Investigating Users' Understanding of Privacy Policies of Virtual Personal Assistant Applications

Baiqi Chen
The University of Queensland
CSIRO's Data61
Australia

Tingmin Wu
CSIRO's Data61
Australia

Yanjun Zhang
University of Technology Sydney
Australia

Mohan Baruwal Chhetri
CSIRO's Data61
Australia

Guangdong Bai
The University of Queensland
Australia

ABSTRACT

The increasingly popular virtual personal assistant (VPA) services, e.g., Amazon Alexa and Google Assistant, enable third-party developers to create and release VPA apps for end users to access through smart speakers. Given that VPA apps handle sensitive personal data, VPA service providers require developers to release a privacy policy document to declare their data handling practice. The privacy policies are regarded as legal or semi-legal documents, which are usually lengthy and complex for users to understand. In this work, we conducted a subjective study to investigate the level of users' understanding of the privacy policies, targeting the VPA apps (i.e., *skills*) of Amazon Alexa, the most popular VPA service. Our study focused on technical terms, one of the greatest hurdles to users' understanding. We found that 84.2% of our participants faced difficulty in understanding technical terms appeared in the skills' privacy policies, even for participants with IT background. Additionally, 64.3% of them reported that explanations for the technical terms are generally lacking. To address this issue, we proposed two principles, i.e., *domain-specificity principle* and *implication-oriented principle*, to guide skill developers in creating easy-to-understand privacy policies. We evaluated their effectiveness by creating explanation sentences for 23 representative terms and examining users' understanding through a second user study. Our results show that using explanation sentences based on these principles can significantly improve users' understanding.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**.

KEYWORDS

Privacy compliance, privacy policy, user study

ACM Reference Format:
Baiqi Chen, Tingmin Wu, Yanjun Zhang, Mohan Baruwal Chhetri, and Guangdong Bai. 2023. Investigating Users' Understanding of Privacy Policies of Virtual Personal Assistant Applications. In *ACM ASIA Conference on Computer and Communications Security (ASIA CCS '23), July 10–14, 2023, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3579856.3590335

ASIA CCS '23, July 10–14, 2023, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0098-9/23/07…$15.00
https://doi.org/10.1145/3579856.3590335

1 INTRODUCTION

The enormous popularity of virtual personal assistant (VPA) services, e.g., Amazon Alexa and Google Assistant, has brought about a promising ecosystem. Third-party developers can create VPA apps (e.g., *skills* in Amazon Alexa and *actions* in Google Assistant), and release them through app stores. Users can then access them by simply saying a command (or *utterance*) like "*Alexa, open <app name>*" to their smart speakers. The openness of this model may breed user privacy violations though. Recent studies [12, 21, 23, 49, 52] have shown that dishonest VPA apps may stealthily collect or even query for user's private data like addresses, names and locations.

Personal data protection has raised global concerns in recent years. Many countries have put in place stringent data protection regulations such as the well-known European Union (EU) General Data Protection Regulation (GDPR) [2] and California Consumer Privacy Act (CCPA) [7]. They impose obligations on entities (i.e., *data controllers* and *data processors*) that collect, use and share user data. Data controllers and data processors are required to disclose their data handling practices to users (i.e., *data owners*) in a transparent manner. Failures to follow these data legislations may lead to a huge penalty. In response, VPA service providers require the developers to release a privacy policy for each app to disclose their data handling practices [5].

Privacy policies are recognized as legal or semi-legal documents, and are often lengthy and formal. This renders their readability weak for users without technical and legal backgrounds. Several prior studies [37, 39, 44] have revealed that due to their difficulty of understanding, most users simply skip them before installing apps. As a result, developers become indolent and provide low-quality documents, or even put in statements that are inconsistent with their data handling practices. Indeed, a recent study [49] finds that many skills fail to disclose the data they collect from the users, including child users that are under special protection by regulations like the Children's Online Privacy Protection Act (COPPA) [1].

In this work, we conducted a subjective study to investigate users' comprehension of the privacy policies, taking as our target skills of Amazon Alexa, the most popular VPA service. Our study

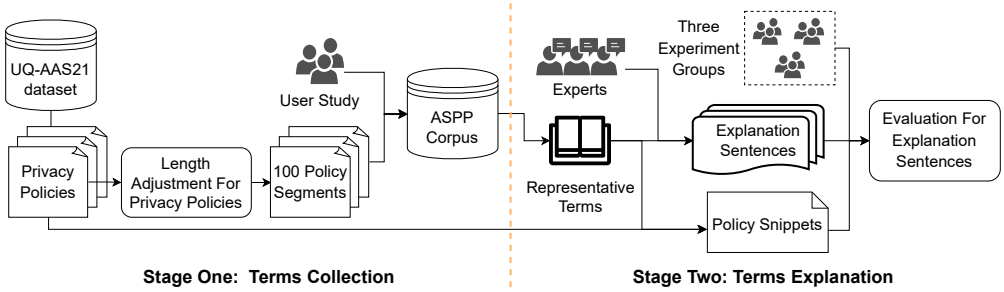


Figure 1: Overview of our approach of two inter-dependent stages.

focused on the effect of the words, phrases and jargons (summarized as *technical terms*), the common hurdles of users' understanding, as revealed by prior studies on general website privacy policies and security-related documents [43, 47]. We followed a two-stage methodology as illustrated in Fig. 1.

Stage One (Terms collection). Stage One aimed to identify technical terms in the context of VPA services, where specific personal data (e.g., voice, audio and sensors) may be involved. We collected technical terms and their difficulty level from users' perspectives, and then studied how the use of identified technical terms affects users' understanding. To this end, we designed a user study that explores the following (first) two research questions (RQs).

RQ1. What are considered as technical terms from the users' perspective? How well do users understand these technical terms?

RQ2. Why are technical terms difficult for users? Does users' understanding of technical terms relate to their IT background?

We obtained the privacy policies of top ranked skills from an existing dataset [48], as the data source of our study. We then recruited 311 participants to take part in this study, where we asked them to annotate and rate technical terms in an assigned policy segment. We finally collected 2,740 technical terms from those 100 policy segments to create the *Alexa Skill Privacy Policy Corpus* (*ASPP Corpus*).

Key findings in Stage One. Our study in this stage has revealed several key findings as summarized below:

- 30.73% of terms in the ASPP Corpus are *broad terms*, while 17.66% terms in the ASPP Corpus can be considered as *domain-specific terms*. Among them, the use of certain terms might violate CCPA and GDPR regulations.
- 62.38% of the participants indicate that the *broad* and *vague terms* in the Alexa skills' privacy policies are difficult to understand.
- Participants with IT background assign a *higher* difficulty level for difficult terms than participants without IT background, especially for terms in the "Special/Region Law" category.
- 64.3% of the participants suggest that skills' privacy policies should include relevant context or explanations for technical terms in order to improve their understanding.

Stage Two (Terms explanation). Based on the last finding of Stage One, we conducted a follow-up study to examine whether adding explanation sentences can improve users' understanding of technical terms, and in turn the privacy policies. We proposed two principles for creating explanations, namely *domain-specificity principle* and *implication-oriented principle*, based on related studies [22, 36, 40] and the textual feedback from our participants in Stage One. To evaluate the effectiveness of the explanation sentences composed based on our principles, we conducted a study of 150 participants with and without IT background, aiming to answer the following two RQs.

RQ3. How well do users understand the provided explanations based on different principles? Do the explanation sentences improve users' understanding?

RQ4. Does users' understanding of explanation sentences relate to their IT background?

Key findings in Stage Two. The Stage Two study reveals some key findings as summarized below.

- Explanation sentences can significantly improve users' understanding of the skills' privacy policies.
- Participants with IT background can understand the explanation sentences and related terms and policies better than participants without IT background.

Contributions. Our main contributions are summarized below.

- **The first study on users' understanding of VPA apps' privacy policies especially for technical terms.** We conducted a subjective study to investigate users' understanding of technical terms within the Alexa skills' privacy policies and the difficulties they face. This is the first study in this domain.
- **A corpus of hard-to-understand terms.** Our study yielded the ASPP Corpus, which includes the terms that users find hindering their understanding of policies. We also characterized them and unveiled the factors that make them hard to understand.
- **Principles for composing explanation sentences.** We proposed two principles for composing explanation sentences, to facilitate developers to enhance users' understanding of their privacy policies.

Ethics consideration and availability. We note that our study has been reviewed and approved by The University of Queensland Research Ethics and Integrity (2022/HE000695) and CSIRO Social and Interdisciplinary Human Research Ethics Committee (Ethics Clearance 098/22). Our artifacts have been made available at https://github.com/UQ-Trust-Lab/VPAPPUsability.

2 RELATED WORK

In this section, we briefly discuss related work on the general readability of privacy policies, misconceptions of technical terms, and privacy policies of VPA applications.

Readability of Privacy Policies. A number of prior studies have evaluated the readability of online privacy policies and concluded that it is an ongoing issue [10, 13, 14, 16, 26, 32, 42]. Fabian et al. [14]

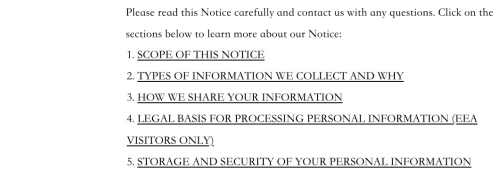
evaluated the privacy policies of 50,000 English-speaking websites and concluded that most privacy policies are difficult to read for general users. Sumeeth et al. [42] reported that most online privacy policies are beyond the capability of general online users, while some can only be understood by users with a post-graduate or higher degree. Singh et al. [32] conducted a user study with 50 participants to investigate their comprehension of privacy policies and found that, contrary to companies' expectations, most users have difficulty understanding them. Das et al. [10] studied privacy policies from 64 applications and proposed that these policies are not comprehensible to most youths, and even some adults.

Privacy policies have undergone several changes after the GDPR came into effect, further affecting users' understanding. Several studies [25–27, 30, 51] have investigated the differences between the pre-GDPR and post-GDPR versions of privacy policies. Linden et al. [25] analyzed 6,278 English-language privacy policies and identified a positive trend in the users' experience following the modification of privacy policies in the EU.

