the voice app testing is complete, developers can submit it for publication through the *distribution* or *deploy* section of the console. In the distribution section, developers provide metadata like description, sample utterances, category, and icons. A privacy policy is required only if the skill collects personal data. Developers can specify country-specific web stores where the skill will be available. Once submitted, skills go through a vetting and certification process.

Skill Certification Process. Newly submitted skills undergo a platform-defined vetting process to meet standards and adhere to policies before being publicly accessible. This process verifies functionality, compliance with platform policies and ensures the absence of prohibited content like profanity, explicit material, or unsolicited ads. Alexa also extends vetting to backend servers, verifying their response,

2. Our code, dataset and other artifacts are available at https://privacy-datahub.csc.ncsu.edu/publication/sabir-sp-2025/

especially for skills hosted on external servers, to ensure they respond only to Amazon's signed requests.

On completion of the certification process, the certification status is displayed on the Alexa developer console (see OSF artifact repository [33]). Certification typically takes a few days; if successful, the skill is promoted to *live* status in the store. If it fails, the skill returns to the *development* phase, and the developer receives an email detailing the reasons for failure, allowing them to make revisions and resubmit. The console also provides usage insights, such as invoked intents and their frequency across users.

Subverting the Vetting Process. While Alexa has a vetting process in place, existing literature has demonstrated how the vetting process can be subverted [28], [29], [30], [34], [35]. As a proof of concept, we also investigated how we can bypass the vetting process to publish policy-violating ads in Alexa skills. We created skills that tell jokes to the users while intermittently running five *fictitious* ads promoting products unrelated to the skills' main functionality, thus violating Alexa's ad policy [24]. We also verified that the ads that we created actually violated the platform's policy by incorporating the ads in the skills and submitting them for certification. The skills failed certification as they contained policy-violating ads (an email response from Alexa is provided is OSF artifact repository [33]. However, the skills were certified once we removed the ads. After the skills were live, we injected the policy-violating ads dynamically by changing the backend code and were able to run ads in skills without getting flagged or detected by the platform. We decided to take down the ad-bearing skills after three months. More details about our policy-violating ads and bypassing vetting process can be found in Appendix A. We obtained the necessary IRB approval for publishing policy-violating ads. We demonstrated live skill interaction to the IRB board members, ensuring our ads were fictitious with an explicit research disclosure at the end.

3. Related Work

Online Advertisements and User Privacy. Online advertising fuels the Internet economy, with US ad spending reaching $322 billion in 2024, a 5.2% increase from 2023 [36]. It operates across various platforms, including websites, mobile apps, and social media [37], [38], [39]. However, it comes with a cost to users, as advertisers utilize user tracking to deliver targeted ads for increased engagement [37], [40], [41], [42]. Many studies have unveiled details of online tracking techniques [43], [44], [45], [46], [47].

Efficacy of Audio Ads. There have been studies emphasizing how effective audio ads can be, despite having a seemingly less capable interface (i.e., not having a display). Johnson et al. found that audio ads may leave an impression even when the listener is engaged in other tasks, such as playing a video game [15]. Park et al. studied the efficacy of interactive vs. non-interactive audio ads and found that interactive ads through smart speakers increase ads' brand and product recognition compared to non-interactive ads [16].

Amazon Alexa has recently launched interactive ads [48]. Research has also been done on how audio ads can be more effective when delivered through voice assistants. Smith et al. proposed three frameworks for advertisements through smart speakers [17] to make it more effective. Lee et al. postulated four ad types for smart speakers: contextual, non-contextual, voice-search recommendation, and voice-search listing ads and showed that non-contextual ads led to less favorable attitudes than the other three types [18]. A recent work by Cho et al. [49] analyzed if users prefer smart speakers as ad sources or as mere mediums (as in radio). They conducted a scenario-based user study in which Siri acted as either a source of ads or a medium for delivering ads. They found that for socially motivated users, Siri as an ad source positively affects user experience, but not for informationally motivated users.

Ads through Smart Speakers. A recent study by Iqbal et al. [20] has looked at determining how user data is harvested through smart speakers to serve targeted ads. Their work showcases that interaction with the Amazon Echo device leads to as much as 30× higher ad bids from advertisers to serve targeted ads. Our work is orthogonal to this work as it focuses on how skill developers can bypass the vetting process to publish ads in skills and how to automatically detect ads in skills, which is manually determined by Iqbal et al. [20] at small scale. Furthermore, Iqbal et al. only manually searched for ads within skills that had contacted some analytics/ad-publisher domain; however, as we show in this paper, ads can also be served within the TTS response from skills that are not served by analytics/ad-publishers.

Policy-violation within Voice Apps. Researchers have been analyzing skills for non-compliance issues. Guo et al. developed "Skill Explorer" to interact with published skills in an automated way and found that 1,141 skills request users' private information without following platform-specific specifications [29]. Young et al. also developed a similar tool "SkillDetective" and identified 6,079 Alexa skills and 175 Google actions violating at least one policy requirement [34]. Shezan et al. did a passive analysis of skills and found that 86.36% of the skills miss disclaimers when providing medical advice [35]. Our work complements existing research by being the first to assess the prevalence of ads in Alexa skills and determine their compliance.

4. System Design

Having demonstrated how developers can publish policy-violating ads (see Appendix A), we propose a dynamic monitoring system that can effectively detect policy-violating ads within the skill responses even after publishing on the official skill store. Toward this goal, we first need to detect ads in the skill responses. Next, for the detected ads, we need to determine whether or not they comply with the platform's ads policy [24]. In this section, we present our automated system capable of detecting ads in skill responses (**RQ1**). Additionally, we perform a compliance analysis against Alexa's ads policy.

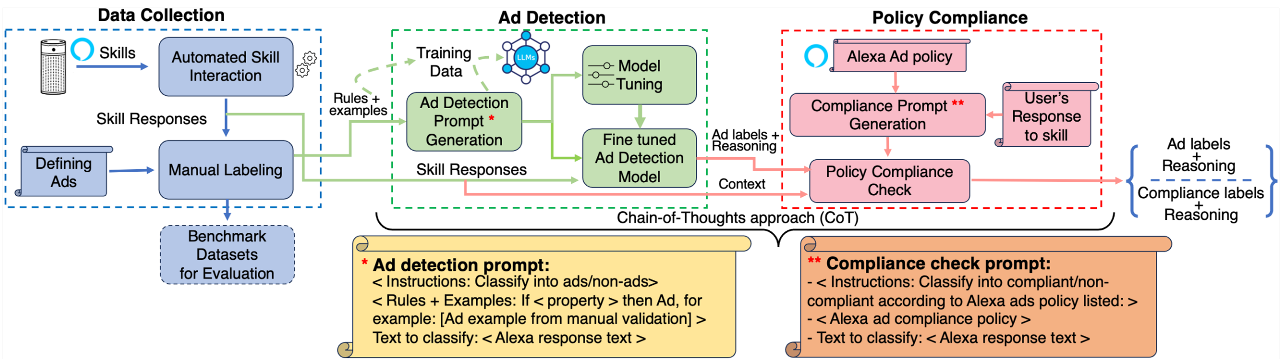

Figure 1: System overview: (i) Data collection: First phase performs automated skill interaction to get Alexa responses for manual labeling to create a benchmarking and fine-tuning dataset, (ii) Ad detection: The second phase consists of an ad detection prompt for LLM and LLM fine-tuning. The LLM takes skill responses as inputs and outputs ad classification labels with reasoning; (iii) Policy compliance: The third phase of our system checks for ad policy compliance for the detected ads. The LLM outputs the ad compliance label and corresponding reasoning.

Our system consists of three major components: (i) automated data collection from Alexa skills for large-scale data analysis, (ii) ad detection procedure that leverages a fine-tuned LLM, and (iii) a Chain-of-Thought (CoT) prompting based technique for checking ad policy compliance. The system overview is shown in Figure 1.

4.1. Automated Data Collection

4.1.1. Automated Interaction with Alexa skills. One of the challenges in detecting ads from voice app responses in the wild is obtaining the responses from different execution paths of the voice apps. Then, we need to repeat this process on a large scale for the voice apps available on the voice app stores. Since there are over 80,000 Alexa skills in the US skill store alone [50], manual interaction is not scalable. Moreover, it is plausible that the ad may not appear after the first invocation phrase but only appear deeper into the interaction. To address this problem, we used the automated skill interaction tool, *SkillDetective* [34], capable of automatically conversing with Alexa skills better than generic chatbots in terms of carrying out more natural and hence longer conversations. *SkillDetective* is the current state-of-the-art skill chatbot surpassing "SkillExplorer" [51] in performance [34]. *SkillDetective* first classifies the response from the skill into distinct question types, then generates an answer based the identified question type. It keeps the conversation going until it reaches a statement that does not contain any question or if it cannot answer a question correctly. Additionally, the interaction tool explores multiple execution paths of the skill by considering different possible responses (e.g., answering a question with 'yes' in the first interaction, and 'no' in the second). *SkillDetective* interacts with Alexa skills using the Alexa developer console and gets the text-to-speech (TTS) of the responses as text through Alexa's skill developer console.

Augmenting SkillDetective. *SkillDetective* interacts with the Alexa developer console using Selenium; however, we modified the tool to be headless and faster (i.e., avoid rendering delays) by recording the web requests of the developer console and replaying them from a script. We benchmarked both vanilla and modified SkillDetective on 100 randomly sampled skills under identical settings. The instrumented version runs $\sim$ 2X faster (7118 sec. vs. 3070 sec.). Since different skills can share the same invocation names and the ones previously enabled may also get activated, we only enable one skill at a time and disable the skill programmatically after interacting with it. We log the conversation with skills and maintain an *interaction graph* where each node represents the questions asked, and each edge represents the corresponding answer. Existing research has explored shallower depths of up to 7 levels [34]. We set the conversation depth limit to a maximum of 25 levels, finding it to strike the right balance between coverage and scalability. We note that we did not alter the execution logic of the tool; rather we made it faster.

Skills Analyzed. To use this tool, we first collected skill IDs and invocation phrases of skills by crawling the US Alexa skill store based on skill categories. We crawled 55,346 skills from the US skill store using our crawler. We exhaustively crawled the US skill store by iterating through tagged categories until pagination ended and were able to find only 55,346 skills. Of these crawled skills, 9,869 were *flash briefing skills* that typically play an audio stream. We do not consider this category of skills in our analysis because Alexa's ads policy allows streaming skills to run ads, as long as they did not use Alexa's voice to deliver them [24]. Moreover, ads in streaming services are usually generic regardless of the medium (e.g., radio vs. skill). Detecting such cases is out of the scope of this paper. We run the interaction tool on the remaining 45,477 skills and saved their responses. Skill interaction, on average, took around one minute for each skill. We obtained a total of 369,509 responses from the skill interaction process.

