American Journal of Qualitative Research
2026, Vol. 10 No. 4, pp. 126-147
https://doi.org/10.66815/ajqr/19234
http://www.ajqr.org

AMERICAN JOURNAL OF QUALITATIVE RESEARCH
ISSN: 2576-2141

Exploring ChatGPT-Assisted Qualitative Data: Comparing Human and AI-Assisted Coding Using Braun and Clarke's Thematic Analysis

Laura Hagan[1] and Alla Chavarga
Touro University, New York, NY, USA

ABSTRACT
The use of artificial intelligence (AI), particularly large language models such as ChatGPT, has significantly expanded across academia and research in recent years. Of growing interest is AI's potential role in qualitative research, especially in expediting tasks such as literature review and synthesis, data management and coding, and manuscript preparation, all while striving to maintain the rigor expected of qualitative inquiry. This paper explores the degree of alignment between AI-generated thematic analysis and a traditional, manual approach to analyzing focus group data and potential uses of AI in qualitative research. Using Braun and Clarke's thematic analysis framework, a publicly available focus group dataset was analyzed in three stages: (1) Data Familiarization, (2) Human-Led Thematic Analysis, and (3) AI-Generated Analysis using ChatGPT. The human coders followed all six phases of the framework prior to engaging with the AI, to prevent influence from AI-generated output. The comparison revealed partial overlap between the human and AI-generated categories, particularly at the level of broader categories and themes. However, ChatGPT's analysis tended to reflect a more superficial reading of the data, lacking the interpretive depth and nuance demonstrated by the human coders. While AI tools like ChatGPT show potential for supporting certain aspects of qualitative analysis, this study highlights key limitations in their ability to fully replicate the complexity and contextual understanding that human-led analysis provides. Further exploration is needed to define appropriate and ethical roles for AI in qualitative research workflows.

KEYWORDS: ChatGPT, Qualitative Research, Thematic Analysis, Artificial Intelligence

In recent years, the use of artificial intelligence (AI) transitioned from a tool for the technology sector to a resource for the average person. Researchers have begun to explore AI's use in qualitative research to expedite the laborious and time-consuming tasks associated with literature searches and synthesis, data management and coding, as well as the writing process (Švab et al.,

[1]Corresponding author; Director and Associate Professor, Doctor of Physical Therapy Program, School of Health Sciences, Touro University, 3 Times Square, New York, NY 10036. Email: Laura.Hagan@Touro.edu

2023). However, there is ongoing debate about whether AI should be used in the qualitative research at all, and if so, what aspects of the research process can AI be integrated especially in methodological approaches that value the researcher's interpretation of the data and reflexivity.

One of the most notable AI tools is ChatGPT. Introduced in 2018 by OpenAI, ChatGPT is based on Natural Language Processing (NLP), which allows the user to input questions or commands in a conversational manner and receive responses in that same format (Ray, 2023). Chat GPT has cycled through multiple versions, each more sophisticated than the last. ChatGPT models are pre-trained, meaning they are exposed to a large body of publicly available data from webpages, books, and other texts to internalize the patterns of human language. This allows them to generate responses to user prompts by predicting patterns and relationships from the source texts (Ray, 2023).

As generative AI platforms have gained widespread popularity with researchers, several analysis tools have already integrated large-language model-driven AI into their qualitative analysis capabilities. Examples of these programs include Nvivo, MAXQDA, and Atlas.ti. These platforms, however, require subscriptions and may not be accessible to all researchers. Due to its minimal to no cost format, popularity, and ease of use, ChatGPT has become a popular application and has expanded into qualitative research analysis. As interest and availability of these platforms has expanded, early research has begun exploring its potential role in the qualitative research process.

Preliminary research has demonstrated the potential benefits of using ChatGPT including its efficiency compared to manual coding (Brown, 2026; Pattyn, 2025; Zhang et al., 2024), level of accuracy with deductive analysis approaches (Jalali & Akhavan, 2024; Morgan, 2023; Pattyn, 2025) and assistance with literature reviews (Mostafapour et al., 2024). Naturally, ChatGPT also raises concerns about accuracy, ethics, and bias (Ray, 2023; Van Manen, 2023) and authors caution against its use as a stand-alone tool for data analysis and suggests it is used to augment, not replace, the researcher in the coding process (Christou, 2024; Jalali & Akhavan, 2024; Mostafapour et al., 2024; Pattyn, 2025; Zhang et al., 2024).

The role of AI in qualitative analysis continues to be a subject of great debate. Recently, Braun and Clarke themselves, along with many other researchers, have argued that generative AI is incapable of reflexive thematic analysis because the reflexive process, by definition, requires a positioned human researcher who actively constructs meaning from the data (Jowsey et al., 2025). However, this position has already been challenged by others who argue that this definition is excessively restrictive and suggest that AI can successfully support qualitative analysis so long as human researchers are responsible for reflexivity and ultimate interpretation decisions (Hitch, 2024; Nguyen-Trung, 2025). These conflicting perspectives suggest that the question of whether and how AI can have an appropriate role in the qualitative analysis process remains unsettled. To address this uncertainly, there may be a need for further empirical investigation to establish the usefulness and compatibility of AI and human researchers for qualitative data analysis.

Despite AI's potential in assisting qualitative analysis, important questions remain regarding the degree to which AI-generated analyses align with or can reproduce human-led thematic analysis, especially when both are carefully guided by the same analytic framework. Existing research in this growing field largely emphasizes efficiency, coding support, or the design and performance of highly engineered prompting strategies, while less attention has been directed toward documenting where AI-generated interpretations converge with, or diverge from, the human analytic process that has defined and characterized thematic analysis from its inception. Understanding these similarities and differences is important for determining the appropriate role of AI within qualitative research workflows, and especially for identifying where guardrails may be appropriate when using technology with a seemingly limitless capability. In this study, we focus deliberately on ChatGPT as the tool of interest because it is currently the most accessible generative AI platform. Individuals can use the free version or a paid subscription that is less cost-prohibitive than acquiring an institutional or individual license to other proprietary software. It is thus ideal for researchers new to qualitative analysis or who otherwise may not have institutional resources in place.

The aim of this exploratory study is to evaluate the alignment of ChatGPT-generated analysis with manual thematic analysis of human coders using focus group data, thereby identifying the stages of Braun and Clarke's thematic analysis in which AI may provide meaningful assistance and those that continue to require human interpretation and judgment. Braun and Clarke's reflexive thematic analysis was selected because it is one of the most commonly used approaches to qualitative data analysis (Braun & Clarke, 2019). Given its popularity, it is likely that qualitative researchers will use AI to assist with this method. Therefore, understanding where AI aligns with, and where it falls short of, this analytic approach is important. The research question that guided this study was: How does AI compare with human coders using Braun and Clarke's reflexive thematic analysis in terms of (1) interpretation of participant meaning, (2) convergence of initial codes, and (3) alignment of final themes? Through this comparison, the researchers sought to examine when human interpretation remains essential as well as opportunities for and limitations of AI use in qualitative analysis.

Author Positionality

As the authors of this study, and by extension, the human coders engaging with reflexive thematic analysis, we bring distinct professional perspectives to the analysis. Author 1 is a Doctor of Physical Therapy (DPT), Director of a DPT Program in a large metropolitan setting and has extensive experience with qualitative methodologies. Author 2 is an experimental psychologist by training and research coordinator in the same DPT program, with specific expertise in data analysis and research methodology. The physical therapist contributed clinical expertise in musculoskeletal rehabilitation and patient–provider interactions, which informed the interpretation and clinical implications of interviewees' responses. The experimental psychologist, with specialized training in statistical analysis and research methodology, contributed a focus on the critical evaluation of coding decisions and transparency in analytic procedures.

