ThemeViz: Understanding the Effect of Human-AI Collaboration in Theme Development with an LLM-enhanced Interactive Visual System

DAYE KANG, Cornell University, USA
ZHUOLUN HAN, Cornell University, USA
JIAHE TIAN, Cornell University, USA
MUHAN ZHANG, Cornell University, USA
JEFFREY M. RZESZOTARSKI, Cornell University, USA

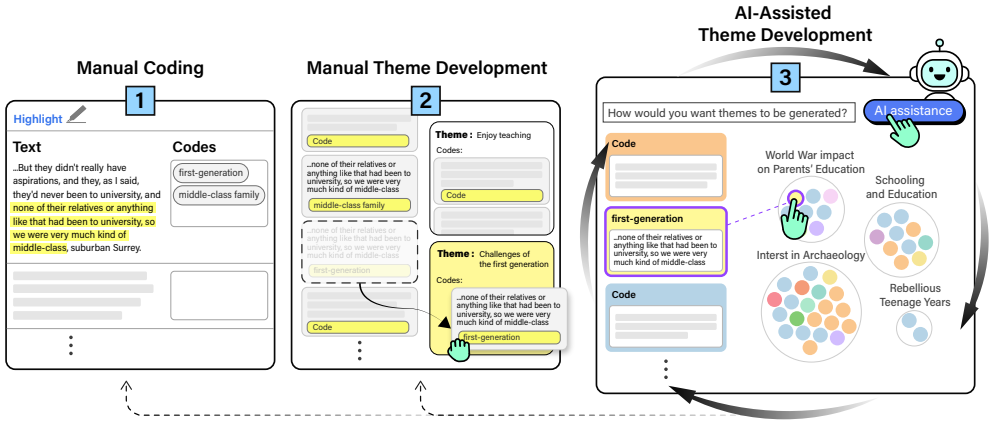

Fig. 1. ThemeViz is a web application that utilizes large language models and interactive visualizations to assist qualitative researchers during theme development in thematic analysis. The system supports manual coding (1) and manual theme development (2) to support researchers' autonomy. The system also supports AI-assisted theme development (3) to foster human-AI collaboration in theme development. Users can generate and explore multiple AI-generated theme versions through interactive prompting and review.

This paper explores the potential role of AI, e.g., large language models (LLMs), in supporting theme development in thematic analysis. While prior applications of AI in qualitative data analysis have focused on supporting coding, we investigate whether LLMs can effectively contribute as collaborators in the more abstract and conceptual phases of qualitative analysis, specifically theme development. Despite growing interest in AI as a collaborator in theme development, there is limited empirical evidence on designing AI-assisted tools while supporting user autonomy and understanding researcher interaction with AI-assisted theme development. To address this gap, we designed ThemeViz, an interactive system that uses GPT-4 to generate and visualize multiple versions of themes based on user input while allowing researchers to maintain control

Authors' Contact Information: Daye Kang, dk564@cornell.edu, Cornell University, Ithaca, NY, USA; Zhuolun Han, zh429@cornell.edu, Cornell University, Ithaca, NY, USA; Jiahe Tian, jt828@cornell.edu, Cornell University, Ithaca, NY, USA; Muhan Zhang, mz574@cornell.edu, Cornell University, Ithaca, NY, USA; Jeffrey M. Rzeszotarski, jeffrzeszotarski@gmail.com, Cornell University, Ithaca, NY, USA.

ACM 2573-0142/2025/11-ARTCSCW494
https://doi.org/10.1145/3757675

through manual coding and theme development. Our study examines the effectiveness of this human-AI collaboration approach in iterative theme development and its implications for future designs.

CCS Concepts: • **Human-centered computing → User centered design**.

Additional Key Words and Phrases: Intelligent system, Iterative theme development, Large language model, Interactive visualization, Human-AI collaboration, Thematic analysis

ACM Reference Format:
Daye Kang, Zhuolun Han, Jiahe Tian, Muhan Zhang, and Jeffrey M. Rzeszotarski. 2025. ThemeViz: Understanding the Effect of Human-AI Collaboration in Theme Development with an LLM-enhanced Interactive Visual System. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW494 (November 2025), 29 pages. https://doi.org/10.1145/3757675

1 Introduction

HCI research on AI-assisted tools has focused on supporting the time-consuming and laborious task of thematic analysis, primarily in *coding*, which entails labeling segments of text data, while other tasks, such as *theme development*, which involves grouping related codes to identify patterns in the data, receive relatively less support [27, 28, 44, 53]. One reason coding received attention for AI assistance is that it involves clear, well-defined tasks, such as labeling data based on a predefined codebook, making it easier for AI to automate and process efficiently.

In contrast, tasks like **theme development**, which occur *after* coding, involve a more complex, conceptual process that computational tools struggle to replicate. Instead of simply tagging pieces of data (coding), theme development requires grouping the labels generated during coding into broader patterns or themes that capture deeper meanings and relationships across the dataset. This process demands that the researcher synthesize and interpret the underlying connections between coded elements, recognize trends, and construct a cohesive narrative that represents the dataset. As a result, theme development is more interpretative and requires higher-level thinking and subjective judgment, making it challenging to automate with computational tools [49]. For example, natural language processing technologies have been criticized for their limited ability to capture the nuances such as subtle distinctions in meaning, tone, and context in text that are essential for thematic interpretation. [9, 10].

With the recent rise of large language models (LLMs) for processing complex text corpora, recent studies suggest new opportunities for LLMs to *collaborate* with researchers in interpretative tasks like theme development by offering alternative interpretations of data that researchers may not have initially considered [38, 49, 65]. In light of this work, there is a possibility that providing new interpretations distinguishes LLMs from conventional analytical tools that merely process data. LLMs can generate novel perspectives and challenge researcher assumptions which are key characteristics of intellectual collaboration [25, 34, 51]. Because of this expanded scope, LLMs may act more as collaborator instead of tool. This raises the question of whether LLMs could assume a collaborative role in interpretive tasks by providing *alternative interpretations* of data.

However, we have a limited understanding of (1) *how* to design effective *human-AI collaboration* for interpretive tasks like theme development and (2) the interaction of researchers with AI-based theme development assistance. Research in CSCW has explored the potential for human-AI collaboration in qualitative data analysis, emphasizing that collaboration could be successful when researchers' autonomy is respected, and the AI does not interfere with their analysis without consent [23, 37]. This underscores the importance of supporting user autonomy in human-AI collaboration within qualitative data analysis.

