ORIGINAL PAPER

Nagoya J. Med. Sci. **86**. 620–644, 2024
doi:10.18999/nagjms.86.4.620

A mixed-methods study comparing human-led and ChatGPT-driven qualitative analysis in medical education research

Takeshi Kondo[1,2], Junichiro Miyachi[2,3],
Anders Jönsson[4] and Hiroshi Nishigori[2]

[1]Department of General Medicine/Family & Community Medicine, Nagoya University Graduate School of Medicine, Nagoya, Japan
[2]Center for Medical Education, Nagoya University Graduate School of Medicine, Nagoya, Japan
[3]Academic and Research Centre, Hokkaido Centre for Family Medicine, Sapporo, Japan
[4]Faculty of Education, Kristianstad University, Kristianstad, Sweden

ABSTRACT

Qualitative research, used to analyse non-numerical data including interview texts, is crucial in understanding medical education processes. However, it is often complex and time-consuming, leading to an interest in technology for streamlining the analysis. This study investigated the applicability of ChatGPT, a large language model, in thematic analysis for medical qualitative research. Previous research has used ChatGPT to explore the deductive process as a qualitative study. This study evaluated thematic analysis including the inductive process by ChatGPT with reference to human qualitative analysis. A convergent design mixed-methods study was used. Using a thematic analysis approach, ChatGPT (model: GPT-4) analysed some interview data from a previously published medical research article. The assessors evaluated the qualitative analysis of ChatGPT using human qualitative analysis as a benchmark. Three assessors compared the human-conducted and ChatGPT-driven qualitative analyses. ChatGPT scored higher in most aspects but showed variable transferability and mixed depth scores. In the integrated analysis including qualitative data, six themes were identified: superficial similarity of results with human analysis, good first impression, explicit association with data and process, contamination by directions in prompts, deficiency of thick descriptions based on context and research questions, and lack of theoretical derivation. ChatGPT excels at extracting key data points and summarising information; however, it is prone to prompt contamination, which necessitates careful scrutiny. To achieve deeper analysis, it is essential to supplement the research context with human input and explore the theoretical framework.

Keywords: qualitative study, medical education, ChatGPT, artificial intelligence, large language models

Abbreviation:
LLMs: large language models

This is an Open Access article distributed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view the details of this license, please visit (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Received: December 5, 2023; accepted: March 19, 2024

Corresponding Author: Takeshi Kondo, MD, MHPE
Department of General Medicine/Family & Community Medicine, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, Nagoya 466-8550, Japan
Tel & Fax: +81-52-744-2951, E-mail: kondo@med.nagoya-u.ac.jp

INTRODUCTION

Qualitative research plays a crucial role in the field of medicine by providing valuable insights for understanding and improving psychosocial aspects of medicine, medical education processes, and practices.[1,2] The history of qualitative research in medicine spans over 20 years, and various research methodologies have been employed.[3,4] Analytical methods, such as thematic and content analyses, have been widely used by researchers.[5-7] Thematic analysis, in particular, is a common method in qualitative research that allows researchers to identify meaningful patterns and themes within data. In addition, clear steps are proposed for thematic analysis, making it easy to validate the analysis process.[7] Although there are supporting tools, such as NVivo, for these analyses, they have a fundamental requirement for humans to manually analyse the language, extract concepts, describe theories, and construct models. Consequently, mastering qualitative analysis requires extensive training, and the analysis process is time-consuming.[8]

In recent years, advancements in large-scale language models have enabled ground-breaking achievements in the field of natural language processing.[9,10] In particular, ChatGPT, based on the GPT-4 architecture developed by OpenAI and released on 14 March 2023, is equipped with advanced text generation and comprehension abilities, demonstrating performance comparable to that of top human scorers on descriptive test questions.[11,12] ChatGPT is also gaining attention in the medical field for its potential. It has been shown that it can generate draft responses to patients' questions that are comparable to those of a doctor.[13] Therefore, there are expectations for its potential to support healthcare services. Such linguistic capabilities of ChatGPT suggest the potential to streamline thematic analysis in qualitative research and alleviate the burden on researchers. Prior research suggests that ChatGPT (GPT-3.5) can assign predefined codes to text.[14] However, no study has examined whether large language models (LLMs), such as ChatGPT, can derive codes and themes from a text on their own in a qualitative analysis process, such as thematic analysis, and whether such codes are equivalent to those generated by human qualitative analysis. Furthermore, there are concerns that the application of LLMs, such as ChatGPT, to academia could lead to misuse or ethical issues.[15] It is also necessary to consider what problems might arise when conducting thematic analysis using ChatGPT.

In this study, we examined the effectiveness of using ChatGPT for thematic analysis in qualitative research and compared its analysis with human analysis. We hope to introduce new possibilities regarding research methods in medicine. Our research investigated to what extent the quality of thematic analysis conducted by ChatGPT is equivalent to that performed by humans. We explored the characteristics of any differences and their implications. Through this study, we aim to propose new methodologies for enhancing the reliability and efficiency of qualitative research in medicine, as well as to deepen the discussion on the potential of utilising LLMs for qualitative research. By elucidating the advantages and limitations of ChatGPT-based qualitative analysis compared with human analysis, we provide insights into the selection and application of qualitative research methods that may benefit future medical studies. Furthermore, the findings of this study advance qualitative research not only in medicine but also in social science and cultural anthropology fields, among others.

METHODS

Design

In this post-positivism-based mixed-methods research study, we used a convergent design to compare human-conducted qualitative analysis in previously published medical education research with qualitative analysis newly implemented using ChatGPT.[16,17] In the previous study, a human-performed thematic analysis was conducted on the transcriptions of interviews with supervisors, which were previously published in an academic paper.[18] Both the full transcripts of the interviews and the process of the thematic analysis were made publicly available. In this research, a thematic analysis of the same data was performed using ChatGPT. The process of this analysis was then compared to the human-conducted thematic analysis from both qualitative and quantitative perspectives. We conducted a quantitative examination using the scales described below to gain a comparative perspective. To consider the meaning of the obtained scores and further examine the depth and creativity of ChatGPT's analysis, assessors also qualitatively evaluated the analysis by ChatGPT and documented their evaluation as comments. Subsequently, analysis and integration were conducted using the quantitative and qualitative data. Using this approach, we thoroughly investigated the effectiveness of qualitative analysis using ChatGPT and the differences between it and human analysis. The overall flow of the research is illustrated in Fig. 1.

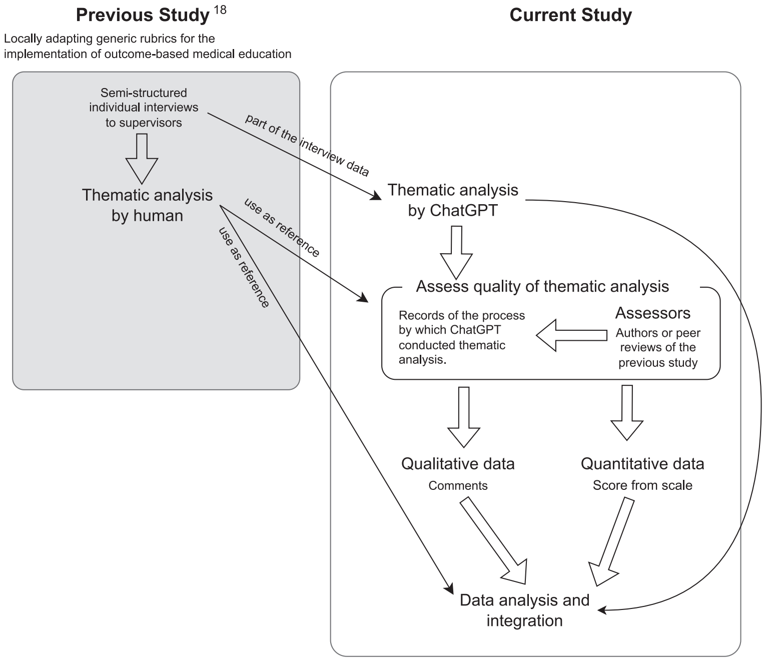
Previous Study 18
Locally adapting generic rubrics for the implementation of outcome-based medical education
Current Study
Semi-structured individual interviews to supervisors
Thematic analysis by human
part of the interview data
use as reference
use as reference
Thematic analysis by ChatGPT
Assess quality of thematic analysis
Records of the process by which ChatGPT conducted thematic analysis.
Assessors
Authors or peer reviews of the previous study
Qualitative data
Comments
Quantitative data
Score from scale
Data analysis and integration

Fig. 1 Overall research flow

The previous study refers to an academic paper that has already been published. Within this paper, the interviews with supervisors that were subjected to thematic analysis, as well as the process of thematic analysis itself, are publicly available. In the current research, a thematic analysis using ChatGPT was conducted on these publicly available interview records. The thematic analysis performed by humans in the previous study served as a reference for qualitatively and quantitatively evaluating the thematic analysis conducted by ChatGPT.