Technical Term Misconceptions. Prior research has studied users' understanding of security documents and found that users have misunderstandings about technical terms related to security [17, 46, 47, 50]. Wolf et al. [46] proposed that users, including security experts, misunderstand the biometrics functionality in the authentication. Furnell et al. [17] found that users have misconceptions when understanding and using security-related functionality and security options within common software applications. Zabba et al. [50] investigated users' understanding of security warning messages. Their findings indicate that most users cannot understand technical terminology and such warning messages should use less of them. Wu et al. [47] collected technical terms appeared in the security articles and found that 61% of technical terms in their corpus are difficult for users to understand. Additionally, some previous studies have investigated the issue of technical term misconceptions in privacy policies [15, 43]. Felt et al. [15] collected technical terms related to Android permissions and found that only few users can comprehend them. Later, Tang et al. [43] collected 22 technical terms in website privacy policies and investigated how well users understand them. They found that users had a common misunderstanding about terms with more than half of the survey respondents unable to correctly define them. They also proposed that technical terms in a privacy policy will affect users' experience, but they have not provided an effective method to enhance users' comprehension of such technical terms.

VPA Privacy Policy. More recently, a few studies [24, 28, 49] have analyzed the privacy policies of Virtual Personal Assistant applications. Liao et al. [24] conducted a basic user study on the current skills' privacy policies with a few general questions and found that the privacy policy is too long for 44% of their participants; 47% of their participants were not aware of the data being collected and shared by the skill. Manandhar et al. [28] collected and analyzed the privacy policies from 596 smart home vendors. They found that some policies used broad terms to discuss specific data types collected by vendors and some policies did not provide a specific description of data sharing.

The transparent principle of privacy policy [2] regulates that any information about data collection and processing of collected data needs to be easily accessible and comprehensible for users. The existing work discussed above has shown that users often struggle to understand privacy policies, particularly technical terms, which can affect their understanding of the privacy practices of online applications. Limited effort has gone into investigating users' comprehension of technical terms in VPA privacy policies. Therefore, our study focuses on VPA app's privacy policies and seeks to investigate the reason why users cannot understand technical terms used in this context.


Please read this Notice carefully and contact us with any questions. Click on the sections below to learn more about our Notice:

1. SCOPE OF THIS NOTICE
2. TYPES OF INFORMATION WE COLLECT AND WHY
3. HOW WE SHARE YOUR INFORMATION
4. LEGAL BASIS FOR PROCESSING PERSONAL INFORMATION (EEA VISITORS ONLY)
5. STORAGE AND SECURITY OF YOUR PERSONAL INFORMATION

Figure 2: An example of the structure of a privacy policy

3 STAGE ONE: TERMS COLLECTION

In Stage One, we conducted a study with 311 recruited participants, to examine users' understanding of technical terms in privacy policies. The aim of the study was to answer RQ1 and RQ2.

3.1 Stage One Setup

Data Sources. We selected the policies of the top 300 skills based on the number of downloads from an existing dataset [48], as the data source for the study. After getting rid of the skills that do not provide a policy (105), those that use duplicate policies (124), and those that use the same template and thus have similar contents (71), 73 unique privacy policies are kept. We then manually examined each policy to further quality assurance. We removed those that do not contain correct details about the privacy policy, and those that are too simple (fewer than 500 words in length). Finally, 18 policies were removed from the list of 73, resulting in 55 skills' privacy policies (55-policy dataset) for Stage One study.

Data Pre-processing. Most skills have lengthy privacy policies (around 4,000 words). Previous research has shown that the average reading speed of an adult is around 275 words per minute [29]. Reading excessively long documents might impose a cognitive load on the users and result in poor-quality responses. To avoid this, 100 representative privacy policy segments were selected from the 55-policy dataset following the process described below.

All 55 privacy policies are well structured, comprising several sections with their own title. Fig. 2 shows the typical structure of a privacy policy, using that of Ultimate HISTORY Quiz [6] as an example. Firstly, each privacy policy was split into different sections with their title, such as data collection and data use (refer to Table 4 in Appendix A for general sections). Next, depending upon their length, sections were either used directly or combined together to form policy segments. This way, each policy was split into three or four segments such that each segment is around 1,375 words in length (between four to seven minutes reading time). Finally, we randomly split policies from the 55-policy dataset and obtain the 100 representative privacy policy segments.

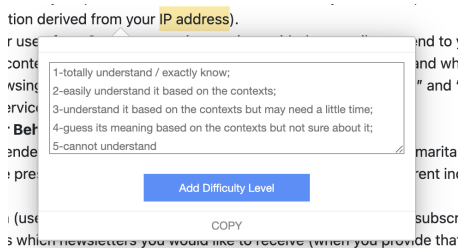


Figure 3: A screenshot of the annotation tool

3.2 Stage One Survey Procedure

Recruitment and Screening. We recruited participants through Amazon's Mechanical Turk (MTurk) platform [3]. Since **RQ2** seeks to examine whether users' IT knowledge level affects their understanding, we first conducted a screening survey to balance the number of participants with and without IT background. We considered participants' education level, major, and current profession to identify whether they might have an IT-related background. Based on the results of screening survey, we invited participants to our main survey. Additionally, participants were required to be 18 years or older, have a HIT approval rate[1] of 95% or higher, and have successfully completed at least 100 tasks (as suggested by [34]). Participants were paid 0.03 US dollars and 2 US dollars for the successful completion of the screening survey and the main survey respectively.

Main survey procedure. We designed a questionnaire to collect the technical terms appeared in skills' privacy policies and their difficulty level from the users' perspective. The questionnaire included three parts: (i) *users' demographics*, (ii) *an annotation task*, and (iii) *two open-ended questions*. We provided a tutorial on how to annotate terms at the beginning of the questionnaire along with a detailed description of the different difficulty levels.

In the demographics section, participants were required to answer questions about their age group, gender, education level, whether they have more than 1 year IT-related background, whether they are native speakers, and whether they have read a privacy policy. The "*prefer not to say*" option was provided for all questions.

In the annotation task, participants were randomly assigned a policy segment and asked to annotate terms. We implemented an annotation tool based on an existing annotator [11], which guarantees that each policy segment was successfully annotated by at least three participants. Participants could select and label a technical term with a difficulty level using their mouse. As shown in Fig. 3, the difficulty level ranges from 1 to 5, with a detailed description in the default area of the text box. After inputting the difficulty level and clicking the "*Add Difficulty Level*" button, the annotated term is highlighted in yellow.

In two open-ended questions, participants were asked to list at least three annotated terms that might significantly affect their understanding. Additionally, they needed to provide the related reasons for why they thought the term was difficult. If participants thought all terms are comprehensible, they needed to describe the reason why they could comprehend such terms.

[1]HIT approval rate represents the proportion of completed tasks that are approved by requestors.

3.3 Stage One Survey Results Overview

We ran our survey for four weeks in July and August 2022. After that, we manually reviewed the responses. Three authors did the majority vote for each submitted assignment and used the following three rules to exclude invalid responses: (i) responses with blank answers on the annotation task and two open questions, (ii) responses with a careless assignment on the annotation task, i.e., all annotated terms are not related to the policy segments, or responses with less than three annotations, and (iii) responses that do not answer the required open-ended questions. In total, we obtained 311 valid responses. Each of the 100 representative privacy policy segments was annotated by at least three participants. The average completion time was around 26.6 min.

User Demographics. Almost half the participants (49.2%) were in the age 18 to 35, 36.0% were in 35 to 50, and the rest 14.8% were above 50. The number of males and females was almost equal. 85.53% of the participants were native speakers. Most participants (82.96%) reported having a bachelor's or higher degree. The screening survey guaranteed that almost half of the participants (43.09%) had at least 1 year IT related background. Additionally, most participants (91.96%) stated that they have read a privacy policy before, either for a website, a mobile app or a VPA app.

Annotation Task Pre-processing. A total of 5,553 technical terms were annotated by the 311 participants. These annotated terms were pre-processed as follows. To begin with, all invalid terms were removed. A term was considered invalid if it has no specific meaning in the privacy policy domain, e.g., "*latter*". Next, duplicate terms were removed. Duplicate terms are possible after lemmatization (e.g., "*third parties*" is the plural form of "*third party*"). The *FuzzyWuzzy* library [4] was used to check the similarity of annotated terms based on fuzzy string matching. We manually identified the terms that were reported as having a similarity larger than 85% (suggested by [38]) and removed one of the terms and assigned the average value of difficulty level to the remaining term. We totally obtained 2,740 unique technical terms, called *Alexa Skill Privacy Policy Corpus* (*ASPP Corpus*).

Data Validity. First, the *mean* and standard deviation (*std*) were calculated for the number of technical terms annotated by different participants within the same segment. Most (82.72%) of the *std*s range from 3 to 15, with an average of 9 across the 100 segments. This relatively low *std* indicates that the number of terms annotated by the participants is relatively balanced across the 100 segments. We then measured the term frequencies in the 100 segments and the 55-policy dataset. We used the Log-likelihood ratio test for the corpus comparison [35]. For each term in the ASPP Corpus, its occurrence frequency in both 100 segments and the 55-policy dataset was compared. The results show that 98.97% of terms in the ASPP Corpus have a similar frequency to the 100 segments and the 55-policy dataset ($p > 0.05$). It indicates that our 100 segments are indeed representative and these contain sufficient terms.

Finally, we assessed the validity of the difficulty level of annotated terms among 311 participants. The *std* range for the difficulty level of each annotated term is between 0.31 and 2, with an average of 0.85. For 87.52% terms in the ASPP Corpus, their *std* value for annotated difficulty level is from 0.31 to 1.35. These results indicate that the annotated difficulty levels are within good quality.

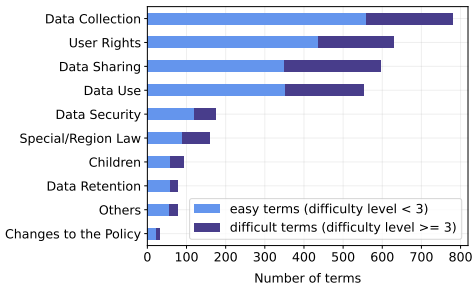


Figure 4: The distribution of difficult and easy terms from the ASPP Corpus in 10 GDPR categories.

Additionally, we checked whether there is an outlier for the annotated difficulty level of each term. The results show that only 2.6% of terms in the ASPP Corpus have outliers. Therefore, we conclude that the annotated difficulty levels are satisfactory from 311 participants and have no significant differences.

3.4 Stage One Survey Results Analysis

This section details the results of the analysis of the ASPP Corpus to answer the first two research questions (**RQ1** and **RQ2**).

RQ1. What are considered as technical terms from the users' perspective? How well do users understand these technical terms?

