scientific reports

Check for updates

OPEN

Evaluation of large language models within GenAI in qualitative research

Supriya D. Mehta[1,2,5]✉, Souvik Paul[2,5], Enid Awiti[3], Sophie Young[2], Garazi Zulaika[4], Fredrick O. Otieno[3], Penelope A. Phillips-Howard[4], Linda Mason[4,6] & Runa Bhaumik[2,6]

Large language models (LLMs) perform tasks such as summarizing information and analyzing sentiment to generate meaningful and natural responses. The application of GenAI incorporating LLMs raises potential utilities for conducting qualitative research. Using a qualitative study that assessed the impact of the COVID-19 pandemic on the sexual and reproductive health of adolescent girls and young women (AGYW) in rural western Kenya: our objective was to compare thematic analyses conducted by GenAI using LLM to qualitative analysis conducted by humans, with regards to major themes identified, selection of supportive quotes, and quality of quotes; and secondarily to explore quantitative and qualitative sentiment analysis conducted by the GenAI. We interfaced with GPT-4o through google colaboratory. After inputting the transcripts and pre-processing, we constructed a standardized task prompt. Two investigators independently reviewed the GenAI product using a rubric based on qualitative research standards. When compared to human-derived themes, we did not find disagreement with the sub-themes raised by GenAI, but did not consider some to rise to level of a theme. Performance was low and variable with regards to selection of quotes that were consistent with and strongly supportive of thematic and sentiment analysis. Hallucinations ranged from a single word or phrase change to truncation or combinations of text that led to modified meaning. GenAI identified numerous and relevant biases, primarily related to the underlying training data and its lack of cultural understanding. Few prior studies have directly compared LLM-driven thematic coding with human coding in qualitative analysis, and our study – grounded in qualitative study rigor – allowed for a thorough evaluation. GenAI implemented in GPT-4o was unable to provide a thematic analysis that is indistinguishable from a human analysis. We recommend that it can currently be used as an aid in identifying themes, keywords, and basic narrative, and potentially as a check for human error or bias. However, until it can eliminate hallucinations, provide better contextual understanding of quotes and undertake a deeper scrutiny of data, it is not reliable or sophisticated enough to undertake a rigorous thematic analysis equal in quality to experienced qualitative researchers.

Generative artificial intelligence (GenAI) within ChatGPT launched in 2022, and an explosion in research application followed. In a PubMed search of "generative AI" in Title or Abstract fields, just 7 occurred prior to 2023, 552 in July 2024, and 825 as of July 2025GenAI can process large amounts of information or data and generate new information – whether as text, images, or audio – that potentially offers rapid and substantial advances in understanding and using data. The goal of using GenAI is to "provide output that is indistinguishable from that of a human", or if possible to exceed it[1]. Thus, it is no surprise that its availability has been seized upon by the research community. A systematic review and meta-analysis found that GenAI in health related research was used most frequently for diagnostic and screening processes (e.g., improving accuracy) in relation to diabetes, radiology, cardiology, and gastrointestinal medicine[2], and its performance compared to physicians has also been assessed in meta-analysis, showing variable performance by physician expertise and specialty[3].

Large language models (LLMs), as implemented within GPT-4o[4], are advanced natural language programming (NLP) systems trained on huge amounts of text. These models use deep learning, particularly transformer architectures[5], to perform tasks such as writing text, translating languages, summarizing information, and analyzing sentiment. By recognizing patterns in language, LLMs can generate meaningful and natural responses,

[1]Division of Infectious Diseases, Department of Medicine, Rush University College of Medicine, Chicago, USA. [2]Division of Epidemiology & Biostatistics, University of Illinois Chicago School of Public Health, Chicago, USA. [3]Nyanza Reproductive Health Society, Kisumu, Kenya. [4]Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK. [5]Supriya D. Mehta and Souvik Paul contributed equally to this work. [6]Linda Mason and Runa Bhaumik jointly supervised this work. ✉email: Supriyad@uic.edu

making them useful for chatbots, content creation, and customer service. In short, LLMs make it easier for humans to communicate with technology more naturally. The application of GenAI incorporating LLMs raises potential utilities for undertaking analysis in qualitative research. The aim of qualitative research is a naturalistic enquiry to understand perceptions, feelings, thoughts or behaviors as experienced by study participants themselves, without attempting to broadly generalize to a wider context as occurs with quantitative research. Rigor and credibility of qualitative study is ascertained by its trustworthiness – with underlying concepts of confirmability, reflexivity, and transferability. This requires reflection and transparency in order to negate or acknowledge any bias, including that of the analysts' own thoughts and actions[6]. Of the different methods of analysis, thematic analysis appears particularly appropriate for use of AI as it offers a systematic and robust framework for analyzing qualitative data which makes it particularly useful in applied health research[7].

Currently, thematic analysis is a lengthy and labor intensive activity. The phases of thematic analysis involve familiarization with the data (transcription, reading, re-reading); systematically generating initial codes; searching for or deducing themes and subthemes by collating codes; reviewing themes; defining and naming themes; and reviewing transcripts to ensure codes and themes are appropriate to the context and represent the range of participant views. The narrative is then produced, with illustrative quotes added to support the text.[7] GenAI may conduct such thematic coding, with time efficiency and without specific biases[8], but may lack nuanced understanding of the data, and an ability to make sense of the complex and often irrational thoughts or behaviors of humans, as well inability for reflexivity on the processes ensuring rigor. Studies have emerged evaluating the use of GenAI in qualitative research. For example, a study compared LLM for inductive thematic analysis of public health-related content from social media posts compared to human thematic analysis, and found that LLM identified several themes consistent with those identified by humans, and that additional themes were relevant and reasonable[9]. Other analyses have been limited to medical records or policy documents[8], or lack formal comparison to human analysis[10]. However, these qualitative analyses involve relatively short or simple texts, as compared to qualitative research studies, such as focus group discussions, which involve simultaneous and interactive conversation among multiple people. As summarized by Lee et al., several studies have applied GenAI to various stages of thematic analysis in qualitative research with varying results, identifying challenges such as prompt-dependency, hallucination (i.e., "made up" information), repetition of output, word and syntax errors, and requirement for human assistance to refine codes and themes[11].

GPT-4o offers significant advances over prior versions, including expanded coherence, ability to generate richer explanations, greater capacity for ambiguous queries and multiple interpretations, a broader knowledge base, and improved recognition of user intent[12]. Given the expanded utilities and building on prior studies, we applied LLM to a qualitative study and evaluated how well it conducted thematic analysis: (1) compared to a formal human analysis, and (2) through a replicable framework to assess the quality of the GenAI output. Our test case makes use of a qualitative study that assessed the impact of the COVID-19 pandemic on the sexual and reproductive health of adolescent girls and young women (AGYW) in rural western Kenya[13]. The main objective of the present study was to compare the thematic analyses conducted by GenAI using LLM to that conducted by humans, with regards to major themes identified, selection of supportive quotes, and quality of quotes; and secondarily to explore quantitative and qualitative sentiment analysis conducted by the GenAI, as a novel potential adjunct to traditional qualitative thematic analysis. The process and results of this study present a replicable framework to evaluate the potential complementarity, augmentation, and errors or bias that may arise when using GenAI to support qualitative study.

Methods

This study was approved by the institutional review boards of Maseno University Ethics Review Committee (MUERC, MSU/DRPI/MUERC/01021/21), University of Illinois at Chicago (UIC, #2022 – 0220), and (Liverpool School of Tropical Medicine (LSTM, #21–087, favourable ethical opinion). Written informed consent was obtained for all participants. All research was performed in accordance with relevant guidelines/regulations. Per OpenAI's safety protocols, they actively monitor for potential misuse and safeguard user data by declaring input is not accessible or shared to anyone else, or the OpenAI team[14]. Complete annotated code is available in Supplementary File 1.

