Demystifying Privacy: Building Tools for Clear and Accessible Data Security Practices

Adalina Ma axma@ucsd.edu
David Yonemura dyonemura@ucsd.edu
Edward New enew@ucsd.edu
Haojian Jin haojian@ucsd.edu

Abstract

In an era where data privacy is a growing concern, but expressing that concern clearly and specifically seems to be a pain point for most, as reflected by the privacy paradox, efficiently understanding and categorizing user privacy preferences is crucial not only for researchers and developers but the users themselves as well. Traditional qualitative coding methods for privacy concerns can be time-consuming and inconsistent, making it difficult to extract meaningful insights from user responses. Our project explores novel ways to streamline and enhance qualitative coding by leveraging AI-driven generative surveys and structured response filtering. By dynamically adapting survey questions and refining response specificity, we aim to make privacy preference elicitation more efficient and scalable. This approach seeks to improve the accuracy and depth of qualitative coding while reducing manual effort, ultimately providing a more structured and systematic way to analyze user privacy concerns.

Website: https://github.com/DataSmithLab/PrIDE-web/tree/survey
Code: https://github.com/DataSmithLab/PrIDE-web/tree/survey

1 Introduction

In today's rapidly evolving digital landscape, data privacy has become an increasingly critical concern. As organizations and service providers collect, process and share more personal data, users are becoming more aware of how their information is used and the potential risks involved. However, despite the growing importance of data privacy, there are significant challenges to ensure that privacy concerns are understood and addressed effectively. One of the main obstacles is the complexity of translating qualitative privacy preferences into actionable insights. Users express privacy concerns in nuanced and subjective ways that are often difficult to capture and analyze in a systematic manner. Moreover, users frequently struggle to articulate their privacy preferences clearly and specifically. A well-known issue is the privacy paradox, where individuals' actions often contradict their stated privacy concerns. This paradox may stem from users' inability to accurately express their privacy preferences, leading to the perceived gap between their words and actions.

Traditional methods of capturing privacy preferences—such as long arduous surveys with a laundry list of predefined choices—may fail to account for the subtle variations in how individuals perceive and prioritize their privacy. These methods often lack the flexibility to capture the full range of user concerns or to adapt to the evolving nature of user preferences. Another common method involves collecting free-text responses from users, which are then manually labeled by researchers or trained qualitative coders to identify underlying privacy concerns. However, this approach is slow, labor-intensive, prone to errors, and costly due to the reliance on manual labeling.

This project aims to explore novel ways to make the qualitative coding of user privacy preferences more efficient and specific. By focusing on improving the methods and tools used to analyze user-generated data, we seek to develop more precise and scalable techniques for capturing and categorizing privacy concerns. This involves leveraging the power of large language models such as ChatGPT to generate context-specific survey response options, combing the two traditional approaches to qualitative coding mentioned previously. Our goal is to streamline the coding process, enhance the specificity of the results, and ultimately provide more accurate insights into what users truly care about when it comes to their privacy.

Through this work, we aim to bridge the gap between the qualitative nature of user privacy preferences and the quantitative analysis required for effective privacy policy implementation. By improving the efficiency and specificity of qualitative coding, we hope to empower organizations to better understand and address user privacy concerns, fostering a stronger sense of trust and alignment with users' expectations. Additionally, this project has the potential to contribute to the broader field of privacy research, helping to shape more effective privacy policies and tools that are in line with the needs of today's digital society. Beyond privacy, the techniques explored in this system can be generalized and applied to any domain requiring qualitative coding, offering valuable insights and a new approach to qualitative coding.

