A Comparative Analysis between AI and Human Coding in Survey Research

Isabela Bertolini Coelho, Frauke Kreuter

Joint Program in Survey Methodology, University of Maryland
icoelho1@umd.edu

Introduction

Privacy is central to discussions surrounding data protection and ethical considerations in both survey methodology and AI. Understanding stakeholders' attitudes, perceptions, and participation levels toward privacy is crucial to identifying the barriers to adopting formal privacy models in sample survey data, especially for official statistics.

In this study, we present a comparative analysis between Large language models (LLMs)-generated codifications and human-coded responses to open-ended questions regarding privacy.

Based on the transcriptions of two focus groups conducted with data privacy experts, our investigation delves into the similarities and disparities between codifications generated by LLMs and those crafted by human coders.

Human Coding

Human Coding involves the manual categorization and interpretation of the data by researchers, aiming to uncover patterns, themes, or relationships that provide deep insights into human experiences and societal phenomena.

The final codebook for this study includes:

- Defining privacy: Participants describing how they define privacy.
- Privacy topics: Privacy topics that participants identify as most concerned about.
- Differential Privacy: Participants describing their familiarity and how they define Differential Privacy.
- Formal Privacy: How do participants define Formal Privacy.
- Privacy breaches: Participants discuss the privacy breaches they are aware of.
- Privatizing data: How often do respondents privatize data, and what type of data is privatized.
- Privacy protection methods: Types of privacy protection methods that respondents know.
- Reasons: Reasons for using privacy protection methods in official statistics.

LLM Coding

LLMs can help automate the initial coding of qualitative data by identifying common themes, sentiments, or patterns within the data. Before coding, qualitative data often needs transcription and preprocessing. LLMs can automate transcription from audio recordings and help clean and prepare text data for coding, saving time and reducing errors.

Codebook Structure

Theme: Definitions of Privacy

- **Code**: Personal Definition
  - **Definition**: The participant's personal or layman understanding of privacy.
  - **Example**: "Privacy is the right not to be disturbed or monitored by government or businesses without consent."
- **Code**: Professional Definition
  - **Definition**: Technical or academic definitions of privacy as understood in the participant's professional context.
  - **Example**: "Privacy involves controlling the access and use of personal data."
- **Code**: Privacy vs. Confidentiality
  - **Definition**: Participants' distinctions between the concepts of privacy and confidentiality, highlighting how these concepts are understood and applied differently.
  - **Example**: "Privacy pertains to the individual's right to control their personal information, while confidentiality concerns the obligation of the data holder to protect personal information from unauthorized access."

Theme: Differential Privacy

- **Code**: Familiarity
  - **Definition**: Participant's level of knowledge and familiarity with differential privacy.
  - **Example**: "Very familiar as I work with differential privacy."
- **Code**: Concerns and Critiques
  - **Definition**: Concerns or criticisms regarding the application or implications of differential privacy.
  - **Example**: "Concerned about adding too much noise and making data useless."

Figure 1. Codebook Structure proposed by GPT-4.

Discussion

Human coding is labor-intensive and subjective, relying on the researcher's ability to discern nuances and context. To ensure reliability and validity, researchers often employ strategies such as inter-coder reliability checks. As a result, the costs associated with human coding are high, and the number of interviews that can be coded is limited [1].

LLMs streamline the coding process by automatically identifying preliminary themes and patterns due to their ability to rapidly process vast datasets, thereby offering a cost-effective alternative.

In this study, the results from human and AI coding are consistent. It is important to note that LLMs might not fully grasp the context or nuances of qualitative data, leading to potential inaccuracies or oversimplifications in coding.

LLMs can inherit biases from their training data, which could skew analysis if not carefully monitored. Thus, while the use of AI tends to save time and costs and minimize coding errors in qualitative research, it does not eliminate the need for researchers to validate the AI-generated results. This ensures that the coding aligns with the research objectives and adheres to ethical considerations [2, 3].

Another important aspect of using AI for coding qualitative data is the potential for privacy breaches. The most significant concern is during model training, where there's a risk that the model may incorporate specific data points into its parameters. This could allow the model to generate outputs related to or reminiscent of the original data, posing a privacy risk if used later for different purposes. To mitigate this, we employed GPT-4 in our study.

References

[1] Anna-Carolina Haensch, Bernd Weiß, Patricia Steins, Priscilla Chyrva, and Katja Bitz. The semi-automatic classification of an open-ended question on panel survey motivation and its application in attrition analysis. *Frontiers in big Data*, 5:880554, 2022.

[2] Bernard J Jansen, Soon-gyo Jung, and Joni Salminen. Employing large language models in survey research. *Natural Language Processing Journal*, 4:100020, 2023.

[3] Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*, 2024.