Study design and participants

Data for this analysis came from the Cups and Community Health (CaCHe) study[15,16], which involved a sub-set of participants within the Cups or Cash for Girls (CCG) trial (ClinicalTrials.gov NCT03051789). The CCG trial has been described in detail[17]. Briefly, CCG was a cluster-randomized controlled trial in which secondary schools in western Kenya were randomized into 4 arms (1:1:1:1): (1) provision of menstrual cups with training on safe cup use and care; (2) conditional cash transfer (CCT) based on ≥ 80% school attendance in the previous term; (3) menstrual cup and CCT; and (4) control (usual practice). For the CaCHe study, we enrolled approximately 20% of girls in the cup only and control arms of the CCG trial. After enrollment, CaCHe participants were followed every 6 months, with the 72-month study visit completed in July 2024. The 24-month study visit window (May through June 2020) coincided with the COVID-19 pandemic and was missed. At the 48-month study visit (April 5 – July 1, 2022), following observed increases in STIs and pregnancy[18,19], we initiated a qualitative study to understand the impact of the COVID-19 pandemic on the sexual and reproductive health of AGYW in the study.

Methods of qualitative study and thematic analysis

We conducted facilitator-led FGDs to explore this topic for its ability to elicit rich data for complex and nuanced topics. In this approach, a structured interview guide was used, with open-ended questions allowing flexibility in discussion flow and ability to explore participants' meanings and interpretations. The structured interview

guides and human-led thematic analysis methods used in this study are published[13]. Briefly, semi-structured FGD guides with overlapping themes were applied for AGYW and males. FGDs lasted 1.5–2 h, and were carried out in English, Swahili, Dholuo or a mixture, with the Kenyan moderator (EA) being fluent in all three. FGDs were audio recorded with participant permission and transcribed verbatim, with Dholuo and Swahili being translated to English. The moderator (EA) compared all transcripts against the original audio for accuracy. For the translated transcripts, EA conducted a quality check on 10% of the content to verify consistency and correctness. General patterns that emerged were identified and noted as initial themes; line by line narratives were ascribed detailed codes, which were assigned to the themes and built into a framework. The framework was edited dynamically until all transcripts were coded and a series of subthemes and themes identified. Further analysis was undertaken to compare opinions, behaviours, language across and within AGYW and community men transcripts. EA and SY coded and analysed the data independently, and then compared results, discussing and resolving with SDM and LM where any disagreement occurred.

GenAI: thematic analysis

We used GPT-4o (November 2023 version) to conduct thematic analyses (i.e., identification of key attributes) separately for transcripts from AGYW and community males. While platforms such as Google's Gemini and Meta AI exist, ChatGPT remains the more widely utilized tool for research[20–23]. Although Gemini Pro has been tested in a limited number of studies, its use has primarily been for comparison against ChatGPT[24–26]. No such research papers have reported utilizing Meta AI for similar purposes.

We interfaced with GPT-4o through Google Colaboratory[27]. We chose to access GPT-4o via the OpenAI API in a Google Colaboratory notebook rather than the ChatGPT web interface for data safety. In this environment, all data exchange occurs over our university's encrypted pipeline: credentials and user inputs never pass through a third-party GUI. Using Colab allowed us to script and version control every step of the analysis (from prompt templates to post processing), which greatly enhances reproducibility. The API driven approach gave us programmatic control over batching, rate limits, and logging, enabling a consistent, automated workflow that would be difficult to achieve via the interactive ChatGPT app.

Our primary task involved prompting the models to generate themes when given the transcripts.There were 6 transcripts from 54 girls and 5 transcripts from 53 men. Each transcript consisted of moderator's questions and participants' answers. We only used the English translation of participants' responses. We pre-processed transcripts in python before providing the texts as input to GPT-4o. We were able to input all the data at once, separately for AGYW and men. The final input size was 41,280 words for AGYW and 59,985 words for men. After inputting the transcripts and pre-processing, we constructed a standardized prompt for this task. The prompt used in this study was structured in plain text and consisted of three main sections (Supplemental Table 1). The first section contains all the FGD questions (verbatim) directed towards male or female participants. This was done so that GPT-4o would generate themes parallel to those being sought by humans. No pre-existing themes were provided to GPT-4o. After GPT-4o generated an exhaustive set of themes, we posed specific questions to it, where we provided two primary themes along with their sub-themes. Lastly, we included detailed instructions for the AI to generate outputs containing key words, descriptions, and relevant supporting quotes. The specific wording of the task prompts was based on best practices in AI prompt engineering[28]: being specific and clear in requesting the desired format (e.g., using key words, asking for direct and exact quotes from the original text), inclusion of specific constraints (e.g., provide 3 supportive quotes), iteratively refining prompts.

To standardize our LLM Application Programming Interface (API) calls, all models were set to a temperature of 0.7 and 4000 maximum output tokens. We experimented with different values of the "temperature" parameter, which can range from 0 to 2. In OpenAI's models, temperature controls the randomness of the AI's output[29–31]. A temperature of 0 results in highly deterministic and repetitive responses, while a temperature of 2 encourages more creative and diverse outputs. In our case, the model did not function properly at temperatures of 1.5 and above (i.e., providing overly fanciful and nonsensical results), but worked at 1.4 and below. We tested a range of values from 0.7 to 1.2 (1.0 is the default value). While the AI provided some well-structured and coherent answers at higher temperatures, the quotes it generated were often altered from the original text, making them difficult to trace back to the source. After reviewing the results, we fixed a temperature setting of 0.7 to ensure that AI-selected quotes came from the original text and maintained a precise connection with the accompanying descriptions.

We repeated the thematic analysis, recording the subthemes with each iteration (Supplemental Table 2), until no new themes appeared. We quantified the stability of the subthemes across different iterations using BERTScore F1. The BERTScore F1 ranges between 0 to 1[32–35], and measures the similarity between two texts based on their semantic meaning. We obtained BERTScores F1 in the range of 0.89–0.99, indicating a high degree of consistency, implying that the AI produced stable outputs over repeated trials. Once an exhaustive list of subthemes was obtained for AGYW and for community males, GenAI was asked to provide an overview of each theme and identify three supporting quotes for each of the themes.

Evaluation of thematic analysis

The exhaustive list of subthemes was taken as the final product from the GenAI and was evaluated by two investigators (EA, SY), who had previously coded the transcripts manually. They independently reviewed this product from the AI using a rubric (Supplemental Table 3). We developed the rubric for this study with attention to consolidated criteria for reporting qualitative research (COREQ)[36], recommendations for standardized reporting of qualitative research[37], and the Critical Appraisal Skills Program (CASP) checklist for reporting qualitative studies[38]. These standards emphasize that qualitative research with high quality should demonstrate critical reflection, with key criteria of: credibility (plausible and trustworthy findings that align between theory, the research question, data collection, and results); confirmability (clear link between the data and the

findings, such as through use of quotes), and reflexivity (self-reflection regarding subjectivity and influence on the research). Given that GenAI had no role in the design of the qualitative study, hypothesis generation, or development and implementation of the interview guides, we evaluated its work primarily on confirmability. Therefore, the three metrics we selected for evaluation were: (1) how well the themes were described (not at all, partially, completely), (2) for each supportive quote selected, whether it was consistent with and supportive of the theme (no, yes), (3) for each supportive quote selected, the extent to which it was consistent with and supportive of the theme (not at all, partially, completely). Evaluators were asked to provide comments where themes or supportive quotes were not completely explained or consistent. Interrater reliability of the evaluators is reported as percent agreement, Cohen's kappa, and Gwet's AC1.