2 Methods

2.1 Generative Choice Survey

To achieve our goals, we employed several key strategies and technical steps:

1. **Survey Component Integration**: Edward contributed significantly by integrating the 'survey-react-ui' package, which forms the backbone of the survey generation functionality. This package provides an intuitive way to programmatically generate and customize surveys in React, a critical step in enabling interactive survey flows for users. This integration also involved adjusting the survey structure to support dynamic content, eventually allowing the survey to change based on user responses.
2. **Development of Survey Generation**: Once the new survey package had been successfully integrated, David worked on adapting the current survey data schema to the new survey structure of 'survey-react-ui', allowing us to generate a unique survey for each decoupled privacy diagram data path.
3. **Survey Page Routes and Export Functionality**: Edward, alongside David, implemented new dynamic routes to handle survey/data path-specific pages. We leveraged session cookies to uniquely identify each user session, allowing developers to send privacy surveys to users with one generalized link. This new direct survey deployment workflow for developers means developers can have immediate access to survey responses, completely eliminating the reliance on third-party systems such as Qualtrics for survey distribution and response collection.
4. **Database and Backend Enhancements**: We redesigned the database schema to support multi-session storage, enabling longitudinal analysis of user responses and improving the adaptability of future surveys.
5. **Generative Survey Options**: We utilized the ChatGPT API to generate adaptive survey response options tailored to users' previous answers, enabling a more efficient and precise identification of their privacy preference profiles.

3 Results

The project resulted in the development of an integrated platform that combines the two traditional approaches to qualitative coding, resulting in a novel system that leverages the power of advanced large language models to dynamically generate context-specific survey response options for each question. The key outcomes include:

1. **Dynamic Survey Integration:**
   - A survey system powered by the `survey-react-ui` package that adapts questions based on user inputs.
   - Metadata-enriched survey nodes for tracking and documenting data interactions.
   - Features for exporting survey results and linking responses to backend storage for transparency.

- Automation of survey creation steps, eliminating reliance on external tools like Qualtrics by using our own programs to dynamically generate and manage surveys.

2. **Deliverables:**
   - **Generative Survey Platform:** Overhauled survey generation, distribution, and collection system allowing for future development of dynamic surveys.

4 Discussion

The ideas and frameworks explored in this project represents a significant step forward in addressing the challenges of understanding and categorizing user privacy preferences. By leveraging AI-driven generative surveys and structured response filtering, we have created a system that not only streamlines the qualitative coding process but also enhances the specificity and accuracy of the insights derived from user responses. This approach directly tackles the limitations of traditional methods, such as manual coding of free text responses and static surveys, which are often time-consuming, inconsistent, and fail to capture the nuanced nature of privacy concerns.

One of the key achievements of this project is the ability to bridge the gap between qualitative user privacy preferences and quantitative analysis. By automating the generation of survey options and refining response specificity, the system reduces the manual effort required for qualitative coding while improving the depth and accuracy of the results. This not only benefits researchers and developers but also empowers users by ensuring their privacy concerns are more accurately captured and addressed.

Despite these advancements, further research is still needed to more rigorously benchmark the effectiveness of this novel approach against established baselines. Specifically, future work should compare the performance of our system to a few key alternatives: (1) the traditional method of manual labeling of free-text responses, (2) a modified approach where a Large Language Model (LLM) is used to directly label free-text responses, and (3) the traditional exhaustive survey. Such comparisons would provide valuable insights into the relative strengths and limitations of each method, particularly in terms of accuracy, efficiency, scalability, and the quality of data collected.

5 Conclusion

This project has laid the groundwork for a novel approach to qualitative coding, focusing on the specific challenge of capturing and analyzing user privacy preferences. By combining AI-driven generative surveys with structured response filtering, the system offers a scalable and efficient alternative to traditional methods, which are often labor-intensive and or lead to vague, general answers. The dynamic generative survey system, powered by survey-react-ui and ChatGPT, demonstrates the potential to improve the specificity and accuracy of qualitative coding while reducing manual effort.

However, as an exploratory effort, this project is admittedly limited in its scope and lacks rigorous experimentation and user testing. Further research is essential to validate the effectiveness of this approach. Future studies should benchmark the system against traditional manual labeling of free-text responses and a modified approach using an LLM to label free-text responses. These comparisons will help determine the relative advantages and limitations of each method, providing a clearer understanding of how this novel approach can be refined and applied more broadly.

Ultimately, this project highlights the potential of AI-driven tools to transform qualitative coding processes, particularly in the context of privacy research. By continuing to build on this foundation, future work can contribute to the development of more effective privacy policies and tools, fostering greater trust and alignment between users and organizations in the digital age.