Furthermore, while much of the existing research that has explored the capability of LLMs in thematic analysis has primarily focused on the use of traditional prompt-and-response interfaces

(e.g., ChatGPT, APIs in Python scripts, and Jupyter notebooks) in thematic analysis [65], these tools do not fully meet the needs of qualitative researchers. When developing themes, researchers often rely on visual aids, such as thematic maps, to visually outline and organize their themes [7]. These visual aids provide a clear overview of candidate themes [7]. Text-based prompt-and-response interfaces (e.g., ChatGPT) fail to offer this level of support, making it more challenging for researchers to maintain an overarching view of the themes. Additionally, managing textual data in prompting can be difficult without system-level support, as researchers must manually input and update raw data in prompt-and-response interfaces that are not designed specifically for qualitative data analysis.

In this paper, we seek to bridge this gap by designing 'ThemeViz,' an LLM-enhanced system that enables researchers to develop themes with AI collaboratively. ThemeViz provides three forms of support to users: 1) **Autonomy support**: Building upon previous CSCW findings on the importance of supporting researchers' autonomy [23, 37], our system facilitates manual coding and theme development rather than fully automating the theme development process through AI. By integrating researchers' manual analysis into theme development, the system preserves and emphasizes researchers' autonomy throughout the process. 2) **Interactive visualization support**: ThemeViz presents developed themes through an interactive bubble chart, allowing users to easily gain an overview of the themes. This interactive design enables users to explore source data and LLM responses by displaying related text extracts and codes through a bubble visualization metaphor. 3) **Prompting support**: ThemeViz scaffolds interactions with its LLM by embedding the model into its system, implicitly automating prompts by managing metadata that contains information about the raw text, code, and text extracts behind the scenes.

To investigate the efficacy of ThemeViz, we conducted a lab experiment combined with semi-structured interviews with 28 qualitative researchers. Through these experiments, we investigated the following research questions:

RQ1 How useful is ThemeViz for theme development compared to conventional prompting interfaces like ChatGPT?
RQ2 To what extent does ThemeViz's design encourage users to view its AI assistant as a collaborative partner in theme development compared to traditional interfaces like ChatGPT?
RQ3 What are the limitations of AI assistance in ThemeViz?

Note that enhancing LLM performance through model training is not the focus of our research. Instead, we investigate the impact of *design and interaction* with existing LLM models in the context of theme development. Specifically, we aim to understand the differences between a refined graphical interface and a traditional text-based prompt-response interface, as well as the implications of integrating LLMs more deeply into interpretative tasks like theme development.

Our findings reveal that ThemeViz effectively aids qualitative researchers in theme development by allowing them to explore more iterations and view data from multiple perspectives. Its features for prompting and visualizing AI responses were beneficial. However, unlike previous research suggesting that LLMs could serve as collaborative partners in interpretative tasks, researchers did not perceive the ThemeViz AI assistant (or ChatGPT) as a collaborator. This perception was attributed to the AI's lack of agency and the absence of reciprocal, interactive discussions. Additionally, ThemeViz's AI assistance faced limitations regarding data privacy and AI bias concerns.

We contribute to the HCI and CSCW communities by: 1) investigating human-AI collaboration in interpretive tasks, specifically theme development, an area that has been largely overlooked. We highlight the potential and limitations of human-AI collaboration in theme development; 2) providing empirical evidence on how tailored human-AI collaboration tools designed for qualitative data analysis impact usability and shape user perceptions of AI as an effective collaborator. 3)

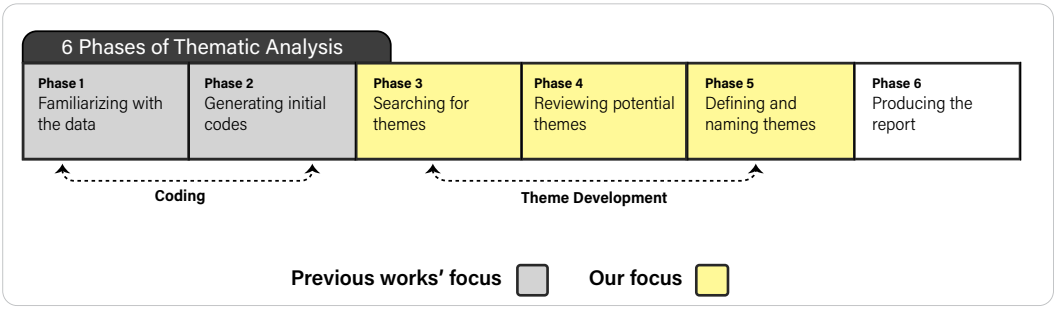

Fig. 2. Within the 6 phases of thematic analysis, we focus on supporting phases 3,4, and 5 where theme development occurs. Prior research has examined phases 1 and 2 [27, 28, 44, 53]. We developed this diagram based on theories of thematic analysis [7, 10].

Finally, our study highlights limitations in current human-AI interaction (e.g., passive conversation, lack of AI customization within qualitative data analysis support tools). Based on our findings, we offer design implications for enhancing future human-AI interactions in interpretative data analysis tasks such as theme development.

2 Related work

In this section, we begin by defining thematic analysis. Then, we outline how researchers developed fundamental knowledge on the impacts of LLMs on qualitative data analysis. Finally, we highlight specific analysis systems.

2.1 Thematic analysis

Understanding qualitative data in human-computer interaction user research often involves thematic analysis (TA), which aims to identify, analyze, and interpret patterns of meaning, known as themes, within the data [6].

In TA there are six phases [7, 11], as illustrated in figure 2. Here, we explain each phase.

Phase 1 **Familiarizing yourself with the data**: This stage involves deep engagement with the data through reading.

Phase 2 **Generating initial codes**: This stage involves assigning brief summaries or labels, known as codes, to data segments relevant to the research questions.

Phase 3 **Searching for themes**: This stage entails grouping similar codes by shared meanings to develop themes. In this paper, we define 'themes' following Braun and Clarke's definition, 'patterns or shared meaning, united by a central concept or idea' [9, 19]. Themes differ from topics because they are based on meaning; this allows themes to be constructed by bringing together data that may seem disparate on the surface to reveal deeper insights. In addition, themes can be developed based on less frequent but important data, depending on the research question [19].

Phase 4 **Reviewing potential themes**: In this stage, researchers refine themes for clarity and coherence. While refining themes, researchers rigorously evaluate themes against codes and the entire dataset to reconstruct themes to better align with their objectives [7]. During this iterative revision process, researchers' understanding of the data deepens, if they find better themes researchers may dissect broader themes into more specific themes or all themes can be reconstructed after eliminating initial themes in light of a new perspective [7, 8, 11].

Phase 5 **Defining and naming themes**: Once researchers develop quality themes in phase 4, they name themes and articulate the meaning of each identified theme.