Thematic analysis using ChatGPT

We conducted a qualitative analysis of interview data using OpenAI's ChatGPT (model: GPT-4). For the analysis, we selected the qualitative analysis section of the authors' published paper 'Locally adapting generic rubrics for the implementation of outcome-based medical education: a mixed-methods approach'.[16] This study compared generic rubrics and localised evaluation forms used to assess medical residents. This paper was selected because transcripts of the interviews are publicly available, peer reviewers could be contacted for collaboration with the current study, ChatGPT does not know about the paper because it was published after ChatGPT's data were collected, and the authors (TK, AJ, and HN) thoroughly understand the qualitative analysis process, which makes it easier to evaluate the differences between ChatGPT's qualitative analysis and human analysis. Because ChatGPT currently cannot accept long inputs, the authors selected easily understandable sections from the published interviews, even if they were short texts.

Thematic analysis was conducted following Braun and Clarke's six-step process for using an inductive thematic analysis approach: (1) familiarisation with the data, (2) generating initial codes, (3) searching for themes, (4) reviewing themes, (5) defining and naming themes, and (6) producing the final report.[7,19] The research question for the analysis was the same as in the original paper: 'What is the effect of such local adaptation on supervising doctors as assessors?' Analysis using ChatGPT was conducted via the following process.

First, ChatGPT was asked to act as an experienced qualitative researcher. We conducted a thematic analysis and confirmed that the answers aligned with the authors' understanding. After briefly explaining the research background in a prompt, we requested that ChatGPT familiarise itself with the interview data and summarise them as part of Step 1. The interview data were input in two parts, and we asked for summaries at each stage to ensure that the generated summaries were consistent with the authors' understanding. Next, as part of Step 2, we asked ChatGPT to generate the initial codes for the read text. After confirming that the codes were appropriately assigned, we requested that ChatGPT perform steps (3) searching for themes, (4) reviewing themes, and (5) defining and naming themes. Finally, considering the progress thus far, we asked ChatGPT to produce a final report (Step 6) regarding the differences between the two evaluation forms. Appendix 1 presents the prompts and generated responses. This process required 45 min to complete.

Assessors

The authors and peer reviewers of the analysed article participated in this study to assess the quality of the ChatGPT analysis. The authors and peer reviewers were experts involved in medical education research and well-versed in qualitative research. Furthermore, their involvement in writing and reviewing the paper gave them a deep understanding of the research data and content.

Quantitative data and scale for assessing qualitative analysis

To assess qualitative analysis with ChatGPT from a quantitative perspective, a custom scale was developed to compare the differences between human- and ChatGPT-driven qualitative analyses. Because existing scales for evaluating the quality of qualitative research primarily focus on the overall research design rather than the specific analysis process,[20-22] a new scale was created for this study by integrating relevant sections from textbooks on qualitative coding and articles on qualitative research quality assessment.[20-26] The assessors, either the individuals who conducted the thematic analysis in previous studies or peer reviewers who validated those evaluations, were chosen based on their in-depth understanding of the process involved in the human-led qualitative analysis. These individuals are suitable for the task as they can utilise human qualitative analysis as a benchmark when evaluating the analysis performed by ChatGPT. Therefore, all scores for human qualitative analysis were fixed at 4 as a reference when evaluating the analysis performed by ChatGPT. Furthermore, by placing the reference point mid-scale, it was possible to assign both higher and lower scores using the same scale. The qualitative analysis performed by ChatGPT was scored on independent perspectives on the scale, which ranged from 1 to 7 (Table 1).

Table 1 Scale to evaluate quality of qualitative analysis

Name	Description	Main studies
Depth	Evaluate the depth and richness of the analyses. Consider if it goes beyond a mere description of the data and provides meaningful interpretations. On the scale, 1 should represent superficial and 7 in-depth.	Tong et al,[22] 2007

Confirmability	Evaluate the confirmability of the analysis, considering its accuracy in relation to the text. On a scale from 1 to 7, 1 represents no correlation and 7 represents a strong correlation.	Braun et al,[19] 2019;
		O'Brien et al,[20] 2014;
		Saldaña et al,[23] 2021

Credibility	Consider whether the analysis accurately aligns with the research question. On a scale ranging from 1 to 7, 1 indicates complete lack of credibility and 7 indicates high credibility.	Stenfors et al,[24] 2020
Transferability	Consider whether the analysis is applicable in other contexts. On a scale ranging from 1 to 7, 1 indicates complete lack of transferability and 7 indicates high transferability.	O'Brien et al,[20] 2014; Lumsden et al,[25] 2022
Dependability	Consider the extent to which transparency is ensured in the analysis and whether the process is clear. On a scale from 1 to 7, 1 represents low dependability and 7 represents high dependability.	O'Brien et al,[20] 2014; Saldaña et al,[23] 2021
Consistency	Consider whether the analysis is free from inconsistencies and remains consistent throughout. On a scale ranging from 1 to 7, 1 indicates complete lack of consistency and 7 indicates high consistency.	Frambach et al,[21] 2013

Three assessors, who had either authored or peer-reviewed the target paper and were thus familiar with human qualitative analysis, were asked to review a Word document containing the ChatGPT qualitative analysis process conducted via thematic analysis. They then assigned scores to the ChatGPT analysis using the developed scale. Owing to the limited sample size, no statistical analysis was performed, and the results are presented as graphs.

Qualitative data

Because of the limitations of quantitative evaluation solely based on a scale, which tends to bias towards content that can be objectively evaluated, it becomes challenging to consider the depth and creativity of the analysis. Therefore, we also qualitatively assessed the thematic analysis performed by ChatGPT. After evaluation based on the scale, the assessors were asked to provide free-text descriptions in the same form.

Data analysis and integration

In convergent design mixed-methods approaches, an integration analysis is performed to consolidate the quantitative and qualitative data that have been collected.[27] We integrated the quantitative and qualitative data through a process inspired by the analysis process of thematic analysis.

1. Familiarisation with the data: TK and JM independently and thoroughly reviewed the quantitative data and qualitative data, the process of human-driven qualitative analysis, and

the ChatGPT-driven qualitative analysis.

2. Generating initial codes: TK and JM independently assigned codes to the qualitative data (comments).
3. Code refinement and theme development: TK and JM brought their codes together for discussion. Taking advantage of TK's position as the party who conducted the human-driven qualitative analysis and JM's expertise in qualitative analysis without being involved in the human-driven analysis, the codes were revised to incorporate richer perspectives through discussion. In addition, during the discussion, quantitative data derived from the scores were also compared and considered. Based on these discussions, TK generated themes.
4. Theme review and refinement: All researchers reviewed and refined the themes, ensuring they accurately represented the dataset and provided a comprehensive understanding of the differences between human-driven and ChatGPT-driven qualitative analyses. Discrepancies were resolved through discussion and consensus.
5. Reporting findings: The final themes and associated codes are presented in the results, highlighting the key differences between human-driven and ChatGPT-driven qualitative analyses.

Reflexivity. Reflexivity was considered and practiced throughout the research process to enhance the trustworthiness and validity of the findings. The research team, including the authors of the paper[16] and the peer reviewers, comprised individuals with diverse backgrounds, experiences, and perspectives related to medical education research.

TK, the primary investigator responsible for the ChatGPT-driven qualitative analysis, is a physician in general practice and a medical education researcher. As the author of the analysed paper, TK is deeply familiar with its content and has prior experience in qualitative research. During the comparison of human- and ChatGPT-driven qualitative analyses, TK actively reflected on how his biases and assumptions potentially influenced the evaluation process. JM, a family physician with a master's degree in medical anthropology and a PhD degree from conducting qualitative research in medical education research, provided a unique perspective to the team. Importantly, although JM was not directly involved in the original paper on which the comparison was based, he has expertise in the subject matter of that paper, having lectured on the topic of rubrics in domestic faculty development programmes. This positions him to provide additional and valuable insight. AJ, the peer reviewer of the analysed paper, offered an external and independent perspective on the quality of the qualitative analysis. TK and JM, who conducted the data analysis and integration, have received training in qualitative research in a master's programme in medical education and a master's programmes in medical anthropology and PhD, respectively. HN, a qualitative research expert in medical education, provided valuable insights into the research process and contributed to reflexivity discussions.