Exploratory sentiment analysis

Analysis conduct

Sentiment analysis in text analysis involves identifying and categorizing the emotional tone or sentiment expressed in a text, typically as positive, negative, or neutral. Several lexicon-based tools exist, such as VADER[39], NRC[40], TextBlob[41]. Lexicon-based approaches rely on a pre-defined dictionary of words labeled with sentiment (positive, negative, or neutral). Sentiment is derived by identifying words in the text that match the lexicon and assigning a score based on the lexicon entries. GPT-4o is a transformer-based language model trained on massive datasets and can generate human-like text, and uses deep neural network analysis to generate sentiment, based on each entire sentences structure, context, and relationship between words. Based on this capability, we tasked it to conduct sentiment analysis of our transcripts.

We provided GPT-4o API with a predefined set of sentiment/emotion categories for the analysis. For sentiment analysis, we first applied word tokenization to the transcripts, breaking each participant's response into a vector of tokens (i.e., numerical representations of words). GPT-4o then mapped this vector to patterns in its training model and data, to predict sentiment. After having sentiments for each participant response, we requested the percentage distribution of sentiments. Additionally, we instructed the AI to extract and highlight relevant keywords and provide supporting quotes from the text.

Two different types of sentiment analysis were conducted: (1) The evaluation of transcript tokens categorized as "Very negative", "Negative", "Neutral", "Positive", and "Very Positive" from VADER lexicon. VADER (valence aware dictionary and sEntiment reasoner) uses a list of words to determine positive or negative valence. Originally designed for shorter pieces of text, it is commonly used for sentiment analysis of social media[42]. (2) Evaluation of more detailed feelings or emotions, including "fear", "anger", "trust", "surprise", "joy", "disgust", "sadness", "positive", and "negative" was adapted from the Circumplex Model based on subjective feelings[43]. The circumplex encompasses numerous emotional states, but there is evidence that certain states are more reliably recognized and assessed, such as sadness and anger, representing both the arousal and deactivation quadrants of negative valence, or surprise and joy for positive valence[44]. Because we did not know a priori the specific sentiments most likely to be expressed in the transcripts, we chose a range of negative sentiments covering both arousal (fear, anger, disgust) and deactivation (sadness) of negative valence, neutral valence sentiment (surprise), and positive valence sentiments (trust, joy), and included "positive" and "negative" as "catchalls" for other states.

Initially we attempted sentiment analysis within sub-themes, but due to excessive repetition of supporting quotes within and across sub-themes we abandoned this exercise. Sentiment analysis was then conducted separately for each of the two objectives, which also produced repetitious results. Therefore, as the depth of data appeared insufficient for these approaches, we proceeded with an overall sentiment analysis, stratified by female and male inputs. Within each analysis, GPT-4o was asked to provide three supportive quotes for each sentiment. BERT F1 scores comparing different sets of transcripts ranged from 0.88 to 0.92, indicating a high level of agreement and reliability between the models. After reliability analyses, we proceeded to evaluate the sentiment analysis based on three randomly selected analyses each for female and male transcripts.

Human evaluation of sentiment analysis

Each of the six sentiment analysis results (3 AGYW FGDs, 3 male FGDs) were reviewed independently by the two reviewers (EA, SY). Reviewers were asked to rate how much they agreed or disagreed with the key indicator words chosen for the sentiment analysis description with a score ranging 1–5: "Completely Disagree", "Somewhat Disagree", "Neither agree or disagree", Somewhat Agree", "Completely Agree". We used a different scale than for thematic analysis given the more subjective nature of sentiment analysis. For each of the three supporting quotes provided for each sentiment, reviewers were asked to determine whether the quote was consistent with or supportive of the sentiment (yes/no), and for each quote, to rate how strongly supportive the quote was (0–2): "Not at all", "Somewhat", "Very supportive". Lastly, stratified by sentiment, we report the percentage of quotes that were consistent with sentiments and a mean score of strength of support.

Bias analysis. GenAI was asked separately for male and female transcripts with three repetitions: "As an AI, what potential biases might you have in conducting this analysis?" We grouped biases into selection biases (representativeness of the study sample/transcripts, representativeness of quotes) and information biases (how quotes were interpreted in relation to generation of themes).

Results

Among 54 AGYW participating in the FGDs, mean age was 20.9 years, few (11%) were employed, with 44% currently in school (Table 1). Among 53 community men participating in the FGDs, mean age was 27.1 years, nearly half (47%) were married, and all were employed.

Participant characteristics	AGYW, *N*=54 *n* (%)	Community men, *N*=53 *n* (%)
Mean age in years (range)	20.9 (18 – 24)	27.1 (19 – 41)
Married	4 (7%)	25 (47%)
Has children	13 (24%)	29 (55%)
Currently in school	24 (44%)	1 (2%)
Employed	6 (11%)	53 (100%)
Occupation (not mutually exclusive, some men had multiple jobs/occupations; AGYW employment was defined as "outside of home chores")	2 (4%) Farming	31 (58%) Motorcycle taxi driver
	1 (2%) Hotel work	13 (25%) Fisherman
	1 (2%) Salon	8 (15%) Miner
	2 (4%) Shop attendant	4 (8%) Mechanic
		3 (6%) Businessman
		2 (4%) Farmer
		1 (2%) Artisan

Table 1. Distribution of FGD participant characteristics.

Summary of human manually coded themes and GenAI coded themes

GPT-4o identified 13 themes from AGYW transcripts, taking 7 repetitions to reach exhaustion (i.e., no new themes emerging), and 11 themes from community male transcripts, taking 10 repetitions to reach exhaustion. The final list of GenAI themes are presented alongside human derived themes in Table 2; results from each repetition are in (Supplemental Table 2). Some themes may be considered closely related, but we decided not to apply too much amalgamation, since we could not have a "back and forth" discussion with GenAI as human investigators would have done. For example, when we would suggest that two themes overlapped and could they be merged into one? (i.e., posed as a question), GenAI always capitulated (e.g., "yes you're correct"; "I see your point"), rather than providing explanation for why they were initially identified as separate themes.

On the surface, human-derived and GenAI-derived themes were somewhat different; however, this stemmed from organization: in our analysis, we grouped sub-themes by themes that were similar for AGYW and male stakeholders. For example, one of our major themes was the *impact of the COVID-19 pandemic on sexual behaviors* (increased frequency and number of partners), with *drivers of increased sexual activity* as a second theme, including a sub-theme of poverty and economic insecurity; in our manuscript, discussion around this encompassed the economic dependency, and transactional exploitative nature[13]. Conversely, GenAI noted the increased pressure for sexual relationships and exploitative relationships as a sub-theme to impacts of the pandemic on men's attitudes and relationships with girls or schoolgirls. Increases in pregnancy (as a result of increased sexual behaviors) was also noted in our analysis as a theme, whereas GenAI indicated increased pregnancy rates as a sub-theme of the impact of the pandemic on schoolgirls. Similarly, increased domestic responsibilities are noted in our analyses as a result of school closures, but were not identified as a sub-theme by GenAI. Overall, we did not find disagreement with the sub-themes raised by GenAI, but did not consider some to rise to level of a theme, identifying them more as context or explanation to a theme or sub-theme.

Results of GenAI thematic analysis and human evaluation

Across all AGYW and male thematic analysis repetitions, there was 100% agreement between the two raters that the themes were completely described and explained, except in one set for males. However, as shown in Fig. 1 (Panel A), performance was low and variable with regards to selection of quotes that were consistent with themes. Both evaluators classified quotes from male thematic analysis as less frequently being consistent with the thematic description, ranging from 33 to 79% of quotes, as compared to 64–87% of quotes from AGYW. GenAI also performed poorly when assessed on the quality of the selected quotes, with AGYW quotes supporting themes 36–67% and male quotes 18–55% of the time. Additionally, several hallucinations were noted (Supplemental Table 4). In some instances, a single word or phrase was changed; in more labile hallucinations, there were examples of combining two quotes, and/or substantial modification (see examples, Fig. 2).

