Comparison of AI and Manual Coding of Medical Students' Reflections: An Asynchronous Discussion Board Exploration

Cynthia Kay, MD, MS; Sarah Vepraskas, MD; Hari Paudel, MD; Marika I. Wrzosek, MD, JD; Leslie Ruffalo, PhD, MS; Kathleen Beckmann, MD, MPH; Carley Sauter, MD; Mary Ann Gilligan, MD; Heather Aschoff, BA

ABSTRACT

Introduction: The use of generative artificial intelligence (AI) for medical education research remains controversial. Although AI has the potential to revolutionize qualitative research analysis, concerns regarding its validity and accuracy remain barriers to its use in research. We sought to compare AI-generated qualitative coding with human coding.

Methods: As part of a mandatory course, third-year medical students watched 2 health care-related videos and responded to prompts on a virtual discussion board. Students' posts were anonymized. Human coders met multiple times over a 4-month period to create a code list, generate themes, and produce an explanatory model. Generative AI also was asked to identify the main themes. Three external reviewers with expertise in qualitative research compared each pair of human-coded and AI-generated themes.

Results: Human coders required nine 60-minute meetings to code students' posts, whereas AI completed the analysis in approximately 30 seconds. Consensus among the external reviewers was that the pairs of themes were not the same or similar.

Conclusions: AI required substantially less time and effort than human coders. Although most reviewers concluded that AI-generated themes were not comparable to human-generated themes, this finding was not unanimous. Generative AI may facilitate and accelerate qualitative research in medical education; however, concerns regarding the validity and contextual interpretation of AI-generated analyses must be addressed before its widespread use in qualitative medical education research.

• • •

Author affiliations: Department of Medicine, Clement J. Zablocki VA Medical Center, Milwaukee, Wisconsin (Kay); Department of Medicine, Medical College of Wisconsin (MCW), Milwaukee, Wisconsin (Beckmann, Gilligan, Kay, Paudel); Department of Pediatrics, MCW, Milwaukee, Wisconsin (Beckmann, Vepraskas); Department of Psychiatry and Behavioral Health, MCW, Milwaukee, Wisconsin (Wrzosek); Department of Family and Community Medicine, MCW, Milwaukee, Wisconsin (Ruffalo); Department of Physical Medicine and Rehabilitation, MCW, Milwaukee, Wisconsin (Sauter); Academic Affairs, MCW, Milwaukee, Wisconsin (Aschoff).

Corresponding author: Cynthia Kay, MD, MS, 5000 W National Ave, Milwaukee, WI 53295; email ckay@mcw.edu; ORCID ID 0000-0003-2649-6078

INTRODUCTION

Qualitative studies involving written reflections and interviews can provide details and deeper understanding of participants' perspectives not otherwise attainable through quantitative methods. The analysis of such qualitative studies, however, is generally labor-intensive. Manually coding written posts and transcripts requires human coders and sufficient time for analysis. For most medical educators with a variety of responsibilities, finding time to meet and code qualitative data can be challenging.

Generative artificial intelligence (AI), such as ChatGPT, offers an alternative to manual coding that could change the way qualitative research is done. Generative AI is defined as "deep-learning models that can generate high-quality text, images, and other content based on the data they were trained on."[1] Research on generative AI in medicine and medical education remains limited. As an emerging area, studies on generative AI have tended to focus on patient care-related outcomes.[2-4] Studies in medical education have generally explored student evaluations rather than educational assignments.[4]

Virtual discussion boards have long been used in undergraduate education and other health professions education, such as dentistry.[5] Their use in medical education has been less common, although it increased during the COVID-19 pandemic. Posts on virtual discussion boards became a straightforward way to track and confirm student engagement with required online activities. Students' reflections contain a wealth of information, making written assignments valuable sources of data for research.[6] However,

examining reflections for research is traditionally both time- and labor-intensive.

Generative AI presents an opportunity to analyze qualitative research more efficiently with fewer personnel.[2] Integrating AI into required coursework for research and educational purposes could enhance productivity without adding to the workload. However, the extent to which AI can understand the context of data compared with human coders remains uncertain. The objectives of this study were to compare themes identified from students' reflections by AI and human coders and to evaluate how each performed the analysis.