Throughout the research process, the team engaged in reflective discussions and continuously examined their roles, biases, and potential influences on the research findings. These reflective practices, combined with the diverse backgrounds of the research team members, contributed to generating a more robust and transparent analysis, ensuring the trustworthiness and validity of the results.

RESULTS

Two of the authors and one of the peer reviewers who had worked on the analysed paper agreed to participate in the ChatGPT qualitative analysis. Thus, three assessors conducted a comparative examination of the human and ChatGPT qualitative analysis.

Quantitative data

As mentioned in the Methods section, human qualitative analysis was assigned a fixed score of 4, whereas ChatGPT qualitative analysis was assessed on a scale of 1 (lowest) to 7 (highest) based on six aspects: depth, confirmability, credibility, transferability, dependability, and consistency. Three assessors conducted analyses based on these criteria. The evaluation was conducted using a web form from April 1st to April 6th, 2023.

The results of the evaluation are shown in Fig. 2. For confirmability, credibility, dependability, and consistency, all assessors assigned higher scores to the ChatGPT qualitative analysis than to the human qualitative analysis. Regarding transferability, one assessor assigned an equal score to both the human and ChatGPT analyses, while the others assigned higher scores to the ChatGPT analysis. However, the evaluation of depth varied more among assessors.

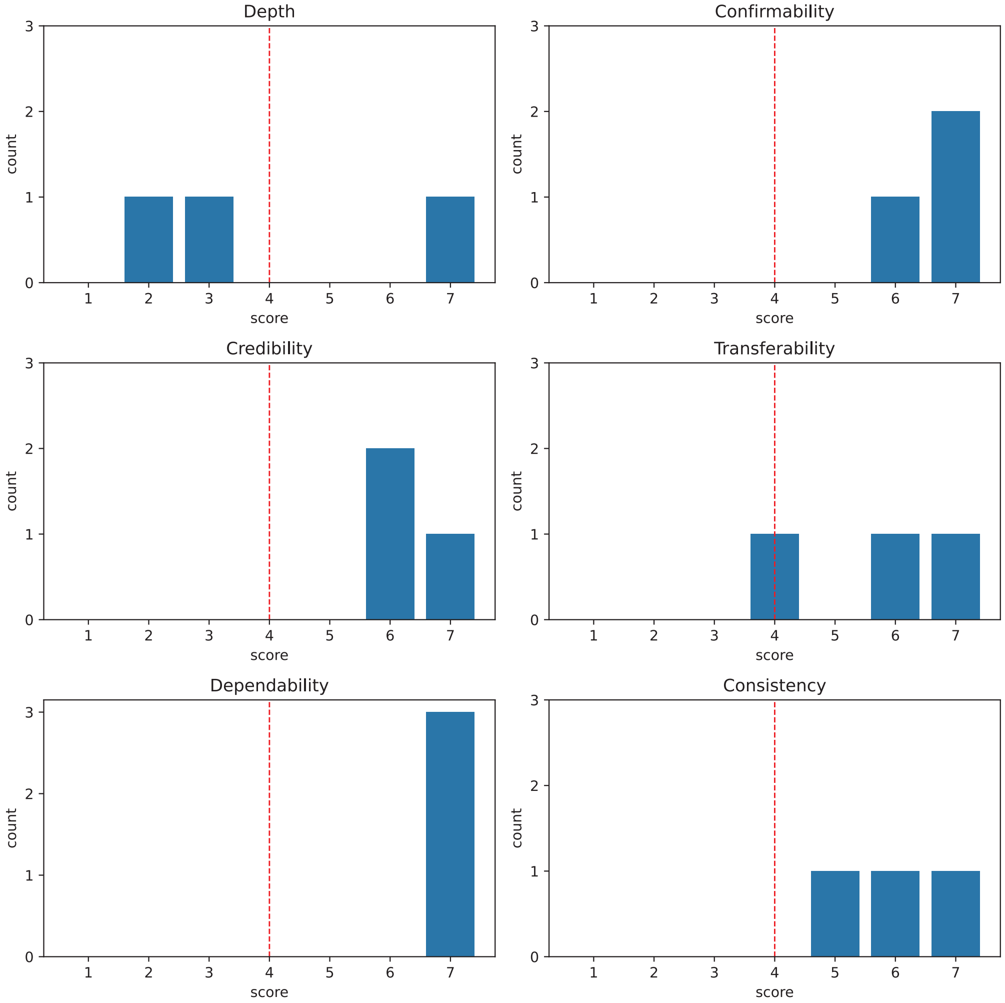

Fig. 2 Qualitative analysis scores of ChatGPT compared with those of humans
Three assessors rated ChatGPT on six aspects, with the human qualitative analysis assigned a fixed rating of 4: depth, confirmability, credibility, transferability, dependability, and consistency. The minimum score was 1, and the maximum score was 7.

Qualitative data

The three assessors provided comments on the human and ChatGPT qualitative analyses via a web form, along with the above scores (Supplemental Table).

Data analysis and integration

Data analysis and integration were led by TK and JM as mentioned in Methods. The obtained codes were tabulated, and themes were identified (Table 2).

Table 2 Codes derived from the thematic analysis

Code ID	Code	Text ID
1	Faithful to sentences in the text	A-1
2	Oversight of contamination	A-1,B-5
3	Lack of analysis of social situation and relationship beyond the text	A-2
4	Analysis of what is said not why and how it happens	A-2
5	Explicit association between interview data and research question	A-3
6	Ensuring third person verification of analysis process	A-4
7	Consistency	A-5,B-3
8	Good impression of ChatGPT	A-6,B-1
9	Ability to trace the steps of thematic analysis explicitly	B-2
10	Explicit process	B-3
11	Transparent process	B-3
12	Succinct report production	B-4
13	Lack of ‘thick’ description	B-4
14	Coherent and precise analytical steps	B-5
15	Better summary of the interview	B-6
16	Lack of explanatory analysis	B-7
17	Absence of theorisation	B-7
18	Good at inductive process	B-8
19	Questionable ability for deductive approach	B-8
20	Lack of consideration of larger context	B-8
21	Similarity in theme	C-2
22	Lack of context-sensitive meaning-making	C-3,C-5
23	Lack of consideration of local context	C-4
24	Sociocultural contextualisation of the data	C-7

Superficial similarity of results with human qualitative analysis (Related codes: 8, 9, 21). Similar processes were followed in the ChatGPT and human thematic analyses, and there were some similarities in the derived themes and their meanings. The theme ‘Invalid assessments’, derived by humans, suggests that abstract descriptions and difficult-to-assess items in the generic rubric lead to inappropriate assessments, which are improved by localised tools. This is similar to the meaning of the 'streamlined assessment process' derived by ChatGPT, which suggested that simple items in the localised tools make assessments easier. Despite these differences, the main finding was that the human qualitative and ChatGPT analyses were similar.

Good first impression (Related codes: 8, 18). ChatGPT's ability to identify key points from interview data and provide summaries, along with its clear analytical process, generally led assessors to have a favourable impression of its analysis.

Explicit association with data and process (Related codes: 1, 5, 6, 9, 11, 14, 15). The ChatGPT analysis made it easier to understand the linkages to the interview data and research questions than the human analysis, and it was also clear which steps of the thematic analysis had been performed in which part of the process. In human thematic analysis, it is difficult to see how the six steps of thematic analysis were used; however, in the analysis by ChatGPT, it was clear which parts of the steps were used.

Contamination by directions in prompts (Related codes: 2). Analysis by ChatGPT included mixed results, with some results derived from prompt directions rather than interview data. In the sixth step of the thematic analysis using the ChatGPT, Theme 1 included the following text: 'the localised tools seemed to better capture the specific context and situations of the hospital under study'. This does not appear in the ChatGPT analysis. ChatGPT was given the following direction outlining the background of the study: we localised the generic rubric and developed localised tools, taking into account the situation of the hospital under study and the situation, and we believe that this information constituted contamination of the analysis results. However, this contamination was not noticed by the assessors involved in this study.