Results of sentiment analysis and human evaluation

The distribution of sentiments identified by GPT-4o for AGYW and community male transcripts is shown in (Fig. 3). Sentiment analysis of AGYW transcripts revealed 50% of sentiments were very negative (20%) or negative (30%), compared to 25% of sentiments being classified as very positive (10%) or positive (15%). Men's sentiments were also more commonly classified as negative than positive. The distribution of the more emotive sentiments were similarly distributed for AGYW and males, though AGYW transcripts were somewhat more frequently classified as expressing "sadness" (15%) compared to men's transcripts (5%).

Human evaluation of GenAI sentiment analysis

Evaluators were largely in agreement that the keywords or descriptions used to define the sentiments (Supplemental Table 5) were consistent (Fig. 1, Panel B). For sentiments ranging from "very negative" to "very positive", evaluators mostly agreed that quotes selected were consistent with the ascribed sentiment for AGYW transcripts (74–100%), and generally indicated that quotes were strongly supportive (87–100%, with

Human: themes, sub-themes	GenAI
In relation to adolescent girls and young women's transcripts	
Impacts on schooling	*Impact of COVID-19 pandemic on schoolgirls*
Uncertainty in returning to education	(1) Education disruption/increased dropout rates/reduced study time and academic performance
Loss of motivation to study	(2) Increased pregnancy rates
Sexual behaviors and practices	(3) Economic hardships/economic hardship and increased workload
Increased number and frequency of sexual partners	(4) Mental health and social issues/psychological stress and anxiety
Using or taking advantage of men	(5) Increased domestic responsibilities
Power imbalance	*Impact of COVID-19 pandemic on men's attitudes and relationships with girls or schoolgirls*
Drivers of increased sexual activity	(1) Increased pressure for sexual relationships/exploitative relationships
Poverty/economic insecurity	(2) Economic exploitation
Opportunity and leisure time	(3) Mixed reactions and attitudes
Peer influence and pressure	(4) Relationships with older men
Increase in pregnancy	(5) Financial dependency and exchange/relationships for financial gain/ transactional nature of relationships
Impact of unintended pregnancy	(6) Breakdown of relationships/reduction in supportive relationships
Social and school impacts, responsibility, blame, consequences	(7) Older men as partners
Rise in abortion due to unintended pregnancy	(8) Exploitation and abuse Breakdown of relationships/reduction in supportive relationships
Knowledge, attitudes, perception, access to family planning, contraception and condoms	
In relation to community males transcripts	
Impacts on schooling	*Impact of COVID-19 pandemic on schoolgirls*
Uncertainty in returning to education	(1) Increase in teenage pregnancies
Loss of motivation to study	(2) Increased reliance on men for financial support/economic hardships and dependence on men
Sexual behaviors and practices	(3) Disruption of education/school dropouts
Increased number and frequency of sexual partners	(4) Increased vulnerability to exploitation
Using or taking advantage of men	(5) Mental health and stress
Power imbalance	*Impact of COVID-19 pandemic on men's attitudes and relationships with girls or schoolgirls*
Drivers of increased sexual activity	(1) Exploitation and sexual relationships/preference for younger, school-going girls
Poverty/economic insecurity	(2) Negative attitudes and blame towards girls/general indifference and blame-shifting/justification of exploitative behavior/attitudes of condemnation and justification
Opportunity and leisure time	(3) Predatory behavior
Rise in pregnancy	(4) Protective yet conditional support
Impact of unintended pregnancy	(5) Blame on external factors
Social and school impacts, responsibility, blame, consequences	
Rise in abortion due to unintended pregnancy	(6) Varied intentions and mixed feelings
Knowledge, attitudes, perception, access to family planning, contraception and condoms	

Table 2. Summary of human manually coded themes and GPT 4o coded sub-themes.

one set of quotes excepting at 53%). Overall, evaluators determined that GPT-4o performed poorly at selecting quotes that were consistent with (53–73%) and strongly supportive (33–67%) of the sentiments ascribed to male transcripts. When examining more complex emotions, evaluators determined that GPT-4o described the sentiments well, but there was variable and lower performance in selection of supportive quotes, and selection of strongly supportive quotes. For example, in one repetition of a male sentiment analysis, the evaluators were in agreement with descriptive basis provided by GenAI regarding the "Positive" classification: sentiments that express optimism, hope, or positive outcomes. However, one of the three supportive quotes provided was "The sponsorship we talked about earlier, he will provide while expecting something in return and that makes it not a better one", which was deemed not supportive of positive sentiment. As another example, GenAI described "surprise" as sentiments that express astonishment, shock, or unexpected outcomes, which evaluators agreed with. Yet a quote deemed non-supportive of this sentiment was "I was in school". In contrast, a quote evaluators found highly supportive of "surprise", included "For me when corona came it really shocked me, it took me around four months to believe that it was with us".

Examining GPT-4o performance stratified by sentiment (Table 3) showed fairly consistent performance for AGYW, in that most quotes were consistent with and strongly supportive of the specified sentiments, except for poor performance on selecting quotes related to "Disgust". For male transcripts, GPT-4o appeared to perform acceptably (mean score ≥ 5, and $\geq 80\%$ of selected quotes being strongly supportive) only for "Fear" and "Negative" sentiments.