We share an interest in the role of AI in higher education and health science research and the desire to better understand the ethical and practical application of AI in qualitative analysis through firsthand experience. We want to contribute to the growing body of knowledge in the area of AI and to help answer the questions of when, how and whether AI should be incorporated in qualitative research. We acknowledge that these diverse backgrounds and our interest in AI had the potential to shape both the coding process and the resulting themes identified in this investigation (Kennedy et al., 2024; Sarfo & Attigah, 2025), and as such, to promote reflexivity and minimize the influence of individual assumptions we engaged in collaborative coding discussions and consensus-building throughout the analytic process in order to mitigate these individual biases.

Methods

Data Source

We specifically utilized a publicly available and anonymized qualitative data set for this study to eliminate privacy concerns that would result if our datasets from previous studies were used (Dejaco, 2024). Because of this, the study was exempt from IRB approval. Given the focus on healthcare with an emphasis on physical therapy, a dataset relevant to these interests was selected. The dataset included transcripts of three focus groups composed of physical therapists who were interviewed about their perspectives on the use of Virtual Reality (VR) in physical therapy for patients with shoulder pain, following an opportunity to test the technology. The researchers chose to include one of the three focus groups transcript in the current study. Prior to the start of coding, all three transcripts were reviewed and determined that the selected transcript was representative of the entire dataset with respect to the topic being discussed. The selected transcript included a focus group over an hour in length with six participants resulting in 16 pages (1,353 lines) of dialogue providing a substantial amount of qualitative data for analysis for the purpose of this study.

Materials

Analyses were conducted using ChatGPT Plus version 5.3 (OpenAI, 2025) to generate AI-based coding. Microsoft Excel (Office 365, v. 2025) was used to prepare and organize transcript data in Comma-separated Values (CSV) format, enabling human coders to input codes and allowing for efficient comparison and calculation of agreement.

Procedure

This investigation used thematic analysis based on the work of Braun and Clarke (Braun & Clarke, 2006) as the method for qualitative analysis. Thematic analysis was selected because it is a commonly used method of identifying and interpreting patterns across qualitative datasets. It also provides a structured analytic process facilitating the comparison between human and AI generated analyses. It is flexible as it can be used for inductive or deductive analysis and acknowledges the role of the researcher in the in the data interpretation and construction of meaning (Braun & Clarke, 2006). It also aligns well the aim of the original study which was to *gain an understanding of the perspectives and experiences of PTs from the Netherlands, regarding the use of immersive VR in treating shoulder pain.* The phases of analysis outlined by Braun and Clarke are:

(1) Dataset familiarization
(2) Data coding
(3) Initial theme generation
(4) Theme development and review
(5) Theme refinement, defining, and naming
(6) Writing up

To execute parallel streams of human and AI coding with minimal bias, we deliberately chose not to read the original article associated with this data set until finalizing the entire analysis to prevent it from influencing our interpretations. Instead, a third-party reviewed the original article to provide the research objective of interest. The coding process was then organized such that each of Braun and Clarke's steps were executed first by human coders and then by ChatGPT. Step-by-step descriptions of each phase of thematic analysis are described below (See Figure 1).

Figure 1
Parallel Application of Braun & Clarke's (2006) Thematic Analysis: Human Coding and ChatGPT Prompt Language

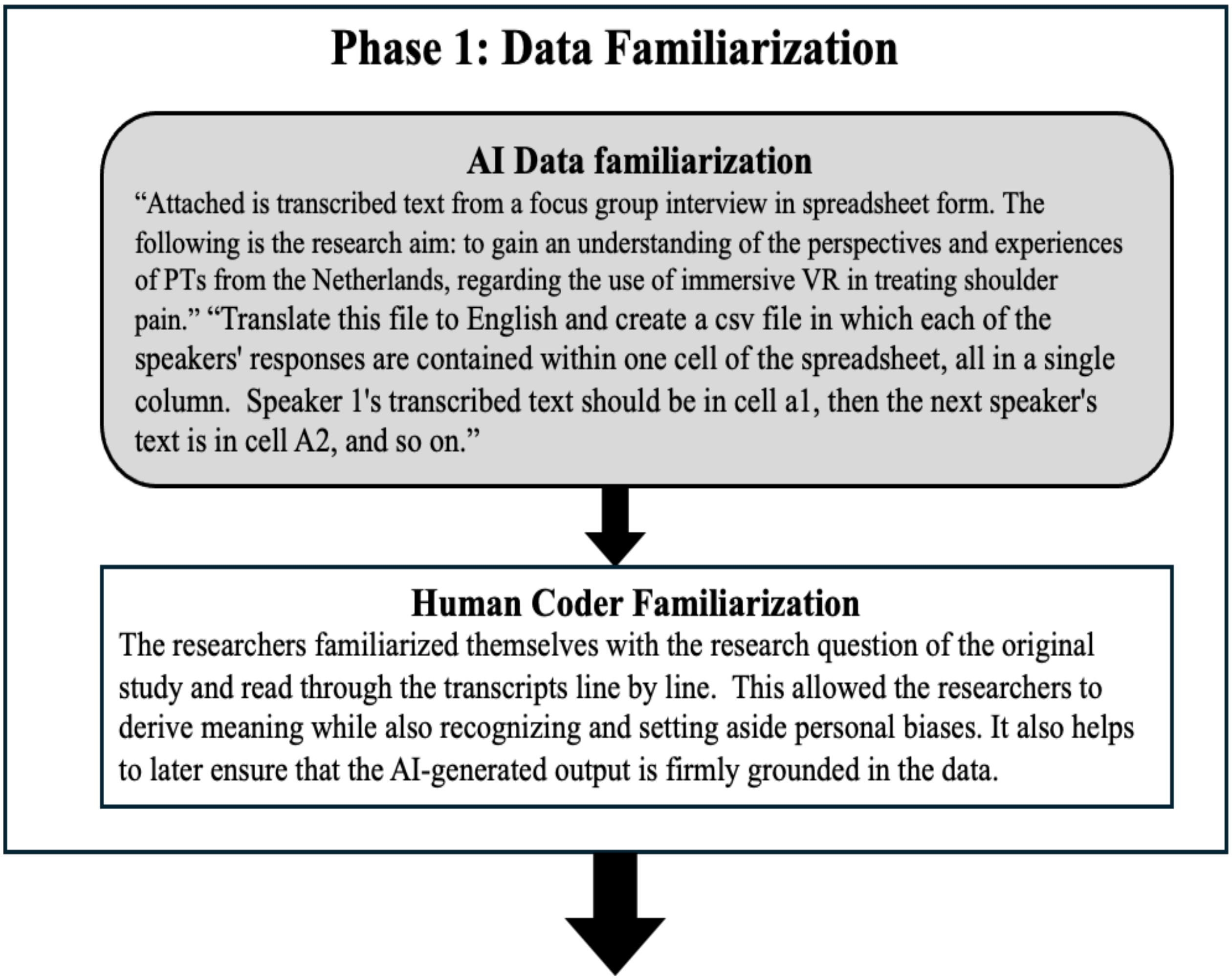


Figure 1 (continued)

Parallel Application of Braun & Clarke's (2006) Thematic Analysis: Human Coding and ChatGPT Prompt Language