Deficiency of thick descriptions based on context and research question (Related codes: 3, 4, 13, 16, 19, 22, 23, 24). While ChatGPT selected and summarised the main points of the interview results, it did not connect them to a thick description that considered the interviewee's context and the background from which the research questions were derived. As pointed out in C-3 (in Supplemental Table), ChatGPT derived the theme 'Enhanced clarity and specificity', which is clear from reading the text alone, whereas humans derived the theme 'Mismatch between the rubrics and clinical context', which considers the context in which the interviewee was placed and the wording of the rubric.

Lack of theoretical derivation (Related codes: 17, 20). ChatGPT did not derive a theory or theoretical framework for textual content analysis. In the analysis by humans, cognitive load theory was devised as a theoretical framework based on the situation in which interviewees were placed and the interviewees' comments on the number of questions asked; accordingly, the analysis was deepened. However, no such theory was derived or used in the analysis by ChatGPT.

DISCUSSION

In this study, we employed a large-scale language model, ChatGPT (model: GPT-4), to conduct thematic analysis of text data based solely on simple, pre-training-free instructions, and compared the results quantitatively and qualitatively with those obtained from human-performed thematic analysis. The qualitative analysis conducted by ChatGPT yielded higher scores than those of the human analysts in multiple domains. However, the qualitative assessment revealed certain challenges associated with the qualitative analysis performed by ChatGPT.

The qualitative analysis conducted by ChatGPT yielded results similar to those derived from the human analysis. In this context, the connection between the text and the analysis performed by ChatGPT was more explicit than that of the human qualitative analysis, and the relationship between the analytical steps in thematic analysis was also clear. This is likely to have contributed to the high scores achieved for numerous indicators during the quantitative assessment, as well as the positive impressions conveyed in the comments regarding the analysis. These findings indicate that ChatGPT accurately processed the main points of interview texts and effectively summarised and synthesised information. Previous studies have demonstrated that LLMs, such as ChatGPT, exhibit superior performance in summarising text content.[28] Meanwhile, such seemingly positive analytical outcomes may also make it more difficult to discern potential issues within the analysis.

This study revealed that one of the less noticeable problems with ChatGPT analysis was the contamination by directions in the prompts. This means that the directive used in interactions with ChatGPT inadvertently influenced the results of the data analysis. It is already known that LLMs, such as ChatGPT, exhibit a phenomenon called hallucination, wherein they generate plausible yet factually incorrect outputs.[29] The inclusion of instructional sentences does not necessarily diverge from the truth; however, it often goes unnoticed by researchers because of the assumptions embedded within instructions. In this study, assessors who authored or peer-reviewed the research papers targeted for analysis by ChatGPT did not detect the inclusion of instructional sentences. This inclusion is likely because a premise was not conveyed to ChatGPT, namely that the analysis should be based on interview data alone, without incorporating the content of the instructions. This discrepancy between the assumptions made by humans and the instructions conveyed to ChatGPT can be difficult to detect and necessitates vigilance. To ensure the reliability of analysis results, it can be argued that the results obtained from ChatGPT require careful and meticulous verification by humans. This point, along with the perspective of data protection, should be noted in future research activities involving ChatGPT. In this study, only already published data were inputted into ChatGPT, so there was no concern about data leakage. However, large-scale artificial intelligence systems, including ChatGPT, are often provided on the cloud. As qualitative research often deals with private matters of the participants, it is necessary to pay close attention to how data are handled on the cloud in future research. In other words, it is essential to be mindful of the characteristics of how artificial intelligence handles data and instructions and to what extent data are shared.

Furthermore, qualitative analysis requires more than merely summarising text; it also requires a deep analysis and application of theories that consider the background and context of the interviewees. ChatGPT's qualitative analysis faces challenges in this regard. Both quantitative and qualitative assessments revealed that ChatGPT struggles with in-depth analysis. This is likely because ChatGPT has not been provided with the vast amount of information that humans consider during qualitative analysis, including non-verbal information obtained by immersing oneself in the interview and participants' clinical and educational settings. Additionally, ChatGPT has not been provided with the theories discussed in previous studies. While making the qualitative analysis by ChatGPT more transparent, the absence of such implicit information serves as a limiting factor in its ability to derive valuable insights. These findings highlight the relative strengths of human analysts and ChatGPT.

Shaw et al explained the strategic approach to research analysis as three distinct logics, namely deduction, induction, and abduction, which connect data and theory.[30] In deduction logic, data are interpreted by applying existing theories, and theories are tested using data. In induction logic, data are described, and theories are generated from the data. In abduction logic, a recursive and iterative analysis is conducted, moving back and forth between data and theory.[30,31] With reference to the diagram of Shaw et al, the potential applications of ChatGPT in qualitative analysis based on the results of prior research and the present study are illustrated in Fig. 3.

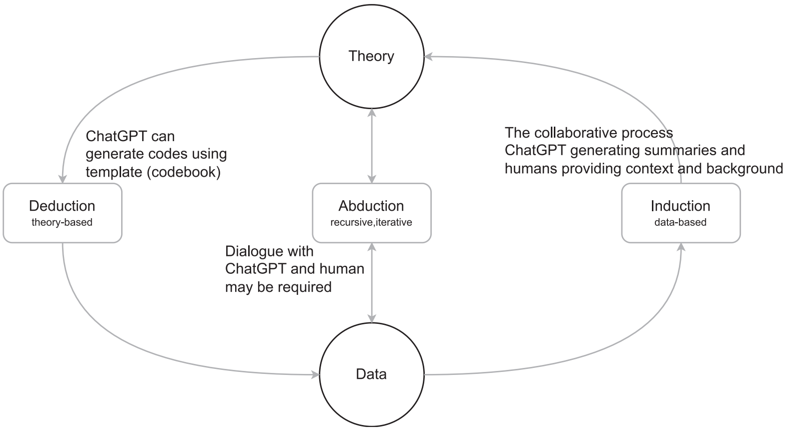

Fig. 3 Potential applications of ChatGPT in qualitative analysis
ChatGPT can utilise templates and assign codes, thus playing a part in the process of deduction. When humans provide context and background, ChatGPT can participate in the process of induction by generating summaries of qualitative data. For the iterative process of abduction, involving a back and forth between theory and data, dialogue between ChatGPT and humans may be required.

Prior research involving text coding based on codebooks has demonstrated the potential for utilising LLMs, including ChatGPT, for deductive analysis.[28] While both inductive and deductive approaches can be employed in thematic analysis,[32] it has been suggested that, in fields such as medicine, in which analysis is conducted in conjunction with existing theories and frameworks, an abductive approach is often adopted.[30,31] The analysis conducted by ChatGPT in this study summarised the main points of the data more clearly than did humans through an inductive approach. However, it was unable to derive theories that considered the background and context of the data. ChatGPT was unable to extract particularly intriguing statements from the text and delve deeper into them. Additionally, merely instructing ChatGPT to follow the six steps of thematic analysis in sequence appears to be insufficient to facilitate an abductive approach in which the exploration of existing theories and frameworks is conducted concurrently with data comparison for a deeper analysis. Furthermore, for more creative approaches, such as the reflective thematic analysis by Braun and Clarke and the abductive thematic analysis by Thompson, it is essential to consider the context and the researcher's position.[19,31] In this study, even when ChatGPT was instructed to proceed from Step 4 to Step 5 in the thematic analysis (Appendix 1), the analysis did not substantially deepen, largely involving only a slight rephrasing of the content. While this underscores a significant limitation of the current analytical method, these steps could potentially serve as a key to facilitating an abductive approach in the thematic analysis by engaging in dialogue with ChatGPT about the context and theoretical frameworks of the research participants and researchers.

Limitations

At present, detailed information on GPT-4 has not been sufficiently disclosed, which impacts the transparency of this study. More detailed data are anticipated to become available in the future. The data in this study were small-scale and the number of assessors was limited to three, which necessitates caution in interpretation. However, the chosen assessors had a deep understanding of the data, and analysis using prior research and theory enhanced the transferability of the findings.

Owing to the current input limitations of ChatGPT, it was only able to analyse a portion of the interview data in this study. This significant restriction, the inability to handle large amounts of text data, which is often required in qualitative analysis, is expected to be improved with future technological advancements.

In this study, we were unable to identify suitable quantitative measures to evaluate the process of thematic analysis for qualitative assessments by ChatGPT. Therefore, we created and evaluated our metrics using existing studies as guides. We hope that in the future, measures that have been examined for reliability and validity will be developed.