Table 1. Video Title, Competency Domain, and Duration

Video Title	Domain	Duration
Making Mistakes	Professionalism; personal and professional development	1:37:45
Emotional Intelligence	Interpersonal and communication skills; personal and professional development; professionalism	1:04:25
Advocacy on the Frontlines: Supporting LGBTQ Patients in Combat and Beyond	Practice, interpersonal and communication skills, professionalism	0:55:00[a]
Healing in the Aftermath of Hate	Personal and professional development, professionalism	0:50:00[a]
The Art of Medicine – Patient-Centered Communication Skills	Interpersonal and communication skills; professionalism	1:00:05
Healthcare Remains the Third Leading Cause of Death: Is There Hope for the Future	Systems-based practice; practice-based learning and improvement	0:55:00
Gallows Humor	Professionalism; interpersonal and communication skills	54:58
I Am Not Your Inspiration	Practice-based learning and Improvement; professionalism, systems-based practice	0:09:02

[a]Approximate duration of video

METHODS

Participants

Third-year medical students at the Medical College of Wisconsin (MCW), a private medical school in the United States, who were enrolled in the required third- and fourth-year Continuous Professional Development (CPD) course were invited to participate. Students who volunteered were asked to complete an assignment that involved viewing 2 of 8 health care-related videos of their choice and posting a corresponding reflection on a virtual discussion board. They also were asked to respond to 2 posts by their peers. Students had 1 week to complete the assignment. Students who declined to participate received an alternative assignment. By posting on the discussion board, students consented to participate in the study. The MCW Institutional Review Board reviewed and approved the study (PRO00049191). All 224 eligible students participated and completed the assignment.

Videos

Eight videos encompassing a variety of health care-related topics were available for viewing. All videos except one were recordings of live presentations delivered between 2019 and 2021. The exception was a TED Talk. Four videos were sessions developed specifically for the CPD course, and 3 were presented as MCW or departmental grand rounds. Speakers included physicians, authors, an activist, and a lawyer. Topics, corresponding medical education competency domains, and video duration are shown in Table 1. The videos and topics focused on the art of medicine rather than knowledge-based competencies. Students were asked to address at least 1 prompt related to the videos. The prompts asked them to discuss their thoughts and reactions, the videos' themes, and the impact and value of the content on their clinical experiences and future practice.

Analysis

Four CPD course directors (KB, MG, LR, and CS) monitored the discussion board, including periodically commenting on posts. CPD coordinators anonymized all posts after the discussion boards closed. Given the exploratory nature of this study, the investigators chose to focus on the video with the greatest number of threads, titled "I am Not Your Inspiration," which focused on how individuals with disabilities are perceived. The remaining CPD course directors (CK, HP, SV, and MW) used grounded theory methodology to iteratively code the posts. Grounded theory aims to generate theory from coding patterns rather than starting with predetermined ideas.[7] Posts were categorized by student and by whether they were an original post or a response. Only original student posts were coded. Themes were generated by clustering similar sentiments and then tested and refined by the coders.[8]

The CPD course directors responsible for coding reviewed each original student post. Using a grounded theory approach, the group first performed open coding, creating preliminary codes for each post. Subsequently, axial coding was performed to identify connections and group similar codes. Iterative discussion among the coders was used to finalize the codes and organize them into themes.

Generative AI and Thematic Analysis

Microsoft Copilot for Microsoft 365, an AI-powered assistant licensed by MCW, was used for AI-based thematic analysis. Copilot operates using large language models (LLMs) in conjunction with the account user's Microsoft 365 data and internet-based information sources.[9] A PDF containing anonymized student reflections from the video with the most responses was attached to the prompt. Copilot was also provided a link to the TED Talk, the 5 discussion prompts, and instructions to identify 3 themes and 3 representative quotes for each theme.