Phase 6 **Producing the report**: The final phase entails preparing a comprehensive document detailing the identified themes, including descriptions, supporting quotes, interpretations, and a clear narrative that addresses the research question, effectively conveying the key insights uncovered through the analysis.

In this paper, we define *theme development* as encompassing phases 3, 4, and 5, as these phases are directly related to the development of themes, and we focus on supporting theme development as this step has been less explored within the context of AI assisted analysis tool. Instead, previous works focused on supporting *coding* phase of the analysis [27, 28, 44, 53]. For instance, one research focused on supporting the collaboration of qualitative researchers to develop codebook with AI coding assistance tool [27]. Another study explored supporting coding by semi-automating the process through interactively defined code rules and supervised machine learning, allowing researchers to extend coding to unseen data [53].

Supporting theme development is important, as it is a time-consuming, labor-intensive process that demands significant cognitive effort to abstract, synthesize, and interpret themes and meanings within data [4, 7, 9, 45]. To be more specific, theme development is laborious because themes do not simply emerge from the data; instead, they are *actively developed* by qualitative researchers. This active theme development entails *iteratively* revising and exploring multiple versions of themes to arrive at those that best represent the data [7]. Even constructing a single version of themes is cognitively demanding as researchers balance abstract interpretation with precise representation of nuances without oversimplification. The repeated revision and exploration of themes require significant mental energy, as researchers work to create a cohesive thematic structure that accurately captures both broad patterns and specific data details. This involves accurately recalling a large body of text as well as finding relationships among disparate elements. If the themes do not fit the data, researchers may need to tweak themes by collapsing, splitting them, or even discarding them and starting again [7]. This iterative nature of theme development makes the work labor intensive and cognitively demanding.

Studies suggest new opportunities for LLMs to *collaborate* with researchers in interpretative tasks like theme development by offering alternative data interpretations that researchers may not have initially considered, potentially supporting an iterative theme development process by enabling exploration of multiple theme versions with AI [38, 49, 65]. However, it remains unclear whether qualitative researchers would find such interaction useful or how to design LLM-enhanced tools to support tasks requiring high levels of abstraction and conceptualization, like theme development, while preserving user autonomy [23, 37]. To address this gap, we designed and evaluated an LLM-enhanced system, "ThemeViz," for its usefulness in theme development.

2.2 LLMs in qualitative data analysis

We now review literature on the use of LLMs in qualitative data analysis to understand their *performance*. It is important to note that this section focuses on LLMs independently, without examining *systems that integrate LLMs*; systems are further discussed in § 2.3.

Recent studies explored the potential of adopting GPT in qualitative data analysis. For instance, one paper examined the capability of GPT in narrative analysis [18]. Other streams of work investigated the capability of GPT for *deductive* coding tasks. Recent studies found GPT-3.5 can often perform deductive coding (applying predefined code to data similar to data labeling) at levels comparable to humans [16, 62] and GPT-4 performs with high intercoder reliability [21, 39]. A self-experimental study (where the first author acts as a participant in the study) investigated

the use of GPT3.5-Turbo in inductive thematic analysis (themes developed from data instead of developed from pre-defined theories), showcasing its potential [49].

Furthermore, a comparison of themes generated by humans and ChatGPT revealed that AI-generated themes are somewhat similar to those created by humans, [32] indicating the viability of LLMs in this context. This study suggests human-AI collaboration for theme development, utilizing an AI's efficiency alongside human expertise in identifying subtle nuances [32]. Zhang et al. crafted a framework that employs ChatGPT for thematic analysis, grounded in semi-structured interviews [65]. Their work goes a step further and suggests the potential of LLMs as a co-researcher.

Prior literature demonstrates the applicability of LLMs in thematic analysis and suggests potential for human-AI collaboration. However, these studies investigated the utilization of bare LLMs without support from interactive systems and explored the capabilities of LLMs only within programming platforms such as Jupyter Notebooks. These platforms do not provide user-friendly interactions to streamline the process of engaging with the model (especially in the case of practitioners who are not usually programmers). Furthermore, this approach demands users to design their own prompt to handle sending data and receiving data, while prompting is challenging for novice users who are not familiar with LLMs [63].

An interactive system with a graphical user interface (GUI) ought to help users avoid interacting with the model at a raw level, allowing them to focus more on the analysis itself. Therefore, it remains unclear 1) how a system designed to support qualitative research can further enhance the theme development process compared to using LLMs outside such a system, and 2) the exact degree of enhancement that such a system design can provide. To bridge this gap, we designed and implemented a system, then conducted a lab experiment with qualitative researchers to measure the effect of the system design compared to a general, non-tailored LLM-embedded system.

2.3 Systems for LLM-assisted theme development

In this section we focus on reviewing ***systems and tools*** which employ LLMs to support qualitative data analysis. A recent study examined the use of LLMs in theme development within a qualitative data analysis tool, offering primary code group suggestions. However, since the tool provided AI support across multiple phases of thematic analysis (e.g., coding) within a human-human collaborative context, the focus was on enhancing human-human collaboration with AI rather than solely assessing the effects of human-AI interaction in the theme development phase. This made it challenging to isolate the specific impact of AI assistance on theme development [27].

Another recent study focused more on the theme development stage of thematic analysis. They introduced a system called QualiGPT, which generates themes in a coherent tabular format based on an imported raw dataset [64]. The system focuses on developing an initial theme *straight from the raw data*, without supporting autonomy of researchers by providing an opportunity for researchers to develop themes on their own before developing themes using AI. Enabling researchers to manually code and group codes is essential, as it supports their *autonomy* [23, 37], and it enables researchers to reflect and understand their data before developing themes [7]. Furthermore, the system does not provide any visual support to support data sense-making, and the authors recommend adopting data visualizations [64] as future work.

Our system overcomes limitations of previous work by supporting researchers' autonomy with manual coding and theme development. We intentionally integrated AI assistance *only* in the theme development phase to gain a clearer understanding of human-AI collaboration in this phase. Additionally, our research interest is understanding the impact of AI assistance on conceptual tasks like theme development, rather than on lower-level coding. In addition, we adopt interactive visualizations to further support sensemaking of AI-generated themes.

3 Design goals

We aim to design an LLM-embedded system which fits seamlessly into qualitative researchers' workflows and follows best practices sourced from literature. Our target is to enhance the efficiency and reduce the workload of theme development through human-AI collaboration. We reviewed theories of thematic analysis and existing qualitative data analysis tools to develop seven design goals for ThemeViz.