4.1.2. Creating Benchmark Datasets. We created ground-truth benchmark datasets comprised of skill responses to train and evaluate our ad detection and compliance analysis models. For both training and evaluation, we wanted to have a similar proportion of responses containing ads and not containing ads; however, we did not know the distribution of ads in real-world skill responses beforehand. Thus, we collected our dataset in an iterative process till we reached a target number of ad and non-ad examples across skills from all categories. We targeted to find a dataset of 1,500 examples with (750 ads and 750 non-ads). To obtain these, we used the augmented SkillDetective tool to interact with sets of skills. Since we wanted a stratified dataset representative of all 23 skill categories on the Alexa skill store, we randomly selected 10 skills per skill category (230 skills in total). We analyzed these 230 skills in batches of 23 skills each (one from each category). The tool provided the skill responses for each batch of skills, which we manually analyzed. This way, we analyzed the skill responses in batches until we hit our desired number of samples.

Manual Annotation Process. To manually classify the responses, two researchers began with the definitions of *advertisement*. The Code of Federal Regulations (CFR) defines advertisement as *"dissemination of information, including but not limited to paid advertisements, that are reasonably calculated to advise the public how to present a claim"* [52]. Amazon's audio ads guidelines stipulate that audio ads must *"state the brand/product name and educate the customer about the main features and benefits of the product or service"* [53]. Alexa does not explicitly define "ad" for skills but extensively states what kind of ads are allowed or prohibited by the policy [24]. Alexa ad policy states that "promotional messaging" is allowed in only certain cases such as "Skills that are specifically designed to promote a product or service may include audio messaging promoting that product or service", which implicitly defines self-promotion as advertisement.

With these definitions as an anchor, we classify skill responses as ads that contain any promotional, publicity-related content of a product, service, event, organization, or venture. Note that we don't differentiate between compliant and policy-violating ads in this ad detection step, this is done as a follow-up. We, therefore, considered any skill responses that promote or announce their own functionality to encourage user engagement as ads. Note that this type of ad is allowed under the Alexa ads policy [24].

On average, we obtained 700 unique responses in each batch of skills we ran through the ad detection step. For the first batch, two experienced coders labeled the responses while discussing with each order to establish a unified mental model and a set of rules (i.e., codebook) defined for labeling a response as an ad. Both the coders were highly qualified for manually validating Alexa skill ads, possessing extensive expertise with the Alexa platform and skill-building. One coder also had experience in online advertising and user profiling. For the rest of the skill responses, both coders independently coded the responses and assigned a code (0 for ad, 1 for non-ad). After every batch, the two coders discussed whenever their codes did not match or were confused about how to code a particular response. After reaching a consensus, they applied the code and updated the codebook for labeling the responses. The overall inter-rater reliability (IRR) across all batches (before reaching consensus), calculated using Cohen's Kappa, was $\kappa = 0.711$ which shows substantial agreement.

We found that the distribution of responses labeled as ads in these samples was from 19% to 24%. Because the responses containing ads were much lower in number than those not containing ads, we continued labeling skill responses until we obtained the desired number of ad samples. After six batches, we obtained 755 responses labeled as ads, and we selected 750 non-ad responses from the rest of the labeled responses, reaching the goal for our dataset. We call this dataset D_0. During the labeling process, we identified five themes in the ads using our codebook definitions. These themes, referred to as "rules," were employed alongside a corresponding skill response for each rule in our prompts to refine the LLM for ad detection. The rules precisely outline the characteristics the model should identify in the text to classify it as an ad within the context of skills. This involves detecting promotions or suggestions related to products, services, brands, or places, accompanied by one example for each rule (the rules are listed in Table 4, Appendix B).

Dataset for Fine-tuning and Evaluating LLM. Given we had a total of 755 ad and 750 non-ad samples in dataset D_0, we randomly selected 55 ads and 50 non-ads from D_0 to construct the fine-tuning dataset (D_f). We limited the fine-tuning dataset to 105 samples as LLMs can be fine-tuned reasonably with 50 to 100 training samples [54]. After removing the fine-tuning samples, we used the remaining 700 ads and 700 non-ads samples from D_0 to create a benchmark dataset containing a total of 1,400 samples (D_b).

Benchmark Dataset for Policy Compliance. To create the benchmark dataset for the policy compliance phase of our system, we evaluated the 700 ads in our ad benchmarking dataset D_b for ad policy compliance. To do this, we looked at the skill response in the context of the skill's metadata. This metadata consists of the skill name, developer name, and the skill description listed on the skill's web page. We also considered the specific user request (emulated by *SkillDetective*) that generated that specific skill response. We then evaluated it according to Alexa's ad policy [24]. If the ad in the skill did not fall under any exceptions defined in Alexa's ad policy, we labeled it as a potentially policy-violating ad; otherwise, we labeled it compliant. From the 700 ads we included in our benchmark dataset D_b, we labeled 559 ads (79.85%) compliant and 141 ads (20.14%) as potentially non-compliant.

4.2. Ad Detection

Model Selection. In the ad detection phase, our system needs to identify ad or promotional content within Alexa skill responses. We chose LLMs over traditional models like BERT for the following reasons: 1) LLMs require minimal training data, advantageous for tasks lacking publicly available labeled data, as in our case; 2) LLMs offer greater flexibility, allowing us to control model behavior through prompts and adapt to future policy changes; 3) LLMs demonstrated better overall performance in our experiments. For instance, using dataset D_b, we observed a precision of 82.05% for LLM compared to 54.20% for BERT.

Prompt Engineering. To perform text classification, an LLM requires appropriate prompts containing information about the task to be performed, metadata about the input(s) it will receive, and the desired format of the output. We first brief the LLM with an input text (i.e., response from an Alexa skill), and specify the task as classifying the text into an *ad* or a *non-ad*, depending on whether or not the skill response contained ad or promotional content.

We provide the LLM with detailed instructions consisting of rules about the characteristics to consider in an input text for determining if it contains an ad or not, and an example corresponding to each rule. We derived these rules from the themes of ad content that emerged from our manual validation of the benchmark and fine-tuning datasets (jointly called D_0), as described in section 4.1. In total, we provided five rules, along with five corresponding examples, that define the characteristics of a skill response that can be classified as an ad. For example, one of the rules was, *"If the text talks about a company, product, or service and talks about how people should use, buy, or engage with it"*. We wanted our model to perform the classification based on the rules we provided, so we wanted our rules to cover a broad themes of ads, and not restricted to a few specific cases. So, we used careful wording when describing the rules to be used as prompts. For example, suppose we encountered a skill response that advertises a services after the welcome message in our dataset. In that case, our rule says *"company, product, or service"* so the model can also capture other entities. The structure of the prompts was empirically refined (inspired by [55]) based on correctness and soundness of LLM reasoning as well as correctness of the output format as specified. We improved the prompt by testing it on 105 randomly selected samples (55 ads + 50 non-ads, as described in 4.1.2) separated for fine-tuning from D_0. Once we achieved optimal performance and stable responses, we fixed the prompt. These samples were used for optimization to avoid bias, as they were not part of the benchmark dataset. Our careful crafting of rules ensured that the model did not miss other variations of ads of the same type. Finally, we instructed the LLM to provide a label based on the predicted class, and to provide a brief *reasoning* of why it classified an input text a certain way. We include the ad detection model prompts in Table 4 in Appendix B.

Fine-Tuning LLMs As pre-trained LLMs are trained on a huge corpus of data from the Internet [56], they have reasonable general knowledge and reasoning capability. LLMs can also perform various tasks without any specialized training just using prompt engineering [57]. Furthermore, LLMs can easily perform a downstream task by providing a few examples in the prompt, known as *few-shot learning* [58]. Lastly, LLMs can be fine-tuned by updating model weights using training data, optimizing their performance [59]. We evaluated and compared four LLMs: Open AI's GPT-4o (`gpt-4o-2024-08-06`) [60], GPT-3.5 (`gpt-3.5-turbo`) [61], and Meta's open-source LLMs Llama 3.2 (`Llama-3.2-3B-Instruct`) [62] and Llama 2 (`llama-2-7b-chat-hf`) [63]. We benchmarked the base and fine-tuned versions of LLMs on our dataset.

For fine-tuning, we used the same prompt we would use for inference. Each fine-tuning example comprises a conversation between the LLM and the user, where the conversation starts with a system message of *"You are a helpful assistant."* followed by the prompt and text to classify and lastly the correct ground-truth label. To fine-tune the GPT models, we used the fine-tuning API provided by OpenAI [64]. For fine-tuning Llama models, we used the Low-Rank Adaptation (LoRA) approach implemented through the Parameter-Efficient Fine-Tuning (PEFT) library [65], [66]. For GPT 4o/3.5, OpenAI provides the recommended number of epochs for the fine-tuning data provided; in our case, our model was tuned for three epochs. For Llama 3.2/2, we fine-tuned it for 10 epochs as it showed less sensitivity towards fine-tuning. The fine-tuning for Llama 3.2/2 was conducted on a 2x RTX 4090 GPUs. The results for our benchmark dataset are described in Section 4.4.

4.3. Ad Policy Compliance

Once a skill response is classified as an ad during the ad detection phase, the response is passed to the next step, which performs ad policy compliance verification. At this stage, we evaluate whether the ad in the skill response complies with the ad policy using the same LLM used for ad classification. The LLM is provided with a new prompt to evaluate policy compliance. We provide the LLM with the ad policy, and the LLM is prompted to assess whether the content in the ad complies with it. Since Alexa's ad policy permits ads for certain skills or upon users' explicit request, we provide the skill name, developer name, skill description, and the user request that generated the response, as shown in Figure 1. This ensures the LLM can evaluate compliance with all necessary information about the skill.

Chain-of-Thought based Prompt Engineering. At this stage, we present the compliance prompt to the LLM. We build upon the previous chat dialogue with the LLM for continuity with the initial ads classification task. This continuity ensures that context is maintained, and the LLM remembers that the initial task was identifying an ad within an input text. The objective of this phase is to assess policy compliance. Therefore, the ad detection phase served as an intermediate task from the perspective of compliance verification, aiding the LLM in reasoning systematically. Our approach follows a step-by-step method, providing explanations for decisions made, a concept referred to as "Chain-of-Thought" (CoT) prompt engineering [67]. Similar to the ad detection phase, the model is prompted to provide the appropriate label and reasoning behind categorizing a specific ad as potentially policy-violating or compliant. The LLM relies on the ad policy text and skill metadata to classify skill responses for compliance. However, we did not consider the case of streaming skills delivering ads

TABLE 1: Ad detection performance for GPT 4o/3.5 vs. performance of Llama 3.2/2 (base and fine tuned) models.