Table 2. Themes by Human Coders and AI With Representative Student Quotes

Themes by Human Coders	Illustrative Example	Themes by AI	Illustrative Example
Emotions	"I thought about how at one point in my youth I would have been surprised by this." "I have not really thought about it the way she did and it makes me question my concepts of disabilities." "As future physicians, I believe that it can become easy for us to view the world through a 'clinical' or 'problem-base' lens. Such a lens often distorts the true reality of a situation and may alter the empathy that we feel for a particular situation or individual."	Objectification and "Inspiration Porn"	"I never realized that feeling inspired or motivated by a disabled person completing an ordinary act is objectifying them." "We often praise individuals with disabilities solely for their existence... this shouldn't be the standard." "The idea that disabled people are often seen solely as sources of inspiration or exceptional due to their disabilities... underscores a societal tendency to view disability through a narrow lens."
Future-oriented Actions	"My second half of the school year will be influence by this perspective in a positive way." "This video taught me that moving forward, whether on rotations or in the future as a practice, to be mindful of what I say...." "...impact my practice by encouraging a reevaluation of biases and a move away from viewing disability through a lens of tragedy or inspiration."	Impact on Clinical Practice and Personal Growth	"This talk will remind me that for patients with disabilities, I might need to work harder to ensure people feel safe enough to let walls down and realize they don't need to inspire, they just need to be." "I want to make sure that I continue to uplift my patients and support them without making them feel patronized." "This video really made me reflect on that experience because everyone is just learning to use their bodies to the best of their capacity."
Social construct	"Our medical system is not designed for individuals with disability....how difficult performing a physical exam on a patient would be if you are blind...." "This can be applied to other marginalized communities as well ranging from age, race, gender, socioeconomic status, sexual orientation, or other identifying features." "Perhaps our recognition of acts that should be deemed regular stems from us trying to recognize our sympathy for struggles that we cannot perceive."	The Social Model of Disability and Societal Barriers	"She is more disabled by the social world we live in compared to her body." "No amount of smiling at stairs has made it turn into a ramp." "Disability is not the result of the restrictions that someone's body or illness imposes on them but rather the effects of how society interacts with the capacity of a disabled person's body."

Comparison of Themes

Three faculty members not involved in the study compared each pair of human-coded and AI-generated themes. Two were PhD-trained qualitative researchers, and 1 was a physician with qualitative research experience. All were provided only the table of human-coded and AI-generated themes and representative quotes. They were blinded to the manuscript and study objectives. They were asked to assess each pair of themes and determine whether they were the same, similar, or different and why. This approach was chosen because of the exploratory nature of the study and the lack of validated tools for comparing themes.[10]

Reflexivity

The study idea was developed collaboratively based on several years of experience conducting the discussion board activity. All CPD directors work with medical students in clinical settings and through the CPD course. The coders are clinicians with experience in qualitative research. To engage in reflexivity, coders reflected on their experiences in medical education, interactions with patients with disabilities, and their identities and roles within medicine and society.[11] Coders acknowledged that their faculty position and experiences might influence their interpretations of students' posts. Coders openly shared their perspectives during analysis, explaining their interpretations and gaining a better understanding of one another's viewpoints.

RESULTS

The video with the most responses was "I Am Not Your Inspiration," which focused on how individuals with disabilities are viewed. There were 101 original posts for this video. Human coders identified 46 unique codes, which were categorized by consensus into 3 themes: emotions, future-oriented actions, and social construct. Human- and AI-coded themes with representative quotations for each category are shown in Table 2.

Human-Created Themes

Our 4-member group of physicians and medical educators identified 46 codes after reviewing the posts, which were developed into the following 3 themes:

Emotions: Students often reflected on their past experiences and ways of thinking, eg, "This talk made me take a look at an unconscious bias that I didn't know I had." Others thought more deeply about the video's message and how their mindset may have contributed to a larger issue, eg, "I think a common tendency, especially with strangers, is to look at others for who they are to us, rather than who they are for themselves" and "I feel it's very important to be empathetic towards someone, while also not allowing your preconceived perception of a diagnosis to define them and your expectations of their future or health."