DG1: Support user autonomy by supporting manual coding and manual theme development. While allowing an AI to generate themes directly might bypass the laborious theme building process, there are significant risks in this approach. Not giving users a chance to review themes against the data could result in a user either deferring completely to the AI's judgment, growing to distrust the AI due to lack of transparency, or discouraging the user from following best practices. In other words, over- and under-reliance both pose issues to user efficacy and autonomy [23]. Previous research highlighted that without providing enough autonomy in analysis, qualitative researchers are reluctant to utilize AI assistance in their workflow [23, 37]. Therefore, our system supports (1) manual coding and (2) manual theme development in addition to AI-supported theme development. Users retain the full capability to edit individual themes and revisit the previous phases to revise codes and re-read raw data.

Since our research interest is to find the effectiveness of AI assistance in theme development, we did not design a special interface for manual coding and manual theme development. Instead, we adopted industry-level standard practices (e.g., Atlas [2] and Dovetail [20]), having basic and widely used features like highlighting, coding with drop-down selection, and drag and drop grouping.

DG2: Promote efficient theme development through LLM-assisted theme suggestions. A considerable number of studies have noted that qualitative data analysis is time-consuming and labor-intensive [16, 26–28, 44, 57], leading researchers to develop tools to make the analysis process more efficient. While previous work focused on improving the efficiency of coding [16, 26–28, 44, 57], our goal is to reduce the time and effort required for theme development while offering AI-generated theme suggestions to inspire new perspectives on the dataset.

Qualitative data analysis tools like Atlas.ti and NVivo support theme development using interactive features like code managers. They enable users to modify codes and group them to develop themes interactively. However, these features often require manual theme development. Since creating additional themes or even discarding existing themes and reconstructing them is common in theme development [7], such a manual process can be time-consuming and laborious. (Please refer to section 2.1 for more details on why theme development is cognitively demanding, time-consuming, and labor-intensive [4, 7, 9, 45].)

Recent studies have also noted that AI theme generation may reduce the time it takes to develop themes and inspire researchers by offering perspectives on the data [27, 64]. AI generated themes might act both as starting seed ideas and prompts for reflection for the researcher. We propose using LLM-assisted theme development to support the recursive theme revising process, enabling researchers to quickly iterate multiple versions of themes with less manual effort.

DG3: Encourage data exploration by prompting while respecting researchers' perspectives. Thematic analysis values researchers' subjectivity and interpretation [10]. Researchers should be able to explore the dataset with their unique perspectives and angles. Auto-generating themes without users' input may result in an undesirable general topic summary [10]. Putting researchers' thoughts and interpretations into AI theme generation may help develop more unique themes tailored toward researchers' views. Prompting is a method used to steer LLMs towards generating

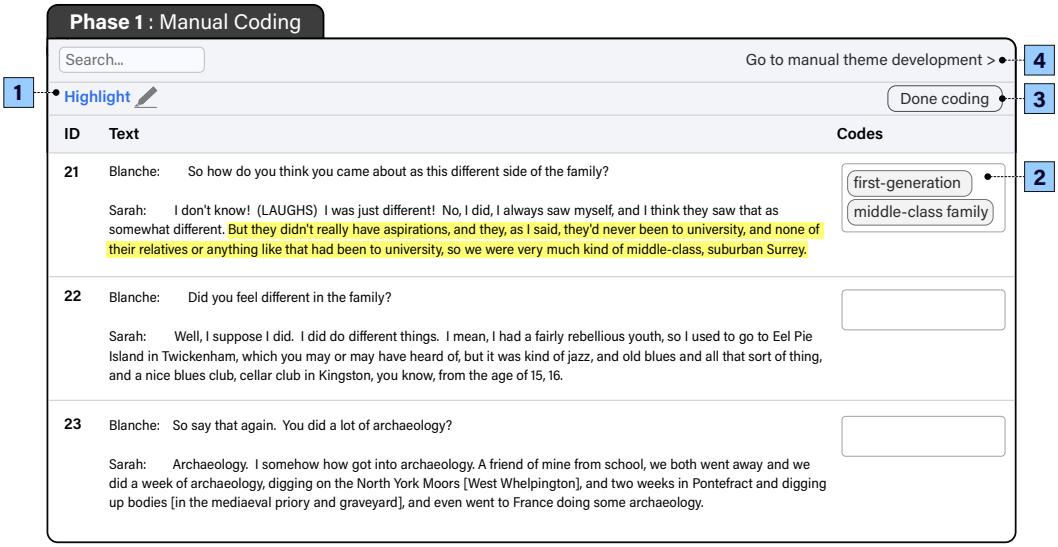

Fig. 3. Manual coding page. On the coding page, users highlight text segments representing main ideas using the highlight feature [1]. New codes can be assigned or existing codes can be selected from the dropdown menu. Codes are assigned for each row [2]. Users save data [3] and can proceed to the next phase [4] via buttons.

desired outputs [52]. Using prompts, researchers can guide the model to generate the output that reflects their thoughts and intentions.

DG4: Reduce the burden of prompting through centralized data management. Numerous studies investigating LLMs' ability in text data analysis concur that prompting poses significant challenges since designing prompts that the AI can understand is difficult and time-consuming [49, 63, 65]. This challenge intensifies in theme development, where researchers repeatedly revise themes through iterative back-and-forth processes. Constantly refining prompts and passing data to the model adds significant burdens, limiting the possible benefits researchers might otherwise get from the system.

The challenges extend beyond just prompting LLMs; obtaining responses in a specific format is also difficult. Researchers may prefer AI-generated themes in visual format like theme map and if these responses need to be integrated into a system, it often requires a specific data format (e.g., JSON). However, the output from generative LLMs may not consistently match these format expectations.

To tackle this issue, we aim to balance free prompting for researchers while also minimizing effort. We simplify the process by offering a centralized data management system for data, codes, and themes from both users and AI. This setup removes the need for manual updates of new data to the LLM and simplifies data handling, letting users focus on creating meaningful prompts without worrying about technicalities. This is a potentially critical benefit of our system over the state of the art, especially versus traditional chat boxes.

DG5: Promote sensemaking with interactive visualizations. The text-based output modality of LLMs may be less than ideal for supporting theme development, as the sheer volume of generated

text, in addition to the original data, can overwhelm researchers, making it challenging to discern, interpret, and effectively utilize the information [43, 61].

Finding a way to organize and visualize such results becomes critical because it allows researchers to make sense of unstructured textual data, the connections of different text, and the hierarchical relationship between the themes and the texts, all of which might lead to important insights [15, 33]. There have been many approaches for visualizing textual data to support sensemaking [14, 22]. We adopt a similar visualization-oriented strategy. Our implementation includes a vibrant hierarchical bubble diagram, allowing users to quickly grasp the breadth of the data set. Through this approach, qualitative researchers can swiftly comprehend AI-generated themes without the need to read a long text response.