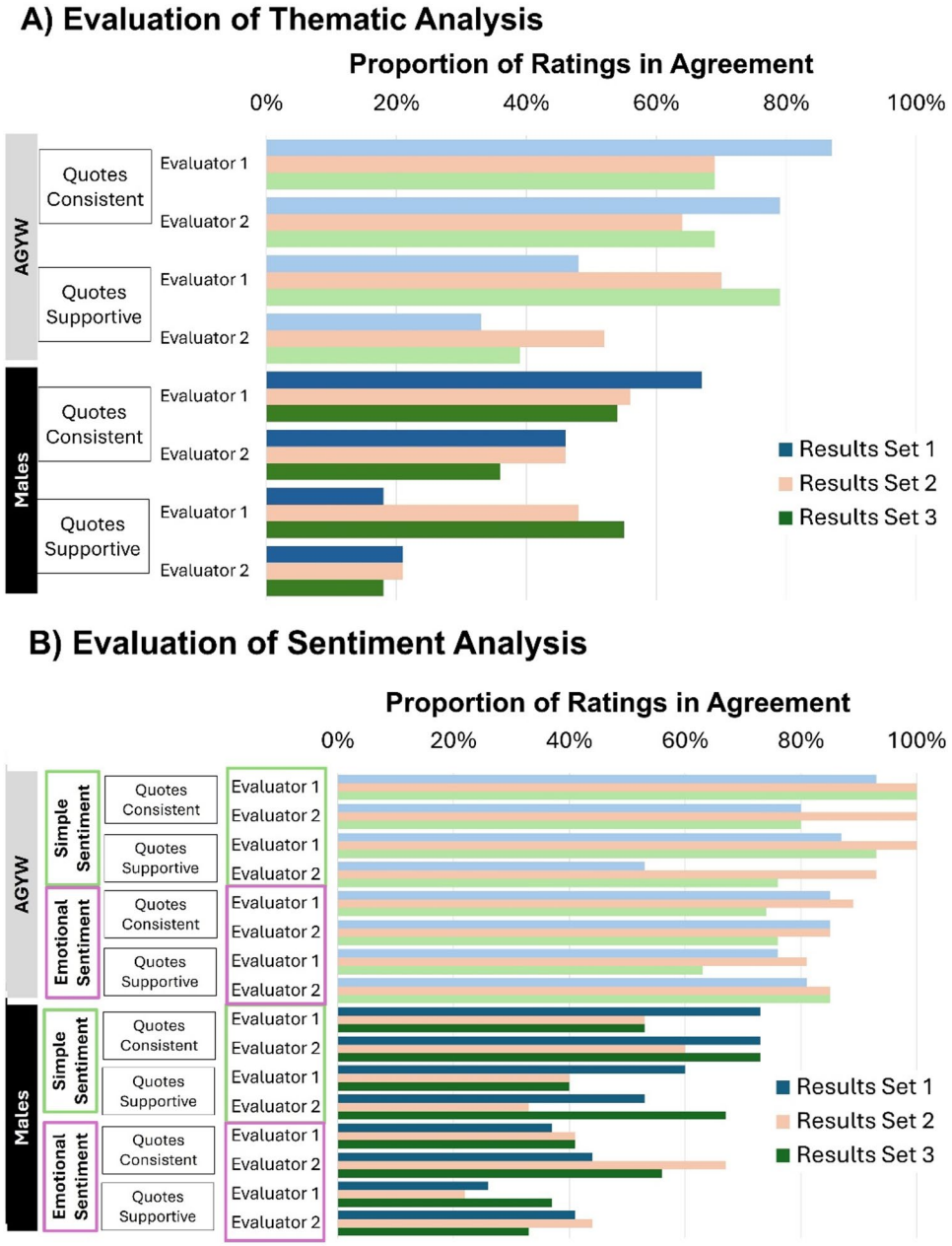

Fig. 1. Evaluation of GPT-4o thematic and sentiment analyses by two independent evaluators. (**A**) Evaluation of thematic analysis. For transcripts from adolescent girls and young women (AGYW), GPT-4o generated 13 themes and selected 3 quotes per theme, resulting in 39 evaluation points for each Results Set. For transcripts from community males, GPT-4o generated 11 themes and selected 3 quotes per theme, resulting in 33 evaluation points per Results Set. The bars represent proportion (x-axis) of GPT-4o selected quotes that evaluators rated as (1) consistent with/supportive of the theme for each of the 3 results sets (light grey, medium grey, dark grey), and (2) the proportion of quotes that were rated completely consistent with/supportive. (**B**) Evaluation of Sentiment Analysis. For transcripts from AGYW and community males, GPT-4o selected 3 quotes for each of 5 sentiments very negative to very positive, resulting in 15 evaluation points per results set. For VADER emotional sentiments (Anger, Disgust, Fear, Negative, Sadness, Joy, Positive, Surprise, Trust) GPT-4o selected 3 quotes per sentiment, resulting in 27 evaluation points per results set.

[1]The consistency of quotes were rated as yes (1 point) or no (0 points). The range for consistency of scores is 0 (across two raters, none of the 3 provided quotes were consistent with the sentiment) to 6 (across two raters, each of the 3 provided quotes were consistent with the sentiment).

[1]The strength of quotes was rated as not at all (0), somewhat (1), or strongly supportive (2). The range for strongly supportive scores is 0 (across two raters, none of the 3 provided quotes were strongly supportive of the sentiment) to 12 (across two raters, each of the 3 provided quotes were strongly supportive the sentiment).

GenAI Quote	Original Transcript
"Men were under more pressure to have sex because if her boyfriend asked her to at least go pay him a visit and spend some time together and the lady was held up, the boy could threaten dumping the lady." (under the theme *Increased pressure for sexual relationships*) -- Modified and Truncated	"I just to say that they were under pressure because if her boyfriend asked her to at least go pay him a visit and spend some time together and the lady was held up, the boy could threaten dumping the lady and because they were vulnerable, they would find themselves going."
"Some girls due to peer pressure, were forced to have sex for money. As in, their families didn't have enough income." (under the theme *Economic Exploitation*) --Modified and Truncated	"Due to peer pressure, most girls were being forced to have sex for money. As in, their families did not have enough income so they couldn't get most of their basic needs like clothing."

Fig. 2. Sample hallucinations involving modification of original transcript.

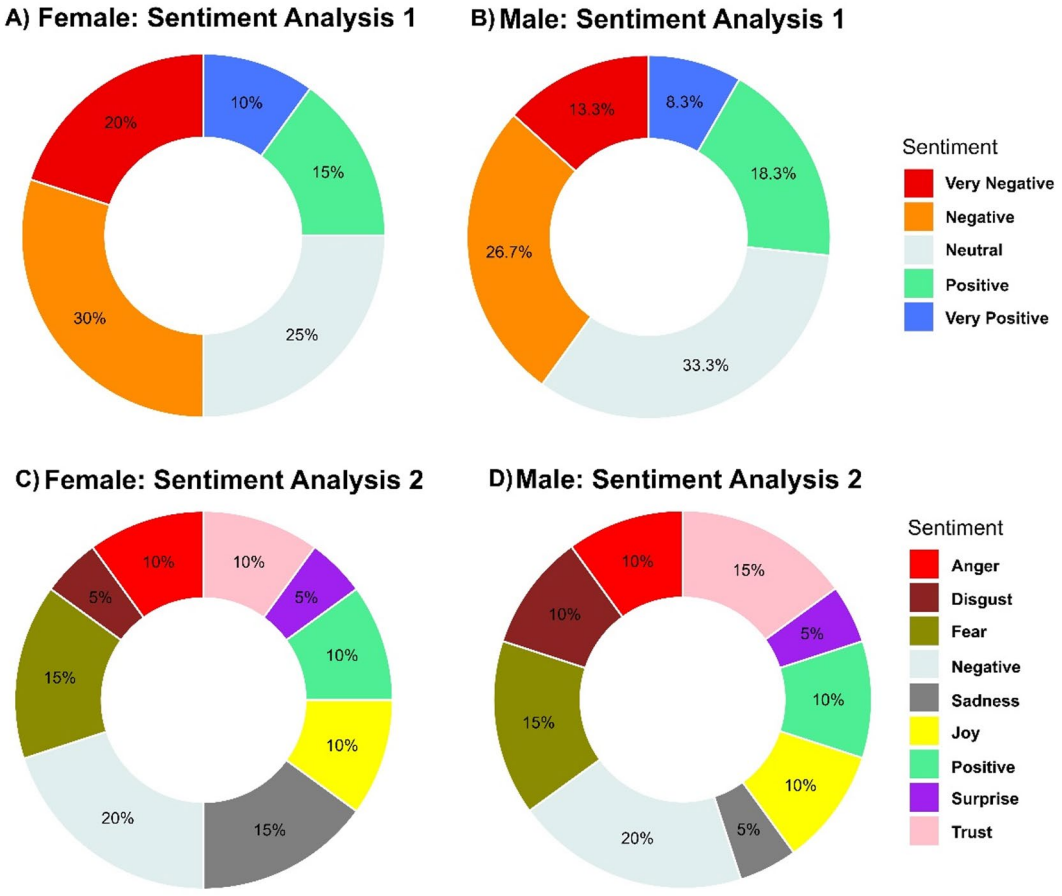

Fig. 3. GenAI sentiment analysis from qualitative study about COVID-19 impacts on AGYW sexual and reproductive health. Sentiment analysis 1 reports the frequency distribution of sentiments ranging "Very Negative" to "Very Positive" for (**A**) Adolescent girls and young women (AGYW), and (**B**) Community males. In sentiment analysis 2, panels represent the frequency distribution of specific emotions listed in the key for (**C**) AGYW and (**D**) Community males.

