Human versus GPT-4 in qualitative analysis: A comparative reanalysis of patient interview data following anterior cruciate ligament injury

The Knee 60 (2026) 104388

Contents lists available at ScienceDirect

The Knee

journal homepage: www.elsevier.com/locate/thekne

Human versus GPT-4 in qualitative analysis: A comparative reanalysis of patient interview data following anterior cruciate ligament injury rehabilitation

Ramana Piussi [a,b,*], Justin F. Schneiderman [c], Yinan Yu [d], Kristian Samuelsson [a,e], Eric Hamrin Senorski [a,b]

[a] Sahlgrenska Sports Medicine Center, Gothenburg, Sweden
[b] Unit of Physiotherapy, Department of Health and Rehabilitation, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[c] Department of Clinical Neuroscience, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[d] Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden
[e] Department of Orthopaedics, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

ARTICLE INFO

Article history:
Received 1 September 2025
Revised 8 January 2026
Accepted 3 February 2026

Keywords:
Language processing
Rehabilitation
Qualitative research

ABSTRACT

Objective: The purpose of this study was to prompt GPT-4 to analyze qualitative data used in a published scientific article where qualitative content analysis was performed by human researchers, and to qualitatively compare results from the published article with the results generated by GPT-4.
Methods: This study was conducted using the full interview dataset from a published qualitative study that aimed to explore experiences of patients treated with rehabilitation alone after an anterior cruciate ligament (ACL) injury. Interview transcripts were analyzed by GPT-4 through iterative prompting to replicate the original six-step content analysis process. Different attempts were conducted to improve GPT-4′s output. GPT-4′s final output was qualitatively compared with the human-generated results.
Results: While the human-made analysis produced one overarching theme supported by three main categories and nine sub-categories, GPT-4′s analysis resulted in four themes, six main categories, and 15 sub-categories. Both analyses captured uncertainty and the impact of knee-related symptoms. GPT-4′s results showed a suspiciously equal distribution of codes across sub-categories, and introduced a theme not grounded in the source data. Multiple prompts were required to produce and organize the material.
Conclusion: The analysis performed by humans and GPT-4 had similarities and differences. The use of GPT-4 for qualitative analysis in its present form is challenging and needs to be performed across several steps. Currently, GPT-4 should not be used as the only tool in a qualitative analysis of interview data.

* Corresponding author at: Institution for Neuroscience and Physiology, Box 430, 405 30 Göteborg, Sweden.
E-mail address: ramana.piussi@gu.se (R. Piussi).

https://doi.org/10.1016/j.knee.2026.104388
0968-0160/

1. Introduction

Evidence-based medicine (EBM) integrates current best scientific evidence, clinical expertise, and patient preferences. The role of patient preferences has grown as patients become more informed participants in care, aided by digital tools and increased access to information [1–3]. Incorporating patient preferences helps address the gap between healthcare supply and demand with shared decision-making and personalized care [4]. Accordingly, qualitative research, that focuses on the patients' experiences, and 'the hows and whys' of a matter instead of 'how many or how much' [5], has flourished in many medical fields, including sports medicine [6].

In the field of sports medicine, to perform a qualitative study, where patients' experiences are collected through recorded interviews that are transcribed and analyzed, is a complex process where large amounts of rich data need to be handled [7].

Medical sciences, such as sports medicine, are currently undergoing a major transformation with the advent of AI. Chatbots based on large language models (LLMs), such as GPT-4, are capable of generating human-like responses and of delivering expert-level information [8]. The LLMs are trained on vast amounts of information [9], and chatbots that use them can in turn be used to improve the efficiency of healthcare delivery [8–10]. In fact, ChatGPT (different versions) has been reported to pass stringent exams such as the United States Medical Licensing Exam [11]. In orthopedics, ChatGPT has been employed to generate treatment plans [12], to accurately answer patients questions [13,14], and to provide accurate information pertinent to the field of anterior cruciate ligament (ACL) injury, for both medical doctors and patients [15]. While LLMs do not continuously learn after deployment, the performance and outputs of LLM-based services may change over time due to model updates, fine-tuning, or the integration of external information sources [16]. However, continuous learning poses challenges, e.g., for reproducibility and for the curation of reliable training data; as the model's knowledge base continuously evolves, the same input may yield different outputs over time, making it difficult to replicate findings. 'Dead internet theory' is a further concern wherein the internet becomes increasingly populated with machine-generated content such that distinguishing high-quality, human-authored sources from automated output becomes increasingly difficult [17].