DG6: Ground AI-generated themes in the data. While it is crucial to develop themes grounded in data, LLMs have a chance of hallucination [49], such as generating themes based on non-existing data [17]. As the stack of prior prompts grows, performance decreases for tasks that involve data or historical context [29]. In the case of thematic analysis, when the raw data and instructional prompting were fed to the model separately, it sometimes neglected certain features of the text, especially if the text is extensive in content [49]. Our prompting mechanism automatically combines the original data corpus with instructional prompts as input to the LLM, without requiring user involvement. In doing so, every prompt sent to the model includes the original dataset, ensuring the LLM consistently generates themes grounded in the actual data, thereby reducing the risk of hallucination.

DG7: Bolster researchers' subjectivity by supporting inductive thematic analysis.
Our goal is to support ***inductive*** thematic analysis, which emphasizes the *researcher's subjectivity* as an analytic resource and their reflexive engagement with data and interpretation [9] or processes that rely on human judgment and deep understanding of textual nuances [10]. Unlike deductive approaches that apply predefined codebooks or theories and reduce subjectivity through structured methods (e.g., intercoder reliability), inductive analysis leverages researcher subjectivity to develop rich textual understanding. Previous computational support for qualitative analysis has primarily focused on deductive methods [26–28, 53, 62] because effectively supporting inductive processes has remained challenging as traditional natural language processing has struggled to capture the textual nuances essential for inductive analysis [9, 10]. However, the advanced language understanding capabilities of LLMs may provide new possibilities for supporting inductive thematic analysis [49].

Since the goal of this study is to understand the effect of AI support on *interpretative* tasks like theme development, we designed our system to support user subjectivity rather than impose a deductive approach. This design enables users to freely code and develop themes based on their own interpretations of the text, refining them iteratively as their understanding evolves.

4 ThemeViz system

We designed and implemented ThemeViz, a GPT-4 powered system that supports an iterative theme development process. In this section, we provide a usage scenario with key features and implementation details.

4.1 Usage scenario and key features

Here, we introduce an example scenario to demonstrate the theme development workflow in ThemeViz (see Figure 5). Although our research primarily focuses on understanding the effects of human-AI interaction in Phase 3, where AI assists in theme development by generating multiple versions of themes, we included Phases 1 and 2 (manual coding and manual theme development) to provide context for the entire process.

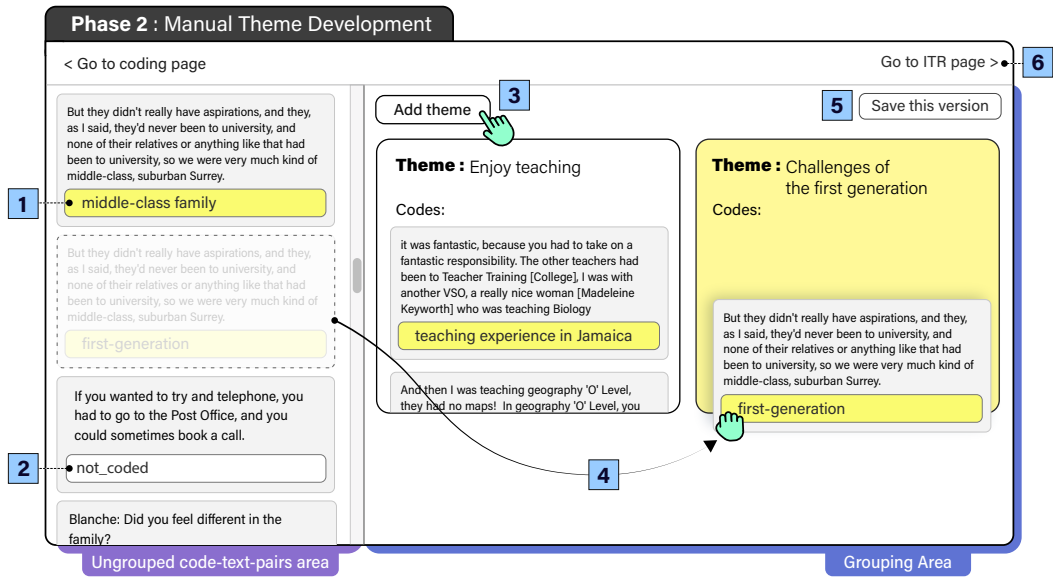

Fig. 4. Manual theme development page. On the left side of this page, there is an area for ungrouped code-text-pairs. Code-text-pairs display both coded and uncoded data. Codes are highlighted in yellow [1], and uncoded data are marked [2]. Users create new themes by clicking the "Add Theme" button [3]. Users can drag and drop a code-text-pair on the left side into a theme box to assign it [4]. They can also save the current themes [5] and proceed to the iterative theme refinement phase [6] via buttons.

Suppose there is a qualitative researcher named **Sam** who's working on analyzing a public dataset that contains interview transcripts of female scholars after World War II in the UK. The dataset is about their lives and how their life experience affected their academic careers.

Phase 1: Manual coding. Using ThemeViz, Sam starts to read through the data and starts to code each row. She highlights the most important part of the text using the highlight tool (Fig.3-1). Then, she adds codes to capture the meaning of the highlighted text (Fig. 3-2). The system's manual coding feature allowed Sam to maintain autonomy as a researcher by giving her the opportunity to read and code the dataset independently before receiving any AI assistance (DG1). After coding data for a while, she decided to develop initial themes with the codes that she developed so far. Sam clicks the "Done Coding" button to save her codes (Fig.3-3), then goes to the next page.

Phase 2: Manual theme development. In the next phase, Sam can manually group codes and name themes (Fig. 4). In the beginning, she sees the page is divided into two main sections. On the left panel, Sam can see the list of codes with their corresponding quotes (which will be referred to as a code-text-pair) that are not yet grouped as themes. In the case of un-coded data, they will be automatically assigned a "not-coded" marker by the system (Fig.4-2).

After skimming the code-text-pair cards, Sam comes up with an idea for one potential theme. She clicks the "Add theme" button (Fig.4-3) to generate a new theme box on the right side of the page. She can drag and drop code-text-pairs into the box to group them together and name the theme title accordingly (Fig.4-4). She continues this manual theme-building process until she feels ready to iterate. She saves a version of the themes (Fig.4-5). Similar to the manual coding phase, the

manual theme development phase allowed her to consider potential themes independently before receiving AI assistance, helping her maintain autonomy as a researcher.