To answer this question, the 2,740 technical terms in the ASPP corpus were split into two groups based on their average difficulty level: *easy terms* (difficulty level < 3) and *difficult terms* (difficulty level >= 3). This results in 34.05% (933/2,740) difficult terms and 65.95% (1,807/2,740) easy terms.

Term Categories. The annotated technical terms were first separated into 10 categories in accordance with the GDPR privacy regulation and then analyzed based on their difficulty level. Each technical term was assigned a category based on its position in the privacy policy. Some terms might appear in several sections of a privacy policy and therefore be assigned to multiple categories. We also achieved automated classification, as detailed in Appendix B. But for our following analysis, we use manually generated categories to make sure we have more accurate analysis results.

Analysis of Categories. Fig. 4 shows the distribution of terms for each category in the ASPP Corpus based on their difficulty level. The results show that the "Data Collection", "User Rights", "Data Sharing" and "Data Use" categories have a relatively large proportion of technical terms. This indicates that developers or practitioners provide more content for these four sections when describing a privacy policy. Among the four categories, 36.23% and 41.37% of terms are difficult in "Data Use" and "Data Sharing" respectively, which is higher than the average proportion of difficult terms in the ASPP Corpus (34.05%). Of the remaining six sections, the "Children" and "Special/Region Law" categories are also difficult for users to comprehend, with 37.23% and 44.30% of terms labeled as difficult. The results illustrate that users might need to spend more time on reading and understanding these four sections of the skill's privacy policy, compared to the other sections.

Broad vs. Domain-specific Terms. Some technical terms were labeled as *broad terms* or *domain-specific terms*. Broad terms do not have a precise definition and can mean more than one thing, e.g., "sensor information" in the "Data Collection" category can refer to a range of information including heat, light, and humidity. Similarly, "analytics providers" in "Data Sharing" can include a wide range of providers for analytics purpose like Google Analytics and Google Assistant-enabled device. Domain-specific terms have a precise and unambiguous definition specific to VPA applications, e.g., "voice recordings" in "Data Collection" and "camera-enabled products" in "Data Sharing". Two of the authors manually labeled the terms in the ASPP Corpus and then jointly discussed the labels to ensure consistency and consensus.

Finding 1: 30.73% terms in the ASPP Corpus are broad terms, while 17.66% terms are domain-specific terms. 842 out of 2,740 terms were broad terms rather than describing specific data types or third parties. Among them, 33.73% (284/842) were difficult terms, implying that participants cannot clearly understand them or only have limited knowledge about them, e.g., a skill's policy uses "motion sensor data" to describe the data type that the developer collects from users. However, motion sensor data might either refer to human motion data (e.g., heat data or movement data) or drive event data (e.g., car movement data). Participants annotating this term with the average difficulty level as 5 indicates that participants might have very limited knowledge about motion sensor data. 484/2,740 of the technical terms are domain-specific terms, and among them, 38.63% (187/484) are difficult terms. This indicates that, compared with broad terms, domain-specific terms are more difficult to be understood by our participants.

Finding 2: The use of certain broad terms and domain-specific terms might violate the CCPA and GDPR. Based on Finding 1, we find violations from both types of terms. Violations of broad terms consist of two aspects. First, the data collection section of the policy, which uses the broad terms, might not follow the CCPA, because CCPA regulates the disclosure of all applicable data categories. For example, "sensor information" for data collection may have several sub-categories' data types. Second, "analytics providers" have sub-categories like analytics services in Google and in Apple HomeKit. Using such a term to describe third parties might not follow the GDPR, because GDPR regulates the disclosure of all recipients or recipients categories. Additionally, for difficult domain-specific terms (refer to Finding 1), they might lead to widespread misunderstanding. This might violate the transparency principle of the privacy policy based on the GDPR, because transparency principle regulates that the privacy policy should provide explicit descriptions and such descriptions should be easy for users to understand.

RQ2. Why are technical terms difficult for users? Does users' understanding of technical terms relate to their IT background?

We collected the answers to two open-ended questions which (i) cover the list of difficult technical terms that significantly affect participants' understanding of policy segments, and (ii) explain the reasons why they provided the corresponding difficulty levels. We used open card sorting [41] to classify the 311 comments provided by participants. More specifically, three authors randomly selected

50 responses and identified the positive or negative reasons, e.g., "can totally understand because" as a positive reason and "cannot understand because" as a negative reason. We then grouped the positive and negative reasons into separate categories representing specific reasons why users did or did not understand the technical terms. Three authors did the majority vote to validate the results.

Finding 3: Only 15.76% (49) of participants can understand all technical terms used in the policy segments. Only 49/311 (15.76%) participants annotated all technical terms with the lower difficulty level (around 1 or 2). Among 49 participants, 21 mentioned that they are familiar with this domain, however, the terms may not be as common to the layman. Other 28 participants mentioned that they could understand terms with the useful context provided by the policy segments. This indicates that a small proportion of non-expert users can understand the policy and its implication if appropriate context are provided. In contrast, the majority of the participants (84.24%) claimed that difficult technical terms affect their understanding and provided related reasons. We summarized reasons in Fig. 5 (except 8 participants provided reasons that are not specific about the policy segments).

Finding 4: Missing/Useless Context - 39.87% (124) of participants think that the context in which the terms are used is either missing or useless. The provided contents of policy segments probably exceed the average reading level of users. **(1)** 32 participants pointed out that the policy segments have no relevant definition or context about terms, e.g., one participant said "*I have never heard of framing technique and am unsure of what it means. It is not defined in the segments*". The author confirmed the original privacy policy does not provide any relevant context for annotated terms. **(2)** 62 participants stated that the policy segments do provide some context for the technical terms, but they do not directly explain what the terms are or only provide limited details about them. For example, one participant stated "*I gave 'binding arbitration' as 4. The text did give more information about the term, but didn't explain directly what the term meant*". This indicates that users are still confused about the term's meaning even though the skills' privacy policies have provided a few descriptions. Additionally, among 62 participants, 9 participants mentioned that it took them more time to understand the terms and related context. These results confirm that the language used in those policies might be difficult for the general public. To raise users' understanding, developers should use plain and general language in such policies and provide specific context for the technical terms and related policies.

Moreover, (3) 30 participants stated that the use of acronyms without providing the full phrase significantly affects their understanding. We double-checked the policy segments mentioned by those 30 participants. The acronyms listed by 29/30 participants do not have the full form at their occurences in the original privacy policy. For example, a participant mentioned that "*I don't know the meaning of these acronyms, like EM&V*". The original privacy policy does not provide the full form of the acronym. This indicates that some developers use technical terms in their policies without explicit elaboration.

Finding 5: Vague Terms - 34.08% (106) of participants think technical terms and related policy segments are inferrable but vague, necessitating more details to help them confirm their guess. Based on the context of the policy segments, most participants were able to guess the meaning of technical terms. However, they might have no idea about the correct meaning of the terms. A participant said "*I am wholly unaware of what a Pixel Tag is. I can infer that it's a visual element but not where it is or what it's used in relation to*". In such situations, users either use the skill with limited understanding or misunderstanding of the terms and related policies. Additionally, 25 participants think some terms are similar in their minds and get confused when reading those terms. It means that technical terms might have different meanings but participants have common misconceptions about such terms, e.g., a participant said "*URL is to be confused with an IP address*". The misconception might lead to unconscious privacy leaks. The results confirm that (1) the detailed context might be useful for users' understanding and (2) practitioners or developers should provide explanation context about technical terms or use other helpful methods to enhance users' understanding.

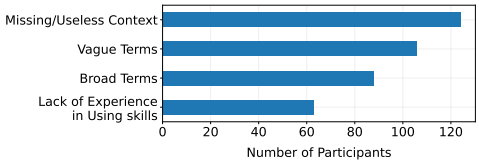


Figure 5: The distribution of reasons why technical terms are considered difficult from users' perspective.

Finding 6: Broad Terms - 28.3% (88) of participants think that technical terms or related privacy segments are too broad, and cannot be understood in the VPA-specific context. 88/311 (28.3%) participants pointed out the unspecific terms or policy context. Among them, **(1)** 60 participants mentioned that the annotated terms were very broad for them. We recorded such terms listed by those 60 participants in open-ended questions. 51 of them labeled at least two terms as "broad". The distribution of terms is not stable with the participants annotating terms with a wide range of differences. 60 participants totally mentioned 92 terms as "broad". Only a few terms were mentioned more than 2 times. We infer that **those 100 segments contain a large number of different broad terms, and these terms do affect users' understanding of the skills' privacy policies**. For example, participants stated that (i)"*cookie is a very general term to describe many different kinds of cookies*"; (ii)"*Tracking Technologies: It is broad and non-specific*"; (iii)"*GDPR is a very broad and deep topic that is only touched on in passing throughout this agreement*". Besides that, **(2)** 28 of 88 participants mentioned that they understand the annotated technical terms in general, however, the terms should be specific to the VPA domain. For example, a participant annotated "HTML5" as the difficult term and explained their decision as follows: "*I have no idea what the difference between that and the HTML used on websites is or how it's used with Sirius radio*".

Finding 7: Lack of Experience in Using skills - 20.25% (63) of participants have never used Alexa skills previously. 63/311 (20.25%) participants annotated technical terms as difficult because they lacked relevant experiences in the VPA domain. 32 of them mentioned that they had never used any services or skills described in the policy segments. For example, a participant labeled terms about opt-out features as higher difficult because he was not familiar with the opt-out features of those advertising companies. To

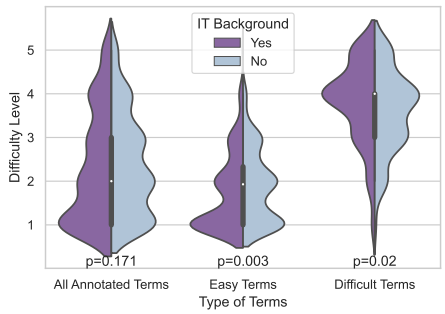


Figure 6: The distribution of difficulty levels for term types including all annotated terms, easy terms (difficulty level < 3) and difficult terms (difficulty level >= 3) with a significant difference (p < 0.05).

improve it, practitioners or developers should provide general and transparent context about such skills to briefly introduce them.

Finding 8: Having IT background does not help participants understand the difficult technical terms appeared in skills' privacy policies. We analyzed the results from the main survey, where 43.09% of the participants have IT background. We analyzed the annotated terms and their difficulty levels to investigate the differences between users with and without IT background. 548/2,470 technical terms are annotated by both participants. We used the Mann-Whitney U test to compare annotated difficulty levels between these two groups.