Another possible application of LLMs within sports medicine concerns the use of such chatbots to perform a qualitative investigation. For example, a survey on 101 qualitative researchers reported that the majority of researchers would embrace the use of AI for transcription purposes, but to a lesser extent for coding, to generate findings, or to write a report [18]. Some attempts have been made to use AI as an aid to analyze qualitative data [19]. While interest in using AI for qualitative analysis is growing, robust empirical evidence remains limited, highlighting the need for further research.

The purpose of this study was to prompt GPT-4 to analyze qualitative data used in a published scientific article where qualitative content analysis was performed by human researchers, and to qualitatively compare results from the published article with results generated by GPT-4.

2. Materials & methods

2.1. Data collection

This study was performed using the full interview dataset from a previously published qualitative study [20]. The purpose of the qualitative study was to "explore the experiences of patients treated with rehabilitation alone after an anterior cruciate ligament (ACL) injury" [20]. The study received ethical approval from the Swedish Ethical Review Authority (registration number 2020-02834) and was performed in accordance with the Declaration of Helsinki. The data was collected from patients who had suffered one ACL injury treated with rehabilitation alone over the course of at least 2 years. A convenience sample of patients was recruited through communication with physiotherapists working at sports rehabilitation clinics in Gothenburg, Sweden. Data was then collected via individual semi-structured interviews that were recorded and transcribed. All interviews were performed digitally (via ZOOM web-based application, version 5.0.8 [21]). The mean length of the interviews was 21 (range 10–35) minutes, which resulted in transcripts of 50,167 words. A detailed description of the methods used for data collection is available in the original publication [20].

Patients were informed that participation was voluntary and that they could withdraw at any time. Interviews were analyzed confidentially. Oral recorded consent was collected at the beginning of each interview. Ethical approval was obtained from the Swedish Ethical Review Authority (registration number 2020-02834).

Fourteen patients were included in the referenced study (Table 1) [20].

The analysis in the original publication [20] was performed using an interpretive constructivist epistemological approach, using qualitative content analysis based on the framework developed by Graneheim and Lundman [22,23]. The first author (R.P.) was primarily responsible for the analysis process, but the full analysis was performed by six different authors, three males and three females, where data was continuously triangulated among them. The analysis was performed according to the following steps [22,23]: (Step 1) Transcripts were first read thoroughly to obtain a general understanding of the collected data. (Step 2) Meaning units were identified, extracted, and shortened to condensed meaning units. The resulting 627 condensed meaning units were then extracted to a spreadsheet. (Step 3) Each condensed meaning unit was assigned a code. (Step 4) Codes were grouped for similarities and differences in sub-categories. While grouping codes in sub-categories, transcripts were read again, and sub-categories were validated against the transcripts. In total, nine sub-categories were produced. (Step 5) Sub-categories were grouped for similarities and differences into three main categories. (Step 6) Via an extensive discussion, and allowing for interpretation of the data, a single theme was created to summarize patients' experiences of rehabilitation alone as a treatment after ACL injury.

Table 1
Demographics for patients included in the referenced study [20].

Sex	
Female number, (%)	8 (57%)
Age, years	
Mean	35.9
Median; range	37; 19–56
Time from ACL injury to interview, months	
Mean	32
Median; range	32; 24–44

Piussi et al. [20] *BMJ Open Sport Exerc Med* 2023;9:e001501. ACL, anterior cruciate ligament.

2.2. Patients and public involvement statement

Neither patients nor the public were involved in recruitment to or the conduct of the study.

2.3. Data analysis

To answer the aim of this study, we input the anonymized transcripts of all the interviews that were used [20] to the most-expensive consumer subscription plan of ChatGPT4 (accessed via the web interface; OpenAI Pro consumer subscription plan) [8]. In order for the transcripts to be fully anonymized, all timecodes, geographical information, and other trace-enabling meta and primary data were permanently deleted; only raw anonymous interview text was copied to GPT-4. The analysis was performed in January 2025.

The analysis was performed in several different attempts. The several attempts undertaken in the analysis were necessary due to issues encountered with GPT-4 with respect to the instructions provided in the prompts. The prompt and description of the issues encountered are reported in a Box for each step. For the full prompt history and issues encountered, please see Supplementary Table S1.