Model	Accuracy	Precision (Ad)	Precision (Non-ad)	Recall (Ad)	Recall (Non-ad)	F1 Score (Ad)	F1 Score (Non-ad)
GPT-4o Base	86.57%	82.9%	91.15%	92.14%	81%	87.28%	85.77%
GPT-4o Fine tuned	88.92%	89.43%	88.43%	88.28%	89.57%	**88.85%**	**88.99%**
GPT-3.5 Base	70.71%	89.40%	64.05%	47%	94.42%	61.60%	76.32%
GPT-3.5 Fine tuned	87.64%	86.75%	88.57%	88.85%	86.42%	87.79%	87.48%
Llama 3.2 Base	74%	84.28%	68.46%	59%	89%	69.40%	77.39%
Llama 3.2 Fine tuned	63.85%	58.69%	84.15%	93.57%	34.14%	72.13%	48.57%
Llama 2 Base	59.21%	59.44%	58.99%	58%	60.42%	58.71%	59.7%
Llama 2 Fine tuned	67.21%	62.53%	77.44%	85.85%	48.57%	72.35%	59.7%

in-stream as they are not platform-specific (e.g., present both in radio and skill). Our model is expected to flag an ad as potentially policy-violating if it does not align with exceptions outlined in the Alexa's ad policy [24]. Table 5 in Appendix B provides the detailed prompts used for ad policy compliance verification.

4.4. Model Performance

We evaluate the base versions of LLMs as well as their fine-tuned versions on our benchmark dataset D_b. To make the results of the LLMs consistent, both OpenAI and Llama LLMs have the `seed` parameter (we set `seed` to 10 for consistency across runs).

Ad detection. We first evaluated the base versions of all the LLMs on the benchmark dataset D_b containing 700 ads and 700 non-ad examples, as described in Section 4.1. Then, we evaluated the same dataset D_b on the fine-tuned versions of both the LLMs. In general, we saw significant improvements in the performance of the LLMs after fine-tuning. Overall, we found that GPT-4o outperformed all other models. The performance results are provided in Table 1. We selected the best performing LLM, i.e., fine-tuned version of GPT-4o, for our large-scale analysis. We see a substantial improvement in F1 scores after fine-tuning and selecting the model with the highest F1 scores for large-scale analysis. However, we also observed a drop in specific performance metrics like recall for certain models (e.g., Llama) after fine-tuning which suggests room for improvement in fine-tuning to boost overall performance for specific models.

Policy Compliance. We assessed policy compliance for the fine-tuned GPT-4o model using the ad samples from our benchmark dataset. The model demonstrated a policy compliance accuracy of 94.52%. The precision and recall for non-compliant ads were 86% and 87.61%, respectively. For compliant ads, the precision and recall were 96.76% and 96.54%, respectively. Additionally, we applied our compliance check LLM to prototype skills designed to test Alexa's skill vetting process with policy-violating ads (please see Section A in Appendix for details). Our model successfully flagged those policy-violating ads in every instance. In fact, during our large-scale analysis, described in Section 5, the skill interaction tool also interacted with our deployed skills, three of which injected ads during tool interaction. All three responses were labeled as ads and flagged as non-compliant by our model.

For some models, like GPT-4o, certain metrics like recall (ad) decreased after fine-tuning, although overall performance improved. This is due to the precision-recall trade-off, where gains in one often reduces the other [68]. Additionally, overall performance may not improve across all classes (e.g., Llama 3.2) if fine-tuning on a specific dataset or parameters fails to enhance general performance [69].

Takeaway. LLMs can be effectively used (with 88.92% accuracy) for detecting ads and promotional text within skill responses. Also, using the chain-of-thought approach, we can capture ads that do not comply with Alexa's ad policy with 94.52% accuracy.

5. Analyzing Ads in Real-world Voice Apps

Given existing research has shown that *audio ads* are highly effective [15], [16], and since Alexa allows skill developers to track users and deliver targeted ads through *advertising IDs*, we hypothesize the ads will become prevalent in skills if they are not already.

In this section, we first analyse the prevalence of ads in Alexa skill ecosystem (**RQ2**). Detecting ads in voice app responses presents challenges, especially since ads in TTS (Text-To-Speech) responses lack distinguishable differences in content sources or contacted domains. All TTS responses originate from Alexa backends, unless a voice app explicitly communicates with a third-party backend. In a recent study, Iqbal et al. [20] examined ads in skills by analyzing network traffic from various third-party skills. While they identified instances of skills connecting to analytic and advertising endpoints, their study focused solely on activating skills and lacked dynamic interaction across multiple execution paths, which we do in our work. Given the potential for ads to appear at various interaction points, their approach offers limited insights into the presence of ads within skills. We detect ads in skills by scrutinizing the content of TTS responses, employing our fine-tuned LLM described in Section 4.2. In this section, we only analyze ads without analyzing their compliance. We determine policy compliance as a next step in the pipeline as discussed in Section 5.4.

5.1. Large-Scale Data Preprocessing

We first removed duplicate skill responses collected in the wild before running our ad detection tool on them. The duplicate responses appeared when some part of the interaction graph overlapped with other branches. After filtering

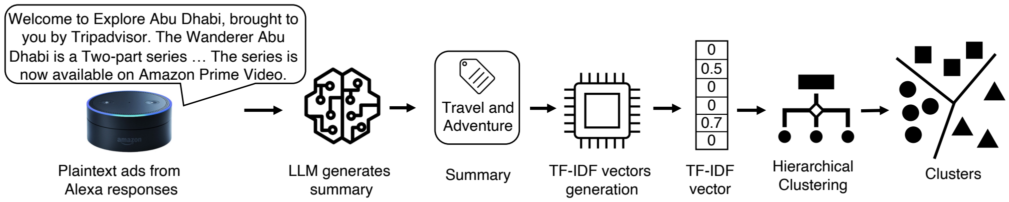

Figure 2: Preprocessing and agglomerative clustering process for ads. Plaintext ads from Alexa skills are first converted to short summaries based on the topic of advertisement by LLM. The summaries are converted to numeric vectors using TF-IDF representation. Finally, the TF-IDF vectors are clustered using agglomerative clustering into various clusters of ads based on similar topics.

the duplicates, we ended up with 103,784 unique skill responses from 45,477 skills. These skill responses might also contain the response from Alexa instead of the skill, which can happen when Alexa cannot launch or find the skill for a particular invocation phrase (typically happens when that skill is not working correctly). Sometimes, it can happen when one invocation phrase matches multiple skills, and Alexa suggests one to confirm. During our benchmark labeling phase, we identified such responses that are in the form of `"Sorry, I'm having trouble accessing your <skill name> skill right now"` or `"Hmm, I found a few skills that might help"`. Moreover, if a skill is not active when requests using the specified invocation phrases are made, Alexa may automatically conduct a generic web search to retrieve a pertinent result from a website. Such responses are also detected as they follow a specific format such as `"From <website domain>: <text>"`. We identify such responses using a *regular expression*. We removed all such responses that do not come from skills. After removing such responses, we were left with 83,472 unique skill responses from 23,518 unique skills on which we ran our ad detection model. We used the same `seed` as in the performance evaluation (Section 4.4) to ensure consistent model performance.

5.2. Prevalence of Ads in Voice Apps

Our ad-detection LLM detected 11,335 of the responses (i.e., 13.58% of all 83,472 unique skill responses) as ads. These responses originated from 6,095 unique skills. We performed manual validation to evaluate the model's correctness on large-scale data and characterized the ads. The same two researchers who had manually coded the benchmark datasets performed the manual validation by reading through 10% of the responses labeled as ads (randomly selected) and an equal number of non-ads. For manual validation, researchers followed the codebook definitions outlined in Section 4 and labeled the skill responses into "Ad" or "Non-ad". The inter-rater reliability Cohen's $\kappa = 0.93$ showed a very high agreement between the coders.

During manual validation, we leveraged the reasoning provided by the LLM to help determine whether the text contained an ad or not. However, LLM explanations were only considered as an additional factor in resolving the labels and were not entirely relied upon, as this could introduce bias from the LLM itself. To ensure accuracy, we manually assessed the alignment of LLM explanations with the initial rules we provided. It is also important to note that we did not factor in Alexa's ad policy or the ad's compliance with it during this validation process. This is because the current phase of the system was solely designed to detect ad or promotional content within skill responses. We found that our ad detection system performed similarly on both the large-scale data and the benchmark dataset. Through manual validation, we evaluated the precision and recall for ad detection as 89.49% and 89.26%, respectively.

5.3. Analyzing Ads Topics

For further insights into the detected ads, we delve into their content, summarizing them based on emerging topics (**RQ3**). To group ads by similar topics, we apply clustering on the ad content. Clustering is a well-known method for grouping similar unlabeled data points together [70]. Since the ideal number of clusters was not known beforehand, we opted for a clustering algorithm that does not require this information, yet offers control over cluster granularity. For this purpose, agglomerative clustering [71] proved suitable for our needs, allowing us to adjust cluster granularity by setting a distance threshold parameter.

To cluster ads by topics, we condensed them into concise summaries that capture their promotional essence. These summaries were created using GPT-4o, as LLMs are extensively used for summarizing tasks [72]. This helped us filter text not related to the ad's intent. Employing the prompt *"You are helping me assign a theme for an advertisement or promotional text sourced from Amazon Alexa skill response. Please assign a short high-level theme or a category to the advertisement. Provide a concise summary or high-level theme in a few words for the following promotional text:"* we derived summaries for each ad. Through empirical evaluation, we found that clusters formed using themes were more effective in describing topics than those based on raw plaintext. Next, we transformed these summaries into numerical vectors to prepare the data for clustering. For vector transformation, we compared the results of different embedding methods, such as OpenAI's `text-embedding-ada-002` [73] and simpler techniques like TF-IDF [74]. We found that OpenAI's `text-embedding-ada-002` embeddings were effective in producing vectors that grouped ads with similar sentence structures and overall semantics. In contrast, TF-IDF tended to create clusters based on similar non-common terms (e.g., objects of the ads), with less emphasis on the overall sentence structure. We empirically found that for our purpose, TF-IDF was better at clustering based on topics because ad topic depends on the object of the ad rather than overall sentence semantics. Hence, we used TF-IDF method for vector transformation. We chose *cosine* as the similarity metric, as it is preferred over Euclidean distance for text analysis [75]. The process of converting plaintext ads to TF-IDF-based vectors is outlined in Figure 2.

Our clustering process comprises three key steps: 1) LLM-driven inductive theme assignment to each ad, 2) vectorization and agglomerative clustering using TF-IDF embeddings and cosine distance, and 3) qualitative cluster summarization through manual review. To derive a representative topic for each cluster and group similar clusters, we qualitatively analyzed 10% of randomly selected ads and their assigned themes within each cluster. This approach allowed us to identify and summarize the prominent themes and topics within our dataset.