Future-Oriented Actions: Many students looked to the future and considered how the video would influence their interactions with

patients and their practice of medicine, eg, "In my future practice, I want to create a safe space where people don't have to feel like they can't have negative feelings or thoughts." Some were specific about what they might do differently, eg, "…in the future, we can pay extra attention to their clinic experience…to see if there is any way we…can make the experience more like any other patients" and "I hope to continue to utilize this moving forward to make a conscious effort to internalize what she said and share what I learned with those around me."

Social Constructs: Students looked beyond themselves and examined society and its role in what is considered the norm, eg, "It underscores a societal tendency to view disability through a narrow lens…." They recognized their own biases and how inequities contribute to societal constructs. Students identified that certain beliefs and ideas are often reinforced for the benefit of some and to the detriment of others, eg, "True inclusivity involves addressing these challenges rather than tokenizing inspirational stories."

Generative AI-Created Themes

The PDF of anonymized posts from the video was attached to the Copilot prompt that read:

> *"Third year medical students watched the following video (https://www.ted.com/talks/stella_young_i_m_not_your_inspiration_thank_you_very_much/transcript) Anonymous I Am Not Your Inspiration- Original Posts Only 1.pdf and then were asked to respond to the following prompts: Please reflect on this video/session by answering at least one of the following questions: What themes resonated with you from this video and why? How will themes you noticed impact your practice? What was/was not meaningful about this video and why? What are your thoughts about this video? You are now at least several clerkships into your third year – how may have your experience been different if you'd seen this video earlier? Why? Please review the attached 'Anonymous I Am Not Your Inspiration- Original Posts Only. pdf' responses from the students and identify 3 themes with 3 representative quotes for each theme."*

Copilot provided the following response:

> *"After analyzing the student reflections on Stella Young's TED Talk, 'I'm Not Your Inspiration, Thank You Very Much,' 3 major themes emerged consistently across the responses. Each theme is supported by representative quotes from the students' reflections."*

(Representative quotations are shown in Table 2).

Copilot then listed the following 3 themes and a short summary of each:

Theme 1: Objectification and "Inspiration Porn." Many students reflected on how society often objectifies people with disabilities by labeling them as "inspirational" for simply living their lives, a concept Stella Young critiques as "inspiration porn."

Theme 2: The Social Model of Disability and Societal Barriers. Students resonated with Young's emphasis that societal structures, rather than individual impairments, often create the most significant barriers for people with disabilities.

Theme 3: ***Impact on Clinical Practice and Personal Growth.*** Students reflected on how the talk challenged their biases and would influence their future interactions with patients, particularly by encouraging them to avoid assumptions and foster more respectful, individualized care.

Comparison

Consensus was reached, although it was not unanimous. Both PhD faculty members concluded that all pairs of themes were different. Their explanations included a perceived lack of contextual depth and the observation that the AI-generated themes were broad and overlapping. The physician faculty reviewer concluded that all theme pairs were either the same or similar, as did the human coders.

Total Time Spent on Analyses

Total time spent by the human coders included the cumulative time devoted to coding during group meetings. This amounted to nine 60-minute meetings conducted over a 4-month period to complete coding and theme development. Total time spent on analysis using Copilot was less than 30 seconds, including the time required to enter the prompt and generate a response.

DISCUSSION

Generative AI identified themes in a fraction of the time required by human coders. Two of the 3 external reviewers did not find the AI-generated themes to be the same or similar to the human-coded themes. Coding and theme development are inherently subjective processes, which may account for the differences in how AI and human coders labeled themes and selected representative quotations. It is noteworthy that the physician reviewer and human coders, who were also physicians, considered the theme pairs comparable. This finding may illustrate differences in qualitative experience and training between qualitative research faculty and physicians who conduct research.

Qualitative research, particularly interviews and written commentary, can provide deeper insights in participants' own words that may not be attainable through other research means. While educators routinely read and review students' written responses, analyzing those responses for research purposes typically requires additional personnel and substantial time.

Generative AI provides a generalized overview of learners' input. However, evaluating and determining the quality of individual posts remains a persistent challenge.[12] At present, manual review of each learner's assignment is still needed to assess individual content and effort. This limitation remains a clear challenge for generative AI. If the purpose of reflections, discussion posts, and other written assignments is to assess learners' understand-

ing or engagement, manual review appears more appropriate than AI-assisted analysis.