The assessors were either authors or reviewers of the papers analysed by ChatGPT, and one of the two individuals who conducted the thematic analysis of the assessors' comments was the author of the analysed paper. While this may introduce bias into the results and analysis process, it also has the potential to lead to a deeper analysis owing to the assessors' familiarity with the background and context.

In this study, ChatGPT was only given thematic analysis and text data without detailed information on the background of medical education research and theories. Consequently, its full potential may not have been accurately assessed. Future research that includes input regarding the research context and dialogue regarding theories may reveal capabilities not evident in this study.

CONCLUSIONS

In this study, we conducted a quantitative and qualitative examination of thematic analysis performed by humans and ChatGPT. ChatGPT demonstrated excellence in extracting and summarising data points, but there was the possibility of direction contamination, which requires careful verification. Furthermore, to achieve deeper analysis, it is necessary to supplement the research context with human input and examine the theoretical framework.

ARTICLE INFORMATION

Ethics statement

Ethical approval was not required.

Contributors

TK assumes responsibility for the overall study design, data collection and analysis, as well as for the composition of the manuscript. JM provided guidance on the research design, collaborated with TK in conducting thematic analysis, and contributed to writing the manuscript. AJ offered advice on the research design and worked alongside TK in the manuscript composition process. HN partnered with TK to devise the comprehensive study design and participated in writing the manuscript.

Competing interests

All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/disclosure-of-interest/ and declare: TK is supported by JSPS KAKENHI Grant Number 21K10372; no other relationships or activities that could appear to have influenced the submitted work exist.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 21K10372. We would like to thank Editage (www.editage.com) for English language editing.

Role of the funding source

The JSPS KAKENHI Grant Number 21K10372 covered the English editing service expenses.

Transparency declaration

The lead author affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study have been explained.

REFERENCES

1 Britten N. Qualitative interviews in medical research. *BMJ*. 1995;311(6999):251–253. doi:10.1136/bmj.311.6999.251.

2 Stalmeijer RE, Mcnaughton N, Van Mook WN. Using focus groups in medical education research: AMEE Guide No. 91. *Med Teach*. 2014;36(11):923–939. doi:10.3109/0142159X.2014.917165.

3 Watling CJ, Lingard L. Grounded theory in medical education research: AMEE Guide No. 70. *Med Teach*. 2012;34(10):850–861. doi:10.3109/0142159X.2012.704439.

4 Rees CE, Monrouxe LV, McDonald LA. Narrative, emotion and action: analysing 'most memorable' professionalism dilemmas. *Med Educ*. 2013;47(1):80–96. doi:10.1111/j.1365-2923.2012.04302.x.

5 Elo S, Kyngäs H. The qualitative content analysis process. *J Adv Nurs*. 2008;62(1):107–115. doi:10.1111/j.1365-2648.2007.04569.x.

6 Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res*. 2005;15(9):1277–1288. doi:10.1177/1049732305276687.

7 Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. 2006;3(2):77–101. doi:10.1191/1478088706qp063oa.

8 Malterud K. Qualitative research: standards, challenges, and guidelines. *Lancet*. 2001;358(9280):483–488. doi:10.1016/S0140-6736(01)05627-6.

9 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arXiv*. Preprint posted online June 12, 2017. doi:10.48550/arXiv.1706.03762.

10 Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. Published online 2018. Accessed March 1, 2024.

11 OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. *arXiv*. Preprint posted online March 15, 2023. doi:10.48550/arXiv.2303.08774.

12 Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 Passes the bar exam. *Philos Trans A Math Phys Eng Sci.* 2024;382(2270):20230254. doi:10.2139/ssrn.4389233.

13 Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589–596. doi:10.1001/jamainternmed.2023.1838.

14 Xiao Z, Yuan X, Liao QV, Abdelghani R, Oudeyer PY. Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In: *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23 Companion)*. Association for Computing Machinery; 2023:75–78. doi:10.1145/3581754.3584136.

15 Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. 2023;11(6):887. doi:10.3390/healthcare11060887.

16 Shannon-Baker P. Making paradigms meaningful in mixed methods research. *J Mix Methods Res*. 2016;10(4):319–334. doi:10.1177/1558689815575861.

17 Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. *Health Serv Res*. 2013;48(6 Pt 2):2134–2156. doi:10.1111/1475-6773.12117.

18 Kondo T, Nishigori H, van der Vleuten C. Locally adapting generic rubrics for the implementation of outcome-based medical education: a mixed-methods approach. *BMC Med Educ*. 2022;22(1):262. doi:10.1186/s12909-022-03352-4.

19 Braun V, Clarke V. Reflecting on reflexive thematic analysis. *Qual Res Sport Exerc Health*. 2019;11(4):589–

597. doi:10.1080/2159676X.2019.1628806.
20 O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. *Acad Med*. 2014;89(9):1245–1251. doi:10.1097/ACM.0000000000000388.
21 Frambach JM, van der Vleuten CP, Durning SJ. AM last page. Quality criteria in qualitative and quantitative research. *Acad Med*. 2013;88(4):552. doi:10.1097/ACM.0b013e31828abf7f.
22 Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care*. 2007;19(6):349–357. doi:10.1093/intqhc/mzm042.
23 Saldaña J. *The Coding Manual for Qualitative Researchers*. 4th ed. London: SAGE Publications Ltd; 2021.
24 Stenfors T, Kajamaa A, Bennett D. How to … assess the quality of qualitative research. *Clin Teach*. 2020;17(6):596–599. doi:10.1111/tct.13242.
25 Lumsden K. Assessing the 'quality' of qualitative research. https://qualitativetraining.com/2022/01/21/assessing-the-quality-of-qualitative-research/. Published January 21, 2022. Accessed March 24, 2023.
26 Yadav D. Criteria for good qualitative research: A comprehensive review. *Asia-Pacific Educ Res*. 2022;31(6):679–689. doi:10.1007/s40299-021-00619-0.
27 Creswell JW, David Creswell J. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 5th ed. London: SAGE Publications, Inc; 2018.
28 Yang X, Li Y, Zhang X, Chen H, Cheng W. Exploring the limits of ChatGPT for query or aspect-based text summarization. *arXiv*. Preprint posted online February 16, 2023. doi:10.48550/arXiv.2302.08081.
29 Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. *Radiology*. 2023;307(2):e230163. doi:10.1148/radiol.230163.
30 Shaw RL, Hiles DR, West K, Holland C, Gwyther H. From mixing methods to the logic(s) of inquiry: taking a fresh look at developing mixed design studies. *Health Psychol Behav Med*. 2018;6(1):226–244. doi:10.1080/21642850.2018.1515016.
31 Thompson J. A guide to abductive thematic analysis. *Qual Rep*. 2022;27(5):1410–1421. doi:10.46743/2160-3715/2022.5340.
32 Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: Striving to Meet the Trustworthiness Criteria. *Int J Qual Methods*. 2017;16(1):1–13. doi:10.1177/1609406917733847.

APPENDIX

Appendix 1 Thematic analysis using ChatGPT

2023/4/1
Using ChatGPT(Model: GPT-4)
The analysis took 45 minutes from start to finish.

—

Input-1

System:
You are an experienced qualitative researcher.

User:
Please tell me about thematic analysis, a method of qualitative analysis.

ChatGPT-1

Thematic analysis is a widely used method of qualitative analysis that involves identifying, analyzing, and reporting patterns (themes) within a dataset. This approach is particularly useful for exploring people's experiences, perspectives, and understanding of a given phenomenon, as it enables researchers to capture and interpret rich and complex data from sources such as interviews, focus groups, and observations.

The thematic analysis process typically involves six key steps:

1. Familiarization: The researcher immerses themselves in the data by reading and re-reading the transcripts, field notes, or other data sources, gaining an in-depth understanding of the material.
2. Generating initial codes: The researcher identifies meaningful segments in the data that relate to the research question or objectives and assigns descriptive codes to these segments. These codes help to organize and categorize the data.
3. Searching for themes: The researcher examines the codes to identify patterns and relationships between them, which may form the basis for potential themes. These themes represent broader, more abstract concepts that capture the essence of the coded data.
4. Reviewing themes: The researcher refines and validates the proposed themes, ensuring that they accurately represent the coded data and are coherent, distinct, and relevant to the research question. This step may involve merging, splitting, or discarding themes as necessary.
5. Defining and naming themes: The researcher provides a clear and concise definition for each theme, describing its scope, significance, and relationship with the research question. The themes are given descriptive and informative names that capture their essence.
6. Writing the report: The researcher presents a detailed and well-structured analysis of the themes, providing evidence from the data to support their interpretations. This report should be accessible, engaging, and persuasive, demonstrating the value and implications of the findings for the broader field of study.