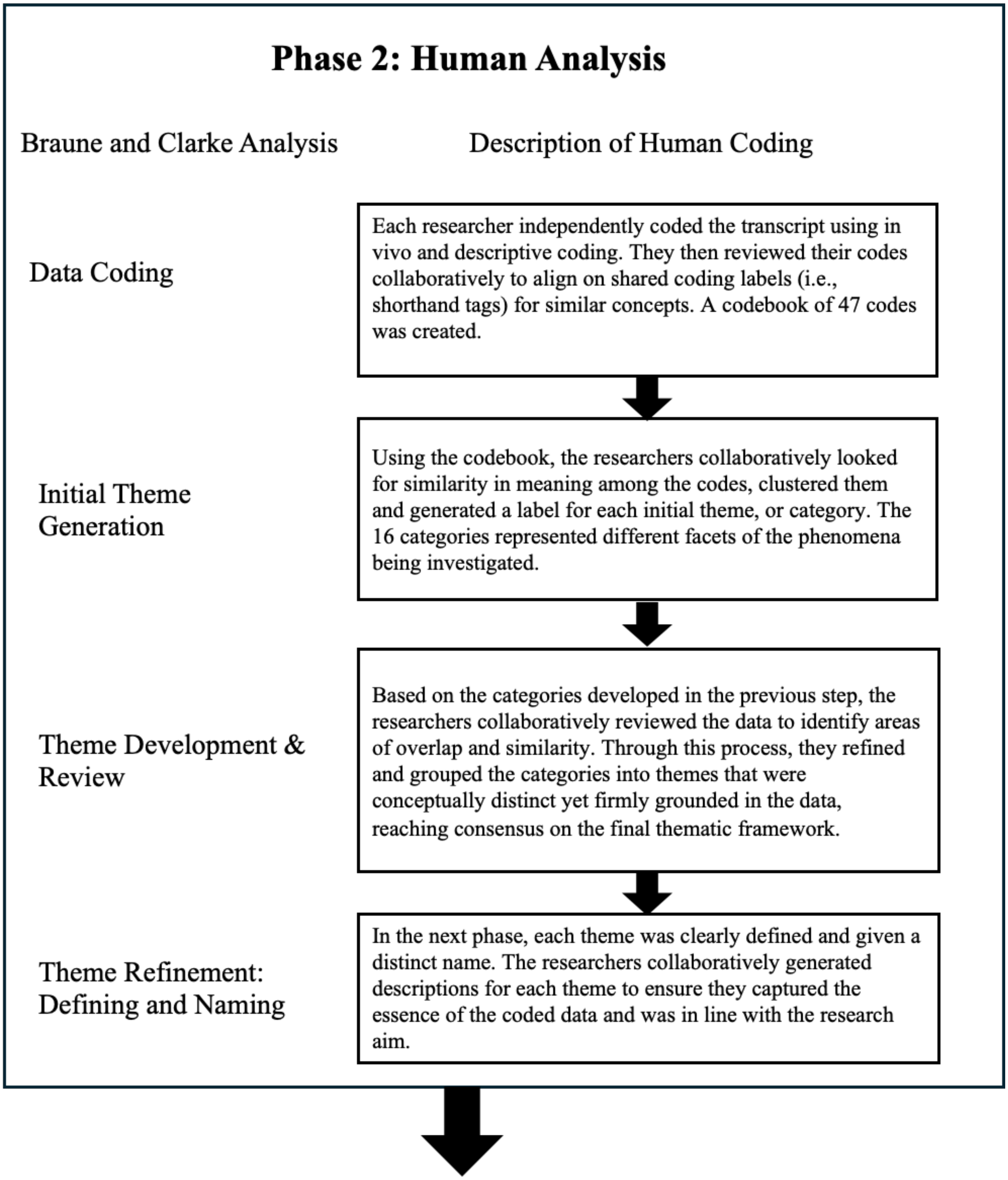
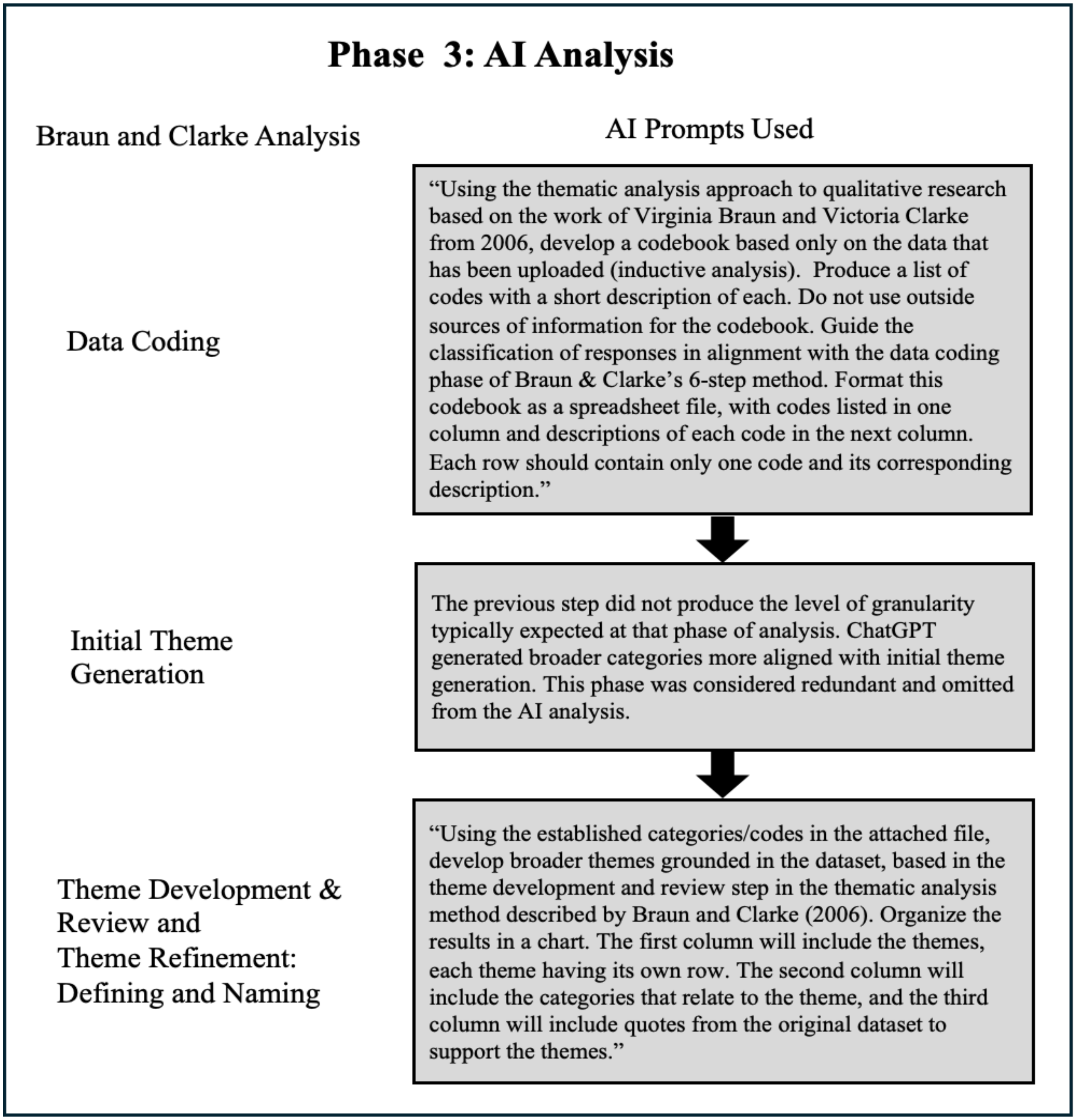


Figure 1 (continued)
Parallel Application of Braun & Clarke's (2006) Thematic Analysis: Human Coding and ChatGPT Prompt Language



Phase 1: Data familiarization

AI. The original interview transcripts were recorded in Dutch. ChatGPT was utilized to translate the transcripts into English and format the transcript in an Excel file so that so that each row a segment of dialogue corresponding to a speaker's turn. This organization was retained because it reflected distinct speakers as they responded to each question. Access to a bilingual individual fluent in both Dutch and English was not available, therefore the accuracy of the translation could not be verified. While translation accuracy is an important factor that could influence qualitative analysis, the focus of this study was to compare AI and human coding using the same dataset and less on the qualitative findings themselves. Any nuances missed or misinterpreted in the dataset would be consistent across both analytical approaches and unlikely to impact the findings.

Human. Once transcripts were translated to English and structured into a spreadsheet, the two coders began the initial phase of data familiarization by reading and re-reading the full focus group transcript independently. During this phase each coder took notes, jotting down initial thoughts and ideas.