Empirically, we determined a distance threshold (i.e., cosine similarity = 0.96) for clustering that achieved the desired granularity, resulting in 17 clusters. Although the text vectors comprised 2,734 dimensions (based on vocabulary), we visualized the clusters in two-dimensional space using t-SNE [76]. For conciseness, we grouped clusters with common themes in Table 2 providing overarching themes along with example ads. The theme for each ad is automatically inferred by the LLM using the contextual instructions provided in the prompt listed in Section 5.3. The prompt guides the model to treat the ad text as promotional content sourced from an Alexa skill and instructs it to assign the most relevant "high-level theme." This approach helps the LLM identify the dominant topic of the ad, even when multiple entities are mentioned.

We also examine the likely causes of false positives and negatives in ad detection. These errors often arise in nuanced cases, such as when the "Soft drink facts" skill (B07LDND6K4) states, "Santa has been featured in Coke ads since the 1920s," which is classified as an ad as it sees the text 'ad' implicitly. For false negatives, the model may overlook promotions when they appear within longer responses with additional details. For example, in the "Quantum Stories" (B088YQT9LP) skill, the response includes "Quantum Stories Gizmo Girl explains why we need quantum algorithms... you can even drop out of school and join Microsoft as a quantum programmer." Here, mention of a company is missed due to surrounding context.

5.3.1. Common Intents of Ads in Voice Apps. Additionally, we aimed to analyze the underlying intents behind the ads or the purpose of their inclusion in voice apps. To accomplish this, we manually examined 10% of the ads and identified common intents based on the content being promoted. We present some of the common intents that we observed with examples of ads belonging to each intent type (quotes include the skill ID at the end). These are the most frequently occurring intents for ads that we found.

❶ **Skills promoting their own functionality.** We found it is common for skills to advertise their own functionality. An example is the "Cashew Demo" skill by Cashew.ai, which give users a demo of conversational ad on Alexa that they created as they can "help your brand engage with Alexa users" as per the description. The response says:

> "The early bird catches the worm, which is why listening to your company's conversational ad in the morning is a productive start to your workday. To listen to the demo that Cashew have just created for you, please say your unique Cashew demo 5 digit pin?" (B09FSTGQR3)

❷ **Skills created to promote a specific product.** These skills typically *only* respond with promotional messages. E.g., "Lexus LX" skill which promoted the 2023 Lexus LX.

> "The twenty-twenty-three Lexus LX is *more powerful than ever before. It features an all-new four-hundred-and-nine-horsepower twin-turbo engine ... Its all-new chassis delivers greater agility on-road, and a full-time four-wheel drive system improves capability off-road. And with four-hundred-seventy-nine pound-feet of torque, it can tow up to eight-thousand pounds.* Want to hear about LX F Sport Handling?" (B09X68P9Y9)

❸ **Skills that recommend multiple products based on user request.** These skills promote multiple related products based on what user requests. For instance, the "itcher" skill recommends books and movies, and also enables user to buy the book.

> "OK. You might like this fiction book set in a small town...*Things We Left Behind (Knockemout Series, 3). by Lucy Score. A print version.. It's $12.80.. With delivery by Friday, September 15.. Interested in learning more?*" (B06W5DW343)

❹ **Some businesses create a skill to advertise themselves.** These skills are developed to promote established businesses. E.g., Pawnbrokers of Rodeo Drive is a skill created by the business "Pawnbrokers of Rodeo Drive" who buy jewelry or other items and also provide loans. We found an advertisement of their business in their skill response shown below:

> "Here's the skill Pawnbrokers of Rodeo Drive, *by Axle Web Tech.. Hello and Welcome to Pawnbrokers of Rodeo Drive. We buy and provide loans against jewelry, watches, diamonds, art, wine, luxury cars, antiques, and other fine personal assets.* How can I help you? You can say things like, Pawn my watch, Sell my watch or How to contact you." (B07Y6FL8LB)

❺ **Skills made to provide suggestions at a specific location.** Some skills are made to provide suggestions about something at a specific location. For example, "Bucara bombi" gives suggestions about places to visit for good food.

> "Welcome to Bucara Bombi!! Here's your fact: If you are planning to go out for drinks, *you should try "El Propio". This restaurant and bar serves fantastic cocktails*" (B07KJHS2F9)

TABLE 2: Clusters of ads grouped together based on common themes along with the cluster size and avg. cosine similarity within each cluster. An example from each cluster group is also provided.

Cluster IDs	Avg. cos simlarity	% of ads	Description	Example
0	0.42	20.16%	Music streaming and audiobooks	You can listen to exclusive music from the broken sheep on spotify, itunes, and amazon music purchase the broken sheeps exclusive N F T on opensea Want some more?
6	0.23	17.37%	Community engag-ament, and religious services	The Awesome Foundation is a global community advancing the interest of awesome in the universe $1000 at a time. Each chapter supports awesome projects through one-thousand dollar micro-grants, usually given out monthly, no strings attached.
3, 13	0.88, 0.92	13.08%	Health and well being	Private yoga therapy sessions empower you to discover simple yoga-based practices ... the yoga therapy model of health takes into consideration every aspect of a person's life, rather than only looking at one symptom or condition in isolation. Private sessions start with an initial consultation, then follow-up sessions, as well as at home practices. Visit our website online to explore pricing, and If price is keeping you from achieving your wellness goals - please do not hesitate to reach out to us. At Explore, we believe that cost should not keep anyone from leading a healthy and happy life.
8	0.68	9.54%	Home improvement, and personal services	Welcome to Loyal Home Services! You can count on our top-notch services and competitive prices to get the job done to your satisfaction. Call us at 210-985-7186 to avail the best HVAC services in San Antonio, TX, and nearby areas.
4, 14	0.81, 0.72	9.34%	Trivia, fun games and entertainment	Welcome to Hit Trap Music Trivia! Let's play a game. One to four players can play. How many are playing?
9, 11	0.78, 0.83	8.66%	Food and dining services	No trip to Atlanta would be complete without visiting a quintessential Midtown restaurant. Please tell me if you would like more information about Steamhouse Lounge, Park Tavern, or The Vortex.
5, 10, 15	0.12, 0.9, 0.84	7.54%	Self-care, meditation and spiritual services	Millennium Yoga is a company in Fremont, California. It is run by Shubhangi Kulkarni, who is a certified yoga instructor and a specialist in therapeutic yoga. Shubhangi is a trained personal coach and enjoys working with people on the topics like wellness, health and relationships.
7	0.56	5.66%	Small and local businesses	Ginos is a family owned and operated restaurant, located in Spring hill and serving Hernando County Florida, Since 2010 we have been recognized as one of the best places for a quick, reasonably priced, and tasty breakfast or lunch.
12	0.9	2.82%	Travel and tourism	Welcome to Seattle Ballooning Come experience epic views of Mt. Rainier from 5,000 feet. We have 25/7 chat available on our website at Seattleballooning.com.
1	0.39	2.28%	Educational tools and resources	Here's a fact: Preparing your students for the star test and feel lost on what to do? Search STAAR Writing Spiral on you tube for helpful videos. Want to hear some more?
16	0.81	1.9%	Real Estate Services	Here's a fact: ... Charles Bianco has scrutinized literally over 500 pre approval letters during his current career. He would easily estimate that 30 per cent of all pre approvals he's scrutinized was insufficient for a buyer to actually purchase the home. To learn more, contact Charles Bianco at 516-444-5341 Want to hear some more?
2	0.76	1.63%	Online dating	Dating can be tricky. But you aren't alone. Our team at Three Day Rule has got your back. For the next 30 days, you will be given tips and activities especially designed by TDR's expert team.

❻ **Skills that present a menu for a service or business.** We also saw skill examples that do not follow a regular flow of a dialog but immediately present a menu of services and even sometimes with prices. e.g., "stylish ambi" is the skills created by a beauty business:

> "*Stylish Ambi SERVICE LIST: Eyelash Extension $200 Eyelash Lifting(perm) $80 Eyelash Tinting $40 EyeBrows Tinting $25 ... for more detail visit stylish ambi dot com*" (B07RMDR1GK)

5.3.2. Skill-Building Agencies. Numerous third-party agencies, including advertising firms, provide skill-building services for brands seeking to promote their products or services through Alexa. Alexa suggests to hire these agencies to "build a voice experience that extends your brand's reach and deepens customer engagement" [77]. The US Alexa Skill Kit website features 43 such agencies [78]. Our analysis also revealed ads in skills developed by these agencies (the number of ads detected is listed in parenthesis), including VoiceXP (10) [79], Wunderman Thompson Mobile (10) [80], Cognizant Technology Solutions (2) [81], and Orbita (1) [82]. Among them are also advertising and marketing agencies such as Bluefin (19) [83], Vixen Labs (12) [31], XAPPmedia (5) [84], Isobar (4) [85], iStrategyLabs (1) [86], Linc Global (3) [87], RAIN (2) [88], Skilled Creative (1) [32], and VaynerMedia (1) [89]. Among these agencies and online advertisers, there are some who specifically focus on ads on voice app platforms such as Bluefin [83], Easy Voice [90], Skilled Creative [32], SpokenLayer [91], Voices.com [92], VoiceXP [79] and Volara [93]. These voice platform-specific advertising agencies indicate the growing popularity of voice-based advertisements through voice assistants. Alexa seeks to expand skill-based advertising and encourages such agencies to join the platform, stating, "Agencies are creating innovative Alexa skills so that brands can reach their customers via Amazon Alexa, Echo, Echo Dot, and

Fire TV. If you think your company has the expertise to assist clients in creating skills, please contact us" [77].

Takeaway. Our analysis reveals that 13.4% (6,095 out of 45,477) of the skills contain ads spanning various topics, with the most common topics being music streaming and audio books, accounting for 20.16% of the 11,335 detected ads. Additionally, we identified ads in skills created by 12 advertising agencies and skill-building companies that collaborate with Alexa, including some agencies exclusively dedicated to voice assistant platform advertising.

5.4. Policy-violating Ads in Voice Apps

In Section 2, we demonstrated that the vetting process of Alexa could be bypassed to run policy-violating ads. We wanted to see if other policy-violating ads exist in the ecosystem (**RQ4**). According to Alexa's ad policy, an ad is considered policy violating if the skill in which it appeared is not tailored to order or promote the product, or if the ad is rendered without an explicit user request [24]. Our system flagged 29.18% (3,308 out of 11,335 in total) of detected ads as potentially policy-violating. These policy-violating ads came from 2,059 unique skills out of the 6,095 total skills containing ads.