Thematic analysis can be conducted in an inductive (data-driven) or deductive (theory-driven) manner, depending on the research question and the researcher's epistemological stance. It is a flexible and adaptable method that can be used across a wide range of disciplines and research contexts. However, it requires a rigorous and systematic approach to ensure the trustworthiness, validity, and reliability of the findings.

Input-2

Thank you very much. Now I would like to ask you to analyze the interviews for the medical education research using thematic analysis. In this study, supervisors were asked to use two different sets of evaluation forms to evaluate residents, one set used nationally and the other developed by the researcher, and asked about the differences. The nationally used evaluation form set is a paper evaluation form (generic rubrics) consisting of evaluation forms 1, 2, and 3. The generic rubric is designed to be used in any hospital in Japan and is not specific as to the circumstances under which it should be used. In this study, we localized the generic rubric and developed localized tools, taking into account the situation of the hospital under study and the situation in which the supervising physician actually performs the assessment. The localized tools is an online response evaluation form. Our research question is "What is the effect of such local adaptation on supervising doctors as assessors?" Before we begin analyzing the interviews, please tell me what you understand about this study.

ChatGPT-2

Based on your description, this medical education research study aims to compare the effectiveness of two different sets of evaluation forms used by supervisors to assess residents. One set is a nationally used, generic paper evaluation form consisting of evaluation forms 1, 2, and 3, which can be applied in any hospital in Japan. The other set is a localized, online response evaluation form developed by the researcher, taking into account the specific context of the hospital under study and the actual assessment situations experienced by supervising physicians.

The main research question is: "What is the effect of such local adaptation on supervising doctors as assessors?" The study seeks to understand the implications of using localized evaluation tools for supervising doctors in their role as assessors and how these adapted tools may impact the assessment process.

To investigate this question, the study will use qualitative data collected from interviews with supervisors who have used both sets of evaluation forms. Thematic analysis will be employed to identify, analyze, and report patterns within the interview data, allowing for a deeper understanding of the supervisors' experiences and perspectives on the use of the two different evaluation forms.

Input-3

Thank you very much. We will now begin our thematic analysis. I will give you the interview data in two parts. In earlier part of the interview below, the researcher first asks about the generic rubrics. If you familiarization with the data, please summarize the main points of the interview. If you have any questions, please ask me.

Interview:
Line # Speaker Content
B-25 interviewer I'd like to ask about the evaluation forms. I have brought this with me, but it is an evaluation form used nationwide.
B-26 interviewee The first part of the resident evaluation form, the general comments, can be written roughly, so I think it is easy to write, but the second part, the one that can be evaluated in more detail on seven levels, can be written roughly, but when you say seven levels, it is difficult to know where to put it. I check all the items for the time being, but sometimes I'm not sure whether I should put them on the list or not, because I haven't had a chance to observe them. I'm not sure if I've observed this or not. ......
B-27 interviewer Can you be a little more specific?
B-28 interviewee For example, recognizing conflicts of interest and responding to them in accordance with the management policy, I think that conflicts of interest are considered only when a situation where a conflict of interest arises, or if there is a case where you think it is likely to happen, you would consider it. But if it doesn't turn out to be such a case in the first place and it just slips away, there is no opportunity to evaluate it, but if there are three or four other evaluation items, for example, three of the five evaluation items in medical ethics are about this level, but one is not evaluated at all, or if it turns out to be a little negative, where do you put it? So I kind of give them a fluffy overall rating and write comments for the ones I really like, but I'm not sure if this is the right way to give them. Also, especially in the fields of health and welfare, if a person is not involved to that extent, but is just hospitalized, discharged, and returns to the same life as before, there may not be a case to evaluate. I'm not sure if it's okay to put it on, but I do it for some reason. I think it's a bit detailed, because there are seven items, and for each item there are three or four more detailed evaluation items, so I appreciate the fact that there are so many items, but in the end I don't know where to put them. I guess that's how I feel.
B-29 interviewer You said that the first one was easy to put on, but can you tell us more

about it?
B-30 interviewee The first one says levels 1, 2, 3, and 4, and you can softly write in four levels with those four questions, and the levels 1, 2, 3, and 4 are as expected, whether they are above, below, or if they are below, by how much. So, it's not a matter of giving them this score because they were able to do this specifically, but rather, it's a matter of giving them a 3 if they were able to meet the standards that other residents, in general, would be able to meet. On the second and subsequent pages, there are seven criteria for 5, and the middle one is 5, but how do you differentiate between 4, 6, and other subtle differences? In the end, I didn't really understand how to differentiate between subtle differences such as a 4 or a 6.
B-31 interviewer So, what are the good points of this evaluation form?
B-32 interviewee Is it the first one, or the second one?
B-33 interviewer It would be helpful if you could tell me everything.
B-34 interviewee The good thing about the first one, the Resident Evaluation Form 1, is that you don't have to worry about it too much and can just put it on intuitively or quickly. It's a good point. So, even though they are both in the middle, it is good to be able to check for minor changes, such as a child who was in the bottom of the middle at first, but by the second term, he or she is in the top of the middle.
B-35 interviewer There are three of these at ....
B-36 interviewee Resident Evaluation Form 3 is the same as 1. There is a level 3, which is the middle standard, and you can either go beyond it or not. For level 3 in particular, whether it can be done under the direct supervision of the supervising physician, or whether it can be done in a situation where it can be handled immediately without direct supervision, the levels 1, 2, 3, and 4 are divided quite concretely, so I think evaluation form 3 is the easiest among the three.
B-37 interviewer On the other hand, are there any problems with this evaluation form?
B-38 interviewee For assessment sheet 1, I don't understand what expectations are. I don't know what the expectations are, so if I were to assign it to first-year residents, I would assume that Level 3 would be the minimum level of misbehavior among first-year residents that I could tolerate, but I don't know if that's how other people assign it. In this sense, both evaluation sheets 1 and 3 are roughly the same level. In that sense, both evaluation sheets 1 and 3 are too rough, or rather, when we say that a child is at Level 3 from Level 3, it looks like the child has not changed at all, which is not good. On the evaluation sheet 2, the detailed one, I can write down the detailed changes, but as I said before, when I wonder where to put the detailed ones, which are good for this one point, but not so good for this one point, I wonder if I should take a break between them or just put them on the good side. I don't know if I should just give it to them and comment on what I like about it.

ChatGPT-3

Based on the interview data provided, the main points discussed can be summarized as follows:

1. The interviewee finds the first part of the generic rubric (Resident Evaluation Form 1) easy to complete because it has a simple 4-level rating scale. However, they are unsure about the expectations for each level and how other supervisors might interpret them.
2. The interviewee has difficulty with the second part of the evaluation form, which has a more detailed 7-level scale. They find it challenging to determine where to place the residents on the scale, especially when they have not observed certain behaviors or when some aspects of the residents' performance are positive while others are negative.
3. The interviewee appreciates the number of detailed evaluation items on the form, but they feel unsure about where to place residents in certain situations, such as when there are no

opportunities to evaluate specific items.
4. Resident Evaluation Form 3 is considered the easiest among the three forms because it has more concrete criteria for the 4-level scale. However, the interviewee also finds it difficult to determine the expectations for each level in this form.
5. The interviewee highlights the benefits of the more detailed evaluation items in capturing subtle changes in residents' performance, but they are unsure about how to balance and weigh these items when assigning an overall rating.

Overall, the interviewee finds the generic rubrics easy to use in some aspects, but they struggle with determining appropriate ratings on more detailed scales and feel uncertain about the expectations for each level. They also express concerns about the applicability of the evaluation forms in certain situations where specific items may not be relevant or observable.

Input-4
Thank you. The next part is the later part of the interview. Here we ask questions for the other set of evaluation forms(localized tools) and further ask about the differences between the two sets of evaluation forms(generic rubrics and localized tools). If you familiarization with the data, please summarize the main points of the interview. If you have any questions, please ask me.

