Special Issue: The Evolving Landscape of Qualitative Research: Exploring the Potential and Navigating the Challenges of Artificial Intelligence

International Journal of Qualitative Methods
Volume 24: 1–13
© The Author(s) 2025
DOI: 10.1177/16094069251338898
journals.sagepub.com/home/ijq
Sage

Leveraging Generative Text Models and Natural Language Processing to Perform Traditional Thematic Data Analysis

Isil Anakok[1], Andrew Katz[1], Kai Jun Chew[2], and Holly Matusovich[1]

Abstract

We explore the possibility of using natural language processing (NLP) and generative artificial intelligence (GAI) to streamline the process of thematic analysis (TA) for qualitative research. We followed traditional TA phases to demonstrate areas of alignment and discordance between (a) steps one might take with NLP and GAI and (b) traditional thematic analysis. Using a case study, we illustrate the application of this workflow to a real-world dataset. We start with processes involved in data analysis and translate those into analogous steps in a workflow that uses NLP and GAI. We then discuss the potential benefits and limitations of these NLP and GAI techniques, highlighting points of convergence and divergence with thematic analysis. Then, we highlight the importance of the central role of researchers during the process of NLP and GAI-assisted thematic analysis. Finally, we conclude with a discussion of the implications of this approach for qualitative research and suggestions for future work. Researchers who are interested in AI-assisted methods can benefit from the roadmap we provide in this study to understand the current landscape of NLP and GAI models for qualitative research.

Keywords

qualitative research, thematic analysis, natural language processing, generative AI, AI-assisted qualitative method

Thematic Analysis With NLP and GAI Technologies

Qualitative research focuses on extracting meaning from data (Hesse-Biber, 2010) and depends on the identification of key ideas that require complex and non-linear processes of coding qualitative data (Hatch, 2023; Saldaña, 2014). Researchers have developed many different approaches to qualitative research. However, the overall goal for data analysis is a "systemic search for meaning" (Hatch, 2023, p. 148). In this study, we focus on TA which is a model of qualitative data analysis that many researchers prefer because it leads them to "the mechanics of coding and analyzing qualitative data systematically" (Braun & Clarke, 2012, p. 58). Traditional TA, while a widely used method as suggested by Braun and Clarke's study (2006) cited over 200,000 times consists of six phases: (1) familiarizing themselves with the data, (2) coding, (3) generating initial themes, (4) developing and reviewing themes, (5) refining, defining and naming themes, and (6) writing up (Braun & Clarke, 2022, pp. 35–36). Each phase is labor-intensive and requires significant time due to its iterative process and cognitive effort for meaningful insights, which makes TA challenging to use with larger qualitative datasets (Bhaduri, 2018; Braun & Clarke, 2012; Saldaña, 2013). This observation raises the question: are there computer-assisted approaches that can mimic TA? Using GAI to enhance traditional TA lies in the potential benefits and innovation that GAI could bring to qualitative research literature and methods.

[1] The Department of Engineering Education, Virginia Tech, Blacksburg, VA, USA
[2] Department of Engineering, The Department of Engineering Fundamentals, Embry-Riddle Aeronautical University, Daytona Beach, FL, USA

Corresponding Author:
Isil Anakok, The Department of Engineering Education, Virginia Tech, 345 Goodwin Hall, 635 Prices Fork Rd, Blacksburg, VA 24061, USA.
Email: ianakok@vt.edu

Recently, the features and abilities of NLP and GAI have introduced new opportunities and potential for conducting TA for research (De Paoli, 2024; Gamieldien et al., 2023; Katz et al., 2024; Mathis et al., 2024; Morgan, 2023; Perkins & Roe, 2024b; Tai et al., 2024). For example, (Perkins & Roe, 2024b, 2024a) did an inductive TA for two different studies with both AI-assisted (ChatGPT-4) and manual processes. They demonstrated the ability of AI on pattern recognition and the importance of the human in the loop process for the interpretation of collected data while finalizing their codebook with the help of both methods. Furthermore, Gamieldien et al. (2023) compared NLP and GAI models with manual TA. They identified similar themes in both AI-assisted models and manual analysis and showed that NLP and GAI have promising performance in qualitative data analysis. Another comparative study conducted by Lixandru (2024) using ChatGPT 3.5 and manual coding found a significant similarity between the findings and emphasized the capability of GAI to interpret qualitative information.

Moreover, Mathis et al. (2024) and De Paoli (2024) conducted inductive TA using the TA phases suggested by Braun and Clarke (2006). While De Paoli (2024) used the GPT 3.5 Turbo model, Mathis et al. (2024) used Llama-2-70B to generate a codebook. They focused on codebook generation, and both studies compared the output of GAI models with their manual coding findings. Their processes had limitations while following the six phases of TA developed by Braun and Clarke (2006). De Paoli (2024) did not use any GAI models to conduct Phase 1, and Mathis et al. (2024) used a GAI model to transcribe the collected data in Phase 1. Thus, their approach to Phase 1 did not align with Phase 1 in the TA approach suggested by Braun and Clarke (2006). In our study, we proposed an approach to mimic Phase 1 defined by Braun and Clarke (2022). Also, we used the most updated process Braun and Clarke (2022) suggested for TA. The differences between these two studies and our study are further discussed in the subsection called Comparing the Previous GAI Models for Thematic Analysis.

Regardless of the type of data analysis, previous literature emphasized that the current capabilities of NLP and GAI tools at the time they conducted their research were far from autonomous, and the role of the researcher is still significant. By having human-in-the-loop continuously, we can enhance the reliability, accuracy, and relativity of research topics and contexts that might help address ethical concerns (Lund et al., 2023) and reduce GAI hallucinations (Ye et al., 2023). Another crucial reason for the significance of a human-in-the-loop process is grasping nuanced findings that may not be achieved by the NLP and GAI models (De Paoli, 2024). NLP and GAI models may struggle to identify the actual human emotions, values, and norms in societies (Arora et al., 2023). Including researchers in the process can infuse the lived experiences that can contribute to robust interpretation of findings. Thus, having a human in the loop can distinguish the nuances that the GAI-driven may oversimplify or misinterpret in collected data. Also, GAI models may have biases that need to be identified by the researchers to help reduce the misinterpretations of findings (Liang et al., 2021; Navigli et al., 2023). Adjusting the outcomes of NLP and GAI models and checking the alignment of outputs with the research purposes and context can contribute to more representative perspectives within the collected data.

Understandably, some researchers may resist, or reject the use of NLP and other approaches to computer-based analytic approaches because of the potential loss of personal touch or the potential for bias when using pre-trained models (e.g., bias coming from their training data). However, it’s important to recognize that traditional qualitative methods also come with their own biases. Qualitative research methods rely on researchers’ subjective interpretations, which may influence the analysis and findings. While it’s natural to be resistant to new approaches, this resistance shouldn’t be absolute, especially when innovative tools can complement and enhance traditional methods.

Furthermore, GAI would provide reliability by applying a consistent algorithm across all forms of data included in the study. This may minimize the variability of codes generated by the researchers. However, we still acknowledge the bias that might be already embedded in the trained datasets in the algorithms (Li et al., 2024). Despite the bias within the GAI algorithms, applying it across the entire dataset could provide us with uniformly emerged codes and themes. Moreover, the capability of GAI to work with large amounts of text can lead to detecting all the patterns that may be overlooked and missed partially by human researchers in traditional TA. This capability of GAI can also allow researchers to collect data from more participants (an even more diverse pool of participants), which may lead to more inclusive qualitative research. However, to be able to achieve a meaningful diverse pool of participants, we should perform intentional recruitment efforts alongside integrating GAI tools so we can ensure a variety of perspectives and representations across the participant pool. When researchers try this new approach and integrate it into their traditional TA, they can contribute to the innovation of qualitative research methods hence paving a new way to understand their studied phenomena.