Phase 3. AI-assisted theme development

(1) Reviewing. On the AI-assisted theme development page, Sam sees the themes she created in a bubble chart. She can easily check the pattern and size of the themes she developed by the outer bubble's radius (DG5). When Sam wants to examine the codes within a theme, she clicks on the smaller bubbles inside the larger theme bubble. Then, the system interactively reveals related data, including codes and quotes associated with that theme, on the left side panel for easy review. (Fig.6-2, 2a). Color coding helps identify recurring codes, making the theme review process intuitive. Such data visualization of themes, combined with interactivity, supports Sam's sensemaking and examination of the developed themes.

(2) AI theme generation. After examining the themes, Sam realizes that the current themes are too granular and that some of them overlap with each other. She decides to reduce the number of themes. She uses the slider (Fig.6-5) on the top of the page to set the number of themes to 8 and clicks the 'AI assistance' button. The slider bar enables Sam to easily adjust the thematic granularity she wishes to explore, eliminating the need to specify theme quantities in each prompt (DG4).

After a moment, the system generated eight new themes based on Sam's request, displaying the results in an updated bubble chart. Sam examined the themes, exploring new perspectives on the data and identifying any gaps within the themes. She realized she could reorganize the themes around the challenges faced by the interviewee as a first-generation student after World War II.

To re-structure themes in a new way without much extra work, she types a query in the text input box (Fig.6-3): 'Generate themes on the effects of world war II on the education of the interviewee and challenges of first-generation student' and then clicks the 'AI assistance' button (Fig.6- 4). In turn, the AI generates new themes in response based on Sam's prompt (DG3).

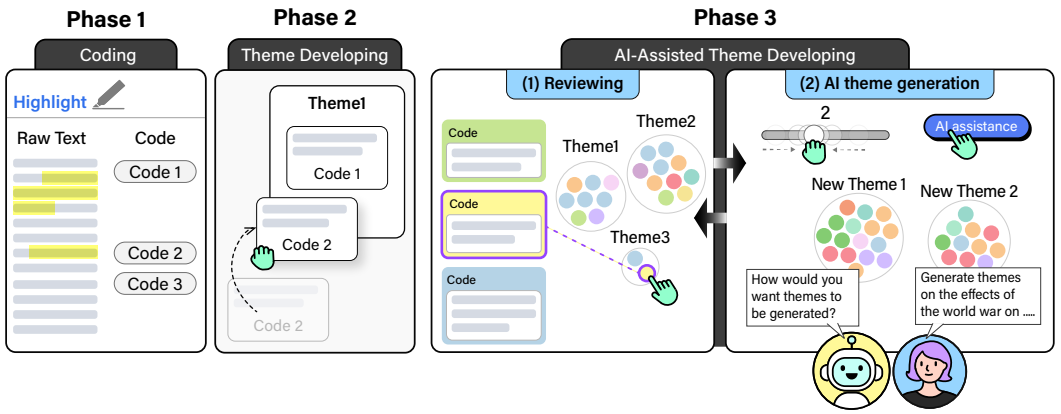

Fig. 5. ThemeViz workflow. Phase 1: Sam highlights text and assigns codes. Then, she moves to Phase 2, where she groups code-text-pairs into themes using drag-and-drop interaction. Next, she moves to Phase 3 - (1) reviewing, where she can review themes using an interactive bubble chart. She can click code bubbles (small bubbles inside the big outer circle) and check their corresponding code-text pair. Once she finishes, she goes to Phase 3 - (2) AI theme generation, where she can generate a new version of themes by prompting AI or giving a number of themes she wishes to receive from AI by moving the slider bar. Sam can review and generate themes multiple times iteratively, and she can revisit Phase 1 and Phase 2 if she needs to.

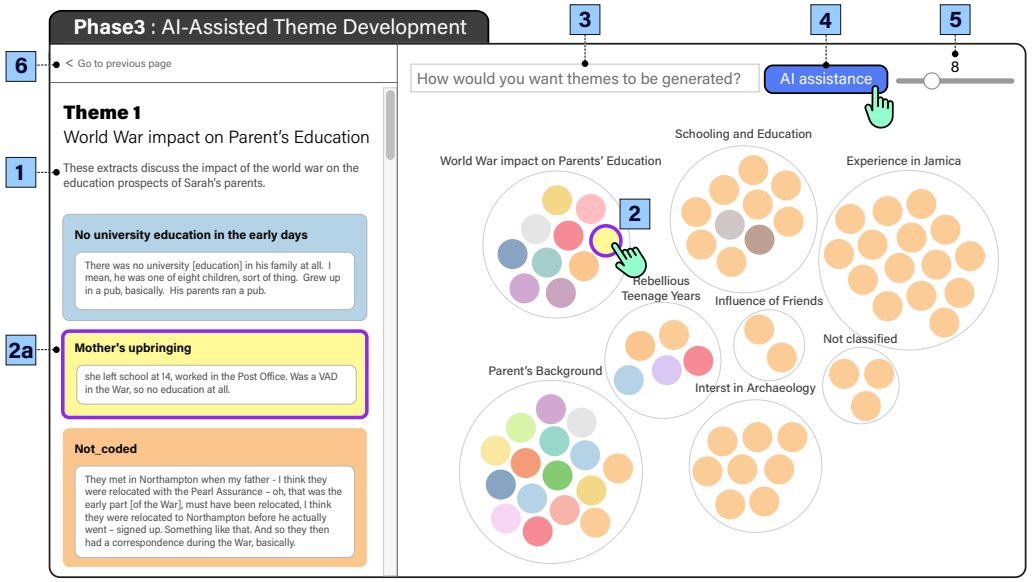

Fig. 6. Iterative theme refinement page. On the left side of the page, there are theme titles, explanations about themes [1], and code-text-pairs that belong to themes. On the right side of the page, themes and code-text-pairs are displayed through an interactive bubble chart. Outer circles in the chart indicate themes and color-coded bubbles indicate code-text-pairs. Linked interactions highlight connections between the panels [2, 2a]. When users want to generate themes with AI, they can provide a prompt in the text input box [3] and then click the "AI assistance" button [4]. They can also adjust their desired theme count with the slider [5]. Users can freely return to the previous phase [6].

Sam appreciates the convenience of not having to manage data within prompts; instead, she can simply specify the type of themes she wants, as ThemeViz handles data management by storing codes, data extracts, and theme metadata within a structured prompt system. This allows Sam to concentrate fully on theme development.

After generating multiple versions of themes with different focuses, Sam likes one version of the theme and decides to revise it more. She goes back to the previous page (manual theme development) to rename some of the themes and moves codes from one theme to the other. She can even go back to the coding page and re-code some of the raw text. Sam can continue this revision process until she reaches a strong set of final themes.