Interview:
Line# Interviewee Content
B-44 interviewee This is an evaluation form that you can fill out online with your cell phone, right? I think online is easier because it's just a click away, and it's not so different from check sheets. As for the evaluation items, there are usually three or four options to choose from, so I don't have to worry too much about it.
B-45 interviewee What I also like about it is that I can write freely. There is a section that asks the residents to write down the specific tasks that they were able to do during their rotations, and there is also a section that asks the residents to tell us what we can do better to help them grow in the future.
B-46 interviewee When I was a resident, that's exactly what I wanted to hear. I thought that if I could receive feedback on how I could do better each time I finished a department, it would make it easier for me to use that information to move on to other departments. I remember that we didn't receive such evaluations every term. It's not too difficult to do.
B-47 interviewer You've covered a lot of ground, but when you compare the evaluation form with this one, what do you think the differences would be?
B-48 interviewee Can you give us an idea? The localized version is more specific and the level of the evaluation is not ambiguous. For example, in the case of item 5 (reporting, communication, and consultation), there are three choices: able to report and consult at an appropriate frequency, not able to report and consult, and never had the opportunity to do so, so there is no confusion.
B-49 interviewee But on the resident evaluation form, whether it's 1, 2, or 3 on paper, to put it bluntly, I feel that my evaluation can vary quite a bit depending on my mood.
B-50 interviewee Also, the comments. I get the impression that the comments are more positive in localized tools. It's easier to write comments because it says "please comment on what would be good to do.
B-51 interviewee In the evaluation forms 1, 2, and 3, it says that if there is a memorable episode, please describe it, so it is okay to write good things, but after that, it says that if the episode falls far short of expectations, please be sure to write it.

B-52 interviewer Do you think this kind of difference has any effect on the growth of residents?
B-53 interviewee I think it would be better to have an evaluation that is simple to understand. If the evaluation is blurred by the senior doctor's mood of the day, it would be difficult to know what to believe.

ChatGPT-4
Based on the later part of the interview, the main points discussed can be summarized as follows:

1. The interviewee appreciates the localized online evaluation form for its convenience and ease of use, as it can be completed on a mobile phone with just a few clicks.
2. They find the localized form's evaluation items less ambiguous and easier to understand, as there are fewer choices and more specific criteria to follow.
3. The interviewee values the opportunity for free-form comments on the localized form, as it encourages feedback on residents' growth and areas for improvement. They believe that receiving such feedback during their own residency would have been beneficial.
4. Comparing the two sets of evaluation forms, the interviewee feels that the localized form offers clearer and more specific evaluation criteria, which reduces potential variation in ratings based on their mood or other factors.
5. The interviewee also observes that the localized form seems to encourage more positive comments, whereas the generic rubrics could potentially lead to more negative feedback.
6. The interviewee suggests that a clearer and more straightforward evaluation system, like the localized form, would be more beneficial for residents' growth, as it is less influenced by external factors and provides more actionable feedback.

Overall, the interviewee prefers the localized online evaluation form due to its ease of use, clearer evaluation criteria, and focus on constructive feedback. They believe that such a form would be more effective in supporting residents' growth and development.

Input-5
Thank you. Now, we will move to second step of thematic analysis, generating initial codes. In this step, you systematically code the data by identifying significant features or segments and assigning a label or code to them. These codes are essentially the building blocks of your analysis and represent concepts that are relevant to your research question. Please output the codes in the format:

Line number of the text: comma-separated codes

Some lines may have no codes, while others may have multiple codes. Please skip lines with no codes in the output. Please put the serial code number for each code in parentheses after the code.

ChatGPT-5
B-26: ease of writing (1), difficulty in detailed evaluation (2), uncertainty in observation (3)
B-28: example of evaluation difficulty (4)
B-30: simple 4-level rating scale (5), uncertainty in expectations (6)
B-34: difficulty differentiating subtle differences (7)
B-36: concrete criteria in Evaluation Form 3 (8)
B-38: difficulty determining expectations (9), limitations in capturing changes (10)

B-44: ease of online evaluation (11), simplicity of choices (12)
B-45: opportunity for free-form comments (13), feedback on growth (14)
B-46: positive impact on residents' development (15)
B-48: specificity of localized form (16), less ambiguity (17)
B-49: influence of mood on evaluation (18)
B-50: more positive comments in localized tools (19), ease of writing comments (20)
B-51: potential for negative feedback in generic rubrics (21)
B-53: benefits of simpler evaluation (22), influence of mood on evaluation (18)

Input-6
Thank you! Now, we will move on to the third step of thematic analysis, searching for themes. The researcher examines the codes to identify patterns and relationships between them, which may form the basis for potential themes. These themes represent broader, more abstract concepts that capture the essence of the coded data. Please use code numbers when citing codes.

ChatGPT-6
After examining the codes, I have identified the following potential themes:

1. Clarity and specificity of evaluation criteria (codes: 2, 4, 5, 6, 7, 8, 9, 12, 16, 17)
   - This theme captures the differences in the clarity and specificity of evaluation criteria between the generic rubrics and localized tools. The generic rubrics were perceived as having more ambiguous criteria, while the localized tools provided clearer and more specific evaluation criteria, making it easier for supervisors to assess residents.
2. Ease of use and convenience (codes: 1, 5, 11, 20)
   - This theme reflects the preferences of the interviewee in terms of the user-friendliness and convenience of the evaluation forms. The generic rubrics were considered easy to write, while the localized tools were appreciated for their ease of online use and simplicity in providing feedback.
3. Impact of evaluation method on feedback quality (codes: 10, 13, 14, 15, 18, 19, 21, 22)
   - This theme highlights the influence of the evaluation method on the quality and nature of feedback provided to residents. The localized tools encouraged more positive and constructive feedback that focused on growth, while the generic rubrics could potentially lead to more negative feedback. The interviewee also acknowledged the potential influence of the assessor's mood on the evaluation, which could impact the consistency and reliability of the feedback.
4. Importance of feedback for residents' growth and development (codes: 14, 15)
   - This theme emphasizes the role of feedback in supporting residents' growth and development. The interviewee believes that receiving regular, actionable feedback during their own residency would have been beneficial and that a clearer and more straightforward evaluation system, like the localized form, would be more effective in promoting residents' growth.

Input-7
Excellent! Now, we will move on to the fourth step of thematic analysis, reviewing themes. The researcher refines and validates the proposed themes, ensuring that they accurately represent the coded data and are coherent, distinct, and relevant to the research question. Our research question is "What is the effect of such local adaptation on supervising doctors as assessors?" This step may involve merging, splitting, or discarding themes as necessary. Please use code numbers when citing codes.

ChatGPT-7
After reviewing the themes in relation to the research question, I have refined them as follows:

1. Clarity and specificity of evaluation criteria (codes: 2, 4, 5, 6, 7, 8, 9, 12, 16, 17)
   - This theme remains relevant, as it directly addresses the effect of local adaptation on supervising doctors as assessors. The localized tools provide clearer and more specific evaluation criteria, which improves the assessment process for supervisors and reduces ambiguity.
2. Assessment process efficiency (codes: 1, 5, 11, 20)
   - This theme has been refined to focus on the efficiency of the assessment process. Local adaptation through the use of localized tools positively affects supervising doctors as assessors by providing an easy-to-use online platform and simplifying the process of providing feedback.
3. Consistency and quality of feedback (codes: 10, 13, 14, 15, 18, 19, 21, 22)
   - This theme highlights the impact of local adaptation on the consistency and quality of feedback provided by supervising doctors. The localized tools encourage more constructive feedback, minimize the influence of mood on evaluation, and promote more reliable assessments.

As the themes are coherent, distinct, and relevant to the research question, no themes were discarded or merged.

Input-8
Excellent! Now, we will move on to the fifth step of thematic analysis, defining and naming themes: The researcher provides a clear and concise definition for each theme, describing its scope, significance, and relationship with the research question. The themes are given descriptive and informative names that capture their essence.
Again, our research question is “What is the effect of such local adaptation on supervising doctors as assessors?” Please use code numbers when citing codes.

ChatGPT-8
After reviewing the themes in relation to the research question, I have refined them as follows:

1. Clarity and specificity of evaluation criteria (codes: 2, 4, 5, 6, 7, 8, 9, 12, 16, 17)
   - This theme remains relevant, as it directly addresses the effect of local adaptation on supervising doctors as assessors. The localized tools provide clearer and more specific evaluation criteria, which improves the assessment process for supervisors and reduces ambiguity.
2. Assessment process efficiency (codes: 1, 5, 11, 20)
   - This theme has been refined to focus on the efficiency of the assessment process. Local adaptation through the use of localized tools positively affects supervising doctors as assessors by providing an easy-to-use online platform and simplifying the process of providing feedback.
3. Consistency and quality of feedback (codes: 10, 13, 14, 15, 18, 19, 21, 22)
   - This theme highlights the impact of local adaptation on the consistency and quality of feedback provided by supervising doctors. The localized tools encourage more constructive feedback, minimize the influence of mood on evaluation, and promote more reliable assessments.