Despite the practicality of NLP and GAI applications in TA, there are several considerations to question and explore, such as the accuracy and reliability of the themes generated by GAI. We, researchers, must engage with GAI-assisted TA and identify the potential benefits and limitations of the method conducted with GAI. It is significantly important to address both the advancements, such as the expedited process of data analysis, and concerns that arise during the process of data analysis, including the reevaluation of research integrity by the researchers, data privacy, and biases (Davison et al., 2024; Elali & Rachid, 2023; Perkins & Roe, 2024a, 2024b). Nonetheless, GAI in qualitative research is evident and rapidly growing in literature as researchers explore more efficient and innovative approaches for TA.

In this paper, we explore the possibility of using NLP and GAI to streamline the process of TA in social science research. Our goal is not only to compare traditional data analysis with GAI-assisted TA. Instead, we aim to explain how NLP and GAI technologies can be leveraged to conduct TA. Our goal is to investigate how GAI tools, particularly large language models (Llama3.1, Whisper3, etc.), can assist in the six phases of TA by Braun and Clarke (2022) in qualitative research. Through this investigation, our study aims to provide insights and practical guidelines for qualitative researchers on leveraging GAI to bring another way of TA, ultimately contributing to the advancement of qualitative research methods in the new era of increased availability of NLP and GAI. We aim to answer the research questions:

RQ1: How can common steps in thematic analysis be performed using GAI and NLP?

RQ2: What are the advantages and limitations of using NLP and GAI tools for thematic analysis, as demonstrated through a case study?

Method and Case Study

We presented the common phases in thematic analysis (TA) suggested by Braun and Clarke (2022) and translated those phases into analogous steps in a workflow that uses NLP and GAI tools. We call this method Generative AI-Assisted Thematic Analysis (GATA). While we explain how data analysis can be done with NLP and GAI, we acknowledge that it requires some level of programming language knowledge. However, generative code completion tools such as GitHub, Copilot, and ChatGPT help users write computer programs and provide a script to do the tasks, and researchers who are new to programming languages can use GAI tools for help while improving their programming skills. We also provided the code used for this entire workflow in a GitHub repository available at https://github.com/andrewskatz/ijqm-gata-2025. Table 1 provides the overview of TA for both the traditional TA method as described by Braun and Clarke (2022), and our GATA method.

To explore how the six-phase TA guided by Braun and Clarke (2022) can be conducted by using GATA, the overall workflow for TA with the NLP and GAI technologies is shown in Figure 1. Each box represents the TA phases, and the abbreviations in each box show the output of each phase (T = transcript, SP = Summary Point, IC = Initial Code, Th = Theme, Ht = Higher-level theme). Figure 1 clarifies the sequential workflow, emphasizing the human-in-the-loop nature of the process. While NLP and GAI tools can assist with tasks such as summarization, initial coding, and theme generation, Figure 1 indicates where researchers need to review and refine the outputs to ensure quality, reliability, and alignment with their research questions. This iterative process combines the strengths of GAI with the rigor of human-driven analysis. Additionally, the figure highlights the necessity of researcher participation in writing the final manuscript, reinforcing the idea that GAI tools should complement traditional methods rather than replace them.

To illustrate our suggested steps in our method, we demonstrated results from an actual research study we conducted on engineering faculty members. We asked participants: “Has the arrival of GAI impacted their thinking on assessment and assessment practices? If yes, how? If not, why not?” This study received ethical approval from the Virginia Tech Institutional Review Board (IRB) (approval # 21-639). In the following subsections, we explained how GATA was conducted with real data. The data was cleaned prior to this manuscript. Moreover, we did not make any changes during the data analysis done by any NLP and GAI models to be able to show and discuss the analysis. All of the examples provided in this section are the product of NLP and GAI models used in this study. However, we emphasize the importance of researchers’ inputs after running models for each phase while conducting qualitative research.

Phase 1: Familiarizing Yourself With the Data

In traditional qualitative research, researchers start to read the cleaned transcripts and re-read them to become familiar with the data and take notes on the ideas of potential codes emerging from the transcripts Braun and Clarke (2022). When GAI models are used for TA, uploading each transcript to the model and generating summary points mimics some aspects of this reading process. Instructions were given to the generative text model (here, we used Llama 3.1-8b) to generate summary points from transcripts (shown from T_1 to T_a in Figure 1). At the end of this phase, we had a list of summary points shown from SP_{11} to SP_{ij} in Figure 1 where i ranges from 1 to aj, and j ranges from the number of points for person i.

In this phase, researchers had the summary points from transcripts that they can read through and get more familiar with. Thus, loading the transcript into the system and generating summary points from each transcript eased the process of getting familiar with the data and moving to the second phase. Researchers can also use the transcripts without the data cleaning process because the llama 3.1-8b model does not require cleaned data and can de-identify the text while summarizing.

In our case, we uploaded the responses from each participant from transcripts. The responses were given in a column in a csv file. The prompt given to the Llama 3.1-8b model (Llama, n.d) is in Appendix A.1 Prompt- Summarization (Katz et al., 2024). In the prompt, we gave the model persona assignment by telling them it is the expert on text analysis. We gave specific data types and data collection contexts. For our case, the data type was a written response, and the data collection context was a study of faculty reactions to GAI. Then, we provided the task of summarizing the given data type in some of the rules we identified for the model. For

Table 1. Overview of Thematic Analysis With Manual Method and NLP and GAI

Task	Manual method	NLP and GAI methods
Thematic analysis		
Phase 1: Familiarizing yourself with the data	Reading and re-reading the transcripts, noting down initial ideas	Upload transcripts (T) and generate summary points (SP) with llama 3.1-8b (Llama, n.d)
Phase 2: Coding	Systematically coding interesting features of the data across the entire dataset	Clustering summary points (SP) with mxbai, UMAP, and Scikit-learn models and identify initial codes (IC) from them with nous-hermes-2-mistral model
Phase 3: Generating initial themes	Grouping codes into potential themes, gathering all data relevant to each potential theme	Categorizing initial codes (IC) and identifying themes (Th) with the mistral-large-2 model
Phase 4: Developing and reviewing themes	Checking if the themes work in relation to the coded extracts and the entire dataset.	Embedding summary points (SP) and themes (Th) and use cosine similarity scores to check if the themes (Th) represent the original excerpts
Phase 5: Refining, defining and naming themes	Defining each theme, generating clear definitions and names for each theme	Using the mistral-large-2 model to organize the themes (Th) and identify higher-level themes (Ht) with their definitions
Phase 6: Writing up	Synthesizing our findings, writing a manuscript and relating the analysis to research questions and literature	We did not let the models produce reports for the authenticity and reliability of findings

example, we asked the model to summarize each idea discussed in the task in a new line and enumerate them (we called each line a summary point (SP)). In case our data still had some identity information of participants, we asked the model not to include any names or pronouns while summarizing the responses. Then, we also let the model know there is no limit on the number of summary points so it could create as many topics as necessary.