As shown in Fig. 6, the annotated difficulty level does not have statistical differences with users' IT background for all annotated terms, but the difference becomes significant when we compare easy terms and difficult terms separately (both $p < 0.05$). We further compared differences in each category. We found participants with IT background were more likely to give lower difficulty level than those without IT background, e.g., significant difference in "Data Collection" category ($p < 0.02$). Participants with IT background preferred to annotate the higher difficulty level than those without IT background, e.g., significant difference in "Special/Region Law" category ($p < 0.015$). Generally, participants without IT background rarely annotated technical terms as "totally understand" (difficulty level = 1). They preferred to annotate the difficulty level as 2 or 3. It indicates that they can guess the meaning of annotated technical terms based on the policies' context, but might not fully understand the terms. Participants with IT background can totally understand technical terms based on their knowledge or policies' context (like terms for "Data Collection"). However, if the term is difficult or outside of their domain, they will regard them as "higher difficult" (like terms for "Special/Region Law"). On the contrary, participants without IT background think that they can understand or guess most terms, even though they might have some misconceptions.

4 STAGE TWO: TERMS EXPLANATION

A key insight we get from Stage One is that participants needed "*more context*" or "*detailed description*" of technical terms when reading the skills' privacy policies (Refer to Findings 4 - 6). Thus, we conducted a second study to examine whether users' understanding of technical terms can be enhanced with explanations of terms.

4.1 Stage Two Setup

Explanation Sentences Guidelines. Micheti et al. [31] have revealed that the readability of the privacy policy is insufficient for general users. We adopt the following guidelines that are appropriate for the VPA context, when composing explanation sentences. The first three points are summarized from the findings of Micheti et al. and the last point is from Stage One's results.

(1) **Use simple words/sentences based on the general users' understanding.** Explanation sentences should have an average or high readability level to ensure that users with different education levels and backgrounds can understand them.
(2) **Avoid using double negatives.** An explanation sentence should avoid using double negatives to explain a technical term as this can confuse the reader.
(3) **Keep explanation sentences short.** Users prefer to read short sentences and might struggle to read large blocks of texts or a sentence including many subordinate clauses. We design that the whole explanation should be within 100 words.
(4) **Avoid using other difficult technical terms.** A technical term might be described using other technical terms. In such situations, users might become more confused by the explanation. Whenever possible, technical terms should not be explained using other (difficult) technical terms.

Explanation Sentences Principles. We proposed two principles to write the explanation sentences of technical terms, namely *domain-specificity principle* and *implication-oriented principle*.

Domain-specificity Principle. As discussed in Findings 4 and 6, participants can better understand technical terms if more specific context is provided and broad or vague language is avoided. Previous studies [36, 43] have also found that technical terms might lead to users' misconceptions of privacy policies. Some policies that use unclear language hinder users' comprehension of company practices even though they provide some definitions and examples of technical terms. Therefore, *the domain-specificity principle aims to provide specific and transparent explanation sentences for technical terms to avoid such misconceptions.*

Implication-oriented Principle. Besides the findings of our Stage One, some related works have also discussed the privacy concerns of users. Lau et al. [22] found that users have few privacy concerns when using smart speakers. However, their knowledge indicates an incomplete understanding of privacy risks and a complicated trust relationship with those practitioners or developers. If users could improve their understanding of the privacy risks associated with using smart speakers by reading the provided privacy policy, they might build a real trust relationship with the practitioners and completely realize the relevant privacy concerns. Moreover, Soumelidou and Tsohou [40] proposed the IPA (information privacy awareness) definition and reported the differences between online users' IPA levels with their behaviors. Therefore, *the implication-oriented principle stipulates that the explanation sentences emphasize the potential privacy risks to improve users' security awareness and understanding.*

Additionally, the explanation sentences should provide a basic and comprehensive definition of technical terms. Therefore, regardless of the used principle, we provided users with a simple definition of terms using concise words and language.

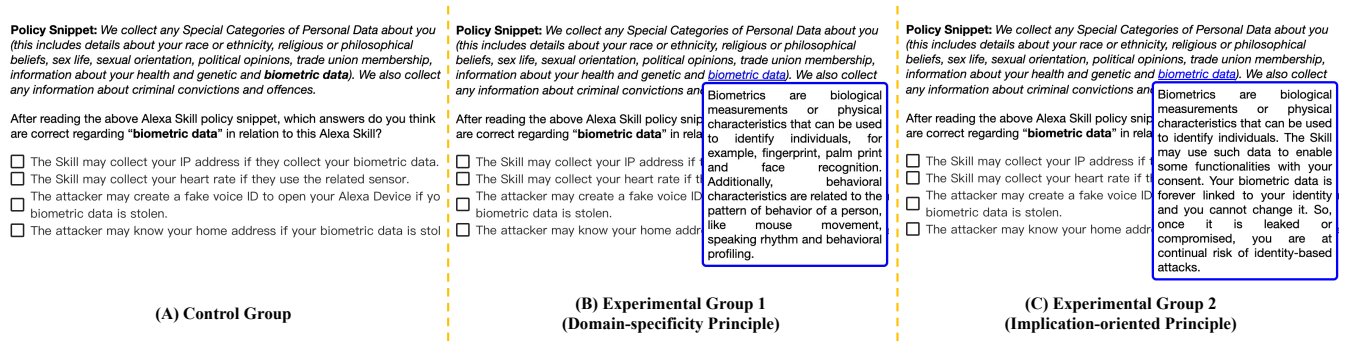


Figure 7: Examples of explanation sentences presented to the participants using a pop-up window when hovering over the term "biometric data" as used in our Stage Two experiment.

Interface Design. Previous research [19] has found that the length of the privacy policy affects users' understanding. Most people are not willing to read the long and cumbersome privacy policies. Thus, the length of the privacy policy might be a factor in examining users' understanding and may lead to poor experiment results. Moreover, [9] proposed that using the pop-up window can improve people's learning when presenting texts. Now, most popular and common dictionary-based tools use the pop-up window to present their dictionary texts, like Google dictionary [20] and Mac dictionary [8]. Therefore, to guarantee the quality of our experiment results, we used the pop-up windows to present explanation sentences of terms in skills' privacy policies, instead of directly adding them to the original policies.

4.2 Stage Two Evaluation Procedure

To evaluate the effectiveness of two principles, we compared participants' understanding of skills' privacy policies when presented with our explanation sentences along with the original texts.

Terms Selection. Too many technical terms and questions in user study might lead to users' poor attention and affect the quality of the experiment results. To mitigate this, we limited the number of examined terms to around 25. We then selected representative technical terms corresponding to each category from the ASPP Corpus using the following criteria: (i) difficult terms with the average difficulty level around 4 or 5, and (ii) difficult terms labeled by most participants. We totally selected *23 representative terms* based on the proportion of difficult terms in each category (shown in Table 1). We did not select terms from "Data Retention", "Others" and "Changes to the Policy", because each of them has less than 22 difficult terms (shown in Fig. 4).

Explanation Sentences Generation. The explanation sentences were generated for 23 selected terms using the following process. To begin with, two of the authors as the writers separately composed the explanation sentences based on the principles presented in Section 4.1. Next, the think-aloud protocol was applied to the sentences evaluation process. For each type of explanation sentences, the other two authors as the reviewers were asked to highlight the words or sentences that did not match the related principle. During this process, two reviewers were free to discuss about the explanation sentences and the possible problems. The writers of the sentences observed the discussion of two reviewers and took notes. If one of the reviewers rejected the explanation sentences, the writer was asked to regenerate or modify them based on reviewers' suggestions. The evaluation process was repeated until both reviewers approved the generated explanation sentences.

Table 1: 23 technical terms selected for Stage Two from different categories.

Category	Examined Technical Terms
Data Collection	biometric data, essential cookie, **pseudonymous data***, IMEI and MEID of your mobile device, **traffic data**, media stamp, **web beacons** (7 terms)
Data Use	anonymized audio recordings, standard contractual clauses, **web beacons**, optimize the heating and cooling algorithms, **PI (personal information)**, **DART cookie** (6 terms)
Data Sharing	**DAA's online resources**, **pseudonymous data**, IDFA, **PI (personal information)**, **traffic data**, **DART cookie**, **web beacons** (7 terms)
User Rights	request erasure of your personal data, GDPR rights, do not track settings, **DAA's online resources**, archive records, request portability of your personal information (6 terms)
Data Security	secure socket layer technology (1 term)
Children	inadvertently collect personal information of a child (1 terms)
Special/Region Law	California consumer privacy act, Alexa communication schedule (2 terms)

*Highlighted terms are duplicated in different categories. A term might appear multiple times in different sections of a skill's privacy policy.

Questionnaire Design. We designed a questionnaire for Stage Two to evaluate the effectiveness of the explanation sentences and related principles. As mentioned in Section 3.1, 37 policies were used from the 55-policy dataset to obtain the 100 segments for Stage One. We used the remaining 18 skills' privacy policies for Stage Two. If a paragraph of a policy contains one of the 23 examined terms, we will select the entire paragraph as the policy snippet for the term. It can guarantee the original structure and contents of the skills' privacy policies with the examined terms. We designed multi-choice questions for each of the 23 extracted policy snippets. Each question was designed with a related policy snippet only for one technical term and had more than one correct answer (see an example in Fig. 7).

To ensure the quality of the experiment results, two attention-check questions were included along with the 23 main questions. These questions were directly related to the questionnaire and all answers could be directly obtained from the provided materials. Participants who carefully complete the other questions are supposed to select the correct choices for these two questions.

Experiment Group. We used three participant groups to evaluate the explanation design, including *CG*, *EG1*, and *EG2* (refer to Table 2 for the detailed descriptions). In *EG1* and *EG2*, explanation sentences were presented using the pop-up window (as explained

Table 2: Three experiment groups in Stage Two Evaluation.

Group	Detailed Descriptions
CG	**Control Group.** Participants in CG were provided with only policy snippets, in which the examined terms were made bold without explanation.
EG1	**Experimental Group 1.** Participants in EG1 were presented with policy snippets and explanation sentences generated by the domain-specificity principle.
EG2	**Experimental Group 2.** Participants in EG2 were presented with policy snippets and explanation sentences generated using the implication-oriented principle.

in Section 4.1). Participants can hover their mouse over the highlighted technical terms to see and close the related explanation sentences (shown in Fig. 7). All three groups were asked the same questions and presented with the same choices.