Sentiment	AGYW, mean scores		Male, mean scores	
	Quotes are consistent[1]	Quotes are strongly supportive[2]	Quotes are consistent[1]	Quotes are strongly supportive[2]
Anger	5.7	9.3	2.7	3.7
Disgust	0.33	0.33	1	2
Fear	5.3	10.3	5	10
Joy	6	12	2	2.7
Negative	6	12	5.7	8.3
Positive	6	12	3.7	6.7
Sadness	5.7	11	0	0
Surprise	4.7	10	3	5.3
Trust	5.3	10.7	2.7	5.3

Table 3. Mean scores of evaluation of GPT-4o GenAI selection of consistent and strongly supportive quotes by specific sentiment.

A higher mean score indicates greater agreement that quotes are consistent or strongly supportive of the specified sentiment.

Rubric performance: interrater reliability of evaluators

Agreement between raters was higher for evaluations related to AGYW than males, and higher for thematic analysis than sentiment analysis (Supplemental Table 6). Agreement was excellent for how well themes were explained; substantial (AGYW) or fair (males) for consistent quote selection; and fair for how strongly supportive quotes were. Agreement was fair-moderate for male sentiment analysis based on circumplex, and was substantial-excellent for AGYW circumplex-based sentiment analysis. There was a similar pattern of agreement for the VADER-based sentiment analyses.

Identifying biases

Several biases were termed differently (e.g., Language and Context Bias vs. Language and Interpretation), but had similar explanations and were grouped together (Table 4). Regarding selection bias with regards to representativeness, GenAI relayed that results may not be generalizable and may not include other important perspectives in how the COVID-19 pandemic affected schoolgirls, in one instance highlighting the potentially important input of parents and teachers. GenAI identified numerous information biases, primarily related to the underlying training data (leading to confirmation bias and potential exclusion of contradictory or more nuanced findings), its lack of cultural understanding (social dynamics and linguistics; in one instance highlighting its Western-biased training data).

Discussion

We conducted formal comparative analysis of GenAI to human thematic coding of qualitative focus group discussion study exploring the impacts of the COVID-19 pandemic on sexual and reproductive health of AGYW, with AGYW and male participants in Kenya. We found that GenAI did an adequate job of thematic analysis. However, selected quotes were unreliable, even after initial feedback, and were subject to hallucination. We often disagreed with the GenAI's performance of sentiment analysis, with general dissent regarding supportive quote selection. In both thematic and sentiment analyses, GenAI's performance was rated more poorly for transcripts generated by male participants. We examine these findings with critical and contextual analysis and recommendations below.

GenAI was tasked with undertaking an inductive approach to thematic analysis in order to compare the output to our human analysis. As stated, it performed adequately in identifying similar themes and subthemes which we identified. Reassuringly, GenAI did not identify themes that did not emerge from our analysis; rather the discrepancies between human evaluation and GenAI resulted from differences in the level of importance that was assigned to them, i.e., as main theme, subtheme, or explanatory theme. The themes and subthemes we identified, and which formed the comparator for this study, were a product of repeated and time-consuming analysis followed by constant refining to form a coherent and logical framework and narrative that was concise enough for publication.

GenAI's sub-optimal performance in quote selection

GenAI perform satisfactorily in identifying the key themes from the transcripts, but performed poorly in identifying and selecting appropriate quotes. This suggests that while GenAI can perform well when tasked with amalgamating this type of data, at a granular level of text analysis it does not have the capacity to perform good scrutinization, and errors become more obvious. While we could have provided this feedback from the evaluators to the GenAI to attempt to improve its performance, we refrained from doing so because the themes, quotes, and feedback are specific to our study and we were seeking to assess whether, and in what aspects, the GenAI could conduct qualitative data analysis on par with human researchers. Given the amount of person-time involved in evaluating the GenAI output and providing feedback, it would not be useful or helpful to iteratively do this until the GenAI met our standard because the "learning" would not transfer to another study: a qualitative study of different topic, or of varying behavioral or ethical complexity, may provide different analytic

Adolescent girls and young women	Community males
Selection bias related to the data source	*Selection bias related to the data source*
Limited to/ specific to the dataset/not representative	Texts provided might not represent all men or girls in Kenya or other regions
Perspectives not in the dataset – parents, teachers, community leaders	If the text provided is not representative of the broader context. If quotes chosen emphasize certain perspectives over others. [*Note: in this response, GenAI is combining selection bias of participants and selection bias in how it selected quotes; it does not reference the training data as underlying the potentially biased quote selection*]
Limited to data provided, which may not cover all aspects/dimensions of the issues faced by schoolgirls during pandemic	
Selection bias related to how GenAI selected quotes may be affected by training data	*Selection bias related to how GenAI selected quotes may be affected by training data*
Termed by GenAI: Confirmation bias Focus on information that confirms existing beliefs or expectations about impact of COVID on girls May highlight quotes that confirm most prominent themes/observations	Termed by GenAI: Selection bias, Representation bias, Confirmation bias Overemphasizing parts of the text based on the AI's training data and algorithms Quotes chosen may reflect more extreme or prominent views, potentially overlooking more nuanced or moderate perspectives GenAI might not represent all perspectives equally, especially viewpoints underrepresented in training
Termed by GenAI: Neglecting positive outcomes Analysis may have focused more on negative impacts, overlooking any positive outcomes or coping strategies [*NB: While this is a selection bias, in this instance the output did not explain why GenAI might have done this – i.e., that it was based on training data*]	Focus on information that confirms or aligns with pre-existing notions or prevalent narratives learned in training
Information biases	*Information biases*
Termed by GenAI: Language and context bias Potential misinterpretation of nuances if the original discussions were conducted in another language and were translated	Termed by GenAI: Contextual bias / limitations Lack of full contextual understanding can lead to misinterpretation of culturally specific nuances May lack the ability to fully grasp broader socio-economic, political, historical context influencing texts
Termed by GenAI: Interpretation bias Interpretation of textual data, e.g., specific words such as "pressure" or "exploitation" may differ in how they are understood within the local context	
Termed by GenAI: Language and context bias Potential misinterpretation of nuances if the original discussions were conducted in another language and were translated	Termed by GenAI: Interpretation bias Limitations to accuracy in interpreting ambiguous statements, human emotion, social dynamics
Termed by GenAI: Language and translation bias If originally in a language other than English, nuances/specific meanings may have been lost or altered	Termed by GenAI: Language and terminology bias AI's understanding and use of language might reflect biases in how certain terms or phrases (e.g., slang, colloquial terms, cultural significance of specific phrases) are interpreted
Termed by GenAI: Language and interpretation bias Colloquial expressions or culturally specific references that AI could misinterpret	Termed by GenAI: Cultural bias May lack a deep understanding of cultural nuances specific to Kenya or the local context, which may affect the interpretations [*AI acknowledges in one response that it is predominantly Western-centric*] Data bias: Training data might be biased reflecting societal biases
Termed by GenAI: Cultural bias Interpretation/ understanding may be influenced by cultural context; AI might lack nuanced understanding of cultural dynamics in Kenya; interpretation influenced by cultural norms and values of training data.	Termed by GenAI: Gender bias Stemming from data on which the AI is trained, biases inherent in data sources regarding gender roles and dynamics
Biases identified from reflexivity analysis of female transcripts that did not emerge in reflexivity of male transcripts	*Biases identified from reflexivity analysis of male transcripts that did not emerge in reflexivity of female transcripts*
Termed by GenAI: Data presentation bias Dataset may have inherent biases on how questions were asked or how responses were recorded [*This could represent information bias, but was the only bias noted that would stem from the investigators rather than the AI itself*]	Termed by GenAI: Ethical and moral bias / lack of human judgment AI responses are influenced by ethical guidelines and moral frameworks embedded in training data, which may not align with local cultural and ethical standards of the community being analyzed
Termed by GenAI: Overgeneralization Overgeneralization based on specific quotes or anecdotes	Temporal bias AI training data cutoff (2023) may not include most recent developments or changes in societal norms
	Termed by GenAI: Implicit biases From training/algorithmic bias
	Mitigation strategies
	Balanced and representative training data
	Continuous learning
	Human oversight with local knowledge and corrective feedback
	Recognition/Transparency about limitations and potential biases

Table 4. Potential biases identified in reflexivity analyses of transcripts from adolescent girls and young women's transcripts and community males transcripts.

goals or results. However, if we were to have done this, one approach may have been to give the GenAI examples of quotes that human researchers judged to be consistent with and strongly supportive of each theme. While this may enhance the LLM if the topic is related, it would be unlikely to help with a different context or question.