To set more clear expectations for the model, we also provided an example of input and output responses that fit exactly the task and rules described in the prompt. To make sure the model understood our rule about the names and the pronouns, we referred back to the example we provided and asked the model to see how our example did not include any names and pronouns. Also, we set up another rule by highlighting how our example did not have any made-up summary points. We warned the model not to make up information that is not in the input text otherwise there is a severe penalty for that. We also provided an example of what to do if there is no meaningful or useful information in responses. For example, we stated that if the text is very short and says “nothing”, do not make up new things. At the end of running the model with responses from participants, it created a list of summary points for each response. In Figure 2, we illustrated how the model works for this phase with an example of input and output. We provided an excerpt from our dataset and the output generated four summary points. In this research study, we had a total number of 104 excerpts (T_1 to T_{104}) and 403 summary points were generated.

Since the text context was given in the prompt Llama 3.1-8b was able to identify and describe the summaries related to faculty reactions. For the given example of excerpts, there were no make-up summary points in the response. However, the role of the researcher is to go through these generated summary points after completing all the responses. If there are any make-up or missing topics related to the context, researchers can consider changing the given prompt to the model. For example, if you do see missing topics, you may

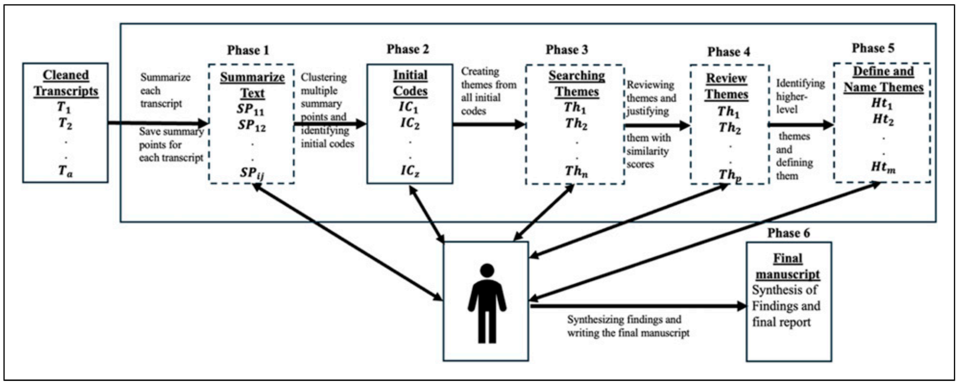
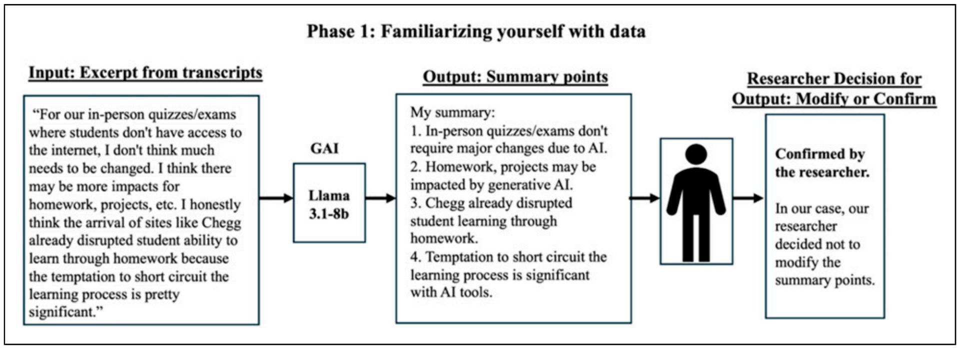


Figure 1. The Overall Workflow for Thematic Analysis With the NLP and GAI Technologies.


Phase 1: Familiarizing yourself with data
Input: Excerpt from transcripts
GAI
Llama 3.1-8b
Output: Summary points
Researcher Decision for Output: Modify or Confirm
Confirmed by the researcher.

"For our in-person quizzes/exams where students don't have access to the internet, I don't think much needs to be changed. I think there may be more impacts for homework, projects, etc. I honestly think the arrival of sites like Chegg already disrupted student ability to learn through homework because the temptation to short circuit the learning process is pretty significant."

My summary:
1. In-person quizzes/exams don't require major changes due to AI.
2. Homework, projects may be impacted by generative AI.
3. Chegg already disrupted student learning through homework.
4. Temptation to short circuit the learning process is significant with AI tools.

In our case, our researcher decided not to modify the summary points.

Figure 2. An Example of How the GAI Identifies Summary Points From Excerpts.

give more detailed instructions and context to the model in the prompt. If they think changing the prompt does not increase the accuracy of the summary points, they should manually edit them until they agree their summary points are reliable. At the end of Phase 1, we had 403 summary points from our dataset. The model was able to go through a large number of excerpts in a short time and the researchers familiarized themselves with the dataset from summary points instead of going through long texts and taking notes or highlighting the lines. This allowed researchers to handle their dataset better in a shorter time since they were able to have the main points of each excerpt. While saving time, they had less cost compared to paying extra people for analysis.

Phase 2: Coding

During coding, researchers systematically code the emerging features of the data across the entire dataset (Braun & Clarke, 2022). This phase takes several hours, depending on the length of the data and the number of participants. Since the goal is to identify the initial codes for rich findings, this may require reading transcripts iteratively multiple times. For the initial codes, researchers write the definition of each emerging code, take a memo, and write down examples from transcripts if necessary. To conduct the second phase from manual coding with the new approach, there were two main steps to follow: clustering summary points with NLP tools and coding the clusters with GAI tools, as illustrated in Figure 3. Summary points were clustered with NLP tools into y clusters (from CL_1 to CL_y) and generated z number of codes (from IC_1 to IC_z in Figure 1) with a GAI model from these clusters.

In the NLP models, we used the summary points (SP) from Phase 1, and the model clusters were then based on recurring patterns in their context. There were three steps to complete this process:

(1) Embed the summary points. To initiate the clustering process, we first embedded the summary points from Phase 1 using a text embedding model. Text embedding models generated high-dimensional vector representations of text based on the notion of distributional semantics. We used the open-source mxbai embedding model for this task, which transforms the input (i.e., each summary point) to a vector of 1,024 dimensions (Li & Li, 2024;Lee et al., n.d).

(2) Dimension reduction: From these high-dimensional representations, we then used dimension reduction techniques because attempting to cluster the vectors in the original 1,024-dimensional embedding space could suffer from the curse of dimensionality. To accomplish this dimension reduction, we first used principal component analysis to reduce the dimensionality to an intermediate embedding space that retains 90% of the original variance in the data. From this intermediate embedding space, we then used Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2020) to five dimensions.

(3) Clustering: We used an agglomerative clustering algorithm as implemented in Scikit-Learn to cluster the lower dimensional data in the previous step (Kramer, 2016). The agglomerative clustering algorithm identifies each data point as a cluster at the beginning and then each cluster is paired with another and merged until the hierarchy is created in between clusters.

Once the clusters were generated from summary points, researchers could examine them for accuracy. Some clusters may be similar or unrelated to the posed research questions. Thus, researchers can further organize these clusters by merging similar ones and removing unrelated clusters. At the end of this step, we have the clusters from CL_1 to CL_y. Coding was done with the GAI model called Nous-hermes-2-mistral and the prompt was given to the model to generate codes (IC), their definitions, and examples from summary points. Once the model structures the initial codes, definitions, and examples, researchers can play a big role in checking if the codes are accurate for the given dataset and research questions, if there are any unrelated or repetitive codes, and make changes accordingly.