Human Phase 2: Data Analysis

Data Coding. Coders separately generated initial codes line-by-line in an inductive, manner, identifying salient features of the data and keywords relevant to the original study's research aim. The unit of analysis for coding was a single speaker turn (i.e., each discrete segment of a single speaker's dialogue), and each unit could be assigned any number of codes to capture the complexity of the speaker's response. At this stage, coders freely applied descriptive codes to meaningful dialogue segments ("feels like a different world") and latent or interpretive codes ("gamification as a facilitator of compliance"). As segments were coded, the coders compared segments to previously used codes to see if they applied or if any new code was needed.

Following independent coding, the coders met to discuss these initial codes and to reconcile their code definitions and labels. This discussion did not alter the original mapping of codes established during the independent coding stage but rather served to establish common code terminology to capture conceptually similar content. This process resulted in the development of a common codebook of 47 code labels, which was necessary to ensure that the subsequent quantitative analyses of agreement were based on conceptual alignment rather than surface-level linguistic differences.

Initial Theme Generation. After dataset familiarization and data coding was complete, we began initial theme generation. Through discussion and consensus building, the researchers grouped the code labels into clusters of larger meaning units, or categories, based on similarity in concepts and meaning. This was completed in Excel to map out the initial themes. Codes that had similar core concepts were color coded and groups together and each initial theme was given a brief description.

Theme Development & Review. Building on the results of the previous step, we continued to refine the categories by clustering those with overlapping sentiments to ensure that they accurately represent the dataset but also are distinct from one another. Through continued consensus building and referencing the original data, we identified potential themes using thematic mapping.

Theme Refinement: Defining and Naming. This final stage involved collaboratively developing titles and definitions for each theme. Theme definitions were designed to capture the core concepts and meaning of that theme while identifying key quotes that exemplified those concepts.

Phase 3: AI Analysis

AI analyses were subject to the same procedural steps as the human coder to evaluate if ChatGPT could perform Braun and Clarke's reflexive thematic analysis rather than through iterative prompting to produce AI output that, superficially, "looked" more like output consistent with Bruan and Clarke's analytical approach. An additional concern associated with re-prompting AI to produce codes that aligned more with expected output was the potential for researcher influence and bias that would create greater agreement with the human analysis. For this reason, our procedure was limited to providing AI with the exact steps elaborated by Braun & Clarke and noting its limitations, rather than introducing researchers' own possibly biased language through iterative prompting. See figure 1 for prompts used for AI analysis.

Data coding. AI was prompted to conduct phase 2 of Braun and Clarke's thematic analysis. The data coding generated by ChatGPT did not produce the level of granularity typically expected at this phase of Bruan and Clarke's analysis. Instead, ChatGPT generated broader categories more aligned with phase 3 or initial theme development.

Initial Theme Generation. Because the previous prompt produced output consistent with generating initial themes or phase 3 of the Bruan and Clarke analytic process rather than phase 2 of generating initial codes, the researchers determined that having ChatGPT independently generate initial themes would be redundant, so a separate prompt for initial theme generation was not performed.

Theme Development and Review, and Theme Refinement. The stages of theme development and review and theme refinement were consolidated for ChatGPT generated analysis. The previous outputs were broad and general in nature, leaving little room for detailed analysis. As a result, we determined that it would be more productive to prompt ChatGPT to generate the final themes based on the previous output along with definitions and exemplar quotes to support those themes.

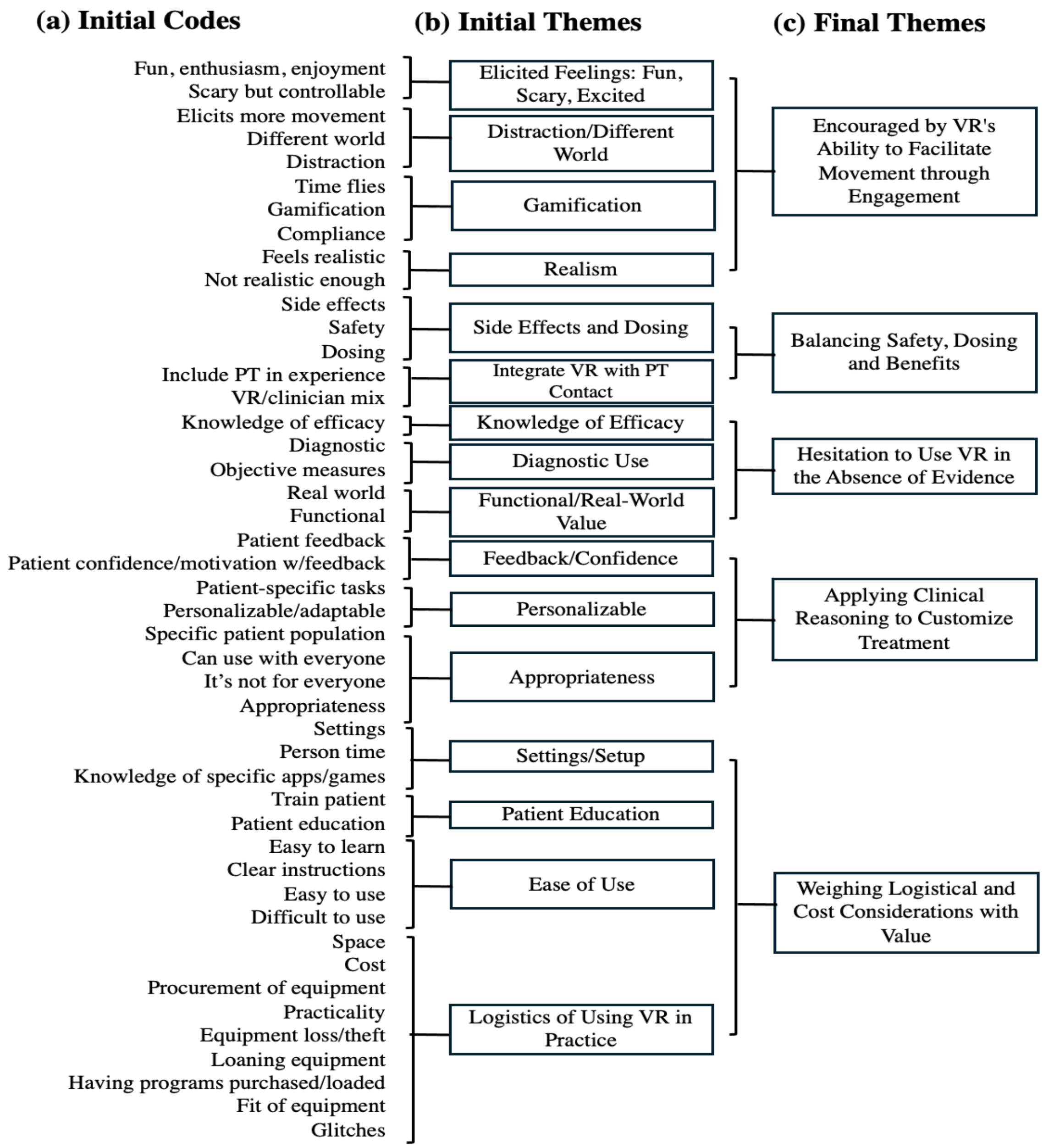
Results

The comparison between the human-led and AI-generated analyses revealed partial overlap in broader categories and themes. However, the underlying processes and depth of interpretation differed significantly.

During phase 2 of Braun and Clarke's framework (Data Coding), the human coders generated a total of 47 codes (See Figure 2). This level of detail is typical of an early-stage, inductive analysis and reflected the coders' nuanced judgment about when surface-level descriptions were sufficient versus when deeper meanings were necessary to address the research question. Human coders demonstrated an ability to contextualize responses and discern meaning beyond literal text, allowing for a more interpretive and flexible coding process.