2.3.1. First attempt: full analysis in one go

After uploading the anonymized transcripts to GPT-4, it was prompted to perform the full analysis exactly as the human analysis was carried out and to produce the relevant results from each of the six analysis steps (Box 1). GPT-4 responded that it required 7 days to produce the full analysis, during which time the browser used for interacting with the chatbot could be (and was) closed. This first report produced by GPT-4 consisted of one table with three items in each of the following output types: meaning units, codes, sub-categories, main categories and themes. Because of the significant discrepancy between the human- and GPT-4-based results, in particular with regard to the quantity of results, a second attempt was deemed necessary.

Box 1. Prompt for the first attempt.
''Conduct a qualitative content analysis of the provided interview transcripts, following the steps outlined below. The analysis should be systematic and comprehensive, adhering closely to the text, except where interpretation is explicitly allowed:

1. *Extract meaning units: Identify all segments of the text that carry meaning relevant to the study's aim, which is to explore the experiences of patients treated with rehabilitation alone after an ACL injury.*
2. *Develop codes: Summarize the essence of each meaning unit with a short, descriptive label.*
3. *Group into subcategories: Combine related codes into subcategories that remain closely tied to the text.*
4. *Group into main categories: Aggregate subcategories into broader main categories, ensuring they reflect the core aspects of the text without interpretation.*
5. *Develop themes: Group the main categories into overarching themes. Themes may include interpretative elements that synthesize and provide deeper insight into the data.*

The results should include a detailed table or structured report showing:
- *Meaning units*
- *Corresponding codes*
- *Subcategories*
- *Main categories*
- *Themes*

Ensure that the interpretative aspects are confined to the themes and that all other steps remain close to the participants' words and context.

2.3.2. Second attempt: serial execution of analysis steps 1–3 followed by steps 4–6 in one go

The prompt provided to GPT-4 was modified to force it to perform Steps 1–3 from each of the 15 interviews separately before continuing to Steps 4–6 (Box 2). GPT-4 took 5 days to produce three meaning units and three codes per interview transcript (i.e., 45 total results for each of Steps 2 and 3; no results for Steps 4–6 were included in the response). While better than the first attempt, the results were still too limited.

2.3.3. Third attempt: fully serial analysis in one go

GPT-4 was prompted with the same prompt as in the second attempt (Box 2), but to fully analyze one interview at a time. This attempt was thought to be necessary to force GPT-4 to more thoroughly and independently analyze each of the interviews. GPT-4 analyzed one interview at a time and reported exactly 10 meaning units, 10 codes, seven sub-categories, five main categories and two themes per interview. While promising, the uniformity of the outputs in terms of results for each interview and analysis step indicated GPT-4 was not treating the entirety of the material equally. For example, longer interviews that were known to have a higher quantity of relevant content resulted in the same number of results as those that were known to have less.

Box 2. Prompt for the second attempt.

Conduct a comprehensive qualitative content analysis of fifteen interview transcripts. For each interview, identify:

- *every meaning unit: extract all text segments that convey information relevant to the study aim*
- *codes: assign descriptive labels to each meaning unit*

then, across all interviews:

- *group codes into subcategories while staying close to the text*
- *aggregate subcategories into main categories that reflect broader concepts*
- *synthesize themes based on main categories. Themes may include interpretative insights*

The final output must include each and every meaning unit and its corresponding code for all fifteen interviews. Show how subcategories and main categories emerge across the dataset, and provide overarching themes with a clear connection to the earlier steps.

2.3.4. Fourth attempt: fully serial meaning unit extraction followed by dynamic coding and steps 4–6

Analysis Steps 1–2: meaning unit extraction from each interview

GPT-4 was prompted to dynamically extract every meaning unit based on the text, even if one interview had significantly more or fewer units than another (Box 3). At this point the instruction provided to GPT-4 referred to the meaning units only due to the difficulties encountered in performing all the steps of the analysis with one prompt. To minimize the possibility that GPT-4 would automatize the process, the same prompt was provided before each transcript, and the transcripts were provided one at a time. GPT-4 then produced 261 meaning units from all the interviews.

Analysis Step 3: dynamic coding of meaning units

Box 3. Prompt for the fourth attempt, Steps 1–2.