The hallucinations were cause for concern that further analysis would produce untrustworthy results. The use of language is central to qualitative research; researchers look not just at the words used, but sentence structure, ordering, emphasis, hesitations, repetitions, tone, connotation and denotation amongst others, which give insight to the meaning, context and the overall message conveyed by the participant. Sociocultural significance is also conveyed by the language used[45]. Any errors in transcripts can alter the meaning and interpretation of results, similar to an incorrect number reported in quantitative research. Taking the example in Fig. 2, the GenAI hallucinations (i.e., quote modifications) change the context by suggesting that men were under pressure

to have sex, rather than the AGYW. While this did not affect the overall themes generated, were we not already familiar with the transcripts and findings from prior analysis, this may have been highlighted as an unusual study finding, or could have led to an unnecessary line of inquiry.

Why male transcripts might have performed more poorly

We hypothesize that GenAI performed worse in selecting quotes from the males' transcripts that were consistent with themes and sentiments because of linguistic differences between the male and AGYW transcripts. We noted that the AGYW were more precise, concise and direct in their responses to our questions and probes. In contrast the speech patterns from the male transcripts were lengthier, with responses to the moderator's questioning often indirect. Men's vocabulary frequently included euphemisms and local vernacular, and their use of grammar was less accurate. As a result, some of their quotes were more difficult to comprehend, increasing the chance of misinterpretation. These linguistic differences may also contribute to the markedly poorer performance in sentiment evaluation of the male supporting quotes.

Limitations

This was secondary analysis of data generated for understanding contextual characteristics of participants and stakeholders during a trial, and FGD were not specifically designed for the purpose of examining the quality of AI technology. We do not know if our processes generalize to another setting, topic, or method of qualitative analysis. However, we used replicable and standardized approaches to (1) conduct the qualitative study[13], (2) conduct the LLM analyses, and (3) applied a literature based rubric to evaluate the AI. Interrater reliability was variable. This may reflect that our rubric criteria were not sufficiently observable or measurable. Conversely, measures of reliability – especially Cohen's kappa - are dependent on the structure of the data (number of subjects, categories, and prevalence). While Gwet's AC1 was also presented and addresses some of these concerns, limitations from data structure remain, as evidenced in some instances by high percent agreement and moderate reliability measures. Therefore, while the rubric we developed provided criteria and replicability to the evaluation, broader application of the rubric is necessary to further refine its utility.

In this study, LLM was not context-specific and relied on pre-existing training data to conduct analysis. As the model may not have been sufficiently trained on topics or contexts inclusive of western Kenya, COVID-19, or sexual and reproductive health, its analysis might lack relevance or accuracy, and this limitation was acknowledged by the GenAI in its assessment of its biases. The selection of lower quality quotes likely occurred because LLM relied on surface-level patterns and lacked contextual understanding. Although GenAI was provided English translations, there are nuances of idiomatic and cultural references that we believe were a challenge to the GenAI and were not resolved. This highlights the importance of building AIs that have been trained in multiple languages and cultures. Hallucinations have been documented in other studies of GenAI thematic analyses[11], and present a substantial threat to its validity and trustworthiness. Statistical approaches to detect hallucinations are emerging[46]; while this can alert researchers to risk, Xu et al. demonstrate LLM hallucination is inevitable and unavoidable[47]. The results of sentiment analysis may have been limited by our choice of circumplex sentiments. This type of analysis should be repeated using the full range of valence and arousal, to determine its utility in more varied and larger datasets.

Customizing GenAI models for specific research contexts and requirements may improve contextual understanding and reduce hallucinations. For example, special training datasets that reflect diverse and contextually rich qualitative scenarios could be used to augment the GenAI model through a retrieval augmentation generated (RAG) approach[48]. This would make it more reliable in producing relevant, context-sensitive outputs. Developing mechanisms for human cross checking would increase reliability. Regarding new capabilities for qualitative research, development should focus on GenAI ability to compare and contrast between and within sources, take into account background characteristics of participants, and look for typical and atypical patterns. GenAI will also need to develop the inherently human trait of being 'curious' with the data – spotting or investigating something that it hasn't been instructed to do, or being able to identify something that emerges from the data that hadn't been thought about beforehand.

Conclusions

Based on our study, GenAI implemented in GPT-4o was unable to provide a thematic analysis that is indistinguishable from a human analysis of focus group study transcripts, due to a combination of errors and a low level of sophistication. We recommend that it can be used currently as an aid to the human analyst in identifying themes, keywords, and basic narrative, and potentially as a check for human error or bias. GenAI may also be useful for rapid appraisal to quickly generate themes and subthemes from which to then hone and refine the direction for a further round of data collection. However, until it can eliminate hallucinations, provide better contextual understanding of quotes, undertake a deeper scrutiny of data, and demonstrate a greater range of reflexivity, it is not reliable or sophisticated enough to undertake a rigorous thematic analysis equal in quality to experienced qualitative researchers.

Data availability

The datasets generated and/or analysed during the current study are available in https://doi.org/10.25417/uic.26495884.v1.

Received: 13 February 2025; Accepted: 4 September 2025
Published online: 07 October 2025