Figure 2

Human Coding and Analysis Process from Initial Coding to Theme Development

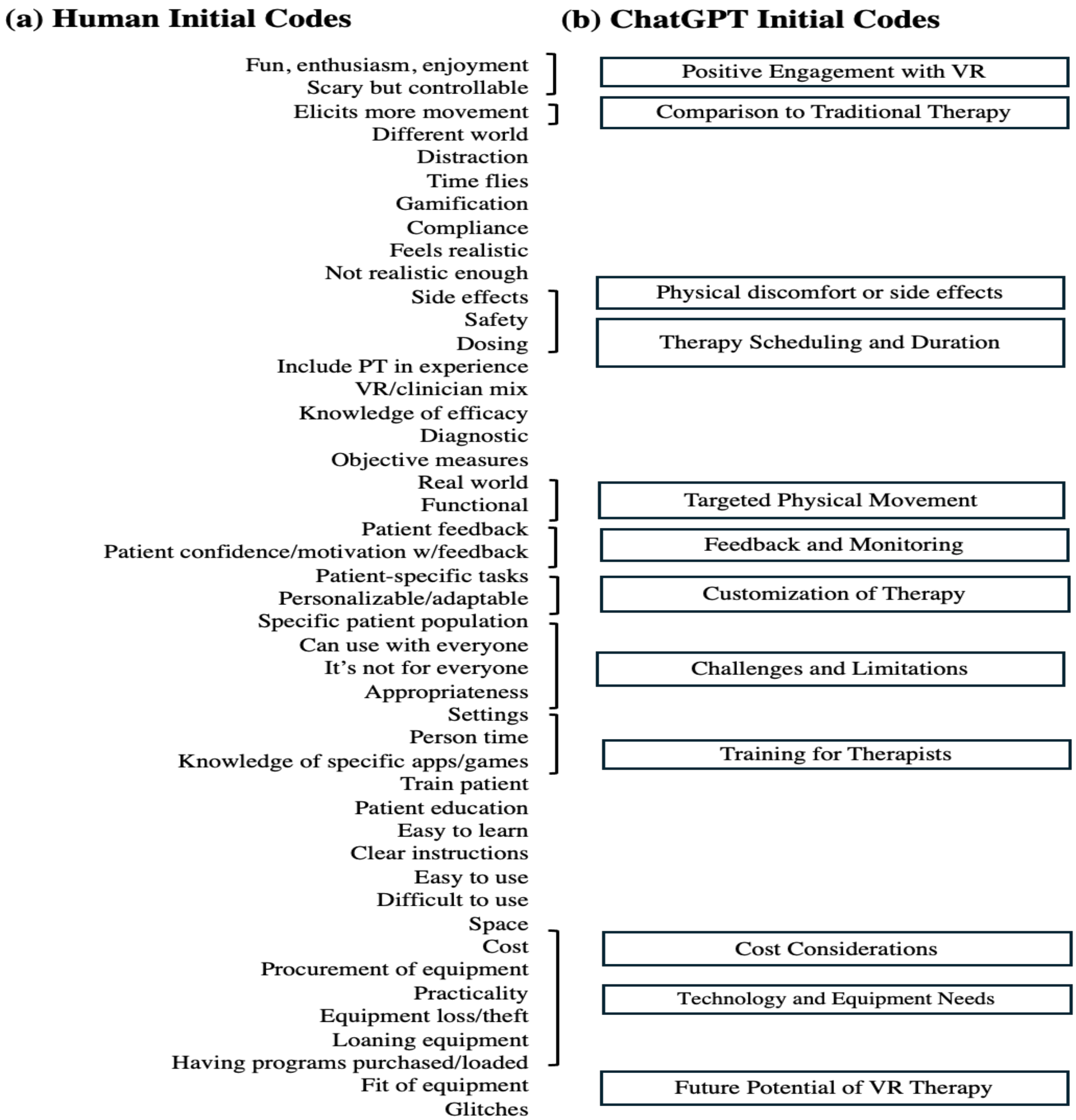


Note: VR= Virtual reality; PT= Physical therapist

In contrast, when ChatGPT was prompted to perform phase 2 of Braun and Clarke's reflexive thematic analysis, it did not produce what Braun and Clarke define as initial coding. The initial round of AI coding on the same dataset produced only 12 codes (See Figure 3). These codes were noticeably broader, resembled initial themes and largely limited to surface-level, semantic features of the text compared to human initial coding. Although some overlap existed between the AI-generated codes and human initial coding codes, ChatGPT failed to capture a significant thread within the dataset related to the gamification of VR. For example, it did not identify the participants' description of being in a "different world" or their perception that "time flies" when using a VR headset, salient concepts that were a prominent feature of the dataset and human analysis. The AI appeared to rely heavily on literal word patterns and lacked the capacity to infer deeper meaning or adapt interpretive strategies based on context. As a result, AI's analysis minimized the complexity of the data, missing the deeper understanding that emerged from the human-led process.

Figure 3

Human Results of Initial Coding vs. AI Initial Coding



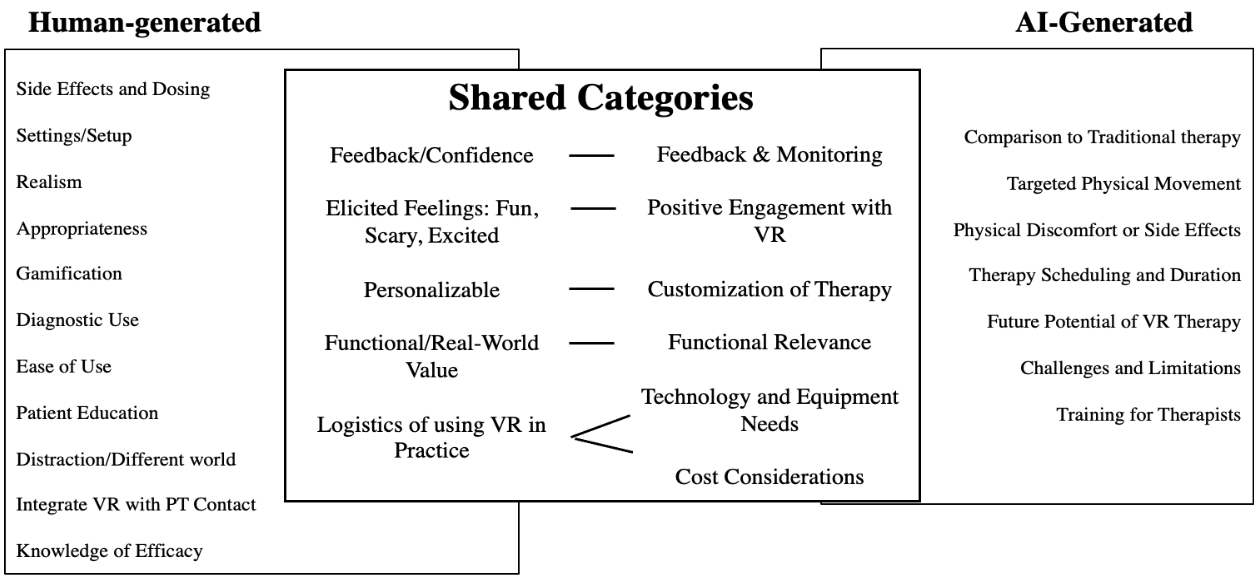
Note: VR= Virtual reality

Because the initial coding phase of AI-generated coding yielded overly broad results resembling high-level categories consistent with phase 3 of Braun and Clarke's analysis instead of nuanced, discrete codes resembling phase 2 coding, we considered the AI output as preliminary themes (phase 3) and proceeded to phase 4 of Braun and Clarke's Thematic Analysis. This adjustment reflects a shift from expecting AI to perform line-by-line coding to instead supporting theme development. From there, we used AI to assist in refining these broad categories into more coherent and distinct themes.

When comparing the human-generated and AI-generated initial themes or category level analysis, we observed a moderate degree of overlap. Both approaches identified similar conceptual groupings, including the feedback provided by VR, users' emotional responses to using VR, and the potential for customization of the VR experience (see Figure 4).