I will provide you with another interview, and I want you to please perform a dynamic qualitative content analysis of the provided interview transcript. The analysis must adhere to the following requirements: extract all meaning units directly from the text, dynamically reflecting the variability and richness of the data. Do not impose any preconceptions, patterns, or fixed numbers of meaning units. For each meaning unit, provide a condensed version that captures the core idea while staying true to the text. No shortcuts: there must be no omissions, reductions, or assumptions about the number or nature of the meaning units. Every extracted unit must be thoughtfully analyzed and condensed. Ensure the analysis is conducted properly, without shortcuts or pre-fixed extraction schemes. If there are any uncertainties or limitations during the analysis, they must be explicitly stated.

GPT-4 was provided with the list of 261 meaning units that it produced in analysis Step 2 and prompted to convert them into codes (Box 4). In this case, GPT-4 returned a list of 60 codes which was deemed insufficient. The rest of the codes were not handled. GPT-4 was then prompted to provide all codes. However, GPT-4 returned a response that did not include all codes. It detected its error and automatically corrected the response to include all codes. This process was automatically reproduced by GPT-4 18 times, without human interaction. The final response included 181 codes in total. Because it had not coded all the meaning units, it seemed apparent that GPT-4 placed too much emphasis on early parts of the meaning unit list. We therefore repeated this step of the analysis, but in a serial fashion, by prompting GPT-4 to analyze batches of independent meaning units (i.e., we manually split the list of meaning units into batches of 50, the last batch containing only 11). We then serially prompted GPT-4 to analyze each batch such that 261 codes were produced.

Analysis Step 4: dynamic sub-categorization of the resulting codes

Box 4. Prompt for the fourth attempt, Step 3.
Please read the following list of condensed meaning units and provide a corresponding code for each one, based on Graneheim and Lundman's approach to qualitative content analysis. Each code should capture the essence of the meaning unit and reflect the core of the content without simplifying or summarizing it. For each meaning unit, I expect you to generate a unique and relevant code that aligns with the data's context and intended interpretation. Make sure the codes are reflective of the data and relevant to the research question, which focuses on understanding the experiences of patients treated with rehabilitation alone after an ACL injury.

GPT-4 was prompted to categorize the 261 codes into sub-categories (Box 5). In the response, GPT-4 created six sub-categories; the sixth/last one was named "uncategorized" and contained 227 codes. GPT-4 then required 44 additional prompts to complete the task of sorting the 261 codes. Each time GPT-4 sorted more codes, e.g., 203, then 215, then 226, and so on. Finally, GPT-4 produced fifteen sub-categories.

Analysis Steps 5–6: main category and theme generation

Box 5. Prompt for the fourth attempt, Step 4.
*I would like you to perform a qualitative content analysis using the following instructions: I will provide you with context and a set of codes. The aim is to group all provided codes based on their similarities and differences into well-defined subcategories according to the description by Graneheim and Lundman. Every single code must be sorted into a subcategory. Do not leave any codes unclassified, and do not assume that remaining codes will fit into already established subcategories without proper analysis. Clearly and explicitly report the results of your categorization. For each subcategory, list the numbers of all codes assigned to it. The output format should look like this: subcategory [name], containing the following codes: [*e.g.*, 1, 2, 4, 46, 121, etc.]. Attention to detail: ensure that subcategories are logically consistent and mutually exclusive, with minimal to no overlaps. There is no limit to the number of subcategories that can be created.*

Upon prompting GPT-4 to categorize sub-categories into main categories (Box 6), it returned six main categories, from which GPT-4 was prompted to create theme(s) (Box 6); it produced four.

Box 6. Prompts for the fourth attempt, Steps 5–6.
Prompt 1: *You now have the context, codes, subcategories, and raw interview texts at your disposal. The next step is to organize the subcategories into main categories based on the qualitative content analysis methodology described by Graneheim and Lundman. Group the provided subcategories into main categories. The purpose is to capture the manifest content of the interviews—that is, focus on what is explicitly being said by the participants. Each main category should encompass subcategories that share similar or related meanings. Ensure the grouping is consistent and reflects the content accurately. There is no restriction on the number of main categories that can be created. Allow the data to guide the structure, ensuring that no subcategory is forced into a group where it does not belong. For each main category, provide: a clear and concise name that reflects the essence of the grouped subcategories, and a description explaining what the main category represents, supported by the explicit content described by the participants. Ensure that your analysis captures what the participants have said directly, without interpreting underlying or latent meanings at this stage.*