In our case, we had 42 clusters in the first step of this phase (CL_1 to CL_{42}) from the initial 403 summary points. To track our sample outcome in the previous phase, we chose the

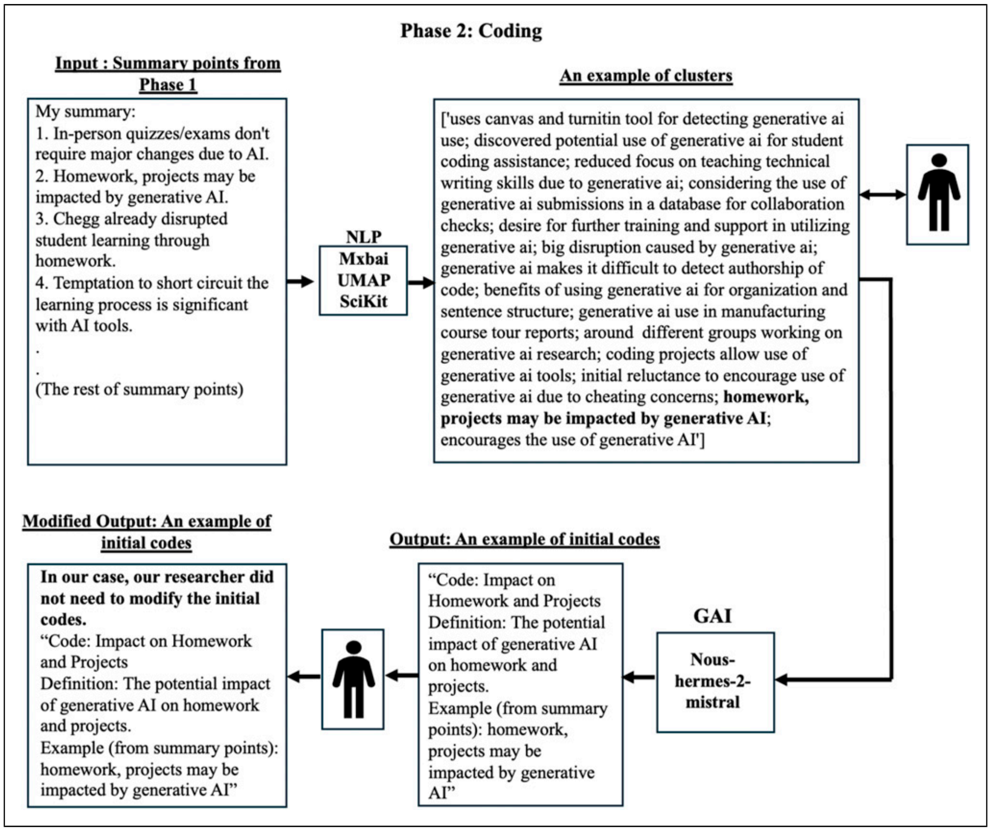

Phase 2: Coding
Input : Summary points from Phase 1
My summary:
1. In-person quizzes/exams don't require major changes due to AI.
2. Homework, projects may be impacted by generative AI.
3. Chegg already disrupted student learning through homework.
4. Temptation to short circuit the learning process is significant with AI tools.
.
.
(The rest of summary points)
NLP
Mxbai
UMAP
SciKit
An example of clusters
['uses canvas and turnitin tool for detecting generative ai use; discovered potential use of generative ai for student coding assistance; reduced focus on teaching technical writing skills due to generative ai; considering the use of generative ai submissions in a database for collaboration checks; desire for further training and support in utilizing generative ai; big disruption caused by generative ai; generative ai makes it difficult to detect authorship of code; benefits of using generative ai for organization and sentence structure; generative ai use in manufacturing course tour reports; around different groups working on generative ai research; coding projects allow use of generative ai tools; initial reluctance to encourage use of generative ai due to cheating concerns; homework, projects may be impacted by generative AI; encourages the use of generative AI']
GAI
Nous-hermes-2-mistral
Output: An example of initial codes
"Code: Impact on Homework and Projects Definition: The potential impact of generative AI on homework and projects. Example (from summary points): homework, projects may be impacted by generative AI"
Modified Output: An example of initial codes
In our case, our researcher did not need to modify the initial codes.
"Code: Impact on Homework and Projects
Definition: The potential impact of generative AI on homework and projects.
Example (from summary points): homework, projects may be impacted by generative AI"

Figure 3. An Example of How the Summary Points are Clustered and Codes are Generated.

second summary point (homework, projects may be impacted by generative AI). The summary points separated by semicolons were clustered together as shown in Figure 3. The cluster of it is shown in the middle box in Figure 3. This cluster had other summary points from excerpts that had some similar patterns between them because the NLP model aimed to categorize them, not to generate similar initial codes multiple times so having them in one cluster would help the GAI model identify the initial codes from the summary points. Once we have our clusters, researchers should read each cluster and see if they see the pattern within the clusters related to their context (in the figure we see the researcher's input for modifications). Researchers may identify some of the clusters that might be similar to each other and merge them together before moving to the next step.

After having the list of clusters, the next step was to make the Nous-hermes-2-mistral model generate the initial codes, their definitions, and examples from summary points. Here, we gave a new prompt (shown in Appendix A.2) to create initial codes from summary points. Similar to the prompt in the previous phase, we assigned a persona to the model. This time we identified the model as an expert in qualitative research methods, especially in TA. We introduced its task, the given data type, and its context. Then, we gave steps: code, definition, and example to generate codes. After giving the steps to structure the initial codes, we also provided some criteria to the model to check the generated information. These criteria asked whether the created codes are related to the given context and whether the labels, definitions, and examples are structured as requested in the provided steps.

The Nous-hermes-2-mistral model created thirteen initial codes from the given cluster above in Figure 3. We chose one of our summary points identified as an initial code labeled "Impact on Homework and Projects" to demonstrate the definition and example given from the model. Overall, the total number of initial codes generated by the model was 154 (from IC_1 to IC_{154}) for our study.

Phase 3: Generating Initial Themes

In manual coding, researchers group initial codes from the previous phase into potential themes (Braun & Clarke, 2022). Similar to generating initial codes in Phase 2, researchers manually review the codes to find themes that require an iterative process. When GAI was used for searching themes, we first clustered the initial codes by embedding them using the mxbai text embedding model, reducing the dimensionality of those embeddings, and then clustering the lower dimensional representations of those embeddings. Once we had the clusters of initial codes, the prompt was given to the GAI model mistral-large-2. The input was the initial codes from Phase 2. First, b number of clusters (CL) was generated and then n number of themes (Th) was generated from these clusters.

In our case, we first clustered the 154 initial codes from Phase 2 by using mxbai, UMAP, and Scikit models (similar to the first step of Phase 2 explained in detail in previous subsection). At the end of the clustering step, we provided a prompt to the Mistral-large-2 model to identify themes. In the prompt shown in Appendix A.3 (Katz et al., 2024), we first assigned a persona and provided our research question following the instructions from the TA instructions by Braun and Clarke (2012, 2022). After giving instructions, we also provided how to structure the findings as outputs. Considering the instructions, the model provided the name of the themes, the reasoning for why it grouped the structure for each theme, and the list of the initial codes for each cluster.

We assigned the persona of an expert qualitative researcher specializing in TA and provided the data type, the initial codes generated in the previous phase, and the context of the data. In addition, we also provided the research question we aimed to explore. Then, we identified the task as removing the redundancies across the labels. Also, we asked to remove the unnecessary codes that are not related to the context of the study and provided examples of how to do these tasks. We specified how the output format should be and asked for a reasoning for the decisions it makes on removing or merging codes and asked for the precise list for its final list of codes by using a chain of thought prompting (Wei et al., 2022). We provided instructions by using the guidance for how to identify themes by Braun and Clarke (2012, 2022).