References

1. Dunn, P., Ali, A., Patel, A. P. & Banerjee, S. Brief review and primer of key terminology for artificial intelligence and machine learning in hypertension. *Hypertension.* **82**, 26–35 (2025).
2. Yim, D., Khuntia, J., Parameswaran, V. & Meyers, A. Preliminary evidence of the use of generative AI in health care clinical services: Systematic narrative review. *JMIR Med. Informat.* **12**(1), e52073 (2024).
3. Takita, H. et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. *Npj Digital Medicine.* **8**(1), 175 (2025).
4. 4. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S. Gpt-4 technical report. (2023).
5 Vaswani, A. et al. Attention is all you need. *Adv. Neural Informat. Process. Syst.* **2017**, 30 (2017).
6. Chowdhury, I. A. Issue of quality in a qualitative research: An overview. *Innovat. Issues and Approach. Soc. Sci.* **8**(1), 142–62 (2015).
7. Braun, V. & Clarke, V. Using thematic analysis in psychology. *Qualitat. Res. Psychol.* **3**(2), 77–101 (2006).
8. 8. Turobov A, Coyle D, Harding V. Using ChatGPT for thematic analysis. *arXiv* (2024).
9. Deiner, M. S. et al. Large language models can enable inductive thematic analysis of a social media corpus in a single prompt: human validation study. *JMIR infodemiol.* **4**(1), e59641 (2024).
10. Morgan, D. L. Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *Int. J. Qualitat. Methods.* **22**, 16094069231211248 (2023).
11. Lee, V. V., van der Lubbe, S. C., Goh, L. H. & Valderas, J. M. Harnessing ChatGPT for thematic analysis: Are we ready?. *J. Med. Internet Res.* **26**, e54974 (2024).
12. 12. Open A. ChatGPT (Mar 14 version)[Large language model]. (2023).
13. Enid, A. et al. âœWhenever i help her I am also expecting her vagina in returnâ: a qualitative analysis to explore menâ€™s and adolescent girlsâ™ perceptions of the impact of the COVID-19 pandemic on the sexual behaviour and health of adolescent girls in rural western Kenya. *BMJ Public Health.* **2**(2), e001214. https://doi.org/10.1136/bmjph-2024-001214 (2024).
14. 14. OpenAI. Enterprise privacy at OpenAI. https://openai.com/enterprise-privacy/ (2024).
15. Mehta, S. D. et al. Analysis of bacterial vaginosis, the vaginal microbiome, and sexually transmitted infections following the provision of menstrual cups in Kenyan schools: results of a nested study within a cluster randomized controlled trial. *PLoS Med.* **20**(7), e1004258 (2023).
16. Mehta, S. D. et al. High prevalence of Lactobacillus crispatus dominated vaginal microbiome among Kenyan Secondary School Girls: Negative effects of poor quality menstrual hygiene management and sexual activity. *Front. Cell. Infect. Microbiol.* **11**, 716537 (2021).
17. Zulaika, G. et al. Menstrual cups and cash transfer to reduce sexual and reproductive harm and school dropout in adolescent schoolgirls: study protocol of a cluster-randomised controlled trial in western Kenya. *BMC Public Health.* **19**, 1–14 (2019).
18. Zulaika, G. et al. Impact of COVID-19 lockdowns on adolescent pregnancy and school dropout among secondary schoolgirls in Kenya. *BMJ Glob. Health.* **7**(1), e007666 (2022).
19. Mehta, S. D. et al. Increased reproductive tract infections among secondary school girls during the COVID-19 pandemic: associations with pandemic-related stress, mental health, and domestic safety. *Sex Med.* **12**(3), qfae045. https://doi.org/10.1093/sexmed/qfae045 (2024).
20. Vaishya, R., Misra, A. & Vaish, A. ChatGPT: Is this version good for healthcare and research?. *Diabetes Metab. Syndrome Clin. Res. Rev.* **17**(4), 102744 (2023).
21. Li, J., Dada, A., Puladi, B., Kleesiek, J. & Egger, J. ChatGPT in healthcare: a taxonomy and systematic review. *Comput. Methods Programs Biomed.* **245**, 108013 (2024).
22. 22. Sallam M, editor. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* (2023).
23. Cascella, M., Montomoli, J., Bellini, V. & Bignami, E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J. Med. Syst.* **47**(1), 33 (2023).
24. 24. Rane N, Choudhary S, Rane J. Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. Performance, Architecture, Capabilities, and Implementation (2024).
25. 25. Lee G-G, Latif E, Shi L, Zhai X. Gemini pro defeated by gpt-4v: Evidence from education. *arXiv* (2023).
26. 26. Carlà MM, Gambini G, Baldascino A, Boselli F, Giannuzzi F, Margollicci F, Rizzo S. Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison. Graefe's Archive for Clinical and Experimental Ophthalmology. 1–15. (2014).
27. 27. Sukhdeve DSR, Sukhdeve SS. Google Colaboratory. Google Cloud Platform for Data Science: A Crash Course on Big Data, Machine Learning, and Data Analytics Services. p. 11–34 (Springer, 2023).
28. 28. Ekin S. Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. Authorea Preprints. (2023).
29. Davis, J., Van Bulck, L., Durieux, B. N. & Lindvall, C. The temperature feature of ChatGPT: modifying creativity for clinical research. *JMIR Human Fact.* **11**(1), e53559 (2024).
30. 30. Akamine A, Hayashi D, Tomizawa A, Nagasaki Y, Akamine C, Fukawa T, Hirosawa I, Saigo O, Hayashi M, Nanaoya M. Effects of temperature settings on information quality of ChatGPT-3.5 responses: A prospective,single-blind, observational cohort study. *medRxiv* (2024).
31. Gilardi, F., Alizadeh, M. & Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci.* **120**(30), e2305016120 (2023).
32. Turchin, A., Masharsky, S. & Zitnik, M. Comparison of BERT implementations for natural language processing of narrative medical documents. *Informat. Med. Unlocked.* **36**, 101139 (2023).
33. 33. Kotelnikova A, Paschenko D, Bochenina K, Kotelnikov E, editors. Lexicon-based methods vs. BERT for text sentiment analysis. International Conference on Analysis of Images, Social Networks and Texts. (Springer, 2021).
34. Alaparthi, S. & Mishra, M. BERT: A sentiment analysis odyssey. *J. Market. Anal.* **9**(2), 118–126 (2021).
35. 35. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. Bertscore: Evaluating text generation with bert. *arXiv* (2019).
36. 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.* **19**(6), 349–357 (2007).
37. O'Brien, B. C., Harris, I. B., Beckman, T. J., Reed, D. A. & Cook, D. A. Standards for reporting qualitative research: a synthesis of recommendations. *Acad. Med.* **89**(9), 1245–1251 (2014).
38. Long, H. A., French, D. P. & Brooks, J. M. Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis. *Res. Methods Med. Health Sci.* **1**(1), 31–42 (2020).
39. 39. Hutto C, Gilbert E, editors. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media* (2014).
40. Mohammad, S. M. *Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. Emotion measurement* 323–79 (Elsevier, 2021).
41. Loria, S. Textblob Documentation. *Release 015.* **2** (8), 269 (2018).
42. 42. Elbagir S, Yang J, editors. Twitter sentiment analysis using natural language toolkit and VADER sentiment. *Proceedings of the International Multiconference of Engineers and Computer Scientists* (2019).
43. Russell, J. A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **39**(6), 1161 (1980).

44. Valenza, G., Citi, L., Lanatá, A., Scilingo, E. P. & Barbieri, R. Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics. *Sci. Rep.* **4**(1), 4998 (2014).
45. 45. Anderson C, Bjorkman B, Denis D, Doner J, Grant M, Sanders N, Taniguchi A. Essentials of Linguistics, (v. 2.2-February 2023). (McMaster University, 2022).
46. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature.* **630**(8017), 625–630 (2024).
47. 47. Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: An innate limitation of large language models. *arXiv* (2024).
48. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Informat. Process. Syst.* **33**, 9459–9474 (2020).

Acknowledgements
In memoriam, we acknowledge Elizabeth Nyothach (EN) of Kenya Medical Research Institute (KEMRI, Kisumu, Kenya), who passed away early December 2024. Ms. Nyothach was integral to the conduct of the qualitative studies and their interpretation; her contributions are summarized in Authors' Contributions.

Author contributions
Conceptualization, RB and SDM; Methodology, RB, SP, SDM, LM; Formal analysis, SP, RB, SDM, SY, EA; Investigation, SDM, SP, EA, SY, GZ, EN, FOO, PAP-H, RB, LM; Resources, EN and FOO; Data curation, SP, EA, SY, FOO, LM; Writing—original draft, SDM, SP, RB, LM; Writing—review & editing, SDM, SP, EA, SY, GZ, FOO, PAP-H, RB, LM; Supervision, SDM, GZ, EN, FOO, PAP-H, LM, RB; Project administration, SDM, EN, FOO, PAP-H; Funding acquisition, SDM and PAP-H. All authors reviewed the manuscript.

Funding
This study was supported by the National Institutes of Health Eunice Shriver National Institute of Child Health and Human Development (R01HD106822 to SDM), and the Joint Global Health Trials Initiative (UK-Medical Research Council/ Department for International Development/ Wellcome Trust/Department of Health and Social Care; MR/N006046/1 to PPH). The funders had no role in the design of the study, the collection, analysis, and interpretation of data, or in writing the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval and consent to participate
This study was approved by the institutional review boards of Maseno University Ethics Review Committee (MUERC, MSU/DRPI/MUERC/01021/21), University of Illinois at Chicago (UIC, #2022 – 0220), and (Liverpool School of Tropical Medicine (LSTM, #21–087, favourable ethical opinion). Written informed consent was obtained for all participants.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-18969-w.

Correspondence and requests for materials should be addressed to S.D.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.