Prompt 2: *The next step in the qualitative content analysis is to develop one or more themes that synthesize the insights from the interviews, codes, subcategories, and main categories. According to Graneheim and Lundman, themes represent a red thread that runs through the data, capturing the latent content—your interpretation of what lies beneath the surface of what is being said. Develop one or more themes that encapsulate the core of participants' experiences, as described in the interviews, codes, subcategories, and main categories. These themes should address the aim of the project. Move beyond what is explicitly stated (manifest content) to interpret the deeper meanings and patterns within the data. Identify underlying ideas, emotional tones, and shared experiences that connect the participants' narratives. Ensure that the themes are directly tied to the aim of the project, providing insightful answers to the research question.*

Results of the analysis, structured as theme, main- and sub-categories, between humans and GPT-4 were described and then qualitatively compared in the results section.

3. Results

3.1. Human results

One theme, supported by three main categories and nine sub-categories, emerged from the previous human-based analysis [20]. The theme which summarized experiences of patients treated with rehabilitation alone after an ACL injury was: "Is the grass greener on the other side? Context characterized by uncertainty" (Table 2).

3.2. GPT-4 results

Four themes, supported by six main categories and 15 sub-categories emerged from the analysis performed by GPT-4 (Table 3). The four themes were: (1) Navigating life after an ACL injury: a journey of adaptation and resilience. (2) Balancing uncertainty and empowerment in rehabilitation. (3) Redefining identity and purpose through recovery. (4) The interplay of support and self-reliance in healing.

3.3. Qualitative comparison

Several similarities were noted between the two analyses: first, uncertainty was lifted to a thematic level. Furthermore, both analyses included decision-making about treatment after ACL injury, but the human-made analysis framed it within a single subcategory "not knowing what to do," whereas GPT-4 analysis expanded this into a broader main category with subcategories like "knowledge needs" and "advocacy". Both analyses put rehabilitation challenges and adaptations to daily life following an ACL injury at a sub-category level.

Differences between the analyses included, but were not limited to, that the human-made analysis resulted in one theme, while GPT-4 returned four. GPT-4 identified several latent contents that were lifted to the theme level such as "a journey of adaptation and resilience" or "redefining identity and purpose through recovery" and "interplay of support and self-reliance" that were put in the main category level (adaptation or support) or not included at all (identity) in the human-made analysis. Main categories were grouped temporally in the human-made analysis as compared conceptually in GPT-4's analysis. Furthermore, the temporal dimension was only apparent in the main and sub-categories of the human-based analysis. Knee-related symptoms after ACL injury were categorized as a main category "the knee: a symptomatic obstacle" in the human-made analysis, but as a sub-category, within "coping and adaptation" in GPT-4's analysis.

Overall, uncertainty was a shared focus, though addressed differently. Both analyses highlighted decision-making and the role of healthcare systems to create or alleviate challenges in patients who suffered an ACL injury. Adaptation and physical limitations were emphasized as central to patients' daily lives. The human-made analysis was given a temporal, narrower approach, while GPT-4′s analysis took a broader psychological perspective with more latent contents. Sub- and main categories in the human analysis are broader and more unevenly distributed compared with the results of GPT-4.

4. Discussion

The main findings in this study were the similarities and differences between a human-made analysis and an analysis made by GPT-4 of the same qualitative interview transcripts. Furthermore, the various steps and iterative prompting undertaken in GPT-4 to produce the analysis should be seen as an important current limitation in the use of LLM within qualitative research.

There were several similarities between the human-made analysis and the analysis made by GPT-4. On a theme level, uncertainty was found to permeate the interviews analyzed, reflecting patients' concerns about treatment decisions in the past, current knee function, and future limitations following ACL injury treated with rehabilitation alone. In the

Table 2
Human-generated main and subcategories supporting the previously published theme.

Main categories	Sub-categories
Past: the ACL injury and its consequences	The terrible impact of the injury Healthcare system: lost without help Treatment choice: not knowing what to do Rehabilitation of an ACL injury: the endless road Lessons: taking responsibility for my body
Present: having knee-related symptoms	My knee does well enough for what I want The knee: a symptomatic obstacle Physical activity: necessary adjustments
Future: what might happen?	The future: uncertainty

ACL, anterior cruciate ligament.

Table 3
Main and subcategories supporting the themes emerged by the GPT-4-made analysis.

