Poster Type: Innovative Idea

Proof of Concept: Leveraging Large Language Models for Qualitative Analysis of Participant Feedback

Andrew B. Crocker[1,2], Marcelo Schmidt[1a], Juan D. Tejeda[1], and Howard Rodríguez-Mori[1]

[1]Texas Tech University – School of Veterinary Medicine, Amarillo, TX, USA.
[2]Texas A&M AgriLife Extension Service – Family & Community Health, Amarillo, TX, USA.

[a]Corresponding Author:

Marcelo Schmidt, PhD
TTU School of Veterinary Medicine
7671 Evans Dr
Amarillo, TX 79106
marcelo.schmidt@ttu.edu

Poster Type: Innovative Idea

Proof of Concept: Leveraging Large Language Models for Qualitative Analysis of Participant Feedback

Introduction

Documenting programmatic successes in Extension education and its impact on the individuals the program serves is an important task. Quantitative data, foundational to program evaluation, helps document program successes, fulfill mandatory reporting requirements, and facilitate professional advancement (Baughman et al., 2012; Vengrin et al., 2018), but quantitative feedback is only part of the story. The value of qualitative feedback, rich in context and nuance, warrants continued acknowledgment in scholarly discourse to support critical evaluation functions. Taylor-Powell and Boyd (2008) discussed evaluation capacity building and defined it as the deliberate effort to establish and maintain processes within an organization that consistently ensure the effectiveness and utilization of high-quality evaluations.

Large Language Models (LLM; e.g., GPT-4, LLaMA, BERT) present an innovative alternative to traditional analysis methods. These models can read and summarize vast amounts of text and, if prompted adequately, generate actionable insights from qualitative responses. LLMs can capture intricate themes, provide a richer, more contextual understanding of a text, and generate human-like language (OpenAI, 2023b). This study aimed to test the capabilities of LLMs, a form of AI (Artificial Intelligence), to analyze qualitative feedback from participants in a community-based Extension education program, comparing results with human analysis utilizing traditional analysis methods. This study advances the AAAE's mission by increasing prosperity through innovation in Agriculture, Food, and Natural Resources (AFNR) systems by exploring the innovative digital tools and practices and their potential applicability in AFNR initiatives.

Methodology

The Texas Tech University Institutional Review Board reviewed and approved this study (IRB # 2023-949). The researchers used a convenience sample and analyzed post-evaluation responses to an open-ended question from program participants ($N = 118$) from AgriLife Extension's *A Matter of Balance* fall risk-reduction program (MaineHealth, n.d.) implemented between January and September 2023 in rural Texas. In particular, the analysis focused on responses to the question: "What other changes have you made as a result of this class?"

Three research team members participated in the qualitative analysis, following analysis recommendations made by Esterberg (2001). The team met to review and discuss their 48 initial codes, engaging in rich discussion about frequency, relevance to the research, and meaning. The team negotiated an agreed-upon set of final codes by reviewing them for overlap, redundancy, and potential relationships. Each team member returned to the comments using a focused coding process to recode the data based on the five agreed-upon codes, ensuring clarity and consistency in their application and establishing themes for changes participants made after attending the program series.

Inductive and deductive processes were used for the qualitative analysis conducted by AI. The researchers developed a protocol (Dai et al., 2023; De Paoli, 2023; Xiao et al., 2023) and followed

tips for efficient prompt engineering based on White et al. (2023). To begin the process, the researchers started a new chat using ChatGPT version GPT-3.5 (OpenAI, 2023a) and followed the protocol, which mimicked the one used by the human researchers. The entire output from ChatGPT was copied into a Word file to document the process and outcomes since there were no reliable ways to export/capture these types of interactions at the time of this study.

Results/Implications

Participants were predominately female (65%) and Caucasian (68%), reporting no Hispanic/Latino heritage (70%) and with an average age of 78 years. Human researchers analyzed the data over several weeks, including at least four face-to-face meetings, resulting in five main themes/codes. As we had multiple raters and categorical variables, we computed Fleiss' Kappa and noted moderate initial agreement ($\kappa = 0.555$, $p < .001$).