Experiment Procedure. Participants were recruited on MTurk and separated into three experiment groups. In order to answer RQ4, we first conducted a similar screening survey (as Stage One) to balance the number of participants with and without IT background in each of three experiment groups. We then invited the same number of participants (excluding participants in Stage One) for three experiment groups to complete the main questionnaire. Participants were paid 0.03 US dollars and 3 US dollars for the successful completion of the screening survey and the main questionnaire respectively.

The first section of the questionnaire is about users' demographics. In the next section, the participants were asked to answer the multi-choice questions about the 23 technical terms and two attention-check questions, as a total of 25 multi-choice questions (available in our repository). Participants in *CG* were simply presented with policy snippets. For *EG1* and *EG2*, an introduction for showing explanation sentences was provided at the beginning of the questionnaire. Participants in these two groups were presented with policy snippets and related explanation sentences. In addition, they were also asked to provide some feedback about the explanation sentences including whether the sentences helped their understanding of policy snippets, suggestions for improving explanation sentences, and other methods or designs they preferred to help their understanding. Participants from three experiment groups were not allowed to use any search engines (e.g., Wiki, Google) to find answers to the questions.

4.3 Stage Two Evaluation Results Overview

We ran our survey for five weeks in October and November 2022. Invalid responses to the multi-choice questions were rejected using the following criteria: (i) responses that had incorrect answers for the two attention-check questions, and (ii) responses that had empty feedback in *EG1* and *EG2*. In total, 50 valid responses were received for each experiment group for a total of 150 valid responses. This includes 25 responses each for participants with and without IT background in each experiment group. The average completion time was 54 min.

User Demographics. 44.67% of the participants were young adults (ages 18 to 35). 38.67% of them were middle-aged (ages 35 to 50) and others were old adults (ages > 50). The number of males and females was almost equal. 79.33% of the participants were native speakers. All participants reported having at least a high school diploma or equivalent. 84.67% had a bachelor's or higher degree. Additionally, 88.66% of the participants stated that they had read a privacy policy before. These proportions are similar to the Stage One's survey results.

Table 3: Detailed Description about our four accuracy levels.

Accuracy Level	Detailed Descriptions
All_Co	**All Correct.** Participants select all the correct choices for a multi-choice question.
Part_Co	**Part Correct.** All selected choices are correct but the participant missed some of them, e.g., the participant only selects choice 2 and 3 when the correct choices as 1, 2, and 3.
Part_In	**Part Incorrect.** Some of the selected choices are correct but others are incorrect, e.g., the participant selects choice 2 and 3 when the correct choices are 1 and 2.
All_In	**All Incorrect.** All selected choices are incorrect.

Measurement Metrics. The main section of the questionnaire consisted of 23 multi-choice questions (excluding two attention-check questions). For further analysis, responses from each participant were assigned to one of four accuracy levels based on the degree of correctness of the responses (details in Table 3). Additionally, in the following analysis, we will conduct the proportion z test [33], which can compare the proportion of data between two groups (like A and B). It first set the H_0 (like $P_A >= P_B$) as its original hypothesis. If the z scores is larger than 1.96 or smaller than -1.96 and the p value is smaller than 0.05, it will reject the H_0 and approve the alternative hypothesis H_a ($P_A < P_B$).

4.4 Stage Two Evaluation Results Analysis

This section details the results and analysis of Stage Two's study to answer RQ3 and RQ4.

RQ3. How well do users understand the provided explanations based on different principles? Do the explanation sentences improve users' understanding?

To answer this research question, we first analyzed the feedback from the participants in *EG1* and *EG2*. Among 100 participants, 96 indicated that the explanation sentences helped their understanding of technical terms and the corresponding policy snippet. The correctness of the participants' responses to the 23 multi-choice questions was calculated using the degree of correctness metric. The "All_Co" and "Part_Co" responses indicate an improved understanding of the technical terms. The "Part_In" and "All_In" responses indicate that participants have misunderstood the technical terms based on the explanation provided. The average number was calculated for each accuracy level within each experiment group for the seven categories listed in Table 1.

Finding 9: Explanation sentences can improve users' understanding of technical terms when reading a privacy policy by a 56.5% increase in correctly answered questions. Fig. 8 demonstrates the distribution of answers for the 23 examined terms, where each subplot shows the comparison among three experiment groups for each term category (except Plot A representing for the total). From Plot (A), we observe overall, the performance of participants in both *EG1* and *EG2* is better than the performance of participants in *CG*, especially for *EG1* with explanation sentences using domain-specificity principle. We first conducted the chi-squared test (X^2 test) as follows. All participants were separated into two accuracy groups: *correct group* ("All_Co" and "Part_Co" participants) and *incorrect group* ("Part_In" and "All_In" participants). We then calculated the number of participants in each accuracy group for each experiment group and conducted the X^2 test. The statistics show that users' correctness is related with the use of explanation sentences between *CG* and *EG1* (p = 0.036), but such relationship is

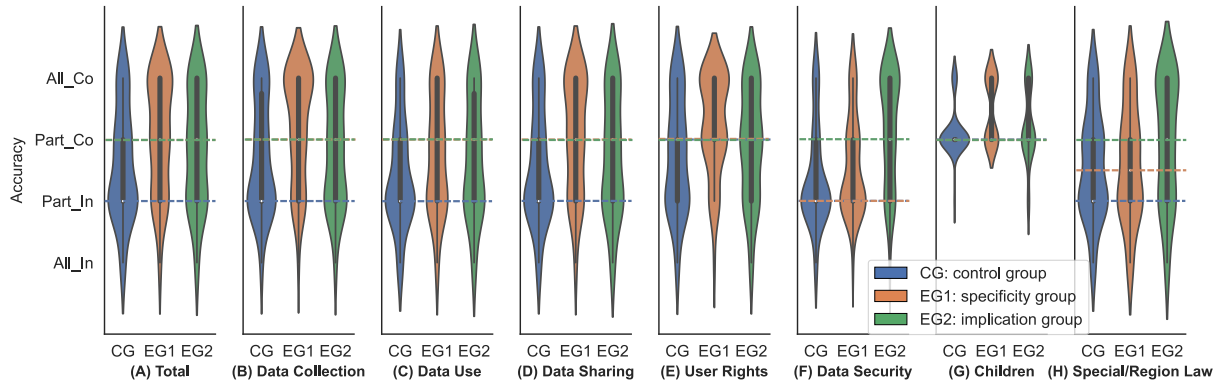


Figure 8: The distribution of accuracy levels of completed responses for each technical term category for three experiment groups. The details about three experiment groups and four accuracy levels are provided in Tables 2 and 3.

not significant between *CG* and *EG2* (p = 0.229). Our result shows that participants in *EG1* have a 56.5% increase in correctly answered questions compared to *CG*, while *EG2* only have a 26.1% increase. We then used the proportion z test to analyze the proportion of four accuracy levels in three experiment groups (detailed analysis in Appendix C.1). The results indicate overall proportion of "All_Co" participants in *CG* is only significantly smaller than the proportion of "All_Co" participants in *EG1*. Therefore, we conclude that in total, the explanation sentences can help users' understanding of technical terms.

Finding 10: Explanation sentences using domain-specificity principle are most helpful in improving users' understanding of terms relating to users' data or rights. We further analyzed the performance of three experiment groups to find out whether these have some differences in any of the term categories. As shown in Fig. 8 (B), (C), (D), (E) and (G), the performance of participants in *EG1* is better compared to *CG*. Thus, we set the related hypotheses for these five categories to compare the performance of participants among three experiment groups. The statistics show that the proportion of "All_Co" or "Part_Co" participants in *EG1* is significantly larger than the proportion in *CG* and *EG2* for terms related to "Data Collection", "Data Sharing", "User Rights" and "Children" and terms related to users' data in "Data Use" (detailed analysis in Appendix C.2). Such terms are all related to users' data or rights. Explanations using domain-specificity principle might provide more explicit and transparent descriptions for users to understand than explanations using implication-oriented principle.

Finding 11: Explanation sentences using implication-oriented principle work better than those using domain-specificity principle for terms relating to security or special laws. To analyze the performance of explanation sentences using implication-oriented principle, we applied the proportion z test again. We set the hypotheses to compare the participants' performance in between *CG* and *EG1* and between *CG* and *EG2* for "Data Security" and "Special/Region Law" category (Fig. 8 (F) and (H)). The statistics show that the proportion of "All_Co" participants in *EG2* is significantly larger than the proportion in *CG*, and the proportion of participants has no significant differences between *CG* and *EG1* (as shown in Table 6.14-17 in Appendix A). Such terms are related to users' privacy security or potential lawful problem, which may implicate possible privacy risks. Explanations using implication-oriented principle can provide more clear descriptions about the potential implications of such terms in policies than explanations using domain-specificity principle.

RQ4. Does users' understanding of explanation sentences relate to their IT background?

Finding 12: Participants with and without IT background have a similar understanding when reading the skill's original privacy policy. We further analyzed the performance of participants with and without IT background. We first analyzed the distribution of accuracy levels of participants in each experiment group (see Fig. 9). We find that the performance of participants with and without IT background is similar in *CG*. However, the distribution shows differences in two experimental groups, especially in the *EG1*. The Mann-Whitney U test was applied to examine whether the performance of different participants has significant differences in each experiment group. We used the number of participants with and without IT background for each accuracy group (correct group and incorrect group mentioned in Finding 9) for each experiment group. The statistics indicate that, in *CG*, the performance of participants with and without IT background does not have significant differences for the correct group and incorrect group (p = 0.16 and p > 0.99 separately).

Finding 13: Participants with IT background can better understand the explanation sentences of terms and related skill's privacy policy. In *EG1*, the statistics show that the performance of participants with and without IT background have significant differences for the correct group and incorrect group (p < 0.001 and p = 0.007 separately). We infer that participants with IT background can better understand the technical terms with explanation sentences. However, the statistics do not show such differences between participants with and without IT background in *EG2* (p = 0.181 and p = 0.909 separately). Thus, based on the Finding 10 and 11, we applied the Mann-Whitney U test for terms in "Data Security", "Children" and "Special/Region Law". The result indicates that the performance of participants with and without IT background only have significant differences in *EG2* for terms in these three categories (p = 0.008 in the correct group).