We performed manual validation for the compliance subsystem to evaluate its accuracy. While validating the ad detection phase accuracy, the same researchers also validated the compliance decision of the model in case the responses were detected to contain an ad. This was done in parallel to ad validation, and researchers marked the responses as "compliant", "non-compliant,". Inter-rater reliability (Cohen's $\kappa = 0.90$) showed a high level of agreement. In contrast to ad detection verification, where the response text itself was enough to validate, compliance validation required the researchers to look at the skill response in the context of the user request (that generated the response) and the skill metadata (e.g., skill name, developer name, and description). This is because, according to Alexa ad policy, the *context* of the skill determines whether an ad is compliant or not. In the compliance check phase, similar to the ad detection phase, we also instructed the model to provide a brief reasoning of why an ad is compliant or not. During validation, we also consider the explanation of the model and determined manually if the explanation is sound based on the ad policy to determine if the label is correct or not. For compliance validation, there were a few cases where the researchers themselves were unsure if the ad may be compliant or not. For example, the skill "Precision Paintless Dent Repair Facts" (B0BG7DG4S3) is supposed to give users facts about dent repair, but rather, it just responds with a predefined message which promotes their business of dent repair and suggests to call them for a free quote. The response text from the skill is as follows: *"Precision Paintless Dent Repair can give instant free estimates over the phone. Know what your dent repair will cost in minutes! Car dent repairs can be repaired inside Precision Paintless Dent Repair's gorgeous repair shop in Port Washington, Long Island, NY. Or they can come to you with their mobile dent repair truck. Want some more?"*. The skill has a similar name as the business, but the purpose was to provide facts and not to advertise. We eventually marked it as compliant during validation because we considered it related to the skill to some extent, and followed a *conservative* approach while labeling potentially policy-violating ads. The validation precision and recall for non-compliant skill responses were 88.33% and 94.58%, respectively. For compliant ads, validation precision and recall were 98.28% and 96.12%, respectively.

We also review false positives and false negatives to understand where the model misclassifies. These errors often occur in nuanced cases. For example, the skill "Cookie Cutter" (B01N0IFGI0) states, "The world's biggest cookie on record was created by the Immaculate Baking Company in 2003," which the model flags as non-compliant due to the brand name mentioned, even though it aligns with the skill's description. Conversely, some promotional skills disguised as general information are not detected, such as "This is the Stoval Center for Entrepreneurship, a hub for innovation and global impact" from the skill "Afternoon Stoval Inspirations" (B0815B5D53), which promotes the center despite an unrelated description.

Characterization of Policy-violating Ads. Here, we characterize the different patterns in which skills deliver potential policy-violating ads. We found these themes in the ads that we manually validated, thus the patterns that we observed here are not exhaustive; however, we found multiple instances of these ads patterns across distinct skills. The skills and the policy-violating ads that we detected from their skill responses are presented in Table 3.

V1 Skills that promote companies, services, or products that are not related with the skill and its functionality. These are the skills that promote products or services in a very unsolicited manner. For example, a skill, "Istanbul Stories" (B07K5F1PKN) provides facts and stories about Istanbul; however, we found that it also advertises a WiFi service. The skill named "Healthy cooking" (B07YNL4LYR) was created to tell users about healthy cooking recipes, but we found it to advertise IT and computer support services. Such advertisements that are unrelated to the skills are not allowed by the Alexa ad policy [24].

V2 Skills that have a particular functionality but they promote a company or product that is loosely related. These are skills that appear to be made for entertainment, but they run ads that are not expected based on the functionality. For example, a skill, "Dating Tips" (B07KD2XWHS) says in the description that it provides "cool tips about dating" but we detected an ad from it promoting a dating website, "match.com". The ad text is presented in Table 3. Interestingly, we also found the same ad in another skill named "Your Happiness Now" (B07JY8DBY3).

Another example of this type came from a skill named "At-Risk Teen Tips" (B08LVZM2RP) which is made for teenagers to give life tips. Along with providing tips, it mentions a book and its author and says that the tips are inspired by it: *"Here's your tip ... Have an optimistic vision*

TABLE 3: Skills found running policy-violating ads. The skill responses containing the policy-violating responses are also shown along with their name and the respective skill IDs. The promotional content in the responses is highlighted in **red**.

Violation Type	Skill ID & Skill Name	Response Text	LLM explanation
V1	B07K5F1PKN Istanbul Stories	Here's your short story: Easiest way to get online in Istanbul is to order wifi service from alldaywifi.com and pickup your hotspot at the airport on-arrival. Enjoy daily unlimited wifi for less than $10 a day.	The ad is not compliant because it promotes a specific product and service (wifi service from alldaywifi.com) that is unrelated to the provided skill name, developer name, or skill description.
V1	B08W2DKBD9 Digital Planner: Mandala Coloring Kit	Get any of this amazing products now Viral Marketing 2.0 Made Easy Buy Viral Marketing 2.0 Made Easy right on https://gainmasteryacademykay.academiyo .com/viralmarketing-2.0/front-end . We got amazing Mandala coloring book kit for you. It is Affiliate Profitz Bazooka Buy it right on this site https://sites. google.com/view/kays-inner-circle-profitz/home. Thanks for...	The ad is not compliant with Alexa's skill policy. The ad in the skill response promotes products and websites that are unrelated to the provided skill, its functionality, skill name, developer name, and skill description.
V1	B07YNL4LYR Healthy cooking	... 3. Hung Curd Sandwich Choose the recipe you want to learn, try it out and enjoy. Say Sandwich, banana walnut or granolabar to continue. sponsored by System Network and Design, LLC 404-968-8288 Call and discover how great local IT services and computer support can be!	The ad is not compliant with Alexa's skill policy because the entity being advertised (System Network and Design, LLC) is not related to the skill, its functionality, skill name, developer name or skill description
V2	B07KD2XWHS Dating Tips	Here is another fact.. Match.com founder Gary Kremen boldly proclaimed that "Match.com will bring more love to the planet than anything since Jesus Christ." Sign me up! Do you want another fact ...	The ad is not compliant because it is promoting Match.com, which is not related to the skill skill "Dating Tips" or its functionality.
V3	B07ZDC1MRD my seven information	Finland and North Korea are separated by only one country The founder of Match.com dating site lost his girlfriend because of a guy she met on Match.com. ant some more?	The ad for Match.com is not compliant because it is promoting a different product/service that is not related to the skill, its functionality, skill name, developer name, or skill description.
V3	B07JY8DBY3 Your Happiness Now	...In their powerful book, Fearless Leaders: Sharpen Your Focus, Greenberg and North explain these principles to analyze some of the greatest leaders of our time and throughout history. For more information about Fearless Leaders please search for Fearless Leaders at Amazon.com.	The ad is not compliant with Alexa's skill policy because it promotes a book called "Fearless Leaders" which is not related to the skill "Your Happiness Now" or its functionality.
V3	B08T3SGWVX WSKG Classical	Ok, Here's WSKG Classical. Welcome to the WSKG Multi Verse. Support the music you love. Make a gift love. Your Public Radio dot org. Now playing WSKG Classical.	The ad is not compliant because it promotes a different entity (Your Public Radio dot org) that is not related to the skill, its functionality, skill name, developer name, or skill description.
V4	B07VBS9ZXW MovieFacts	because you used this skill you should also use spacefacts.,,,,, Hello, ask me to give you a fact about any different movie	The ad is not compliant because it promotes a skill ("space facts") that is not related to the skill, its functionality developer name, or skill description.

A skill webpage has the following URL format: `amazom.com/dp/<skill-ID>` for a given skill ID.

for your future. Dream of the life you desire and believe you can achieve it. These tips are inspired by the book Self-Help for At-Risk Teens by Author Ben Povlow. Would you like another tip?"

V3 Skills that promote another type of media such as other websites or content platforms. These skills are made for a specific purpose, but they also request users to go visit a website or have other calls to action. For example, a skill named "Andys Accordion Lesson" (B07P7MZLBT) is made to tell facts to users. The skill response says *"Here's a fact: If you go to andrew the malibu on you tube you can hear and see him playing a song or 2 Ready to hear some more?"*.

V4 Skills that promote other skills. We observed that skills sometimes request users to use other skills as well. For example, a skill named "MovieFacts" (B07VBS9ZXW) is designed to provide movie facts in response, but it requests users to "check out the Space Facts" skill. The skill and responses are presented in Table 3.

Other than these potential violations, we also observed skills asking users for a positive rating on Amazon, which is a clear violation of Alexa's skill policy [24].

Distribution of Non-Compliance Across Skill Category. The highest prevalence of non-compliant ads was observed in the "Health & Fitness" category (13.3% of total violations), followed by the categories of "Education & Reference " (12.96%), "Games & Trivia " (10.42%), "Lifestyle" (9.97%), and "Business & Finance" (7.98%).

Popularity of Skills with Policy-violating Ads. Alexa skills need to be "enabled" on an Alexa account before they can be utilized. However, the Alexa skills store does not of-

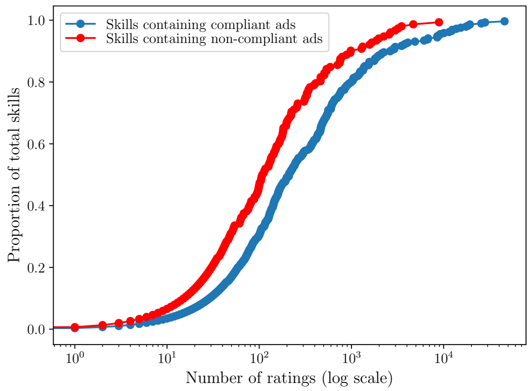

Figure 3: CDF of the number of ratings for skills containing compliant ads and non-compliant ads.

fer any usage metrics for skills, such as the number of users who have enabled a particular skill. Users can, however, rate skills on the skills store. Recent study on online consumer behavior revealed that the number of ratings received by a product is directly proportional to its popularity [94]. Therefore, we use the number of ratings as a proxy to estimate the popularity of skills that our models detected to contain ads, in order to assess their impact. We observed that the median number of ratings for skills containing ads and non-compliant ads is 235 and 107, respectively. Figure 3 illustrates the cumulative distribution function (CDF) of the number of ratings for skills containing only complaint ads, as well as skills containing only non-compliant ads. Interestingly, we found that skills containing non-compliant ads have comparable ratings, although they are statistically lower as determined by a *t-test* with $p < 0.05$. Nonetheless, skills containing non-compliant ads are present across all rating levels similar to compliant ones.

Takeaway. We discovered that a significant portion (2,059 out of 45,477) of skills potentially run non-compliant ads, amounting to 29.18% of the 11,335 ads detected. Skills featuring policy-violating ads also garner a comparable number of ratings to those with compliant ads, albeit statistically lower. Non-compliant ads commonly promote loosely related or unrelated brands, skills and other media types.