Main categories	Sub-categories
Injury circumstances and immediate impact	Injury context and circumstances Unique experiences and complications
Rehabilitation journey and challenges	Rehabilitation process and challenges Practical adjustments and solutions Financial and logistical challenges
Emotional and psychological adjustments	Emotional and psychological impact Social and identity impacts
Decision-making and knowledge needs	Decision-making around surgery Patient knowledge and awareness Patient advocacy and healthcare system navigation
Coping and adaptation in daily life	Adaptation and lifestyle changes Pain and physical limitations Long-term concerns and future outlook
Support and positive reflections	Support systems and resources Positive outcomes and reflections

human-made analysis, uncertainty was lifted as the 'only' and main theme that was found through the interviews. Researchers who performed the analysis interpreted uncertainty as the unifying narrative thread: that is, the dominant characteristic of life after an ACL injury treated with rehabilitation alone. Importantly, this theme is not only generated by words but relies as well on humans trying to understand other humans' situations. For instance, if a patient says, ''I do not know what to do with all these knee symptoms, whether it would be better for surgery or not," we might interpret this as uncertainty. However, GPT-4 might categorize it as ''knowledge needs," as it cannot perceive emotional nuances or interpret human experiences. Conversely, GPT-4's four themes reflected that patients experience diverse emotional, physical, and social dynamics during their recovery, and captured broader psychological, social, and emotional dimensions. This major difference can be the result of the weight and importance given to factors during the analysis. In the human-made analysis, the sub-category ''the knee a symptomatic obstacle" was the predominant category. A spider diagram was created, in the original publication, to illustrate that this sub-category contained the vast majority of all the codes: 163, as compared with the second sub-category for size: ''treatment choice, not knowing what to do" which contained 92 codes. These two sub-categories heavily informed the uncertainty theme; that is, patients reported to suffer many knee-related symptoms and not to know whether opting for surgery would alleviate symptoms, which in turn was interpreted as uncertainty. GPT-4's uniform distribution of codes and categories suggests an artificial structure not grounded in data variability. GPT-4 is a blackbox tool [24]: what we know, as users, is that we give the model a textual input (prompt), and we will receive a textual output (response), but we do not know how the response is generated. We believe that to organize the analysis, GPT-4 created schemes and systematic workflows, which resulted in an even distribution of weight and importance across sub-, main categories, and themes. Furthermore, it remains unclear whether GPT-4 considered the broader context of the original interview transcripts when converting meaning units into codes, codes into sub-categories, sub-categories into main categories, and main categories into themes, or if it processed them as isolated text segments without reference to the full narrative. Furthermore, the third theme identified by GPT-4 was ''redefining identity". In the human-based analysis no result about identity was found. We believe this might be a hallucination by GPT-4. Challenges such as hallucination, that is, GPT-4 incorrectly assigned codes to themes or made up information in summary, have been previously reported [25]. Importantly, the hallucinated data was not included at a sub-category level, but placed in a theme. The placement in a theme suggests that GPT-4 interpreted ''redefining identity" as a central thread through all the interviews.

4.1. Clinical/research implications

One important implication of this study is that the analysis was iterative, and GPT-4 needed 19 prompts to convert all meaning units to codes, and a further 45 prompts to sort all codes into sub-categories. GPT-4 was reported not to be able to manage multiple transcripts and not to be able to capture nuanced data essential to the research question [25].

At the time of the present study, it was unclear why LLMs such as GPT-4 demonstrated these limitations when handling large or multi-part qualitative datasets. While such difficulties were initially speculated to be related to context length [26], subsequent disclosures indicate that limitations observed in consumer-facing ChatGPT services are more likely attributable to system-level factors rather than strict token limits. These include dynamic routing of users to different model variants, each with distinct long-context behaviors, as well as attention-weighting and truncation mechanisms that may prioritize recent or salient inputs and limit full utilization of lengthy transcripts. As model selection and internal processing are not

transparent to the end user in consumer-facing interfaces, such factors may substantially affect the consistency and completeness of qualitative analyses.