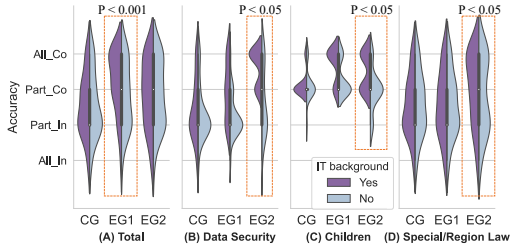


Figure 9: The distribution of accuracy levels of completed responses for participants with and without IT background across three experiment groups.

Participant Feedback. As mentioned in Section 4.2, participants in two experimental groups were required to leave feedback and possible suggestions for improving their understanding of terms and skills' privacy policies. We extracted and summarized the common keywords from the provided suggestions and present the key suggestions here. 17 participants stated that providing some real-life examples would help improve the understanding of technical terms. 11 participants suggested to use graphics or flow charts to explain the terms and policies. 8 participants wanted the important information in the policy to be highlighted. 7 participants stated that they would prefer a video of the policy. Some common feedback has been displayed in Appendix D.

5 DISCUSSION

5.1 Implications

For users. Our findings suggest that most users, including those with IT background, do not understand the skill's privacy policies. This aligns with previous research on the comprehensibility of privacy policies among general users, users with legal experience, and privacy policy experts [36]. That study found that even experts do not have consistent understanding of certain sections of a privacy policy such as data sharing, and their interpretations of technical terms can vary. Our Stage Two's findings also indicate that without explanation sentences, there is no significant difference in understanding between users with and without IT background. However, explanation sentences can improve users' understanding of technical terms, with those with IT background benefiting more.
For developers or practitioners. Our findings suggest that developers or practitioners can improve the comprehension of privacy policies by highlighting the difficult terms and providing explanations and meaningful context. Additionally, based on users' feedback in our Stage Two, they can consider using different media forms, in addition to text, such as audio, video, graphics and animations, for providing explanations.
For researchers. Our findings indicate that providing explanation sentences for technical terms in privacy policies can help improve users' understanding. Future research could focus on creating a unified dictionary of explanation sentences for technical terms commonly used in Alexa skills' privacy policies, automatically identifying and highlighting difficult terms, and automatically summarizing important information in privacy policies for users. These efforts could help users better understand the potential privacy risks and take appropriate steps to protect their personal data.

5.2 Limitations

User study. The ASPP corpus is limited in size and diversity of participants, as well as diversity of skills. This is mainly due to budget limitations. We only considered 300 skills' policies from over 60,000 skills' policies in the UQ-ASS21 dataset [48], and extracted 100 policy segments from them. We recruited 311 participants on MTurk to annotate technical terms in these policy segments using a single criteria - whether they had an IT background or not. It would be beneficial to consider a wide range of skills, and recruit participants from a more diverse population by considering multiple factors such as age, gender, cultural background, and education level, among others. This would allow for a more comprehensive and representative analysis of the skills' policies and also increase the reliability and validity of the results.
Explanation Sentences. The current study only considered two principles for composing the explanation sentences for technical terms. Future work could explore additional principles, and analyze which principle is the most effective for general users. The current study used a pop-up window to display the explanation sentences. Future work could explore other interface designs to determine which design is the most effective for general users. Options could include the prompt window or drop-down window, among others. Additionally, participants in Stage Two suggested some improvements for the explanation sentences, such as adding graphics, using audio summaries or video recordings, and providing some real-life examples for the technical terms, etc. Future work could investigate the effectiveness of such approaches.

6 CONCLUSION

In this paper, we examined users' understanding and comprehension of Alexa skills' privacy policies. We found that most participants struggled to understand the technical terms used in the skills' privacy policies. In Stage One of our research, we found that most users thought that the skills' privacy policies lacked clear explanations or context for the technical terms and related policies. In Stage Two, we generated explanation sentences for 23 representative terms from the ASPP Corpus (resulting from Stage One) using two principles, to evaluate whether such explanations can help users' understanding. Our results show that such sentences can indeed improve users' understanding of technical terms and related policies, particularly if they have IT background.

Based on our findings, we propose two future research directions. First, we will improve the diversity of the participants' population by considering multiple factors such as age, gender, cultural background, education level, etc. Second, we will investigate and evaluate additional principles for composing explanation sentences, using the findings from our first proposed extension.

ACKNOWLEDGMENTS

Baiqi Chen is supported by the University of Queensland and CSIRO's Data61 PhD scholarship. This work is supported in part by UQ Cyber Research Seed Funding.

REFERENCES

[1] 1998. Children's Online Privacy Protection Rule ("COPPA"). https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa

[2] 2016. General Data Protection Regulation (GDPR) - Official Legal Text. https://gdpr-info.eu
[3] 2018. Amazon Mechanical Turk. https://www.mturk.com/
[4] 2020. fuzzywuzzy PyPI. https://pypi.org/project/fuzzywuzzy/
[5] 2022. Amazon Alexa Console. https://developer.amazon.com/alexa/console/ask
[6] 2022. Amazon.com: Ultimate HISTORY Quiz : Alexa Skills. https://www.amazon.com/Television-Networks-Mobile-Ultimate-HISTORY/dp/B075ZS916K
[7] 2022. California Consumer Privacy Act (CCPA) | State of California - Department of Justice - Office of the Attorney General. https://oag.ca.gov/privacy/ccpa
[8] Apple. 2022. Dictionary User Guide for Mac. https://support.apple.com/en-au/guide/dictionary/dic34880/mac
[9] Mireille Bétrancourt and André Bisseret. 1998. Integrating textual and pictorial information via pop-up windows: An experimental study. *Behaviour & information technology* 17, 5 (1998), 263–273.
[10] Gitanjali Das, Cynthia Cheung, Camille Nebeker, Matthew Bietz, Cinnamon Bloss, et al. 2018. Privacy policies for apps targeted toward youth: descriptive analysis of readability. *JMIR mHealth and uHealth* 6, 1 (2018), e7626.
[11] dvnc. 2015. dvnc/annotator. https://github.com/dvnc/annotator
[12] Jide Edu, Xavi Ferrer Aran, Jose Such, and Guillermo Suarez-Tangil. 2021. SkillVet: Automated Traceability Analysis of Amazon Alexa Skills. *IEEE Transactions on Dependable and Secure Computing* (2021).
[13] Tatiana Ermakova, Annika Baumann, Benjamin Fabian, and Hanna Krasnova. 2014. Privacy policies and users' trust: does readability matter?. In *AMCIS*.
[14] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. 2017. Large-scale readability analysis of privacy policies. In *Proceedings of the international conference on web intelligence*. 18–25.
[15] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. 2012. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the eighth symposium on usable privacy and security*. 1–14.
[16] Leah R Fowler, Charlotte Gillard, and Stephanie R Morain. 2020. Readability and accessibility of terms of service and privacy policies for menstruation-tracking smartphone applications. *Health Promotion Practice* 21, 5 (2020), 679–683.
[17] Steven M Furnell, Adila Jusoh, and Dimitris Katsabas. 2006. The challenges of understanding and using security: A survey of end-users. *Computers & Security* 25, 1 (2006), 27–35.
[18] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2020. Overview of the Transformer-based Models for NLP Tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 179–183.
[19] Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. 2016. How short is too short? implications of length and framing on the effectiveness of privacy notices. In *Twelfth symposium on usable privacy and security (SOUPS 2016)*. 321–340.
[20] Google. 2021. Google Dictionary (by Google). https://chrome.google.com/webstore/detail/google-dictionary-by-goog/mgijmajocgfcbeboacabfgobmjgjcoja?hl=en
[21] Zhixiu Guo, Zijin Lin, Pan Li, and Kai Chen. 2020. {SkillExplorer}: Understanding the Behavior of Skills in Large Scale. In *29th USENIX Security Symposium (USENIX Security 20)*. 2649–2666.
[22] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–31.
[23] Suwan Li, Lei Bu, Guangdong Bai, Zhixiu Guo, Kai Chen, and Hanlin Wei. 2022. VITAS: Guided Model-based VUI Testing of VPA Apps. In *37th IEEE/ACM International Conference on Automated Software Engineering*. 1–12.
[24] Song Liao, Christin Wilson, Cheng Long, Hongxin Hu, and Huixing Deng. 2021. Problematic Privacy Policies of Voice Assistant Applications. *IEEE Security & Privacy* 19, 6 (2021), 66–73.
[25] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. 2020. The Privacy Policy Landscape After the GDPR. *Proceedings on Privacy Enhancing Technologies* 2020, 1 (2020), 47–64.
[26] Yuxi Ling, Kailong Wang, Guangdong Bai, Haoyu Wang, and Jin Song Dong. 2022. Are They Toeing the Line? Diagnosing Privacy Compliance Violations among Browser Extensions. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering (ASE)*.
[27] Kulani Mahadewa, Yanjun Zhang, Guangdong Bai, Lei Bu, Zhiqiang Zuo, Dileepa Fernando, Zhenkai Liang, and Jin Song Dong. 2021. Identifying privacy weaknesses from multi-party trigger-action integration platforms. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 2–15.
[28] Sunil Manandhar, Kaushal Kafle, Benjamin Andow, Kapil Singh, and Adwait Nadkarni. 2022. Smart Home Privacy Policies Demystified: A Study of Availability, Content, and Coverage. In *31st USENIX Security Symposium (USENIX Security 22)*. 3521–3538.
[29] Medium. 2014. Read Time and You. https://blog.medium.com/read-time-and-you-bc2048ab620c
[30] Mark Huasong Meng, Qing Zhang, Guangshuai Xia, Yuwei Zheng, Yanjun Zhang, Guangdong Bai, Zhi Liu, Sin G Teo, and Jin Song Dong. 2023. Post-GDPR threat hunting on android phones: dissecting OS-level safeguards of user-unresettable identifiers. In *The Network and Distributed System Security Symposium (NDSS)*.
[31] Anca Micheti, Jacquelyn Burkell, and Valerie Steeves. 2010. Fixing broken doors: Strategies for drafting privacy policies young people can understand. *Bulletin of Science, Technology & Society* 30, 2 (2010), 130–143.
[32] J Miller, M Sumeeth, and RI Singh. 2011. Evaluating the Readability of Privacy Policies in Mobile Environments. *International Journal of Mobile Human Computer Interaction* 3, 1 (2011), 55–78.
[33] Douglas C Montgomery and George C Runger. 2010. *Applied statistics and probability for engineers*. John wiley & sons.
[34] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods* 46, 4 (2014), 1023–1031.
[35] Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *The workshop on comparing corpora*. 1–6.
[36] Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, Rohan Ramanath, et al. 2015. Disagreeable Privacy Policies: Mismatches Between Meaning and Users' Understanding. *Berkeley Technology Law Journal* (2015), 39–88.
[37] Julie M Robillard, Tanya L Feng, Arlo B Sporn, Jen-Ai Lai, Cody Lo, Monica Ta, and Roland Nadler. 2019. Availability, readability, and content of privacy policies and terms of agreements of mental health apps. *Internet Interventions* 17 (2019), 100243.
[38] Rusne Sileryte, Alexander Wandl, and Arjan van Timmeren. 2022. The responsibility of waste production: comparison of European waste statistics regulation and dutch national waste registry. *Waste Management* 151 (2022), 171–180.
[39] Ravi Inder Singh, Manasa Sumeeth, and James Miller. 2011. A user-centric evaluation of the readability of privacy policies in popular web sites. *Information Systems Frontiers* 13, 4 (2011), 501–514.
[40] Aikaterini Soumelidou and Aggeliki Tsohou. 2021. Towards the creation of a profile of the information privacy aware user through a systematic literature review of information privacy awareness. *Telematics and Informatics* 61 (2021), 101592.
[41] Donna Spencer. 2009. *Card sorting: Designing usable categories*. Rosenfeld Media.
[42] M Sumeeth, RI Singh, and J Miller. 2010. Are Online Privacy Policies Readable? *International Journal of Information Security and Privacy* 4, 1 (2010), 93–116.
[43] Jenny Tang, Hannah Shoemaker, Ada Lerner, and Eleanor Birrell. 2021. Defining Privacy: How Users Interpret Technical Terms in Privacy Policies. *Proc. Priv. Enhancing Technol.* 2021, 3 (2021), 70–94.
[44] Matthew W Vail, Julia B Earp, and Annie I Antón. 2008. An empirical study of consumer perceptions and comprehension of web site privacy policies. *IEEE Transactions on Engineering Management* 55, 3 (2008), 442–454.
[45] Shufan Wang, Laure Thompson, and Mohit Iyyer. 2021. Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 10837–10851.
[46] Flynn Wolf, Ravi Kuber, and Adam J Aviv. 2019. " Pretty Close to a Must-Have" Balancing Usability Desire and Security Concern in Biometric Adoption. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[47] Tingmin Wu, Rongjunchen Zhang, Wanlun Ma, Sheng Wen, Xin Xia, Cecile Paris, Surya Nepal, and Yang Xiang. 2020. What risk? I don't understand. An Empirical Study on Users' Understanding of the Terms Used in Security Texts. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*. 248–262.
[48] Fuman Xie, Yanjun Zhang, Hanlin Wei, and Guangdong Bai. 2022. UQ-AAS21: A Comprehensive Dataset of Amazon Alexa Skills. In *International Conference on Advanced Data Mining and Applications*. Springer, 159–173.
[49] Fuman Xie, Yanjun Zhang, Chuan Yan, Suwan Li, Lei Bu, Kai Chen, Zi Huang, and Guangdong Bai. 2022. Scrutinizing privacy policy compliance of virtual personal assistant apps. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering (ASE)*.
[50] Zarul Fitri Zaaba, Steven Furnell, and Paul Dowland. 2011. End-User Perception and Usability of Information Security.. In *HAISA*. 97–107.
[51] Razieh Nokhbeh Zaeem and K Suzanne Barber. 2020. The effect of the GDPR on privacy policies: Recent progress and future promise. *ACM Transactions on Management Information Systems (TMIS)* 12, 1 (2020), 1–20.
[52] Nan Zhang, Xianghang Mi, Xuan Feng, XiaoFeng Wang, Yuan Tian, and Feng Qian. 2019. Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1381–1396.