Regarding time commitment, all coders had prior experience with qualitative research and coding; therefore, no additional training time was required. However, coders struggled to find times when all participants were available to code together. Like many clinical educators, each coder balanced varying administrative, educational, and clinical responsibilities, with research conducted during limited discretionary time. From an efficiency standpoint, generative AI does not require coordinating multiple schedules and can analyze qualitative responses in less than a minute. In this study, AI completed a task that took human coders more than 4 months to complete in approximately 20 seconds.

Available literature comparing generative AI with human coders in qualitative research has similarly demonstrated substantial gains in efficiency.[13] In terms of thematic accuracy, one study found fair to moderate agreement between AI-generated and human-generated themes in qualitative analysis of a text-messaging intervention for patients.[14] Nonmedical studies comparing generative AI and human coders have reported similar findings, including efficiency gains but differences in thematic development.[15,16] Others have also noted that AI may struggle to detect emotion and tone.[17] Sarcasm, for example, can be difficult for AI to detect because contextual understanding is important.[18] Cultural differences and slang further complicate these challenges. Admittedly, humans are not perfect at interpreting written tone either.[19] Additionally, some quotations identified by both the human coders and AI in our study could reasonably have fit within multiple themes, a point raised by one of the external reviewers.

The differing conclusions reached by the external PhD reviewers compared with the physician reviewer and human coders serve as an important reminder that professional background can shape qualitative interpretation. Qualitative research experts often emphasize conceptual depth and clear conceptual boundaries. In other words, they carefully consider whether each theme represents a distinct concept and avoids overlap with other themes. Consequently, the PhD reviewers viewed AI-generated themes as superficial and lacking deeper understanding and insight. In contrast, the physician coders evaluated the themes largely on the basis of practical relevance and how the themes resonated with their clinical experiences.

Many qualitative software programs, including NVivo (Lumivero), Dedoose (SocioCultural Research Consultants, LLC), and MAXQDA (VERBI Software), now incorporate AI-enabled functions.Available feature include autocoding, theme suggestion, summarization, and sentiment detection. For researchers considering AI to support qualitative analysis, software with specific AI capabilities is recommended over the use of general AI tools such as ChatGPT and Copilot so that the analysis is integrated and supported by qualitative methodological rigor.

It is also important to recognize that the appropriateness of either by hand or AI coding depends on the purpose of research. For studies requiring deep interpretive analysis, systematic and consistent procedures, clear analytic coherence, and transparent decision-making, human coders remain essential. When the primary goals are efficiency, rapid summarization, content analysis, or preliminary pattern identification, AI may be a useful analytic tool.

Limitations

This was a single-site study; therefore, the generalizability of the findings is uncertain. Additionally, allowing students to select which videos to watch may have introduced selection bias. Social desirability bias also may have been present, as students may post certain responses to create favorable impressions. Theme comparisons were performed only for the video with the largest number of posts. Whether similar results would have been observed for the remaining videos is unknown. Additionally, students may have used generative AI to help create their discussion posts. Finally, comparison of human- and AI-generated themes could have been strengthened by the inclusion of additional external reviewers.

CONCLUSIONS

Our findings suggest that AI-powered assistants can provide unmatched efficiency for qualitative data analysis. However, limitations remain regarding the validity of AI-generated interpretations and the ability to appreciate context, tone, and nuance. Addressing these limitations could transform how qualitative analysis is conducted in medical education research.

Financial disclosures: None declared.

Funding/support: None declared.

Acknowledgments: The authors would like to thank Heather Aschoff and Kate Hohl for formatting and anonymizing the posts. Generative AI was used to review students' posts, produce 3 themes and 3 representative quotations to compare resultant themes and quotations identified from human coders. The opinions expressed in this article are those of the authors and do not necessarily reflect the position of the Department of Veterans Affairs or the US government.