4.2 System implementation

The front end of ThemeViz was built using HTML, CSS, Javascript, and D3.js. The system saves user changes, such as codes and themes, via a FastAPI back-end and a SQL database. The back end also handles prompting the GPT-4 model, formatting data structures for model requests, and processing its responses.

4.2.1 Prompt engineering. ThemeViz makes use of OpenAI's GPT-4 model. While its API is performant and the model extremely flexible, it can still be challenging for users to correctly prompt [63]. Moreover, the model requires raw data and codes in prompts to develop themes, which may be cumbersome if users need to feed such data. For this reason, when users enter Phase 3 of ThemeViz the back-end will do much of the heavy lifting through prompting and parsing.

Prompts include several components: First, they contain the initial dataset (truncated as necessary for window limits). Second, we include general instructions to the model based on the action performed. These were iteratively developed. This also includes a template which instructs the model to deliver results in a manner our JSON parser can understand. Third, we append a template based on the number of themes a user selects on the slider. Fourth, we include some additional instructions to avoid common data errors, such as adding captions. Finally, we include any user-provided prompts. More details are included in supplemental materials.

4.2.2 Data visualization. For data visualization, we primarily used D3.js and JavaScript to create bubble charts. We mapped codes to the smaller bubbles and themes to the larger outer bubbles, allowing users to easily identify the size of the themes and see which codes belong to each theme. The D3.js force simulation was used to dynamically arrange the bubbles, ensuring that the bubbles are well-spaced and visually clear. Our system includes a parser for model responses in order to power front-end visualizations. To develop a flexible parser, we first iterated on the prompts until we had settled on a template that delivered generally stable behavior (given the variability of models). Then, we engaged in logged testing where we used a prototype to encourage breakdowns in parsing. We corrected errors by improving the prompts, developing logic for common formatting issues (e.g., strings before a JSON), and incorporating retry logic with prompt variations as a final effort. In practice, we have found this sufficiently robust in user studies.

5 User study

Recall that we aim to answer three research questions centered around 1) the usefulness of ThemeViz, 2) users' perception of AI as a collaborative partner, and 3) the limitations of AI support in ThemeViz.

5.1 Hypotheses

To ground our exploration of our three research questions, we have developed a series of hypotheses which we will investigate through our controlled user study outlined later in this section.

5.1.1 ***H1****: ThemeViz users will have greater usefulness than traditional chat-based LLMs for theme development.* In RQ1, we investigate whether ThemeViz is useful in offering improvements for theme development compared to chat-based LLM platforms such as ChatGPT. In this hypothesis, we examine the usefulness of ThemeViz in three folds.

Usefulness in theme development. First, we assume that the support of ThemeViz around autonomy, prompting, and sensemaking using interactive visualizations will help users conduct theme development more effectively than LLM tools without such support. This will manifest in terms of more theme generations and higher self-reported data understanding. We triangulate the notion of efficacy through the following sub-hypotheses related to H1:

H1.1 ThemeViz users will engage in *more theme generations* compared to traditional chat-interface users.
H1.2 ThemeViz users will self-report that it helps them *understand data from different perspectives* better than ChatGPT.
H1.3 ThemeViz's scaffolding will lead to higher acceptance.

Usefulness of prompting support. In ThemeViz, we provide scaffolding for prompting through centralized data management. ThemeViz's backend handles metadata for raw data and user-generated codes, formatting and updating information automatically when a user submits a prompt, reducing the need for manual input and data synchronization. We assume that by managing data

formatting and storage, ThemeViz will ease the burden of prompting from scratch. We developed the following sub-hypothesis:

H1.4 ThemeViz users will self-report ThemeViz *reduces* the burden of prompting compared to chat-interface users.

Usefulness of interactive data visualization support. ThemeViz facilitates the understanding of AI-generated responses through interactive visualization. We assume that compared to ChatGPT's text-heavy responses, ThemeViz will improve users' comprehension of AI-generated themes and enables more effective identification of patterns within the data. To evaluate the usefulness of ThemeViz's interactive visualizations, we subdivide as follows:

H1.5 ThemeViz users will self-report that it helped them understand data more effectively compared to chat-interface users.

H1.6 ThemeViz users will use a self-report interactive bubble chart more useful compared to the chat interface's text-based output.

5.1.2 ***H2****: ThemeViz's AI assistant will be viewed more as a collaborator than as a traditional chat-based LLM assistant for theme development.* In RQ2, we examine the extent to which ThemeViz's design encourages users to perceive its AI assistant as a collaborative partner in theme development, compared to traditional interfaces like ChatGPT. Although both interfaces use the same GPT-4 model, ThemeViz offers tailored support for theme development, including access to manual analysis to support user autonomy [23, 37] and interactive visualization of AI-generated themes. We believe that this specialized support, closely aligned with thematic analysis methodologies, will lead researchers to prefer AI assistance in ThemeViz over ChatGPT as a more effective collaborator.

H2.1 ThemeViz users will consider ThemeViz's AI assistant more as a collaborator compared to ChatGPT.

H2.2 ThemeViz users will self-report that ThemeViz offers better ability as a collaborator compared to ChatGPT.

5.2 Study design

To assess the effectiveness of ThemeViz, we conducted a between-subject study. The aim of the between-subject study is to compare the effect of ThemeViz's interaction on theme development with another LLM tool that is lacking the scaffolding and other support we introduce in this paper. In our study, we chose ChatGPT, a widely used LLM-embedded tool with chat interaction as a baseline. ThemeViz and ChatGPT both utilize GPT-4, but they support different interactions, which can be a good comparison to measure the effect of ThemeViz's interactions compared to the conventional prompting interface. For this between-subject study, we recruited 28 qualitative researchers. These participants were randomly split into two groups, with 14 participants in each group. This study size is comparable to other studies of qualitative investigation tools [27, 28]. One group used ChatGPT to conduct theme development, which we refer to as the 'ChatGPT condition.' The other group used ThemeViz to conduct theme development, which we labeled as the 'ThemeViz condition.'

5.3 Procedure

Figure 7 describes the study procedure. Participants in both conditions followed the same procedure for an approximately one-hour session. On arriving at the session, participants were randomly assigned to work with either *Reddit data* or *interview transcript* data (these datasets are described in detail in Section 5.6). We implemented this random assignment to reduce potential bias that could result from relying on only one type of dataset in our study. A facilitator briefly outlined

the design of the study and then participants received training on the interface(s) they would use based on a common script. As all participants were familiar with thematic analysis, we provided no additional qualitative analysis training.