6. Discussion

Integration of Ad-compliance Module. We view our system as an enhancement to the current skill vetting process, aiming to improve robustness, particularly in ad policy compliance. Our ad detection and compliance checking system is not proposed as a replacement for the existing ad compliance checking process. Determining the compliance of an ad in a skill response is intricate, given the diverse forms and types of ads and the potential interpretations of ad policies. Our LLM-based detection system considers the Alexa ad policy in its compliance check, but its effectiveness relies on the clarity and detail in articulating the ad policy. When our model flags a response as potentially containing a policy-violating ad, a human operator familiar with the ad policy can make the final judgment and take necessary actions if a policy violation is indeed identified. As the compliance aspect of the model relies solely on the provided policy prompt (and not training), it can also easily adapt to future policy changes without the need for additional training. Moreover, further improvements to the pipeline can focus on minimizing human involvement, particularly in clear-cut cases. One potential approach is to incorporate a confidence score and define a threshold above which the model's predictions would not require human verification. However, our current pipeline does not adopt this strategy because we use LLMs, which do not natively provide confidence scores like traditional machine learning models. While it is possible to ask an LLM to estimate its own confidence, prior research has shown that such self-reported scores tend to be unreliable and often inflated, limiting their practical utility [95].

Voice app manufacturers can deploy our system in various ways. One approach is to retrospectively run the system on random skill responses currently active in the ecosystem, leveraging the manufacturer's vantage point in delivering skill responses to users' Alexa devices. Another method involves using an automated chatbot, similar to the one we employed to arbitrarily invoke skills published on the store periodically, running the model on conversation responses. Both approaches extend skill vetting beyond certification, ensuring voice apps maintain compliance throughout their lifecycle. However, it's crucial to note that our system is not suitable for live monitoring of Alexa responses during user interactions with a skill. The relatively slow prediction and text generation by large language models could adversely impact the user experience in real-time conversations. Our LLM takes 2 seconds on average for ad detection and compliance checking, and SkillDetective completes interaction in around one minute per skill on average, making it feasible for background periodic vetting.

Using Open-Source Large Language Models. This study compared OpenAI's GPT 4o/3.5 (proprietary) and Meta's Llama 3.2/2 (open source) LLMs for ad detection and compliance validation, finding GPT-4o to have superior performance. While GPT-4o offers higher accuracy, its proprietary nature and usage-based costs may not be ideal for commercial use. Open-source LLMs are preferred for large-scale commercial deployment due to their adaptability and task-specific training capability. Although hardware constraints often pose challenges in the use of open-source LLMs in a private setup, commercial deployment is not limited by such constraints. Since, larger open source LLMs can perform comparable to proprietary ones, such deployments are more practical in a commercial setting [96].

Better Ad Transparency on Voice Interface. Web and social media advertising has been extensively researched, but advertising in voice apps is a burgeoning field, especially with voice assistant manufacturers testing audio ads on their devices [97], [98], [99]. As ads make their way into the voice assistant ecosystem, accompanied by user tracking and ad targeting, a new advertising platform is emerging. Unlike the established ad transparency measures seen on the web and social media platforms (e.g., visual cues highlighting ad content), the appearance of such cues in voice apps remains unclear. This presents an opportune moment to establish guidelines for clearly distinguishing ad content from other content in voice apps before they become mainstream. Implementing such cues not only enhances user awareness but also lays the groundwork for developing an automated system to regulate and vet ads.

Ethical Disclosure. We sent the list of skills where we found potential policy violations to the Alexa team via their bug bounty program [100]. The report was acknowledged and we received a bug bounty [33].

Limitations. We identify a few limitations in our ad detection approach. First, while our model performs well in ad detection and utilizes the Chain-of-Thought (CoT) method to identify policy violations, it is not entirely accurate. The complexity of determining policy compliance necessitates human validation as a final step before any action can be taken regarding detected violations. Nonetheless, our system serves as an initial step, directing attention to a subset of app responses within voice app ecosystems that may potentially contain policy-violating ads. It can also contribute to the development of future policy-compliance validation systems. Second, despite enabling one skill at a time during automated skill interaction, we observed that it is still possible that a similar skill, sharing the invocation phrase as the one we enable, gets invoked. Hence, ads from those skills are labeled as non-compliant by our system since the intended skill's functionality does not match it. However, such cases were negligible as we explicitly enabled and disabled one skill at a time. Moreover, some skills were unavailable or did not work correctly during our interaction. Lastly, to detect ads in the wild, we leveraged a tool developed to interact with voice apps; however, it has its own limitations, such as not being able to identify certain question types or answer correctly, breaking the conversation flow as it has imperfect accuracy in carrying out skill interaction. So, it can miss specific execution paths of the skills and may not expose the full interaction graph. Furthermore, because voice apps can run ads at any point in the interaction, the automation tool we used may miss some of those ads. While our framework determines ad compliance with Alexa's policies (i.e., ground truth), it does not evaluate the actual maliciousness or misleading nature of the ad content for users. Identifying such content is complex due to the absence of definitive signatures of maliciousness and reliable ground truth data. Creating a robust maliciousness detection mechanism or obtaining suitable ground truth is a significant research endeavor, exceeding the scope of this paper.

7. Conclusion

We show that Alexa's skill vetting process is unable to detect policy-violating ads when injected dynamically after the voice app gets published. To tackle this, we develop an LLM-based system to detect ads in Alexa skill responses and use Chain-Of-Thought prompting to flag if the ad violates the platform's ads policy. Further, we find a significant number of skills running promotions and identify several examples of policy-violating ads running on live skills.

Acknowledgment

We thank our anonymous reviewers for their valuable feedback. This research is partially supported by the National Science Foundation (NSF) under grants CNS-2350075. The opinions, findings, conclusions, or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the funding organization.

References.

[1] F. Laricchia, "Us: Smart speaker installed base 2018-2022," Mar 2022. [Online]. Available: https://www.statista.com/statistics/967402/united-states-smart-speakers-in-households/

[2] "Number of digital voice assistants in use worldwide from 2019 to 2024 (in billions)*," 2024. [Online]. Available: https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/

[3] "How google assistant works with your data," 2022. [Online]. Available: https://support.google.com/assistant/answer/11091015

[4] B. Xiao, "How alexa can use song-playback duration to learn customers' preferences," Dec 2021. [Online]. Available: https://www.amazon.science/blog/how-alexa-can-use-song-playback-duration-to-learn-customers-preferences

[5] "5 ways consumers interact with smart speakers," Mindstream Media Group, 2021. [Online]. Available: https://mindstreammediagroup.com/introduction-smart-speakers-voice-search-brand-advertisers/

[6] T. Ammari, J. Kaye, J. Y. Tsai, and F. Bentley, "Music, search, and IoT: How people (really) use voice assistants," *ACM Transactions on Computer-Human Interaction*, vol. 26, no. 3, pp. 1–28, 2019.

[7] I. Lopatovska, K. Rink, I. Knight, K. Raines, K. Cosenza, H. Williams, P. Sorsche, D. Hirsch, Q. Li, and A. Martinez, "Talk to me: Exploring user interactions with the amazon alexa," *Journal of Librarianship and Information Science*, vol. 51, no. 4, pp. 984–997, 2019. [Online]. Available: https://doi.org/10.1177/0961000618759414

[8] A. Sciuto, A. Saini, J. Forlizzi, and J. I. Hong, ""hey alexa, what's up?": A mixed-methods studies of in-home conversational agent usage," in *Proceedings of the 2018 Designing Interactive Systems Conference*, ser. DIS '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 857–868. [Online]. Available: https://doi.org/10.1145/3196709.3196772

[9] "The smart audio report," Edison Research, 2020. [Online]. Available: https://www.nationalpublicmedia.com/uploads/2020/04/The-Smart-Audio-Report-from-NPR-and-Edison-Research.pdf

[10] Alexa ai — build LLM-powered Alexa experiences, 2024. [Online]. Available: https://developer.amazon.com/en-US/alexa/alexa-ai

[11] "Previewing the future of alexa," Amazon Alexa, 2023. [Online]. Available: https://developer.amazon.com/en-US/alexa/alexa-ai

[12] "Incredible amazon alexa statistics you need to know in 2021," 2021. [Online]. Available: https://safeatlast.co/blog/amazon-alexa-statistics/

[13] "Amazon alexa skills policy," Amazon Alexa, 2022. [Online]. Available: https://developer.amazon.com/en-US/docs/alexa/custom-skills/policy-testing-for-an-alexa-skill.html

[14] "Certification requirements," 2021. [Online]. Available: https://developer.amazon.com/en-US/docs/alexa/custom-skills/certification-requirements-for-custom-skills.html#submission-checklist

[15] V. Johnson, Z. Zhu, R. Anguera, J. Bollinger, J. Eccles, D. Hardtke, M. Breza, and T. P. Zanto, "Increasing brand awareness: Memory for short audio ads," *Psychology & Marketing*, vol. 38, no. 11, pp. 1960–1972, 2021.

[16] K. Park, Y. Park, J. Lee, J.-H. Ahn, and D. Kim, "Alexa, tell me more! the effectiveness of advertisements through smart speakers," *International Journal of Electronic Commerce*, vol. 26, no. 1, pp. 3–24, 2022.

[17] K. T. Smith, "Marketing via smart speakers: what should alexa say?" *Journal of Strategic Marketing*, vol. 28, no. 4, pp. 350–365, 2020.

[18] H. Lee and C.-H. Cho, "Uses and gratifications of smart speakers: Modelling the effectiveness of smart speaker advertising," *International Journal of Advertising*, vol. 39, no. 7, pp. 1150–1171, 2020.

[19] "Digital Audio Advertising - Worldwide," Statista, 2023. [Online]. Available: https://www.statista.com/outlook/amo/advertising/audio-advertising/digital-audio-advertising/worldwide

[20] U. Iqbal, P. N. Bahrami, R. Trimananda, H. Cui, A. Gamero-Garrido, D. J. Dubois, D. Choffnes, A. Markopoulou, F. Roesner, and Z. Shafiq, "Tracking, profiling, and ad targeting in the alexa echo smart speaker ecosystem," in *Proceedings of the 2023 ACM on Internet Measurement Conference*, ser. IMC '23, 2023, p. 569–583.

[21] "Alexa advertising id," 2023. [Online]. Available: https://developer.amazon.com/en-US/docs/alexa/advertising-id/overview.html

[22] "Audio ads. Your brand. Turned up," https://advertising.amazon.com/en-ca/solutions/products/audio-ads, 2024.

[23] A. Tang, "Alexa Blog. Advertising and Alexa," https://developer.amazon.com/blogs/alexa/post/54c3a0f8-5b29-4071-acd7-2b832b860c83/advertising-and-alexa, 2017.