A TABLES

Table 4: 10 categories of technical terms based on GDPR

Category	Description	Examples	Source (GDPR [2])
Data collection	Companies need to describe what data they will collect from users. This category contains any terms relating to data collection, like what type of data the developers will collect and where they obtain the data from.	action tag, voice id, Google AdWords	Art. 5 (1); recitals 32 - 50, 58, 60
Data use	Companies are required to describe how the collected data will be used. This category contains terms about how the developers will use the collected data and for what purpose.	analysis purposes, anonymize collected data	Art. 6 (1), 7 (1), 9, 10, 11, 13 (2) lit. b, 14 (2) lit. b, 15 - 18; recitals 50 - 52, 63
Data sharing	Companies are allowed to share the collected data with third parties for some specific purposes. This category contains terms about what third parties the companies will share the collected data with and for what purpose.	auxiliary products, bot detection	Art. 13 (1) lit. e, 14 (1) lit. e, 44 - 50; recitals 61
User rights	GDPR regulates that users have some rights to control the data collected by developers. This category contains terms about what rights the users have and what they can do to the data collected by developers.	delete your voice id, disable location-based services	Art. 7 (3), 12 - 23, 77 - 79, 82; recital 63 - 70
Data retention	Companies are allowed to retain the data collected from users for a period. This category includes terms about what data the companies will retain, the storage period of retained data, and the purpose of retention.	backup archives, anonymized audio recordings, retention time	Art. 13 (2), 14 (2), 17; recitals 64
Data security	Companies are required to implement appropriate techniques to protect users' data. This category includes terms about what security mechanisms the companies will use and its relevant description.	cryptographic protocol, data integrity	Art. 32; recital 78
Children	The privacy policy has a specific section for children to describe some special rules in such situation. This category includes terms about what special rights the children will have, what data the companies will not collect from children, and what special consents the companies need to obtain from children and their guardians.	child Apple id, Spotify kids account	Art. 6 (1) lit. f, 8, 12 (1); recitals 38, 58, 65
Special/Region Law	The privacy policy has a specific section to describe the special laws or region laws the companies have to follow. This category contains terms about such laws or rules for the special region or special services.	Nevada Revised Statutes, Alexa communication schedule	Art. 85 - 91; recitals 23, 24, 110
Changes to the policy	Companies need to update their privacy policy if they change some policies for their services. This category contains the terms about describing their changes and notification of users.	update the notice's effective date, prominent notice	Art. 12
Others	Besides the sections mentioned above, some companies describe the advertisements or other services they use in their Alexa skill. This category contains terms about the special sections the companies mention but are not commonly regulated by GDPR.	digital environment, browser's audio reader	

Table 5: 5 Classes of technical terms based on the Clustering algorithm.

Assigned Class Name	Size	The closest terms with the centroid point	Class Description
(0) Privacy/ Skill Settings	402	privacy setting, right to request the deletion of your personal information, right to rectification, right to limitation to processing, procedural measures, transient use	Any settings that enable users to control their data collection or data privacy, such as disabling some tracking techniques or only permitting transient use. Any lawful rights that the skills and users have to control the collected data. This class is more related to the "User Rights" category.
(1) Alexa Skill Data Usage	546	delete or anonymize data, transfer your information, retain your anonymized information, analyze our services, access to your information, provide transactional information	Any specific purposes or behaviors that the Alexa skill has for the collected data from users. This class is more related to the "Data Use" category. It contains the specific behaviors and legal rights that the Alexa skill will do with the users' data. Any behavior or legal rights that the developers or operators can have or apply for their applications.
(2) Security-Related Duty	352	duty of confidentiality, principal rights under data protection law, compliance with a legal obligation, anti-fraud and security purposes, fraud prevention information, authorized agent of a consumer	This class includes any terms about the security technique or purposes that the Alexa skill will use or work for. Additionally, it also includes some terms about data protection law or the contents the data protection law for security purposes. This class is more related to the "Data Security", "Special/Region Law" category.
(3) User Data	1370	service data, user identification information, opt out of the data collection, internet activity information, in-app tracking methods, automatic data collection tools	Any data type that the Alexa skills might collect from users, use or retain for a certain purpose and share with others. It also includes some special tools or methods that can track users' data about online activities and the selections about such tools that users can opt out of them or continue to use them. This class is more related to the "Data Collection", "Data Use", "Data Sharing" and "Data Retention" category.
(4) Lawful Vendors and Related Clauses	70	quasi-governmental agencies, government entities, Corporate Transactions, Belkin affiliated companies, california consumer privacy act, European commission's standard contractual clauses	This class includes two sections: 1. the third parties or other vendors that the skill might share the collected data with; 2. terms about any laws or regulations that the privacy policies mention and follow, especially some specific clauses that the third parties use and follow. This class is more related to the "Data Sharing" and "Special/Region Law" category.

Table 6: Overview of the results of proportion z test for Stage Two analysis. "Category/Terms" means that the hypotheses are for the corresponding categories or the corresponding terms. There are three experiment groups: the control group using original policy snippet (CG), the first experimental group using domain-specificity principle (EG1), and the second experimental group using implication-oriented principle (EG2). We have four accuracy levels: all correct (All_Co), part correct (Part_Co), part incorrect (Part_In), all incorrect (All_In).