To track the codes from the previous phase (Impact on Homework and Projects), we provided the example of how the mistal-large-2 model explained the observations and reasoning in Figure 4.

The model grouped some of the codes from the previous phase (z = 154) and explained their concept because the participants mentioned reconsidering traditional assessment methods due to the GAI tools. The NLP model generated seven clusters, including these 154 codes from the previous phase and identified 25 (from Th_1 to Th_{25}) themes. After defining the concept of the codes, the model identified the theme as "Shift in Assessment Methods." Researchers can go through each theme with their concepts and merge them further if necessary.

Phase 4: Developing and Reviewing Themes

In this phase, researchers check if the themes work in relation to the emerging codes, their relevant extracts, and the entire dataset manually (Braun & Clarke, 2012). This process can be repeated iteratively by the researcher until they agree on the themes. With the NLP model, we took each theme generated in Phase 3 (from Th_1 to Th_n shown in Figure 1) and assigned the summary points and their original excerpts that fit under each theme. We accomplished this by calculating the cosine similarity (as a proxy for semantic similarity) between the embeddings for each summary point with the embeddings for the themes by using the NLP model: dolphin mistral. This should identify summary points that were similar to the themes and, thus, potential summaries relevant to that theme. We then matched the theme with the original data by tracking where the summary point originated and linking the theme with that original observation.

Researchers reviewed the output and checked their first eight cosine similarity scores to identify how well the output of the model found the examples from the summary points of each theme from Phase 3. Researchers could manually examine the generated data in this phase, and checked if the themes and excerpts from transcripts underlying specific themes made sense and represented the entire dataset.

In this case, the dolphin-mistral model calculated the cosine similarity (as a proxy for semantic similarity) between the embeddings for each summary point (from Phase 1) with the embeddings for the themes (from Phase 3). The model identified the summary points similar to the themes and then matched the theme with the original data. For example, the theme in the previous step is tracked in this step to illustrate how it was linked back to the original data with the cosine similarity score. The output of the model is shown in Figure 5.

The model provided the most similar original summary and original text from collected data. The NLP model considered the first eight similarity scores calculated between zero and one. The first eight similarity scores for the given example ranged between 0.72 and 0.84. The researchers went through the similarity scores and checked if the original excerpts fit under the identified themes. The researchers should make the changes if the model mismatched the original data with themes. At the end of this phase, the list of the codes is finalized by checking if they fit with the original dataset while they answer the research questions. By using cosine similarity scores, researchers can identify the themes that do not fit the original data and make changes to the themes accordingly.

Phase 5: Refining, Defining and Naming Themes

In traditional TA, researchers review the codes and themes and define them while grouping the themes together to identify the higher-level themes when necessary. Their goal is to make sure all themes and codes are consistent and informative while answering the research questions (Braun & Clarke, 2022; Byrne, 2022). When we used the GAI model, we asked the models to cluster the themes generated in Phase 3 to bring similar ones together. The workflow for this phase is shown in Figure 6 below.

The higher-level themes were represented from Ht_1 to Ht_m. The same model in Phase 3 was used to finalize the themes in this phase. We asked the NLP model to cluster similar themes and the GAI model to identify higher-level themes (Th). The concept of each higher-level theme was also explained by the model. In this way, we also made sure the themes from Phase 3 were not repetitive and these higher levels of themes

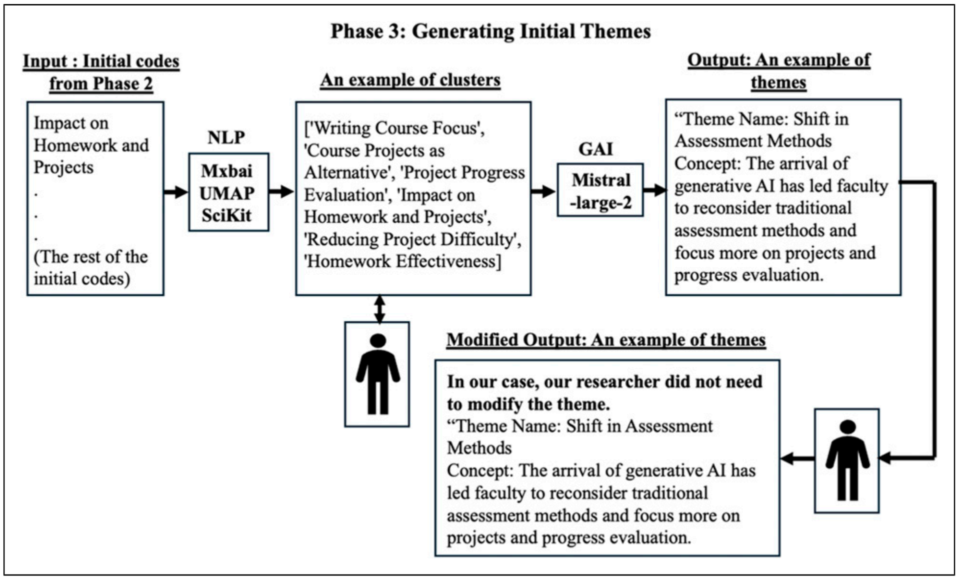

Phase 3: Generating Initial Themes
Input : Initial codes from Phase 2
Impact on Homework and Projects
.
.
.
(The rest of the initial codes)
NLP
Mxbai
UMAP
SciKit
An example of clusters
['Writing Course Focus', 'Course Projects as Alternative', 'Project Progress Evaluation', 'Impact on Homework and Projects', 'Reducing Project Difficulty', 'Homework Effectiveness']
GAI
Mistral -large-2
Output: An example of themes
"Theme Name: Shift in Assessment Methods
Concept: The arrival of generative AI has led faculty to reconsider traditional assessment methods and focus more on projects and progress evaluation.
Modified Output: An example of themes
In our case, our researcher did not need to modify the theme.
"Theme Name: Shift in Assessment Methods
Concept: The arrival of generative AI has led faculty to reconsider traditional assessment methods and focus more on projects and progress evaluation.

Figure 4. An Example of How Themes (Th) are Generated by Grouping Initials Codes (IC) With the GAI Model.

represented different aspects of the data. Researchers could go through these final themes, make changes if necessary, and write the definitions for each theme based on the explanation of GAI models. The prompt to generate the final themes with mistral-large-2 can be found in Appendix A.3. We used the same prompt as the one in Phase 3 since it was a similar process to generate higher-level themes.

In our case, we used the mistral-large-2 model to group the themes from Phase 4, the concept of the higher-level theme for each group. An example from the output of the model is shown in Figure 6.

The model was able to group the themes related to the evolution of assessment methods. In our case, the NLP model generated three clusters from 25 themes from Phase 3 and the GAI model identified 11 higher-level themes (from Ht_1 to Ht_{11}). After going through the list of the grouped themes under higher-level themes, their concept, and the model's observation, the final codebook can be finalized by the researchers.