Figure 4

Category-level Overlap between AI and Human Coders

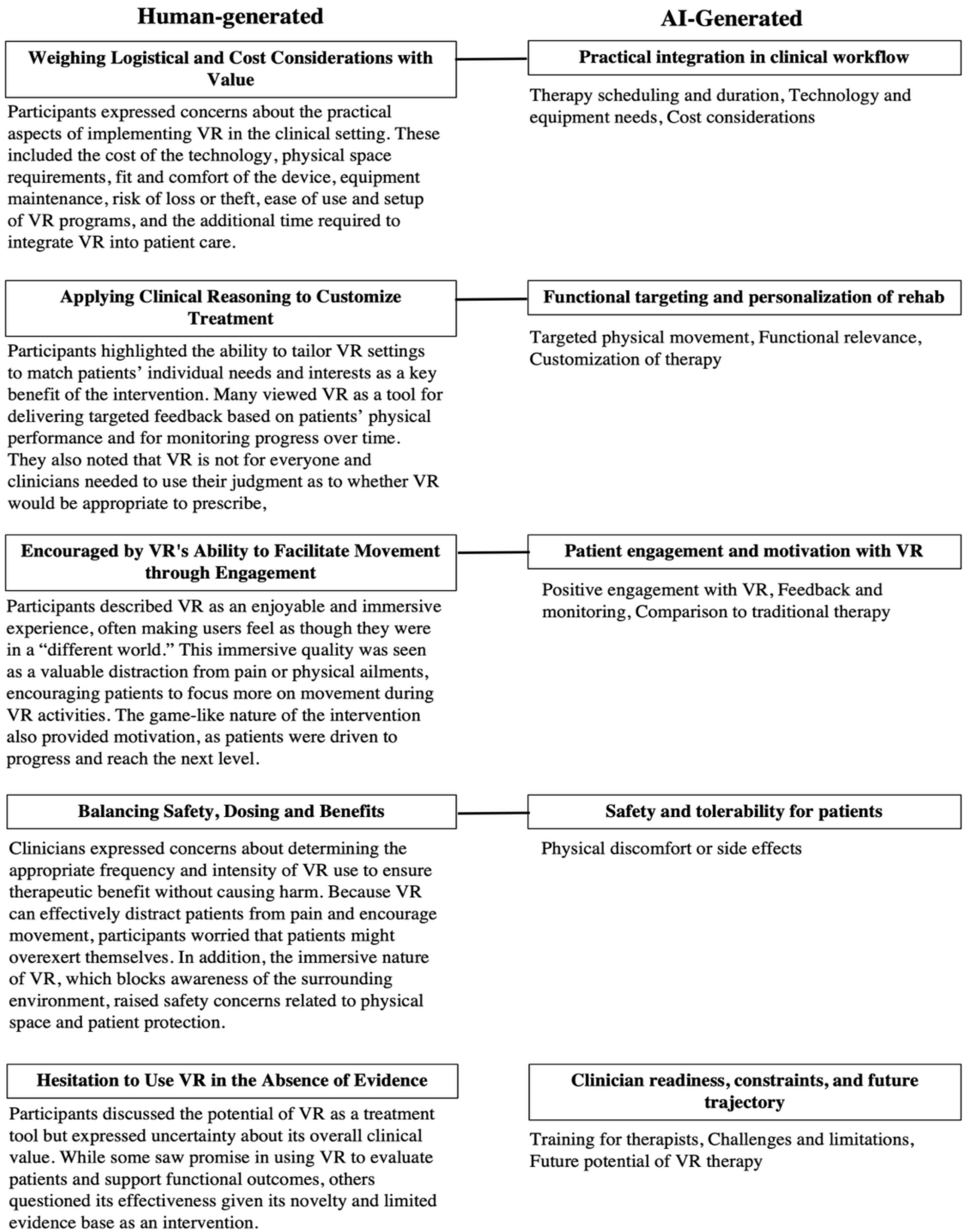


Note: VR= Virtual reality; PT= Physical therapist

However, certain nuanced insights surfaced only through human-led analysis. For example, participants in the focus group described the immersive nature of VR as being a distraction from the shoulder by making them feel as though they were in a 'different world' with the potential to encourage movement in individuals who may otherwise be reluctant to engage in physical activity. They also highlighted the need to use their own clinical reasoning when dosing VR based exercise to maintain the safety of the patient. These subtleties grounded in participants' language and lived experience did not emerge in any stage of the AI-generated analysis. As the ChatGPT analysis progressed into theme development and refinement, we observed increased alignment between the human and AI-generated themes with titles of the ChatGPT generated themes showed similarity to the human generated themes. However, the themes generated by the human coders were more interpretive, offering more insight into the participants' experience with VR (See Figure 5).

Figure 5

Human vs. AI-generated Final Themes with Descriptions



Note: VR= Virtual reality

For this set of final themes, AI was prompted to provide a sample of dialogue exemplars from the data for each theme, several of which are presented in Table 1. Even in this limited set of quotes, a variety of discrepancies were noted: (1) The predominant use of keywords to guide theme development, in which the phrase 'compared to' was included in a theme describing general comparison of VR to traditional PT. The dimension of comparison or specific strengths and weaknesses were excluded entirely from interpretation, (2) decontextualization of speech, as in a description of patients who have chronic pain being distracted to the point of forgetting their pain being classified in the "Physical discomfort or side effects" theme and (3) incorrect semantic interpretation, as in the case of dialogue classifying a quote about the fact that reward-based games provide a physical challenge to patients to facilitate their rehabilitation process as "Challenges and limitations" of VR.

Table 1
ChatGPT-generated Final Themes with Corresponding AI-Identified Exemplar Text

ChatGPT Generated Final Themes	Theme Description Generated by ChatGPT	Quote identified by ChatGPT	Discrepancy
Comparison to traditional therapy	Comments comparing VR-based therapy to standard or traditional methods.	"I actually thought it was fun because you really do feel like you're in a different world. So, what I was wondering, especially with shoulder patients, is—if you have someone moving with VR... There were a few of those more active things where you had to raise your arm above shoulder height, for example. I was curious if someone with shoulder issues would be able to do that. Compared to if they were just given that as a regular instruction. So, I was curious about that." "The difference depends on whether you're just observing. In that case, you can be in a movie or a photo of an environment, but you can't interact with it. Mm."	Code category highlights the aspect of this text that contains the words "Compared to" to define individual quote as a comparison between VR and traditional PT. Human coders categorized this response not in respect to the exact words used, but rather in service to capturing the essence or function of said comparison. Namely, to the engaging aspect of VR that would elicit more movement than traditional PT.

Table 1
Continued

Physical discomfort or side effects	Mentions of adverse physical reactions to VR use.	"Yeah, well, I'm not technical, so uh, I'd never worked with it before, and uh, well, I was really enthusiastic about it. Also, about the fact that you're in a completely different world and that you, uh, make movements because you're so immersed in the game or, uh, in that other world that you, yeah, move without even noticing. So, I think it's a really great tool, especially for chronic pain patients. I don't really care whether they're shoulder or, uh, back patients—but I was really enthusiastic, yeah."	Code category as determined by AI relates to physical discomfort. An illustrative quote contains the words "chronic pain". Human coders identified the use of chronic pain in this context as an example of the benefit or usefulness of VR for treating patients with chronic pain, not an adverse effect.
Challenges and limitations	Reported difficulties or barriers in using VR for therapy.	"You want to use the reward system, right? Yes, that challenges you to definitely take that extra step. But then you also might want the medical, uh, physiological perspective, maybe even limit things a bit so you don't— Mmm, exactly what you said, just going further and further and further." "I think especially that you take away the focus from how they move, give them a bit of distraction. But also indeed, what happens in the brain? That if someone has pain? Mmm no. Yes, we know of course that it can be influenced centrally. So, if you work with VR and you are in such a different experience world. The same as you can get nauseous, maybe the pain is dampened more. Mmm. Uhm. So, in that way you can still challenge the patient with it."	AI category includes challenges and limitations, presumably reflecting difficulty with VR. The word "challenge" appears in the text and was "interpreted" to signify said difficulty. Human coders identified that in this context, "challenge" referred to the nature of the VR task, in the sense that it was challenging, and thus beneficial, to the user's rehabilitation.