[24] "Amazon alexa ads policy," Amazon Alexa, 2022. [Online]. Available: https://developer.amazon.com/en-US/docs/alexa/custom-skills/policy-testing-for-an-alexa-skill.html#5-advertising

[25] "Alexa advertising id policy," 2023. [Online]. Available: https://developer.amazon.com/en-US/docs/alexa/advertising-id/alexa-advertising-id-policy.html

[26] H. Hu, L. Yang, S. Lin, and G. Wang, "Security vetting process of smart-home assistant applications: A first look and case studies," 2020.

[27] D. Su, J. Liu, S. Zhu, X. Wang, and W. Wang, ""are you home alone?" "yes" disclosing security and privacy vulnerabilities in alexa skills," 2020.

[28] L. Cheng, C. Wilson, S. Liao, J. Young, D. Dong, and H. Hu, "Dangerous skills got certified: Measuring the trustworthiness of skill certification in voice personal assistant platforms," ser. CCS '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1699–1716. [Online]. Available: https://doi.org/10.1145/3372297.3423339

[29] Z. Guo, Z. Lin, P. Li, and K. Chen, "Skillexplorer: Understanding the behavior of skills in large scale," in *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020, pp. 2649–2666.

[30] C. Lentzsch, S. J. Shah, B. Andow, M. Degeling, A. Das, and W. Enck, "Hey alexa, is this skill safe?: Taking a closer look at the alexa skill ecosystem," in *28th Annual Network and Distributed System Security Symposium, NDSS*, 2021.

[31] "Vixen labs," Vixen Labs, Jan 2024. [Online]. Available: https://vixenlabs.co/

[32] "Skilled creative," Skilled Creative, Jan 2024. [Online]. Available: https://www.skilledcreative.com/

[33] "Alexa skill ad artifact," 2024. [Online]. Available: https://osf.io/cgp3j/?view_only=ebcf0b112ef24fa28a46a60f4f65d36b

[34] J. Young, S. Liao, L. Cheng, H. Hu, and H. Deng, "SkillDetective: Automated Policy-Violation detection of voice assistant applications in the wild," in *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/young

[35] F. H. Shezan, H. Hu, G. Wang, and Y. Tian, "Verhealth: Vetting medical voice applications through policy enforcement," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 4, dec 2020. [Online]. Available: https://doi.org/10.1145/3432233

[36] "Media advertising spending in the united states from 2020 to 2024," Statista, 2021. [Online]. Available: https://www.statista.com/statistics/272314/advertising-spending-in-the-us/

[37] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan, "Adscape: Harvesting and analyzing online display ads," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 597–608. [Online]. Available: https://doi.org/10.1145/2566486.2567992

[38] "Mobile ads: Here's what works and what doesn't," The Wall Street Journal, 2012. [Online]. Available: https://www.wsj.com/articles/SB10000872396390444083304578016373342878556

[39] E. Mao and J. Zhang, "What drives consumers to click on social media ads? the roles of content, media, and individual factors," in *2015 48th Hawaii International Conference on System Sciences*, 2015, pp. 3405–3413.

[40] M. A. Bashir and C. Wilson, "Diffusion of user tracking data in the online advertising ecosystem," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 4, pp. 85–103, 2018.

[41] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," ser. CCS '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1388–1401. [Online]. Available: https://doi.org/10.1145/2976749.2978313

[42] A. Razaghpanah, R. Nithyanand, N. Vallina-Rodriguez, S. Sundaresan, M. Allman, C. Kreibich, P. Gill *et al.*, "Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem," in *The 25th Annual Network and Distributed System Security Symposium (NDSS 2018)*, 2018.

[43] M. D. Ayenson, D. J. Wambach, A. Soltani, N. Good, and C. J. Hoofnagle, "Flash cookies and privacy ii: Now with html5 and etag respawning," *Available at SSRN 1898390*, 2011.

[44] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, "The web never forgets: Persistent tracking mechanisms in the wild," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014, pp. 674–689.

[45] K. Mowery and H. Shacham, "Pixel perfect: Fingerprinting canvas in html5," *Proceedings of W2SP*, vol. 2012, 2012.

[46] L. Olejnik, C. Castelluccia, and A. Janc, "Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns," in *5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2012)*, Vigo, Spain, Jul. 2012. [Online]. Available: https://hal.inria.fr/hal-00747841

[47] F. Roesner, T. Kohno, and D. Wetherall, "Detecting and defending against Third-Party tracking on the web," in *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*. San Jose, CA: USENIX Association, Apr. 2012, pp. 155–168. [Online]. Available: https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/roesner

[48] "New product alert: Interactive audio ads." Twitter, 2021. [Online]. Available: https://twitter.com/AmazonAds/status/1453030190867746823

[49] E. Cho and S. S. Sundar, "Should siri be a source or medium for ads? the role of source orientation and user motivations in user responses to persuasive content from voice assistants," in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3491101.3519667

[50] F. Laricchia, “Amazon alexa: Skill count in selected countries 2021,” Jan 2022. [Online]. Available: https://www.statista.com/statistics/917900/selected-countries-amazon-alexa-skill-count/

[51] Z. Guo, Z. Lin, P. Li, and K. Chen, “SkillExplorer: Understanding the behavior of skills in large scale,” in *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, Aug. 2020, pp. 2649–2666. [Online]. Available: https://www.usenix.org/conference/usenixsecurity20/presentation/guo

[52] 2023. [Online]. Available: https://www.ecfr.gov/current/title-33/chapter-I/subchapter-M/part-136/subpart-D/section-136.303

[53] “Ad guidelines,” Amazon, Jan 2024. [Online]. Available: https://advertising.amazon.com/resources/ad-specs/audio-ads

[54] “Openai platform - preparing your dataset,” OpenAI Platform, 2024. [Online]. Available: https://platform.openai.com/docs/guides/fine-tuning/preparing-your-dataset

[55] C. Shah, “From prompt engineering to prompt science with human in the loop,” 2024. [Online]. Available: https://arxiv.org/abs/2401.04122

[56] “Openai help center,” OpenAI Help Center, 2024. [Online]. Available: https://help.openai.com/en/articles/6783457-what-is-chatgpt

[57] “What is prompt engineering?” 2023. [Online]. Available: https://aws.amazon.com/what-is/prompt-engineering/

[58] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[59] “Openai platform - finetuning,” OpenAI Platform, 2024. [Online]. Available: https://platform.openai.com/docs/guides/fine-tuning

[60] “Gpt 4o,” OpenAI Platform, 2024. [Online]. Available: https://openai.com/index/hello-gpt-4o/

[61] “Gpt 3.5,” OpenAI Platform, 2024. [Online]. Available: https://platform.openai.com/docs/models/gpt-3-5

[62] “Meta llama,” 2024. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[63] “Ai at meta,” 2024. [Online]. Available: https://ai.meta.com/llama/

[64] “Gpt 3.5 finetuning,” Open AI, 2024. [Online]. Available: https://platform.openai.com/docs/guides/fine-tuning

[65] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” 2021.

[66] “Peft,” Hugging Face – The AI community building the future., 2024. [Online]. Available: https://huggingface.co/docs/peft/index

[67] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[68] “Precision recall curve.” [Online]. Available: https://encord.com/blog/classification-metrics-accuracy-precision-recall/

[69] S. Barnett, Z. Brannelly, S. Kurniawan, and S. Wong, “Fine-tuning or fine-failing? debunking performance myths in large language models,” 2024. [Online]. Available: https://arxiv.org/abs/2406.11201

[70] “What is clustering?” Google, June 2024. [Online]. Available: https://developers.google.com/machine-learning/clustering/overview

[71] F. Nielsen and F. Nielsen, “Hierarchical clustering,” *Introduction to HPC with MPI for Data Science*, pp. 195–211, 2016.

[72] H. Jin, Y. Zhang, D. Meng, J. Wang, and J. Tan, “A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods,” *arXiv preprint arXiv:2403.02901*, 2024.

[73] “Openai emebddings,” OpenAI, Jan 2024. [Online]. Available: https://platform.openai.com/docs/guides/embeddings/embeddings

[74] J. Ramos *et al.*, “Using tf-idf to determine word relevance in document queries,” in *Proceedings of the first instructional conference on machine learning*, vol. 242, no. 1. Citeseer, 2003, pp. 29–48.

[75] C. Emmery, 2024. [Online]. Available: https://cmry.github.io/notes/euclidean-v-cosine

[76] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.” *Journal of machine learning research*, vol. 9, no. 11, 2008.

[77] “Alexa skill kit,” 2024. [Online]. Available: https://developer.amazon.com/en-US/alexa/alexa-skills-kit

[78] “Find experienced skill builders in the united states,” 2024. [Online]. Available: https://developer.amazon.com/en-US/alexa/skill-building-agencies/agencies-us

[79] “Voicexp,” VoiceXP, Jan 2024. [Online]. Available: https://www.voicexp.ai/home

[80] “Wunderman thompson mobile,” Wunderman Thompson Mobile, Jan 2024. [Online]. Available: https://www.vml.com/

[81] “Cognizant technology solutions,” Cognizant Technology Solutions, Jan 2024. [Online]. Available: https://www.cognizant.com/us/en

[82] “Orbita inc.” Orbita Inc., Jan 2024. [Online]. Available: https://orbita.ai/

[83] “Bluefin,” Bluefin, Jan 2024. [Online]. Available: https://www.bluefintechnologypartners.com/alexa-skill-development-services/

[84] “Xappmedia, inc.” XAPPmedia, Inc., Jan 2024. [Online]. Available: https://xappmedia.ai/

[85] “Isobar,” Isobar, Jan 2024. [Online]. Available: https://www.dentsu.com/us/en/who-we-are/our-agencies/isobar

[86] “istrategylabs,” iStrategyLabs, Jan 2024. [Online]. Available: https://isl.co/

[87] “Linc global,” Linc Global, Jan 2024. [Online]. Available: https://www.linc.cx/

[88] “Rain,” RAIN, Jan 2024. [Online]. Available: https://rain.agency/

[89] “Vaynermedia,” VaynerMedia, Jan 2024. [Online]. Available: https://vaynermedia.com/

[90] “Easy voice,” Easy Voice, Jan 2024. [Online]. Available: https://easy-voice.com/

[91] “Spokenlayer,” SpokenLayer, Jan 2024. [Online]. Available: https://www.spokenlayer.com/

[92] “Voice.com,” Voice.com, Jan 2024. [Online]. Available: https://www.voices.com/projects/voice-applications

[93] “Volara,” Volara, Jan 2024. [Online]. Available: https://uniguest.com/volara/

[94] D. Powell, J. Yu, M. DeWolf, and K. J. Holyoak, “The love of large numbers: A popularity bias in consumer choice,” *Psychological Science*, vol. 28, no. 10, pp. 1432–1442, 2017. [Online]. Available: https://doi.org/10.1177/0956797617711291

[95] Y. Pawitan and C. Holmes, “Confidence in the reasoning of large language models,” *Harvard Data Science Review*, vol. 7, no. 1, 2025. [Online]. Available: https://hdsr.mitpress.mit.edu/pub/jaqt0vpb/release/2

[96] W. Xiong, J. Liu, I. Molybog, H. Zhang, P. Bhargava, R. Hou, L. Martin, R. Rungta, K. A. Sankararaman, B. Oguz *et al.*, “Effective long-context scaling of foundation models,” *arXiv preprint arXiv:2309.16039*, 2023.