An additional consideration concerns how GPT-4 was accessed in this study. The analysis was performed using a paid, consumer-facing version of ChatGPT, which is designed to be easy to use but provides limited transparency and control. As AI tools continue to be developed, companies increasingly reserve more controlled and stable implementations of LLMs for enterprise or programming-based use, that can be accessed on a per-token basis. These setups are intended to support reliability and reproducibility but require technical expertise and infrastructure that were beyond the scope of the present study. Whether the use of such enterprise-level configurations would have altered the qualitative analytic performance observed here remains unknown. The findings should therefore be interpreted as reflecting the capabilities and limitations of GPT-4 as deployed through a consumer-facing interface.

Our results confirm what others have indicated: that in the fast-developing world of AI, at the moment, human touch and creativity remain invaluable within qualitative research [27]. This highlights the need for human oversight when using AI tools in qualitative health research, especially when interpreting nuanced patient experiences that influence clinical decision-making and care delivery.

Researchers in other fields have attempted to compare human-made analysis with GPT-4′s analysis on the same data [28,29]. Authors conclude that human-made analysis brought a depth of analysis, sensitivity to nuances, and interpretive flexibility that AI models like GPT-4 lack. Hamilton and colleagues [28] suggested that LLMs such as GPT-4 might not be used as the sole tool to perform a qualitative analysis, but rather as an extra researcher for triangulating the analysis [28].

Finally, it should be noted that the specific results observed in this study relate to the version of GPT-4 available at the time of analysis and may not be directly transferable to newer models released thereafter. However, the need for human oversight and the types of challenges identified remain important considerations for researchers using LLM more broadly. As such, while model performance may evolve, the findings highlight aspects of AI-assisted qualitative analysis that warrant continued human attention regardless of model version.

5. Conclusion

The analysis performed by humans and GPT-4 had similarities and differences, where the analysis performed by GPT-4 presented more evenly balanced results, as opposed to the specific results presented in the human-made analysis. The use of GPT-4 for qualitative analysis in its present form is challenging and needs to be performed across several steps. Currently, GPT-4 should not be used as the only tool in a qualitative analysis of interview data.

Ethics statement

Ethical approval was obtained from the Swedish Ethical Review Authority (registration number 2020-02834). All patients who participated in the original study provided informed consent before participation.

CRediT authorship contribution statement

Ramana Piussi: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Justin F. Schneiderman:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization. **Yinan Yu:** Writing – review & editing, Methodology. **Kristian Samuelsson:** Writing – review & editing, Methodology. **Eric Hamrin Senorski:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Kristian Samuelsson reports a relationship with Getinge AB that includes: board membership. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank all patients who participated in the referenced study.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.knee.2026.104388.