Discussion

This study attempted to utilize and document the performance of a widely available generative language model, ChatGPT, in executing a thematic analysis of a publicly available qualitative dataset. Compared to human coders, AI tended to categorize qualitative data more broadly, advancing to the construction of general themes earlier in the analysis process. AI also exhibited patterns in categorization and interpretation that suggest it relies heavily on surface-level semantic features of text. In contrast, human coders demonstrated the ability to determine when a surface-level analysis was appropriate or when the surface-level reading did not adequately capture the essence of the response in relation to addressing the research question or a meaningful clinical context.

Several observations emerged from this exploration of AI-assisted qualitative data analysis:

(1) the use of AI may have the ability to assist the process of traditional human coding by helping to refine conceptual boundaries and labels, not undertake the entire process unchecked

(2) AI assistance is most effective when used after human readers/coders have developed a comprehensive familiarity and understanding of the raw data. This last observation emphasizes the need for qualitative researchers to critically examine the output of each analytic step completed by AI agents.

The Positionality of AI

An inevitable and important consideration for all qualitative analytic approaches is the positionality of the researcher. In the context of thematic analysis, positionality includes the researcher's values, prior experiences, and theoretical orientations that influence the selection of codes, the framing of categories, and the weighting of some narratives over others, rendering the researcher an active, rather than neutral, interpreter of raw data (Kennedy et al., 2024; Sarfo & Attigah, 2025). In reflexive thematic analysis, researchers' topical expertise and cultural familiarity may help interpret data within appropriate social or historical contexts, in ways that a neutral observer might fail to consider. When incorporating generative AI into qualitative analysis, the discussion of whether AI has a positionality is particularly necessary. As it is often framed to the widespread public audience, AI may be assumed to be a positionless agent or algorithm-executing "robot" incapable of developing the kind of systemic biases that result from having personal experiences. However, because generative AI is trained on massive amounts of human-generated text from articles, books, and websites, its outputs can reasonably be expected to retain the same linguistic, cultural, and conceptual patterns as the population of individuals from whom its training data are drawn. There is also the concern that AI systems may be disproportionality influenced by majority perspective in fields such as healthcare or politics where particular views may be more prevalent, and minority voices may be overlooked impacting AI outputs (Hendawy, 2025). This is particularly concerning for qualitative researchers who look to understand the experience of underrepresented or marginalized populations, where the use of AI in the analytical process may be overshadowed by the majority viewpoint limiting the reliability of the output.

AI output may similarly be shaped by bias that emerges from its interactional history, including prior prompts, user feedback, and the selective information a user may have provided about themselves in its many interactions with the software over time. For example, it has already been documented in the research literature that AI agents acquire and retain information about users so personal that they reliably infer personality traits, user background, and general worldview from text excerpts (Derner et al., 2024; Santy et al., 2023). Perhaps this mechanism of influence creates an AI agent whose positionality mirrors at least some facets of the positionality of its human user, however, it is unlikely that AI can fully capture the complexity of a researcher's lived experiences, values, and background and it should be cautioned against replacing the researcher's unique perspective that is valued in analytical approaches like Bruan and Clarke's thematic analysis with AI driven analysis

In this study, ChatGPT failed to capture several of the nuances within the data identified by the human coders, particularly the coder with clinical experience as a physical therapist. This clinical background provides the lens through which the data and participant responses were interpreted within the context of rehabilitation rather than a superficial description of the VR technology use. This highlights how researcher positionality can shape qualitative analysis and the importance of keeping the human in the process. Recognizing this dynamic is essential for assessing the effectiveness and implications of AI-assisted qualitative analysis and its impact of the trustworthiness of the research findings.

AI as a Collaborator

An important question that surfaced in the thematic analysis process was whether AI could be akin to a true collaborator. This is a challenging notion because AI agents tend to align with their users with repeated feedback, especially in the context of facilitating agreement towards converging ideas or themes. Perhaps AI should best be thought of as an assistant to sketch the landscape of the dataset, constructing a generalized overview or summary of the topical clusters identified by human coders. In this role, AI primarily performs categorization of the data, rather than interpretation.

In our study, ChatGPT did not demonstrate the reformulation, contextual sensitivity, or reflexive decision-making that characterized the human coders' approach to thematic analysis. This is evident in the early stages of AI data analysis. When prompted to conduct phase 2 of Braun and Clarke's thematic analysis, it was with the assumption that the model has the methodological knowledge to execute that phase as intended. When the output results in broader categories reflecting a superficial analysis of the data, we did not iteratively refine the prompt to elicit more detail coding because the purpose of the study was to examine the performance of ChatGPT using standard methodological instructions.

Engineering the prompt may have resulted in more detailed coding, but there is still an underlying limitation of the absence of the reflexive engagement germane to Braun and Clarke's approach. Coding is an iterative and interpretive process that extends the researcher's familiarization of the data and encourages engagement through reflexive observations, memoing and the identification of potential patterns and themes. In this sense, the AI could not replace the deeper analytic and interpretive work required to generate meaningful themes, nor could it disambiguate the core intentions reflected in nuanced.

As the data analysis progressed beyond coding to theme development, there was notable overlap between the AI and human generated categories and themes suggesting that AI can recognize common patterns within participant responses. However, the similarities became less apparent when AI was asked to provide definitions and exemplar quotes for each of the themes it generated which further supports the notion that AI has difficulty moving beyond descriptive categorization into reflexive interpretation.

Although not directly investigated in this study, these finding suggest that AI may be better suited for qualitative approaches that use a more descriptive and objective approach to data analysis rather than approaches that focus on positionality, reflexivity, and substantial interpretation of the data. Similar to the findings of this study, Morgan (Morgan, 2023) found that ChatGPT was more successful at descriptive themes and missed nuances in the data cautioning that the use of AI is only a tool in the larger analytical process. Additional research needs to be conducted on the use of AI in approaches such as deductive content analysis or rapid qualitative analysis where the aim is to organize data into predefined categories or to summarize findings rather than to generate interpretive themes. Even in these approaches, however, AI should function as an assistant rather than an independent analyst with the human researcher taking the time to familiarize themselves with the data, verify any AI output for accuracy, identify hallucination, and ensure the analysis remains grounded in the data.

Limitations and Future Directions

Several recent publications document the use of an array of AI-based platforms in both quantitative and qualitative data analysis (De Paoli, 2024; Goyanes et al., 2025; Nguyen-Trung, 2025). The level of AI involvement and prompt engineering across studies varies from basic to complex, with highly engineered prompts containing instructions that ask the AI Agent to mimic human abilities (i.e., familiarizing itself with the data in explicit steps) and testing its knowledge of the intricacies of the methods it will be prompted to use (Goyanes et al., 2025). This study did not attempt such a comprehensive and fine-tuned a prompt engineering process. We focused our efforts on assessing how AI performs each of the tasks primarily executed by human coders for insights about its practical use in its default configuration. This implies no prior "testing" or adjustment of settings. The prompts use the wording of the original method paper (Braun & Clarke, 2006), which also guides the human coders. There was no real-time micro-adjustment or correction of prompts and/or output after they were initially generated.

An additional consideration is the nature of the dataset itself. The focus group discussions centered around the participants' use of VR for shoulder rehabilitation and was a straightforward experiential account, albeit still requiring a level of interpretive engagement. It is possible that the differences between AI coding and human coding may be even more pronounced with data that reflects highly complex lived experiences with more nuances and is emotionally laden. Future research should explore AI assisted analysis across different qualitative datasets that vary in complexity to better understand under which conditions AI may or may not be appropriate to use.

It must also be acknowledged that the observations and insights gained from this investigation are subject to re-examination in light of the rapid evolution of AI technology we observe in real time. AI agents may continue to improve their comprehension and interpretation abilities and be further specialized in performing different tasks with more targeted training datasets or algorithmic improvements. Popular GPT "wrappers," for example, improve the workflow and ease of use of ChatGPT, and can thus appear inherently smarter or more accurate. We suggest, however, that caution still be taken when using the output of even the most advanced platforms, as users may be less inclined to question and critically evaluate them (Schildt, 2026).