The researchers asked ChatGPT to identify an initial set of codes, resulting in 11 codes “based on common themes and patterns.”. Then they asked ChatGPT to refine its codes, providing five main codes based on the participants’ responses to align with the human researchers’ findings. The human researchers compared the ChatGPT-generated codes with their own by visually reviewing them side-by-side (Hamilton et al., 2023). Further, the researchers used Cohen’s Kappa for two raters to determine the agreement rate between the final human output and the final ChatGPT-generated output. The result denoted minimal agreement ($\kappa = .392$, $p < .001$).

ChatGPT's involvement added a novel dimension to the analysis. The qualitative analysis process in this study involved extensive human effort, with three researchers dedicating several weeks and multiple face-to-face meetings. In contrast, the ChatGPT analysis required only about ten minutes for protocol preparation and less than five minutes of interaction to yield results.

Advice

In the Extension context, AI-assisted qualitative data analysis may be most beneficial for program outcome and impact summaries and interpretation to stakeholders. As Agency-specific databases for reporting and commercially available online survey platforms have streamlined the collection and analysis of quantitative data, it seems only natural for an LLM-based platform or the like to take qualitative analysis beyond the realm of the simple word cloud. However, just as the online survey platform cannot perform more sophisticated analysis, this “quick-to-market” type of qualitative analysis may not be suited for more rigorous endeavors such as peer-reviewed research.

Costs/Resources

The research team had no costs associated with this project other than its professional time. The team explored how an API (Application Programming Interface) can enable more efficient and scalable data analysis, replacing manual inputs into ChatGPT with a streamlined, automated system.

References

Baughman, S., Boyd, H. H., & Franz, N. K. (2012). Non-formal educator use of evaluation results. *Evaluation and Program Planning*, *35*(3), 329–336. https://doi.org/10.1016/j.evalprogplan.2011.11.008

Dai, S.-C., Xiong, A., & Ku, L.-W. (2023). *LLM-in-the-loop: Leveraging large language model for thematic analysis* (arXiv:2310.15100). arXiv. http://arxiv.org/abs/2310.15100

De Paoli, S. (2023). *Can large language models perform an inductive thematic analysis of semi-structured interviews?* https://doi.org/10.48550/arXiv.2305.13014

Esterberg, K. G. (2001). *Qualitative methods in social research* (1st ed.). McGraw-Hill Humanities/Social Sciences/Languages.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *_irr: Various coefficients of interrupter reliability and agreement_* (Version 0.84.1) [Computer software]. https://CRAN.R-project.org/package=irr

Hamilton, L., Elliott, D., Quick, A., Smith, S., & Choplin, V. (2023). Exploring the use of AI in qualitative analysis: A comparative study of guaranteed income data. *International Journal of Qualitative Methods*, *22*, 16094069231201504. https://doi.org/10.1177/16094069231201504

MaineHealth. (n.d.). *Fall Prevention | A Matter of Balance*. Retrieved October 18, 2023, from https://www. mainehealth.org/Services/Aging-Senior-Care/Matter-of-Balance

OpenAI. (2023a). *ChatGPT*. https://chat.openai.com

OpenAI. (2023b). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. http://arxiv.org/abs/2303.08774

Taylor-Powell, E., & Boyd, H. H. (2008). Evaluation capacity building in complex organizations. *New Directions for Evaluation*, *2008*(120), 55–69. https://doi.org/10.1002/ev.276

Vengrin, C., Westfall-Rudd, D., Archibald, T., Rudd, R., & Singh, K. (2018). Factors affecting evaluation culture within a non-formal educational organization. *Evaluation and Program Planning*, *69*, 75–81. https://doi.org/10.1016/j.evalprogplan.2018.04.012

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT* (arXiv:2302.11382). arXiv. http://arxiv.org/abs/2302.11382

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. *28th International Conference on Intelligent User Interfaces*, 75–78. https://doi.org/10.1145/3581754.3584136