Discussion

Using NLP and GAI models to conduct thematic data analysis is a developing approach and it brings significant advancement in qualitative research methods. The combination of both traditional and automated data analysis may bring various advantages to the researchers especially while working with

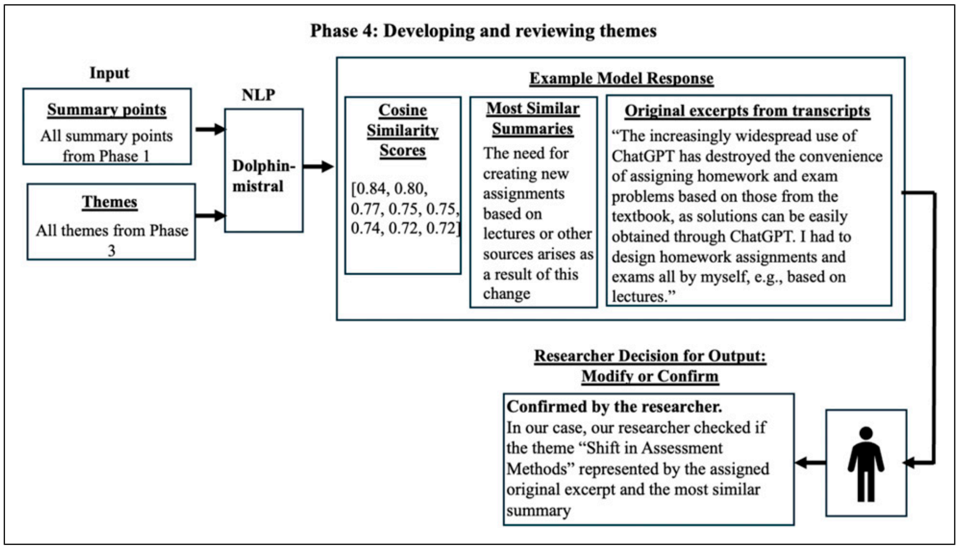

Phase 4: Developing and reviewing themes
Input
Summary points
All summary points from Phase 1
Themes
All themes from Phase 3
NLP
Dolphin-mistral
Example Model Response
Cosine Similarity Scores
[0.84, 0.80, 0.77, 0.75, 0.75, 0.74, 0.72, 0.72]
Most Similar Summaries
The need for creating new assignments based on lectures or other sources arises as a result of this change
Original excerpts from transcripts
"The increasingly widespread use of ChatGPT has destroyed the convenience of assigning homework and exam problems based on those from the textbook, as solutions can be easily obtained through ChatGPT. I had to design homework assignments and exams all by myself, e.g., based on lectures."
Researcher Decision for Output: Modify or Confirm
Confirmed by the researcher.
In our case, our researcher checked if the theme "Shift in Assessment Methods" represented by the assigned original excerpt and the most similar summary

Figure 5. An Example of How the Cosine Similarity Scores are Calculated to Review the Themes With Summary Points and Their Related Excerpts.

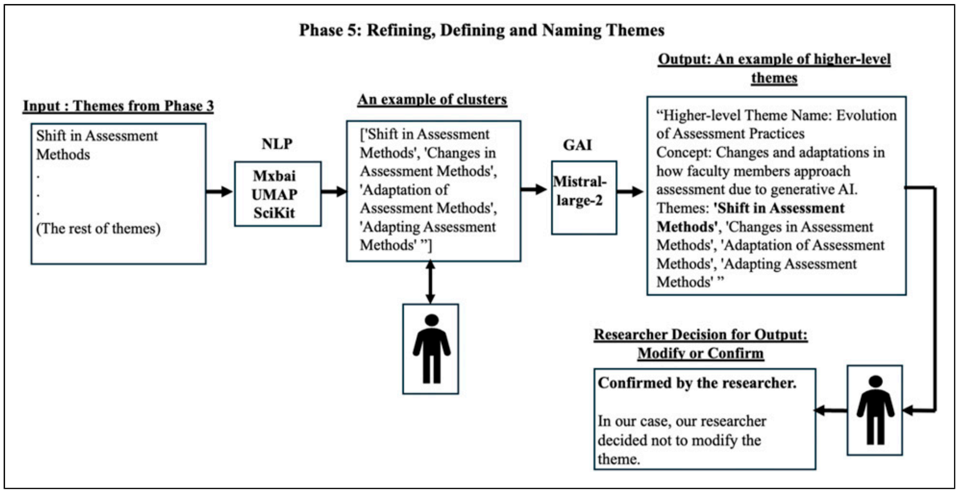

Phase 5: Refining, Defining and Naming Themes
Input : Themes from Phase 3
Shift in Assessment Methods
.
.
.
(The rest of themes)
NLP
Mxbai
UMAP
SciKit
An example of clusters
['Shift in Assessment Methods', 'Changes in Assessment Methods', 'Adaptation of Assessment Methods', 'Adapting Assessment Methods' "]
GAI
Mistral-large-2
Output: An example of higher-level themes
"Higher-level Theme Name: Evolution of Assessment Practices
Concept: Changes and adaptations in how faculty members approach assessment due to generative AI.
Themes: 'Shift in Assessment Methods', 'Changes in Assessment Methods', 'Adaptation of Assessment Methods', 'Adapting Assessment Methods' "
Researcher Decision for Output: Modify or Confirm
Confirmed by the researcher.
In our case, our researcher decided not to modify the theme.

Figure 6. An Example of How the Higher-Level Themes (Ht) are Generated From Themes (Th) by the GAI Model.

large datasets. However, we must acknowledge that it also brings some limitations and disadvantages. In this section, we aim to evaluate the pros and cons of using this hybrid method (we call it human in the loop process) as an innovative and ongoing approach to TA.

Advantages of Using GAI Models in Thematic Analysis

One of the advantages of NLP and GAI models is the shorter time it takes to process large datasets. This feature makes them time efficient for researchers instead of manually analyzing data. For example, familiarizing yourself with the data requires reading multiple times which is time-consuming for researchers. With smaller datasets this is not a significant hurdle; however, as datasets grow in size so, too, does the challenge of familiarization. GAI models such as llama3.1 rapidly process large datasets and can create summary points from the chunks of text in a short time compared to the amount of time a person could spend. Moreover, NLP algorithms can accelerate the process of grouping similar patterns in data and create clusters and GAI models can take these clusters to generate codes, define them, and provide examples. Saving time and collaborating with text analysis tools can also help to reduce the cost of analysis. Most of the NLP and GAI models are open source which means anyone who would like to use them has access to work with them. Using them might reduce the number of people who need to be involved in the data analysis process or people can focus on the different aspects of the research with the time they could spend on analysis. For example, a researcher could use this time with more meaning-making with the different codes generated, or understanding the participants' living experiences and background to interpret findings that would not misrepresent cultural context.

Another advantage of using NLP and GAI for TA is having a systematic and comprehensive output from analysis. Sometimes, researchers get overwhelmed with the process of organizing and sorting data and miss some insights from data. Since the models work in a data-driven approach for identifying codes and themes, they are able to handle large datasets and find patterns to generate codes and identify themes. Therefore, NLP and GAI models are a great aid to support researchers in seeing the scope and direction of their data that might not be immediately apparent to human researchers when they conduct TA manually.

Limitations of NLP and GAI Models in Thematic Analysis

While NLP and GAI models bring advantages to TA such as time and cost efficiency and the ability to handle big datasets, these technologies still have limitations researchers take into consideration seriously while using them. We share some based on our observations while doing data analysis for our case study.

One of the limitations is the replicability of data analysis with NLP and GAI models. While these models have algorithms and trained datasets, they have been improved continuously, and the output of the model may differ. For example, if we use the same model (e.g., Llama-3.1-8b), the model and its weights should be kept frozen. If we set the temperature to 0, we should get the same output for the same input. The given case in the previous section was conducted in April 2024, and we must acknowledge that the output for the themes may differ if different versions of llama or other models are used with the same datasets.