As the themes are coherent, distinct, and relevant to the research question, no themes were discarded or merged.

Input-9
Excellent! Now, we will move on to the fifth step of thematic analysis, defining and naming

themes: The researcher provides a clear and concise definition for each theme, describing its scope, significance, and relationship with the research question. The themes are given descriptive and informative names that capture their essence.
Again, our research question is “What is the effect of such local adaptation on supervising doctors as assessors?” Please use code numbers when citing codes.

ChatGPT-9

1. Enhanced clarity and specificity
   - Definition: This theme explores how local adaptation improves the clarity and specificity of evaluation criteria, making it easier for supervising doctors to assess residents effectively and accurately.
   - Scope and significance: This theme captures the benefits of local adaptation in terms of providing clearer and more specific evaluation criteria, which in turn leads to a more reliable and consistent assessment process (codes: 2, 4, 5, 6, 7, 8, 9, 12, 16, 17).
   - Relationship with the research question: This theme directly addresses the research question by demonstrating the positive effects of local adaptation on supervising doctors as assessors through enhanced clarity and specificity of evaluation criteria.
2. Streamlined assessment process
   - Definition: This theme investigates how local adaptation improves the efficiency of the assessment process by simplifying the evaluation forms and making them more user-friendly.
   - Scope and significance: The theme covers the advantages of local adaptation in terms of streamlining the assessment process, making it more efficient and accessible for supervising doctors (codes: 1, 5, 11, 20).
   - Relationship with the research question: This theme is relevant to the research question by illustrating the positive impact of local adaptation on supervising doctors as assessors through a more efficient and user-friendly assessment process.
3. Consistent, high-quality feedback
   - Definition: This theme examines the influence of local adaptation on the consistency and quality of feedback provided by supervising doctors to residents, ensuring a more reliable and constructive assessment.
   - Scope and significance: The theme encompasses the benefits of local adaptation in promoting more consistent and high-quality feedback from supervising doctors, leading to more effective assessments (codes: 10, 13, 14, 15, 18, 19, 21, 22).
   - Relationship with the research question: This theme directly addresses the research question by highlighting the positive effects of local adaptation on supervising doctors as assessors through the provision of more consistent, high-quality feedback to residents.

Input-11

Great! Now, please conduct the final step of thematic analysis, writing the report. The researcher presents a detailed and well-structured analysis of the themes, providing evidence from the data to support their interpretations. This report should be accessible, engaging, and persuasive, demonstrating the value and implications of the findings for the broader field of study. Please use code numbers when citing codes.

ChatGPT-12

Title: The Effects of Local Adaptation on Supervising Doctors as Assessors in Medical Education

Introduction:
The current study aimed to investigate the effects of local adaptation on supervising doctors as assessors in a medical education setting. The research question was: "What is the effect of such local adaptation on supervising doctors as assessors?" Through a thematic analysis of interview data from supervising doctors, the study identified three main themes that provide insight into the benefits and implications of local adaptation in assessment practices.

- Theme 1: Enhanced clarity and specificity
  - The first theme highlighted the increased clarity and specificity of evaluation criteria provided by the local adaptation of assessment tools (codes: 2, 4, 5, 6, 7, 8, 9, 12, 16, 17). Supervising doctors reported that the localized tools offered more concrete and specific evaluation criteria, which in turn led to a more reliable and consistent assessment process. For example, the localized tools provided clearer distinctions between different performance levels, reducing the ambiguity and subjectivity often associated with generic rubrics (code 6). Furthermore, the localized tools seemed to better capture the specific context and situations of the hospital under study, making it more relevant and easier to use for the supervising doctors (codes 7, 9, 17).

- Theme 2: Streamlined assessment process
  - The second theme focused on the improved efficiency of the assessment process as a result of local adaptation (codes: 1, 5, 11, 20). Supervising doctors mentioned that the localized tools were more user-friendly and less time-consuming to complete. One significant factor contributing to this was the transition from paper-based forms to online response evaluation forms (code 11). Additionally, the localized tools offered fewer evaluation items, simplifying the assessment process and making it more manageable for supervising doctors (codes 1, 5, 20).

- Theme 3: Consistent, high-quality feedback
  - The third theme examined the influence of local adaptation on the consistency and quality of feedback provided by supervising doctors to residents (codes: 10, 13, 14, 15, 18, 19, 21, 22). The localized tools appeared to promote more constructive and positive feedback by encouraging supervising doctors to focus on residents' strengths and areas for improvement. This led to more reliable and effective assessments, ultimately benefiting residents' professional development (codes 10, 13, 15, 19, 21, 22). Supervising doctors also reported feeling less influenced by their mood when using the localized tools, resulting in more consistent evaluations (code 14).

Conclusion:
This study provides valuable insights into the effects of local adaptation on supervising doctors as assessors in medical education. The findings suggest that local adaptation leads to enhanced clarity and specificity, a streamlined assessment process, and more consistent, high-quality feedback. These benefits have significant implications for improving assessment practices and promoting residents' professional growth. Future research could explore the implementation of localized assessment tools in different contexts and investigate the long-term effects on residents' performance and development.

SUPPLEMENTARY INFORMATION

Supplemental Table Assessors' comments

ID	Text
A-1	In terms of depth of analysis, the ChatGPT analysis is faithful to the interview data and captures the interviewees' intentions well.
A-2	However, it does not take into account the situation of the supervisor or use a theoretical framework in its analysis.
A-3	Regarding Confirmability, ChatGPT clearly relates the interview data to the analysis, and the analysis of Credibility also fully considers the research question.
A-4	Regarding Dependability, if the analysis process could be made public, it would ensure transparency more clearly than for humans.
A-5	Consistency is also not a problem, although it is a short text.
A-6	In general, it can be said that the quality of qualitative analysis by ChatGPT is considerably high.
B-1	I was really impressed by the analysis made by the ChatGPT.
B-2	To contextualize the numbers provided in the quantitative evaluation, the ChatGTP was able to (a) summarize the main points made by the individual interviewees, (b) generate codes, (c) identify themes with an explicit connection to the codes, (d) reviewing themes in relation to the research question and providing arguments for their relevancy, as well as (e) defining and naming the themes.
B-3	All of these steps were performed consistently and with precision, and the process was also explicit and transparent.
B-4	In the end, ChatGTP even produced a succinct report on the whole process.
B-5	The human analysis also includes the steps mentioned above, but my impression is that the analysis performed by the ChatGTP is more coherent and precise.
B-6	The themes generated by the ChatGTP provide a better summary of the main points in the interviews.
B-7	An important difference, however, is the use of theory in the human analysis (eg, cognitive load theory), which is (for obvious reasons) not present in the ChatGTP analysis.
B-8	So, while ChatGTP may appear superior when using an inductive approach to thematic analysis, there are still questions regarding a deductive approach and how to place the findings from an inductive analysis in a larger context.
C-1	Similarities and differences were found between the results derived from the thematic analysis conducted by the human researchers and the results derived from the thematic analysis conducted by Chat GPT.
C-2	For example, the theme of "Streamlined assessment process" in the Chat GPT is almost identical to the theme of "Invalid assessments" in the human analysis.
C-3	However, the theme of "Enhanced clarity and specificity" derived from the Chat GPT corresponds to the "Mismatch between the rubrics and clinical context" of the human analysis.
C-4	However, the human analysis is more in-depth in that it even refers to the "clinical context.
C-5	The theme of "Consistent, high-quality feedback" derived by Chat GPT is closest to the theme of "Inhibition of reflection" by the human analysis, but the human analysis is one step deeper.
C-6	Other themes, such as "Learning about competencies" and "Decreased cognitive load resulting from local adaptation," were not observed in the Chat GPT analysis.
C-7	From these comparisons, it can be seen that the major difference is the depth of analysis. Chat GPT may still lack in some areas when it comes to more insightful contextualization of the research questions.

Supplemental glossary

Thematic analysis: Thematic analysis is one of the methodologies for analysing qualitative data. It follows clear and distinct steps, allowing for an easily verifiable analytical process.

Rubric: A rubric is a table that displays assessment scales and criteria, used for evaluating performances such as in reports.