REFERENCES

1. Martineau K. What is generative AI? IBM blog. April 20, 2023. Accessed January 27, 2026. https://research.ibm.com/blog/what-is-generative-AI

2. Prescott MR, Yeager S, Ham L, et al. Comparing the efficacy and efficiency of human and generative AI: qualitative thematic analyses. *JMIR AI*. 2024;3:e54482. doi:10.2196/54482

3. Li KD, Fernandez AM, Schwartz R, et al. Comparing GPT-4 and human researchers in health care data analysis: qualitative description study. *J Med Internet Res*. 2024;26:e56500. doi:10.2196/56500

4. Khan SA, Taiyara J, Zary N, Otaki F. Artificial intelligence in narrative feedback analysis for competency-based medical education: a review. *Stud Health Technol Inform*. 2025;327:1423-1427. doi:10.3233/SHTI250637

5. Donlan P. Use of the online discussion board in health professions education: contributions, challenges, and considerations. *J Contin Educ Health Prof*. 2019;39(2):124-129. doi:10.1097/CEH.0000000000000252

6. Wong A, Trollope-Kumar K. Reflections: an inquiry into medical students' professional identity formation. *Med Educ.* 2014;48(5):489-501. doi:10.1111/medu.12382

7. Glaser B, Strauss A. *Discovery of Grounded Theory: Strategies for Qualitative Research*. Routledge; 2017.

8. Hanson JL, Balmer DF, Giardino AP. Qualitative research methods for medical educators. *Acad Pediatr.* 2011;11(5):375-386. doi:10.1016/j.acap.2011.05.001

9. Ashbaugh, C. Microsoft 365 Copilot overview: transforming collaboration and content creation. Compass365. January 16, 2024. Accessed August 11, 2025. https://compass365.com/microsoft-365-copilot-overview-transforming-collaboration-and-content-creation

10. Licorish SA, Bajpai A, Arora C, Wang F, Tantithamthavorn K. Comparing human and LLM generated code: the jury is still out!. arXiv. Preprint posted online January 28, 2025. Accessed January 27, 2026. https://arxiv.org/html/2501.16857v1

11. Berger R. Now I see it, now I don't: researcher's position and reflexivity in qualitative research. *Qual Res.* 2015;15(2):219-234. doi:10.1177/1468794112468475

12. Hamadi H, Tafili A, Kates FR, Larson SA, Ellison C, Song J. Exploring an innovative approach to enhance discussion board engagement. *TechTrends*. 2023; 67:741-751. doi:10.1007/s11528-023-00850-0

13. Pattyn F. The value of generative AI for qualitative research: a pilot study. *J Data Sci Intell Syst.* 2024; 3(3):184-191. doi:10.47852/bonviewJDSIS42022964

14. Prescott MR, Yeager S, Ham L, et al. Comparing the efficacy and efficiency of human and generative AI: qualitative thematic analyses. *JMIR AI*. 2024;3:e54482. doi:10.2196/54482

15. 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. doi:10.1177/16094069231201504

16. 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*. 2024;42(4):997-1019. doi:10.1177/08944393231220483

17. Frankfeldt V. Beyond the algorithm to emotional communication: why AI lacks the heart of psychoanalysis. *Psychoanal Self Context.* 2025; 20(3):1-23. doi:10.1080/24720038.2025.2483922

18. Frontdesk Team. When machines understand tone: AI's grasp of emotional subtext. Frontdesk blog. September 8, 2024. Accessed June 2, 2025. https://www.myaifrontdesk.com/blogs/when-machines-understand-tone-ai-s-grasp-of-emotional-subtext

19. Winerman L. E-mails and egos. American Psychological Association. February 1, 2006. Accessed October 7, 2026. https://www.apa.org/monitor/feb06/egos

advancing the art & science of medicine in the midwest

WMJ

WMJ (ISSN 2379-3961) is published through a collaboration between The Medical College of Wisconsin and The University of Wisconsin School of Medicine and Public Health. The mission of *WMJ* is to provide an opportunity to publish original research, case reports, review articles, and essays about current medical and public health issues.

© 2026 Board of Regents of the University of Wisconsin System and The Medical College of Wisconsin, Inc.

Visit www.wmjonline.org to learn more.