学 位 論 文 の 要 旨

氏名　坂口　公太

学　位　論　文　名　Evaluating ChatGPT in Qualitive Thematic Analysis with Human Researchers in the Japanese Clinical Context and its Cultural Interpretation Challenges: Comparative Qualitive Study

発　表　雑　誌　名　J Med Internet Res
（巻，初頁～終頁，年）　(27, e71521, 2025)

著　　者　　名　Kota Sakaguchi, Reiko Sakama, Takashi Watari

論 文 内 容 の 要 旨

INTRODUCTION

Qualitative research is vital for understanding the underlying values, beliefs, and emotional experiences of individuals, particularly within the social sciences and complex health care domains. This methodology traditionally relies on meticulous manual analysis by experienced human researchers, a process that is inherently time-consuming and resource intensive. The recent and rapid emergence of sophisticated artificial intelligence (AI) technology, specifically large language models (LLMs) such as OpenAI's ChatGPT, presents a compelling opportunity to potentially streamline and enhance the efficiency of qualitative data analysis. Existing scholarly literature examining the application of AI in this context, however, has primarily centered its focus on English-language datasets. Consequently, the applicability and effectiveness of these advanced AI tools for qualitative analysis of non-English languages, especially those with complex structural and high-contextual features like Japanese, remain significantly underexplored and warrant detailed investigation. This study was therefore designed to address this critical gap.

The primary objective of this research was to conduct a direct comparative evaluation of the performance of ChatGPT against that of experienced human qualitative researchers when analyzing Japanese qualitative data specifically sourced from a clinical setting. A secondary, yet

equally important, objective was to systematically identify and characterize the specific cultural and contextual interpretation challenges that are encountered by the large language model during its analytical process on such specialized, high-context data. This comparative study aims to provide crucial evidence regarding the responsible and effective integration of LLM technology into non-English qualitative research methodologies, particularly within sensitive clinical and health-related contexts.

MATERIALS AND METHODS

This study was designed as a comparative qualitative research project. The data utilized for the analysis consisted of one English-language transcribed interview and one Japanese-language transcribed interview, both derived from a pre-existing, ethically approved study focusing on patient-doctor communication dynamics. Two highly experienced human qualitative researchers, both proficient in thematic analysis, independently performed the analytical process on both the English and Japanese transcripts. Subsequently, the identical Japanese-language dataset was submitted for thematic analysis to ChatGPT version 3.5. To establish a comprehensive basis for comparison, the entire thematic analysis process was rigorously compared across three distinct analytical conditions: (1) Human analysis of the English transcript, (2) Human analysis of the Japanese transcript, and (3) ChatGPT 3.5 analysis of the Japanese transcript. The primary comparative metrics included the model's efficacy in theme identification, the depth of nuance capture achieved, and the overall quality and depth of the resulting thematic interpretations.The study protocol was approved by the Research Ethics Committee of Shimane University.

RESULTS AND DISCUSSION

The comparative analysis yielded several notable findings. ChatGPT demonstrated a basic level of success in identifying and coding surface-level, manifest themes within the Japanese qualitative data. However, a significant performance gap emerged when comparing the LLM's output to that of the human researchers, particularly concerning the deeper, culturally nuanced, and context-dependent interpretations. Specifically, the LLM encountered considerable difficulties in accurately interpreting culturally rich or context-dependent Japanese expressions and understanding the implied meanings that are commonplace within the subtleties of the clinical setting. These interpretation failures consistently resulted in a thematic structure that was notably shallower and less comprehensive than the robust analysis produced by the human researchers. Interestingly, the LLM’s analytical performance on the English transcript was found to be more closely aligned with the human-generated English analysis, suggesting a language-specific degradation in performance for complex, high-context non-English data.

These results underscore a fundamental limitation of the current generation of LLMs: a dependency on direct linguistic translation and a notable deficiency in grasping intricate cultural

and professional contexts. While the model could process the language, it often failed to interpret the meaning and significance behind the words, which is paramount in qualitative research. This finding strongly suggests that, for non-English, high-context qualitative data, LLMs are not yet suitable as standalone tools for comprehensive thematic analysis.

CONCLUSION

ChatGPT exhibits demonstrated potential as a valuable supplementary or assistive tool for undertaking the initial, large-scale coding and surface-level identification of themes within Japanese qualitative datasets. Nevertheless, its current inability to effectively grasp and incorporate critical cultural and intricate clinical contexts severely limits its overall utility for conducting comprehensive, nuanced, and methodologically sound qualitative analysis. Moving forward, future research efforts must be strategically focused on developing and implementing culturally adaptive LLMs and concurrently establishing robust, field-validated quality assessment and assurance frameworks for all aspects of AI-assisted qualitative research, particularly those conducted within non-English and high-context clinical research environments. The responsible integration of AI in this domain mandates that human researchers maintain full oversight and responsibility for the final, critical interpretation of findings.