[97] “Amazon ads,” 2023. [Online]. Available: https://advertising.amazon.com/solutions/products/audio-ads

[98] E. H. Schwartz, “Amazon makes alexa audio ads more interactive,” Oct 2021. [Online]. Available: https://voicebot.ai/2021/10/29/amazon-makes-alexa-audio-ads-more-interactive/

[99] T. Kerns, “Google assistant now shows ads in some answers,” Apr 2019. [Online]. Available: https://www.androidpolice.com/2019/04/06/google-assistant-now-shows-ads-in-some-answers/

[100] “Amazon vulnerability research program - devices,” Hackerone, Jan 2024. [Online]. Available: https://hackerone.com/amazonvrp-devices?type=team

Appendix A.
Publishing Policy-violating Ads

Research has shown gaps in the Alexa's vetting process, revealing instances of skills collecting sensitive user data [29], [30] or violating content guidelines [28], [34], [35]. As a proof-of-concept, we illustrate how malicious skills can persistently run policy-violating ads. We created skills that incorporate both our *fictitious ad* and legitimate content. We obtained IRB approval for deploying skills running policy-violating ads (by demoing to the IRB). Our skills also included a message at the end of ads stating they are fictitious and used only for research purposes.

A.1. Curating Policy-Violating Ads

Alexa has policies prohibiting certain types of content, including ads. The policy states, *"Your skill will be rejected or suspended if it includes or otherwise surfaces advertising or promotional messaging in skill responses, notifications, or reminders"* [24]. However, there are some exceptions where ads are allowed, such as: (i) music/audio streaming services, (ii) if the skill allows customers to order products or services, it can contain ads for those products or services, (iii) a skill can include audio messaging informing customers of promotional offers or deals in response to specific requests explicitly made by customers, (iv) if the skill is specifically designed to promote a product or service. More details of such exceptions are stated in Alexa's policy web page [24].

To test Alexa's ad validation, we first created a Joke telling skill that violated Alexa's advertising policy. We then curated five *fictitious* ads that did not meet any exception criteria. The five ads used for this work are:

Ad-1: By the way, are you planning to buy a new phone? Samsung has recently launched all new Galaxy S22 Series. Head to samsung.com to order yours.

Ad-2: Hey! Are you looking for a new car? There is no better option than the Toyota Corolla. Head to the closest dealership to order yours.

Ad-3: By the way, do you want to be updated about latest tech news? Follow The Verge on Facebook and Twitter.

Ad-4: By the way, are you looking to host a cloud solution to host your app? Microsoft Azure have got you covered! Host your app for free on Azure platform.

Ad-5: Hey, do you know that you can get phone screen protector for as low as five dollars? Order ABC protectors through Amazon now if you need one.

Verifying violating ads. We leveraged the current vetting process of Alexa to ensure that our fictitious ads violated the Alexa ad policy by submitting five skills, each including one of the ads within the skill responses. In all cases, we received an email stating that our skill included an ad not allowed by Alexa (see OSF artifact repository [33]). Note that we consider an ad to be policy-violating if the skill

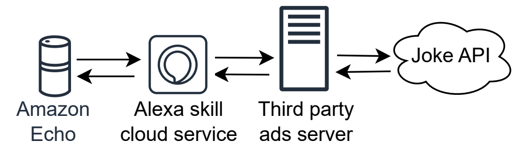

Figure 4: The skill requests jokes through the third-party server that also injects ads.

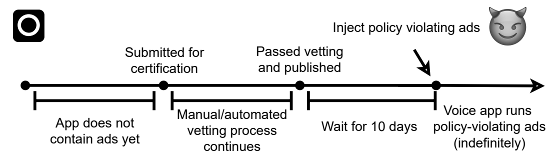

Figure 5: The process of evading skill vetting process to run policy-violating ads after passing the certification.

fails the vetting process and the reason for failure states the presence of policy-violating ads in the skill response.

A.2. Injecting Policy-Violating Ads

We resubmitted the skill after removing ads and also duplicated the skill five times with identical functionality for consistency. All of the new five skills passed the certification process and were published on the US Alexa skill store. To subvert the vetting process, we dynamically controlled when to inject ads within skill responses. Although the skill backend code is not visible to Alexa [30], we still did not hardcode the policy-violating ads in the backend code. Instead we hosted another server that our skill's backend contacted to get its content. Figure 4 shows our experimental setup to inject ads in skill responses. All skills with hidden ad injection capabilities passed the certification process. Following a 10-day waiting period (empirically chosen), we initiated ads on live skills, excluding the control app.

We continued to run ads on all copies of skills. After *three months* of running the policy-violating ads continuously, we concluded that our skills successfully evaded the vetting mechanism. Then we removed all skills from the live stores for ethical reasons so that no real users continue to use these skills despite the fact that they are harmless as they promote existing products (not misleading users) and were followed by a disclaimer. Figure 5 shows the timeline of injecting ads to evade Alexa vetting process.

Appendix B.
Prompts Used

TABLE 4: Prompt for ad detection

Description	Prompt components
Task description	I will provide you some rules to classify an Amazon Alexa skill response text into "ad" or "no-ad" where ad means the text contains an advertisement and no-ad means the text does not contain any advertisement. Here are the rules:
Rules	1. If the text talks about a company or product or a service and describe its functionality or qualities then it is an advertisement . For example, "Try AF Devs, which is right in poonta caughna village a full service multimedia group, that distributes high end audio and video brands to tailor your home with. Making homes smarter and more enjoyable since 2008., their phone number is, 829-899-2757, I have sent these details to the Alexa App on your phone. Time to get some gadgets" contains an advertisement. 2. It contains the wording such as "opening <skill name>by a <developer name>" or "handing off to <skill name>by <developer name>" or "<skill name>from <developer name>". A text which tells an app is being made "by" or "provided by" an entity when entity name is provided. For example "Handing off to Unofficial WhatsApp Jokes by Wisdom Is Principal Apps. Hello there. Welcome to unofficial whatsapp jokes. What is you first name please?" contains an advertisement. 3. If a text asks for an action such as to visit a website or download an app, or call a number. For example "Evergreen Wealth Advisory voice application does not know the answer to that question. Please visit https://evergreenwealthadvisory.com/ or give us a call at (587) 480-7476 and we will gladly answer your question. What else can I help with?" contains an advertisement. 4. If a text has a welcome message from a company or app and also describe its services or products and advertises itself. For example, "Welcome to Prisom Technology LLP. A UX & UI design, Game App & Amazon Alexa Skills Development Global IT Company. What do you want to know about Prisom? Select option from. First, Service. Second, About Us. Third, Contact Us." contains an advertisement. 5. If a text appears to promote or advertise a product, service, company, a place or a brand then it contains an advertisement. 6. Any text not containing an advertisement is a "no-ad".
Output format	If there is an advertisement say "Positive". Then provide a one line brief explanation of why the text contains an advertisement. But if there is no advertisement then say "Negative". Then provide a one line brief explanation of why the text does not contain an advertisement.
Response marker	Here is the skill response: <Response text>

TABLE 5: Prompt for compliance check

Description	Prompt components
Task description	Since the skill response contains an advertisement, I will now provide you a name, developer name, and description of the Alexa skill and your task is to identify if the advertisement that is there only in the previously provided "skill response" complies with the Alexa skill ads policy which is stated below:
Ads policy	Alexa Ads policy: Skill and the ads is not complaint if it includes or otherwise surfaces advertising or promotional messaging in skill responses, notifications, or reminders. There are specific exceptions we will allow in skill responses: 1. Streaming music, streaming radio, podcast, and flash briefing skills may include audio advertisements 2. Skills that allow customers to order products or services may include audio messaging promoting those products or services. 3. Skills may include audio messaging informing customers of promotional offers or deals in response to specific requests from customers. 4. Skills that are specifically designed to promote a product or service may include audio messaging promoting that product or service. 5. Skills may use Alexa for Apps to promote products or services in response to specific requests from customers but cannot accompany responses with unsolicited promotions or direct marketing.
Policy end marker	Alexa Ads Policy ends here.
Output format	Instructions to answer: Say "No" if the advertisement is not compliant with Alexa's skill policy. Say "Yes" if the advertisement is compliant with Alexa's skill policy.
Additional instructions	Please mark the advertisement as non-compliant if and only if an entity that is being advertised in the skill response is strictly not related to the provided skill, its functionality, skill name, the developer name and the skill description.
Explanation output instructions	If it is promoting or advertising the skill name or developer or skill in general then it is compliant. After saying the answer, explain as well in one line why the advertisement is compliant or why it is not compliant. Providing a one line brief explanation is necessary.
Skill metadata markers	Here is the skill name, developer name and description in quotes, respectively: Name: "{name}" Developer name: "{developer}" Description: "{description}"
Request/response marker	The response was generated in result of this user request: "{request}" Here is the response again: "{response}"

Appendix C. Meta-Review

The following meta-review was prepared by the program committee for the 2025 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

C.1. Summary of Paper

This paper presents an automated technique for identifying advertisements in Amazon Alexa skill responses and flagging potential policy violations. The authors build on a large-scale dataset of Alexa skill interactions, use a fine-tuned LLM for ad detection, and then apply chain-of-thought prompting to assess policy compliance. Their study reveals that over 13% of skills contain ads, and nearly 29% of these ads potentially violate Amazon's policies.

C.2. Scientific Contribution

- Creates a New Tool to Enable Future Science
- Addresses a Long-Known Issue
- Identifies an Impactful Vulnerability
- Provides a Valuable Step Forward in an Established Field

C.3. Reasons for Acceptance

- Identifies an Impactful Vulnerability: multiple reviewers highlight that the paper uncovers policy-violating ads in Alexa skills, underscoring a potential security, privacy, or trust risk for users.
- Creates a New Tool to Enable Future Science: the authors' fine-tuned LLM-based framework for ad detection is novel, adaptable, and can be extended to related problems.
- Provides a Valuable Step Forward in an Established Field: as voice assistants grow in popularity, understanding and mitigating non-compliant advertising remains a pressing concern. The paper's measurement study is the first large-scale effort in this domain.
- Addresses a Long-Known Issue: reviewers note that advertisements and policy noncompliance on voice platforms have been suspected but not systematically quantified until now.
- Provides a New Data Set for Public Use: the authors' commitment to releasing their dataset and code will allow independent confirmation and follow-on research.