Another limitation of using NLP and GAI models in TA is the possibility of misinterpretation of the context of the data. There are various reasons for this, such as the bias in training data (Li et al., 2024; Liang et al., 2021). The training data may not capture the cultural differences, sarcasm, or analogies. This may cause the models to generate wrong codes and themes in TA. Also, the models work with algorithms, and they rely on the frequency of the occurrence of the patterns when creating outputs as codes and themes. In qualitative research, sometimes reoccurrence of patterns does not completely uncover the depth of insights that can be gained from data analysis. Therefore, the models may not be able to reflect on the data as much as researchers. The level of understanding and intuition of the models may not be at the same level of researchers.

The availability of the NLP and GAI models and the idea of replacing people with them have raised ethical concerns due to the concerns about losing the authenticity of qualitative research (Lund et al., 2023). One of the concerns is the ownership of data analysis and how original the findings can be with the aid of the NLP and GAI models. These models provide coherent and structured outputs. However, the need for reflections and evaluation of codes in TA still requires the engagement of researchers with the data and the authenticity of researchers for synthesizing findings. Therefore, relying on the models may only bring concerns about maintaining ethical standards for qualitative research.

The limitations of using the NLP and GAI models for TA may not be restricted to what we mentioned in this section. Researchers should be actively involved in each step for TA and take notes, make decisions on how to conduct the research with the aid of these models, and take action and make changes in the process when it is necessary. Researchers should be open to using these advanced tools but should not think they are the replacement for their roles in TA. They should make sure the rich rigor of data is uncovered while the integrity, reliability, and validity of their findings are still maintained.

Prompt Design for GAI Models in Thematic Analysis

The prompts provided guidance for GAI models in various phases of TA process, including summarizing, generating initial codes, identifying themes, and organizing them hierarchically. Well-constructed prompts are essential, as they ensure that the model outputs align with research goals (Giray, 2023; Lu et al., 2022). Clear instructions help the model follow the intended task effectively. Based on experience throughout the research we conducted in our case, here is the information on how to structure prompts.

Persona assignment: The positionality of the model is important to understand the context of the data you provide for data analysis. By assigning a persona, the model can enable more human-like interactions and provide more accurate information and contextually relevant outputs (Araujo & Roth, 2024; Salewski et al., 2023). Consistency in responses is crucial for TA to structure the emerging codes and group them under themes.

Data background: We provided accurate and sufficient information about the data to help the model grasp the nuances of the text it was analyzing. Providing the data type (e.g., written responses to open-ended survey questions; interview transcript segments) and the context in which the data were collected (e.g., a study of faculty beliefs about assessment) informs the model about what kind of data it will analyze.

Clear and precise tasks: When working with language models, it can be helpful to treat them like simple programs that understand instructions to a limited extent and may become confused easily. Therefore, the user's expectations of the model should be clearly defined, and instructions should be precise to help the model understand the requirements of each task. To that end, instructing the model to use a chain of thought in prompts improves the model's understanding the complex reasoning (Wei et al., 2022).

Guiding examples: To make the tasks clearer to the models, we should provide examples for each requirement. This prompting technique is called few-shot prompting (Dang et al., 2022). Demonstrating how the model should generate the codes and themes with examples can serve as a reference to improve the performance of the model and generate more accurate outputs.

Comparing the Previous GAI Models for Thematic Analysis

Recently, there has been an increasing number of studies for AI-assisted TA. In this section, we focus on some of them that are the most similar to our study regarding research purpose and the use of the methods. We chose the study conducted by Mathis et al. (2024) and De Paoli (2024) to discuss further. While the majority of existing literature used the versions of Chat GPT, these two studies followed similar processes regarding the GAI models such as Llama and Whisper. Unlike these two studies, we did not aim to compare the analysis done by the GAI models and human researchers but to guide the researchers who are new to the integration of NLP and GAI models to the TA phases developed by Braun and Clarke (2022).

De Paoli (2024) used the GPT 3.5 Turbo model for inductive TA and compared its output with the manually coded findings. They further discussed the replicability of manual coding by using GAI models. They highlighted that identifying the patterns for themes via GAI tools was efficient and useful. However, to identify the nuanced themes that require interpretation and insight, the engagement of researchers with the data analysis was a necessity. They recommended focusing on building established procedures and prompting to ensure the quality and validity of the qualitative analysis. In our study, we presented how the prompts can be structured and explained in previous subsection, and the actual prompts we used for our case study are shown in the Appendix. The prompt generation and how it affects the model output should be further studied as a part of the development of qualitative research methods with NLP and GAI models. Our work also varied from their work in our approach to the temperature parameter setting. Whereas their work set a model temperature value above 0.5, we elected to set the temperature to 0 to improve replicability. Setting the temperature to 0 makes these inherently probabilistic models more deterministic, which we preferred to ensure other researchers could generate the same output as we did when given the same input.

Mathis et al. (2024) used the GAI model Llama-2-70B model to create codebooks related to healthcare interviews. Similar to most existing literature, they compared human researcher coding with the output of GAI models. Their main focus was Phase 2 where Braun and Clarke (2022) suggested identifying the initial codes. Additionally, they merged Phases 3 and 5 in their method. They used cosine similarities to compare the GAI model's output with the manually generated codebook. They suggested the collaboration of researchers and GAI models to maintain the validity of qualitative research and highlighted the importance of providing coherent prompts for the GAI models to generate the themes.

Overall, there are differences in how each paper interpreted and followed the phases for TA suggested by Braun and Clarke (2006, 2022). For example, De Paoli (2024) did not use any GAI models for the Phase 1 (getting familiar with data) phase and recommended preparing the raw data to the next phase by cleaning and converting the data formats to txt. Mathis et al. (2024) showed Phase 1 as only converting mp3 recordings to transcripts to use in the following phases. Thus, both studies did not use any AI-assisted approach in Phase 1. We generated summary points from cleaned data in Phase 1 where we aimed to make researchers not go through the whole dataset, as well as to reduce the amount of data the NLP and GAI models in the following phases. Thus, our study brought a different approach to autonomic Phase 1 as well as the rest of the steps in TA.

We commonly suggest that GAI and NLP can reduce the cost and labor of TA, especially with large data. Also, the importance of researcher oversight was emphasized in both studies. However, we showed how the researchers can be involved in the process in more detail. We offered structured guidance on how researchers can be involved while integrating NLP and GAI models in TA phases which provides a practical road map for those who are interested in this innovative qualitative research method.

Implications and Conclusion

Our study provided a road map to researchers for implications of GATA and presented the limitations, advantages, and significant points to consider during the process of TA. Our method and case study suggest that NLP and GAI technologies can streamline the steps for TA suggested by Braun and Clarke (2022). The nature of TA is iterative and time-consuming and GATA helps save time and reduce the intensity of coding and generating themes (De Paoli, 2024; Mathis et al., 2024). The efficiency of GATA is undeniable especially for the studies including large datasets. The increasing accessibility of AI-assisted tools presents both advantages and challenges. One benefit of GAI models is the potential decrease in labor costs and increased automation in data analysis, which may encourage more data collection for qualitative studies. Using the advantages of GATA, researchers can spare their time to collect more data from diverse participant pools leading to rigorous and more inclusive findings (Gamieldien et al., 2023; Lixandru, 2024). However, the high cost of necessary hardware poses a barrier, risking inequalities between researchers who can afford it and those who cannot. As these models become more accessible, GAI technology may be widely adopted, reshaping the roles of researchers and computational tools in qualitative research methods.