References

[1] Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312:71–2. doi: https://doi.org/10.1136/bmj.312.7023.71.
[2] Kelly MP, Heath I, Howick J, Greenhalgh T. The importance of values in evidence-based medicine. BMC Med Ethics 2015;16:69. doi: https://doi.org/10.1186/s12910-015-0063-3.
[3] Madanian S, Nakarada-Kordic I, Reay S, Chetty Th. Patients' perspectives on digital health tools. PEC Innov 2023;2:100171. , 10.1016/j.pecinn.2023.100171.
[4] Tringale M, Stephen G, Boylan AM, Heneghan C. Integrating patient values and preferences in healthcare: a systematic review of qualitative evidence. BMJ Open 2022;12:e067268. doi: https://doi.org/10.1136/bmjopen-2022-067268.
[5] Moser A, Korstjens I. Series: practical guidance to qualitative research. Part 1: introduction. Eur J Gen Pract 2017;23:271–3. doi: https://doi.org/10.1080/13814788.2017.1375093.
[6] Evans AB, Natalie B-R, Joanna B, Georgia C, Fiona D, Stine F, et al. Qualitative research in sports studies: challenges, possibilities and the current state of play. Eur J Sport Soc 2021;18:1–17. doi: https://doi.org/10.1080/16138171.2021.1899969.
[7] Secules S, McCall C. What research can do: rethinking qualitative research designs to promote change towards equity and inclusion. Stud Eng Educat 2023;4:26–45. , https://doi.org/10.21061/see.96.
[8] OpenAI. Introducing ChatGPT. 2024. Available at: https://openai.com/blog/chatgpt.
[9] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56. doi: https://doi.org/10.1038/s41591-018-0300-7.
[10] Bubeck S, Chandrasekaran V, Eldan R, Gehrke JA, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. ArXiv 2023;abs/2303.12712.
[11] Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024;26:e60807. doi: https://doi.org/10.2196/60807.
[12] Dagher T, Dwyer EP, Baker HP, Kalidoss S, Strelzow JA. "Dr. AI Will See You Now": how do ChatGPT-4 treatment recommendations align with orthopaedic clinical practice guidelines? Clin Orthop Relat Res 2024;482:2098–106. doi: https://doi.org/10.1097/corr.0000000000003234.
[13] Villarreal-Espinosa JB, Berreta RS, Allende F, Garcia JR, Ayala S, Familiari F, et al. Accuracy assessment of ChatGPT responses to frequently asked questions regarding anterior cruciate ligament surgery. Knee 2024;51:84–92. doi: https://doi.org/10.1016/j.knee.2024.08.014.
[14] Smith AM, Jacquez EA, Argintar EH. Assessing the efficacy of an AI-powered chatbot (ChatGPT) in providing information on orthopedic surgeries: a comparative study with expert opinion. Cureus 2024;16:e63287. doi: https://doi.org/10.7759/cureus.63287.
[15] Kaarre J, Feldt R, Keeling LE, Dadoo S, Zsidai B, Hughes JD, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. Knee Surg Sports Traumatol Arthrosc 2023;31:5190–8. doi: https://doi.org/10.1007/s00167-023-07529-2.
[16] Andriollo L, Picchi A, Sangaletti R, Perticarini L, Rossi SMP, Logroscino G, et al. The role of artificial intelligence in anterior cruciate ligament injuries: current concepts and future perspectives. Healthcare (Basel) 2024;12:300. doi: https://doi.org/10.3390/healthcare12030300.
[17] Muzumdar P, Cheemalapati S, RamiReddy SR, Singh K, Kurian G, Muley A. The dead internet theory: A survey on artificial interactions and the future of social media. arXiv preprint 2025, arXiv:250200007.
[18] Marshall DT, Naff DB. The ethics of using artificial intelligence in qualitative research. J Empir Res Hum Res Ethics 2024;19:92–102. doi: https://doi.org/10.1177/15562646241262659.
[19] Morgan DL. Exploring the use of artificial intelligence for qualitative data analysis: the case of ChatGPT. Int J Qual Methods 2023;22:16094069231211248. doi: https://doi.org/10.1177/16094069231211248.
[20] Piussi R, Simonson R, Kjellander M, Jacobsson A, Ivarsson A, Karlsson J, et al. When context creates uncertainty: experiences of patients who choose rehabilitation as a treatment after an ACL injury. BMJ Open Sport Exerc Med 2023;9:e001501. doi: https://doi.org/10.1136/bmjsem-2022-001501.
[21] Inc. ZVC. Zoom Video Communications Inc. 5.8.0 ed.: https://zoom.us; 2021.
[22] Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004;24:105–12. doi: https://doi.org/10.1016/j.nedt.2003.10.001.
[23] Graneheim UH, Lindgren BM, Lundman B. Methodological challenges in qualitative content analysis: a discussion paper. Nurse Educ Today 2017;56:29–34. doi: https://doi.org/10.1016/j.nedt.2017.06.002.
[24] Rai A. Explainable AI: from black box to glass box. J Acad Mark Sci 2020;48:137–41. doi: https://doi.org/10.1007/s11747-019-00710-5.
[25] Lee VV, van der Lubbe SCC, Goh LH, Valderas JM. Harnessing ChatGPT for thematic analysis: are we ready? J Med Internet Res 2024;26:e54974. doi: https://doi.org/10.2196/54974.
[26] Liu J, Zhu D, Bai Z, He Y, Liao H, Que H, et al. A comprehensive survey on long context language modeling. arXiv preprint 2025, arXiv:250317407.
[27] Naqvi WM, Shaikh SZ, Mishra GV. Large language models in physical therapy: time to adapt and adept. Front Public Health 2024;12:1364660. doi: https://doi.org/10.3389/fpubh.2024.1364660.
[28] Hamilton L, Elliott D, Quick A, Smith S, Choplin V. Exploring the use of AI in qualitative analysis: a comparative study of guaranteed income data. Int J Qual Methods 2023;22:16094069231201504. doi: https://doi.org/10.1177/16094069231201504.
[29] De Paoli S. Performing an inductive thematic analysis of semi-structured interviews with a large language model: an exploration and provocation on the limits of the approach. Soc Sci Comput Rev 2023;42:997–1019. doi: https://doi.org/10.1177/08944393231220483.