	Category	Original Hypothesis: H_0	Alternative Hypothesis: H_a	z-score	p-value
1	Total	$H_{0_1} : P_{CG_{All\_Co}} >= P_{EG1_{All\_Co}}$	$H_{a_1} : P_{CG_{All\_Co}} < P_{EG1_{All\_Co}}$	-2.182	0.015
2	Total	$H_{0_2} : P_{CG_{All\_Co}} = P_{EG2_{All\_Co}}$	$H_{0_2} : P_{CG_{All\_Co}} \neq P_{EG2_{All\_Co}}$	-1.155	0.248
3	Data Collection	$H_{0_3} : P_{CG_{All\_Co}} >= P_{EG1_{All\_Co}}$	$H_{a_3} : P_{CG_{All\_Co}} < P_{EG1_{All\_Co}}$	-2.278	0.011
4	Data Sharing	$H_{0_4} : P_{CG_{All\_Co}} >= P_{EG1_{All\_Co}}$	$H_{a_4} : P_{CG_{All\_Co}} < P_{EG1_{All\_Co}}$	-2.182	0.015
5	User Rights	$H_{0_5} : P_{CG_{All\_Co}} >= P_{EG1_{All\_Co}}$	$H_{a_5} : P_{CG_{All\_Co}} < P_{EG1_{All\_Co}}$	-2.111	0.017
6	Children	$H_{0_6} : P_{CG_{All\_Co}} >= P_{EG1_{All\_Co}}$	$H_{a_6} : P_{CG_{All\_Co}} < P_{EG1_{All\_Co}}$	-3.676	< 0.001
7	Data Collection	$H_{0_7} : P_{CG_{All\_Co}} = P_{EG2_{All\_Co}}$	$H_{a_7} : P_{CG_{All\_Co}} \neq P_{EG2_{All\_Co}}$	-0.661	0.508
8	Data Sharing	$H_{0_8} : P_{CG_{All\_Co}} = P_{EG2_{All\_Co}}$	$H_{a_8} : P_{CG_{All\_Co}} \neq P_{EG2_{All\_Co}}$	-0.937	0.349
9	User Rights	$H_{0_9} : P_{CG_{All\_Co}} = P_{EG2_{All\_Co}}$	$H_{a_9} : P_{CG_{All\_Co}} \neq P_{EG2_{All\_Co}}$	-0.676	0.499
10	Children	$H_{0_{10}} : P_{CG_{All\_Co}} >= P_{EG2_{All\_Co}}$	$H_{a_{10}} : P_{CG_{All\_Co}} < P_{EG2_{All\_Co}}$	-2.139	0.016
11	Children	$H_{0_{11}} : P_{EG1_{All\_Co}} = P_{EG2_{All\_Co}}$	$H_{a_{11}} : P_{EG1_{All\_Co}} \neq P_{EG2_{All\_Co}}$	1.633	0.102
12	Children	$H_{0_{12}} : P_{EG1_{Part\_Co}} = P_{EG2_{Part\_Co}}$	$H_{a_{12}} : P_{EG1_{Part\_Co}} \neq P_{EG2_{Part\_Co}}$	-0.806	0.42
13	Children	$H_{0_{13}} : P_{EG1_{Part\_In}} >= P_{EG2_{Part\_In}}$	$H_{a_{13}} : P_{EG1_{Part\_In}} < P_{EG2_{Part\_In}}$	-2.041	0.021
14	Data Security	$H_{0_{14}} : P_{CG_{All\_Co}} >= P_{EG2_{All\_Co}}$	$H_{a_{14}} : P_{CG_{All\_Co}} < P_{EG2_{All\_Co}}$	-3.491	< 0.001
15	Special/Region Law	$H_{0_{15}} : P_{CG_{All\_Co}} >= P_{EG2_{All\_Co}}$	$H_{a_{15}} : P_{CG_{All\_Co}} < P_{EG2_{All\_Co}}$	-2.673	0.004
16	Data Security	$H_{0_{16}} : P_{CG_{All\_Co}} = P_{EG1_{All\_Co}}$	$H_{a_{16}} : P_{CG_{All\_Co}} \neq P_{EG1_{All\_Co}}$	-0.28	0.779
17	Special/Region Law	$H_{0_{17}} : P_{CG_{All\_Co}} = P_{EG1_{All\_Co}}$	$H_{a_{17}} : P_{CG_{All\_Co}} \neq P_{EG1_{All\_Co}}$	-0.521	0.603
	Terms	**Original Hypothesis: H_0**	**Alternative Hypothesis: H_a**	**z-score**	**p-value**
18	Anonymized Audio Recordings	$H_{0_{18}} : P_{CG_{All\_Co}} >= P_{EG1_{All\_Co}}$	$H_{a_{18}} : P_{CG_{All\_Co}} < P_{EG1_{All\_Co}}$	-2.182	0.015
19	Web Beacons	$H_{0_{19}} : P_{CG_{All\_Co}} >= P_{EG1_{All\_Co}}$	$H_{a_{19}} : P_{CG_{All\_Co}} < P_{EG1_{All\_Co}}$	-2.417	0.008
20	DART Cookie	$H_{0_{20}} : P_{CG_{All\_Co}} >= P_{EG1_{All\_Co}}$	$H_{a_{20}} : P_{CG_{All\_Co}} < P_{EG1_{All\_Co}}$	-2.701	0.003
21	PI (personal information)	$H_{0_{21}} : P_{CG_{All\_Co}} = P_{EG1_{All\_Co}}$	$H_{a_{21}} : P_{CG_{All\_Co}} \neq P_{EG1_{All\_Co}}$	-0.459	0.646
22	PI (personal information)	$H_{0_{22}} : P_{CG_{Part\_Co}} >= P_{EG1_{Part\_Co}}$	$H_{a_{22}} : P_{CG_{Part\_Co}} < P_{EG1_{Part\_Co}}$	-2.811	0.005

B AUTOMATIC TERM CLASSIFICATION

In RQ1, technical terms were grouped into ten categories based on their position in skills' privacy policies. However, categories have the overlap and terms within a category may not have a similar characteristic. Thus, we used an unsupervised clustering algorithm to cluster terms into classes based on their meanings.

First, each term in the ASPP corpus was converted to a vector of array values. The BERT model is widely applied to process any sentences or documents as one of the most common NLP techniques [18]. It can transfer the words in sentences to the vector based on their meaning. Thus, a pre-trained BERT phrase model [45] was used to obtain the embedding of every term as its vector, where this model has been examined and performed better in phrases. Next, the clustering algorithm was used to categorize all terms in the ASPP Corpus. The clustering algorithm calculates the distance between the vectors and then selects some centroids to generate several clusters. The number of centroids is considered as the K value of the clustering algorithm.

Considering the size of the ASPP Corpus and the distribution of the vectors, we used the AgglomerativeClustering algorithm from the sklearn library. We customized the distance metrics and used the cosine distance to calculate the distance (similarity) of two vectors (term's embedding). Then we used the Elbow method to determine the K value. Setting K from 2 to 30, the Elbow method can plot the variation of the clustering algorithm with the different K values and pick the elbow of the curve as the K value (number of clusters). After applying the elbow method, we identified the K value for AgglomerativeClustering as 5. Table 5 in Appendix A shows the results of the clustering algorithm and lists the closest terms with the centroid in a cluster. After identifying the closest terms in each cluster and also observing the remaining terms, we assigned a name for each cluster (class) and summarized the description of a class. Each class contains terms from one or multiple categories.

Users might have some misunderstandings or lack related awareness about their rights and security-related knowledge.

As shown in Table 5 in Appendix A, Class 3 (User Data) has the most terms (50%) in the ASPP Corpus. Such terms are related to a broad or a specific data type that the skill will collect, use, share or retain. In this class, 407/1370 terms (around 30%) are difficult terms, indicating that users can comprehend most terms related to their data. However, for class 0 (Privacy / Skill Settings), 2 (Security-Related Duty) and 4 (Lawful Vendors and Related Clauses), the proportion of difficult terms is 40.05%, 44.03% and 47.14%. These results indicate that users might underestimate privacy risk due to lack of privacy awareness, and fail to take appropriate actions, resulting in unconscious acceptance of privacy risk when using VPA applications.

C DETAILED ANALYSIS

C.1 Detailed analysis for Finding 9

We first analyzed the overall performance of "All_Co" participants in *CG* and *EG1*. As shown in Table 6.1 in Appendix A, the statistics show that the overall proportion of "All_Co" participants in *CG* is smaller than the proportion of "All_Co" participants in *EG1*. We then analyzed the overall performance of "All_Co" participants in *CG* and *EG2*. However, the statistics show that the overall proportion of "All_Co" participants should have no significant differences between *CG* and *EG2* (shown in Table 6.2 in Appendix A). Therefore, we can conclude that in total, the explanation sentences using domain-specificity principle can significantly help users' understanding.

C.2 Detailed analysis for Finding 10

We first analyzed the performance of participants in each category between *CG* and *EG1*. Thus, we set the related hypotheses for those categories, which can compare the proportion of "All_Co" participants for each category in *CG* and *EG1*. The statistics (shown in Table 6.3 - 6 in Appendix A) indicates that for these four categories, the proportion of "All_Co" participants in *EG1* is significantly larger than the proportion of "All_Co" participants in *CG*. Therefore, we can infer that *the performance of participants in EG1 is better than the performance of participants in CG for terms related to "Data Collection", "Data Sharing", "User Rights" and "Children".*

We then analyzed the performance of those four categories between *CG* and *EG2*. The statistics show that the proportion of "All_Co" participants has no significant differences between *CG* and *EG2* for "Data Collection", "Data Sharing" and "User Rights" category (shown in Table 6.7 - 9 in Appendix A). The proportion only has the significant differences in "Children" category between *CG* and *EG2* (shown in Table 6.10 in Appendix A). Thus, we further compared the proportion of participants with different correctness categories between *EG1* and *EG2* for "Children" category. The proportion has no significant differences in "All_Co" and "Part_Co" participants between *EG1* and *EG2*. However, the proportion of "Part_In" participants in *EG1* is smaller than the proportion of "Part_In" participants in *EG2* (shown in Table 6.13 in Appendix A). We can infer that *explanation sentences using domain-specificity principle can help users understand more than sentences using implication-oriented principle for terms in those four categories.*

Moreover, our overall statistics present that the performance of participants between *CG* and *EG1* has no significant differences for "Data Use" category. However, "Data Use" category contains some terms about users' data. Thus, we further analyzed each term within this category (terms shown in Table 1). We used the proportion z test again to first compare the participants' performance of those terms about users' data in "Data Use" category. We set the related hypotheses for anonymized audio recordings, web beacons, PI and DART cookie. The statistics (shown in Table 6.18 - 20 in Appendix A) approve that users' understanding in *EG1* should be significantly better than those in CG. Additionally, for PI, even though the proportion of "All_Co" participants has no significant differences between *CG* and *EG1*, the proportion of "Part_Co" participants in *EG1* is larger than proportion in *CG* (shown in Table 6 .21 and 22 in Appendix A). Therefore, we concluded the Finding 10.

D USERS' COMMON FEEDBACK IN STAGE TWO

We list the most common feedback from users' responses in Stage Two.

- *I found the pop-up window very helpful. The only thing I can think of that would make it easier is if there were more examples of exactly what the skill can or can not do. Perhaps listing actual events that have happened as an example - a real-life situation would make me understand how it actually works and is applied to the situation.*
- *I would like to see some images as examples. I am a visual learner, so I think that that would help me a little.*
- *Highlighting/bolding the most important phrases or sentences might help. Also, examples are always helpful, like in Question 2, so adding more would be nice.*
- *I think a video or an audio recording of the policy will help people understand it very fast.*
- *The pop-ups are helpful, but the information is still thick and technical – the only way around that, it seems, would be to have one version which is legally binding and another provided just as an explanation written in more layman's terms.*
- *Having the option of a "plain English" version of each policy snippet would make things easier to understand, even if such a version lost some of the finer details.*
- *Provision of links to demos that could explain concepts in lay terms, using appropriate examples from the real world.*