While the NLP and GAI models are improved, the newer versions can provide more robust coding, ensuring uniform applications across the collected datasets, and leading to more reliable findings (Li et al., 2024). However, with the current abilities of the NLP and GAI models, researchers should remain engaged with the process of TA phases developed by Braun and Clarke (2022). In the current era, maintaining a human-in-the-loop approach can address the current limitations of GATA, such as biases and, thus, potential misinterpretations of collected data (Davison et al., 2024; Li et al., 2024; Perkins & Roe, 2024b). To mitigate some of the concerns about reliability and biases, researchers should also keep improving the NLP and GAI models and establish ethical guidelines for the GATA process.

Acknowledgments

The authors are thankful for the participant's time and insights during the data collection process.

ORCID iDs

Isil Anakok https://orcid.org/0000-0002-1572-8024
Andrew Katz https://orcid.org/0000-0002-3554-9015

Ethical Statement

Ethical Approval

This study received ethical approval from the Virginia Tech Institutional Review Board (IRB) (approval # 21-639) on December 22, 2021.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 2113631 and 2107008. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material for this article is available online.

References

Araujo, P. H. L. d., & Roth, B. (2024). Helpful assistant or fruitful facilitator? Investigating how personas affect language model behavior (arXiv:2407.02099). arXiv. https://doi.org/10.48550/arXiv.2407.02099

Arora, A., Kaffee, L.-A., & Augenstein, I. (2023). Probing pre-trained language models for cross-cultural differences in values (arXiv:2203.13722). arXiv. https://doi.org/10.48550/arXiv.2203.13722

Bhaduri, S. (2018). NLP in Engineering Education—Demonstrating the use of Natural Language Processing Techniques for Use in Engineering Education Classrooms and Research [Virginia Tech] https://hdl.handle.net/10919/82202

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

Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57–71). American Psychological Association. https://doi.org/10.1037/13620-004

Braun, V., & Clarke, V. (2022). *Thematic analysis: A practical guide*. Sage.

Byrne, D. (2022). A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity*, *56*(3), 1391–1412. https://doi.org/10.1007/s11135-021-01182-y

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). How to prompt? Opportunities and challenges of zero- and few-shot learning for human-AI interaction in creative applications of generative models (arXiv:2209.01390). arXiv. https://doi.org/10.48550/arXiv.2209.01390

Davison, R. M., Chughtai, H., Nielsen, P., Marabelli, M., Iannacci, F., Van Offenbeek, M., Tarafdar, M., Trenz, M., Techatassanasoontorn, A. A., Díaz Andrade, A., & Panteli, N. (2024). The ethics of using generative AI for qualitative data analysis. *Information Systems Journal*, *12504*(5), 1433–1439. https://doi.org/10.1111/isj.12504

De Paoli, S. (2024). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, *42*(4), 997–1019. https://doi.org/10.1177/08944393231220483

Elali, F. R., & Rachid, L. N. (2023). AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns*, *4*(3), 100706. https://doi.org/10.1016/j.patter.2023.100706

Gamieldien, Y., Case, J. M., & Katz, A. (2023). *Advancing Qualitative Analysis: An Exploration of the Potential of Generative AI and NLP in Thematic Coding* (SSRN Scholarly Paper 4487768). *Social Science Research Network*. https://doi.org/10.2139/ssrn.4487768

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, *51*(12), 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Hatch, J. A. (2023). *Doing qualitative research in education settings* (2nd ed.). State University of New York Press.

Hesse-Biber, S. (2010). Qualitative approaches to mixed methods practice. *Qualitative Inquiry*, *16*(6), 455–468. https://doi.org/10.1177/1077800410364611

Katz, A., Fleming, G. C., & Main, J. (2024). Thematic analysis with open-source generative AI and machine learning: A new method for inductive qualitative codebook development (arXiv:2410.03721). arXiv. https://doi.org/10.48550/arXiv.2410.03721

Kramer, O. (2016). Scikit-learn. In O. Kramer (Ed.), *Machine learning for evolution strategies* (pp. 45–53). Springer International Publishing. https://doi.org/10.1007/978-3-319-33383-0_5

Lee, S., Shakir, A., Koenig, D., & Lipp, J. (n.d) *Open source strikes bread—new fluffy embedding model—blog*. Mixedbread. Retrieved November 1, 2024, from. https://www.mixedbread.ai/mxbai-embed-large-v1

Li, M., Enkhtur, A., Yamamoto, B. A., Cheng, F., & Chen, L. (2024). Potential societal biases of ChatGPT in higher education: A scoping review (arXiv:2311.14381). arXiv. https://doi.org/10.48550/arXiv.2311.14381

Li, X., & Li, J. (2024). AnglE-optimized text embeddings (arXiv:2309.12871). arXiv. https://doi.org/10.48550/arXiv.2309.12871

Liang, P. P., Wu, C., Morency, L.-P., & Salakhutdinov, R. (2021). Towards understanding and mitigating social biases in language models. In Proceedings of the 38th international conference on machine learning (pp. 6565–6576). ACM. https://proceedings.mlr.press/v139/liang21a.html

Lixandru, I.-D. (2024). The use of artificial intelligence for qualitative data analysis: ChatGPT. *Informatica Economica*, *28*(1/2024), 57–67. https://doi.org/10.24818/issn14531305/28.1.2024.05

Llama. (n.d.). Llama. Retrieved May 4, 2025, from. https://www.llama.com/

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity (arXiv:2104.08786). arXiv. https://doi.org/10.48550/arXiv.2104.08786

Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, *74*(5), 570–581. https://doi.org/10.1002/asi.24750

Mathis, W. S., Zhao, S., Pratt, N., Weleff, J., & De Paoli, S. (2024). Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Computer Methods and Programs in Biomedicine*, *255*, 108356. https://doi.org/10.1016/j.cmpb.2024.108356

McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold approximation and projection for dimension reduction (arXiv:1802.03426). arXiv. https://doi.org/10.48550/arXiv.1802.03426

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *International Journal of Qualitative Methods*, *22*(1), 16094069231211248. https://doi.org/10.1177/16094069231211248

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality*, *15*(2), 1–21. https://doi.org/10.1145/3597307

Perkins, M., & Roe, J. (2024a). Academic publisher guidelines on AI usage: A ChatGPT supported thematic analysis. F1000Research, 12(390-395), 1398. https://doi.org/10.12688/f1000research.142411.2

Perkins, M., & Roe, J. (2024b). The use of Generative AI in qualitative analysis: Inductive thematic analysis with ChatGPT. *Journal of Applied Learning and Teaching*, *7*, 390-395. https://journals.sfu.ca/jalt/index.php/jalt/article/download/1585/753/5729

Saldaña, J. (2013). *The coding manual for qualitative researchers* (2nd ed.). Sage.

Saldaña, J. (2014). Coding and analysis strategies. In P. Leavy (Ed.), *The oxford handbook of qualitative research*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199811755.013.001

Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., & Akata, Z. (2023). In-context impersonation reveals large language models' strengths and biases (arXiv:2305.14930). arXiv. https://doi.org/10.48550/arXiv.2305.14930

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, *23*(1), 16094069241231168. https://doi.org/10.1177/16094069241231168

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing system* (Vol. 35, pp. 24824–24837). MIT Press.

Ye, H., Liu, T., Zhang, A., Hua, W., & Jia, W. (2023). Cognitive mirage: A review of hallucinations in large language models (arXiv:2309.06794). arXiv.