Conclusion

This study addresses both the promise and the cautions of incorporating generative AI into qualitative data analysis. While ChatGPT was effective in rapidly generating broad categorizations and drawing the overall conceptual boundaries of the content of our dataset, it lacked the interpretive depth, contextual sensitivity, and reflexive judgment that characterize human thematic analysis as described by Braun and Clarke. Its reliance on surface-level semantics and implicit adoption of biases from prior user interactions suggests that AI may be better utilized as an assistant whose outputs require critical review by human researchers or possibly with analytical approaches that employ descriptive analysis. AI systems are not neutral but instead reflect both the cultural imprints of their training data and the interactional histories they accumulate. As such, integrating AI into qualitative research requires careful methodological transparency, explicit acknowledgement of these dynamics, and a recognition that the responsibility of thorough data familiarization and meaningful interpretation still belongs to human analysts. Future work should explore the usefulness of AI towards alternative qualitative analysis strategies, such as content analysis, documenting the critical stages of analysis where human oversight is necessary. As AI continues to evolve, establishing best practice for the protocols for its integration into qualitative research will help to ensure these tools are used to complement, not replace, the analytic capacities of human researchers.

Declarations

Author Contributions

L.H. conceived the study. L.H. and A.C. collaboratively identified and obtained the publicly available dataset, conducted the reflexive thematic analysis, reviewed and refined AI-assisted coding outputs, developed the tables and figures, interpreted the findings, and jointly wrote, reviewed, and approved the final manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

This study involved a secondary analysis of a publicly available, de-identified dataset and therefore did not constitute human subject research, and therefore Institutional Review Board approval was not required.

Informed Consent Statement

This study analyzed a publicly available, de-identified dataset collected by the original investigators. No direct interaction with human participants occurred as part of the present study, and thus informed consent procedures were not applicable.

Data Availability Statement

The publicly available dataset utilized in this study is available from the original study authors through their public repository. The coding data, codebook, and other analytic materials generated during the present study are available from the corresponding author upon request.

Acknowledgements

The authors thank Dejaco and colleagues (2024) for making the replication dataset associated with their study, Experiences of Physiotherapists Considering Virtual Reality for Shoulder Rehabilitation: A Focus Group Study, publicly available through the DANS Data Station Life Sciences repository. The availability of these data enabled the secondary analysis performed in this study.

Conflicting Interest

The authors declare no conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Reference

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Brown, C. (2026). Exploring parental vaccine hesitancy: Reflexive thematic analysis—insights from traditional vs. AI methods. *American Journal of Qualitative Research*, *10*(2), 115–138. https://doi.org/10.29333/ajqr/18007

Christou, P. (2024). Thematic analysis through artificial intelligence (AI). *The Qualitative Report*, *29*(2), 560–576. https://doi.org/10.46743/2160-3715/2024.7046

De Paoli, S. (2024). Further explorations on the use of large language models for thematic analysis. Open-ended prompts, better terminologies and thematic maps. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *25*(3). https://doi.org/10.17169/FQS-25.3.4196

Dejaco, B. (2024). *Replication data for the study on experiences of physiotherapists considering VR for shoulder rehabilitation: A focus group study* (Version 2.1) [Dataset]. DANS Data Station Life Sciences. https://doi.org/10.17026/LS/COWIPU

Derner, E., Kučera, D., Oliver, N., & Zahálka, J. (2024). Can ChatGPT read who you are? *Computers in Human Behavior: Artificial Humans*, *2*(2), 100088. https://doi.org/10.1016/j.chbah.2024.100088

Goyanes, M., Lopezosa, C., & Piñeiro-Naval, V. (2025). The use of artificial intelligence (AI) in research: A review of author guidelines in leading journals across eight social science disciplines. *Scientometrics*, *130*(7), 3725–3741. https://doi.org/10.1007/s11192-025-05377-0

Hendawy, M. (2025, January 13). *The violence of the majority: Rethinking AI positionality in decision-making.* Internet Policy Review. https://policyreview.info/articles/news/ai-positionality/1820

Hitch, D. (2024). Artificial intelligence augmented qualitative analysis: The way of the future? *Qualitative Health Research*, *34*(7), 595–606. https://doi.org/10.1177/10497323231217392

Jalali, M. S., & Akhavan, A. (2024). Integrating AI language models in qualitative research: Replicating interview data analysis with ChatGPT. *System Dynamics Review*, *40*(3), e1772. https://doi.org/10.1002/sdr.1772

Jowsey, T., Braun, V., Clarke, V., Lupton, D., & Fine, M. (2025). We reject the use of generative artificial intelligence for reflexive qualitative research. *Qualitative Inquiry*, 10778004251401851. https://doi.org/10.1177/10778004251401851

Kennedy, E., Castanelli, D. J., Molloy, E., & Bearman, M. (2024). Addressing positionality in qualitative research: Significance, challenges and strategies. *The Clinical Teacher*, *21*(6), e13820. https://doi.org/10.1111/tct.13820

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *International Journal of Qualitative Methods*, *22*, 16094069231211248. https://doi.org/10.1177/16094069231211248

Mostafapour, M., Fortier, J. H., Pacheco, K., Murray, H., & Garber, G. (2024). Evaluating literature reviews conducted by humans versus ChatGPT: Comparative study. *JMIR AI*, *3*, e56537. https://doi.org/10.2196/56537

Nguyen-Trung, K. (2025). ChatGPT in thematic analysis: Can AI become a research assistant in qualitative research? *Quality & Quantity*, *59*(6), 4945–4978. https://doi.org/10.1007/s11135-025-02165-z

OpenAI. (2025). *ChatGPT* (Version GPT-5.3) [Large language model]. https://chatgpt.com

Pattyn, F. (2025). The value of generative AI for qualitative research: A pilot study. *Journal of Data Science and Intelligent Systems*, *3*(3), 184–191. https://doi.org/10.47852/bonviewJDSIS42022964

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Santy, S., Liang, J., Le Bras, R., Reinecke, K., & Sap, M. (2023). NLPositionality: Characterizing design biases of datasets and models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9080–9102. https://doi.org/10.18653/v1/2023.acl-long.505

Sarfo, J. O., & Attigah, D. K. (2025). Reflecting on reflexivity and positionality in qualitative research: What, why, when, and how? *Journal of Advocacy, Research and Education*, *12*(1). https://doi.org/10.13187/jare.2025.1.75

Schildt, H. (2026). The case for artificial intelligence in qualitative research. *Organization Studies*, 01708406261452993. https://doi.org/10.1177/01708406261452993

Švab, I., Klemenc-Ketiš, Z., & Zupanič, S. (2023). New challenges in scientific publications: Referencing, artificial intelligence and ChatGPT. *Slovenian Journal of Public Health*, *62*(3), 109–112. https://doi.org/10.2478/sjph-2023-0015

Van Manen, M. (2023). What does ChatGPT mean for qualitative health research? *Qualitative Health Research*, *33*(13), 1135–1139. https://doi.org/10.1177/10497323231210816

Zhang, H., Wu, C., Xie, J., Rubino, F., Graver, S., Kim, C., Carroll, J. M., & Cai, J. (2024). *When qualitative research meets large language model: Exploring the potential of QualiGPT as a tool for qualitative coding* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2407.14925

Notes on Contributors

Laura Hagan, PT, PhD, DPT, MS, is a licensed physical therapist, and serves as the Director and Associate Professor in the Department of Physical Therapy at Touro University, New York, NY. She is an experienced qualitative researcher with expertise in qualitative methodology and physical therapy education.

Alla Chavarga, PhD, is an experimental psychologist whose areas of expertise include data analysis and research methodology. She serves as an assistant professor and Research Coordinator in the Department of Physical Therapy at Touro University, New York, New York.

ORCID

Laura Hagan, https://orcid.org/0000-0003-0875-7000
Alla Chavarga, https://orcid.org/0000-0002-7169-4607