In ThemeViz condition, participants started with coding using ThemeViz. In ChatGPT condition, they were asked to use Atlas.ti Web since ChatGPT does not support manual coding and theme development. In every condition, participants were asked to code for 10 minutes. After that, they formulated an initial theme for 5 minutes. Once the time was up, participants moved to the 'AI-assisted theme development' phase.

Participants were asked to conduct AI-assisted theme development for 20 minutes using their assigned tool (ThemeViz or ChatGPT). While ThemeViz directly allowed storage of versions of themes, ChatGPT offered no such affordance. To provide this capability, we used the SingleFile browser extension which quickly saves a mirror copy of a page (including the ChatGPT transcript) which loaded into a new tab. During the 20-minute AI-assisted theme development session, participants could revisit previous stages to review raw data, code, or manually develop themes.

After 20 minutes of AI-assisted theme development, participants finalized their themes for 5 minutes and answered a short survey. Finally, at the end of the session, the facilitator conducted a 10-minute semi-structured interview with each participant about their experiences.

5.4 Survey design

We created a common survey for all conditions (ThemeViz and ChatGPT) with slight question variations based on the system. Our survey included the 'Modified Technology Acceptance Model (mTAM)' [41] to enable participants to evaluate the usefulness and ease of use of the system. In addition, we included 7-point Likert scale questions (1 = Extremely disagree, 7 = Extremely agree) to evaluate usefulness in theme development. Participants were asked to rate statements such as *I believe that output modality (visualization or text, depending on the condition) helped me understand text data more effectively by providing an overview*, and *It helped me to view data from different perspectives*. To understand users' perceptions of the AI assistant as a collaborator, we asked participants to evaluate sentences such as *The AI assistant felt like a collaborator/co-worker*. Questions assessing the AI collaborator's abilities were adapted and modified from measures used in studies on collaboration [3, 36] and can be found in supplemental materials.

5.5 Participants

We recruited participants through university postings and snowball sampling [30]. Participants were required to be adults fluent in English who have conducted thematic analysis in the past.

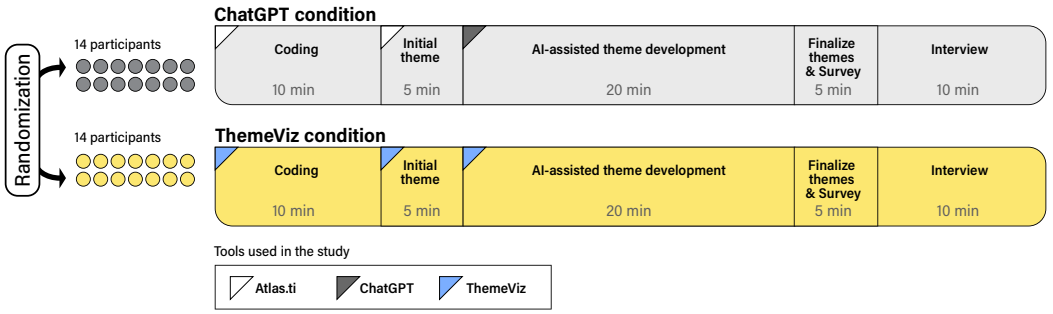

Fig. 7. Study procedure

Table 1. **Participant demographics** The participant demographic table is organized to show that P1–P14 participated in the ThemeViz condition, while P15–P28 participated in the ChatGPT condition.

	Occupation	Field of expertise	Experience
P1	Graduate Student - PhD	Computer-mediated communication	6 years
P2	Undergraduate Student	Interactive Technology, HCI	2 years
P3	Graduate Student - PhD	Participatory policy and AI	8 years
P4	Graduate Student - PhD	Gender and technology	6 years
P5	Graduate Student - PhD	Social computing	2 years
P6	Graduate Student - PhD	Design for low-resource communities	3 years
P7	Graduate Student - PhD	Fan Studies, HCI	2 years
P8	Graduate Student - PhD	Science and technology studies	7 years
P9	Graduate Student - PhD	Technology for frontline health workers	10 years
P10	Graduate Student - PhD	AI-mediated communication	4 years
P11	Master's Student	Educational equity	1 year
P12	Undergraduate Student	Interactive Technology, HCI	3 years
P13	Graduate Student - PhD	Human-AI interaction and design	3 years
P14	Graduate Student - PhD	Hybrid meetings	4 years
P15	Graduate Student - PhD	Algorithmic harms	3.5 years
P16	Graduate Student - PhD	Co-writing system	1 years
P17	Graduate Student - PhD	Infrastructure studies	10 years
P18	Graduate Student - PhD	Online harassment in social media	2 years
P19	Graduate Student - PhD	Accessibility, HCI Research	4 years
P20	Undergraduate Student	HCI	2 years
P21	Graduate Student - PhD	Environment, people, and computing	7 years
P22	Graduate Student - PhD	Food networks	15 years
P23	Graduate Student - PhD	HCI	1.5 years
P24	Graduate Student - PhD	AI in communication	5 year
P25	Graduate Student - PhD	Safety net programs in the US	8 years
P26	Undergraduate Student	AR to facilitate strangers	2 years
P27	Graduate Student - PhD	Ecosystem of language data	1 year
P28	Graduate Student - PhD	Ethics in changing scientific practices	10 years

Participants were compensated with $20 Amazon gift cards for each hour of participation. This study was approved by our university ethics review board. Of the 28 total participants, they reflected an average age of 29 (Mdn: 29, SD: 5), a majority had an education of at least undergraduate level, and the gender distribution was 9M, 19F, and 0NB. Participants had an average of 4.75 years of experience (Mdn: 3.75, SD:3.51) in thematic analysis (Table 1). We observed no differences between the groups assigned to each condition due to random assignment.

5.6 Datasets

As thematic analysis relies heavily on contextual interpretation of textual data, it is potentially risky to focus solely on one dataset in our study design. To mitigate such potential biases, we employed two data sets for thematic analysis: one from the social media platform Reddit and another from public interview transcripts [58]. We selected these datasets because they represent two common sources of data for qualitative analysis: social media comments and interview transcripts [47]. These data sets also comply with the privacy and ethical guidelines of GPT-4 use, avoiding unauthorized

data access or usage. Both datasets were randomly sampled to be of comparable size and, in preliminary testing, yielded similar amounts of codes and themes.

Reddit Data: Our first dataset was from Reddit's r/aiwars subreddit, a community established in December 2022. Utilizing the Reddit API, we procured 816 top posts and 21,205 comments by May 2023. We focused on the comments because they contained detailed discussions among users with different opinions.

Interview Transcript: The second dataset comprises detailed life story interviews with pioneers in British social research, primarily active between the post-war years of 1950 and 1990 [58]. This dataset provides insights into researchers' academic journeys, research processes, and challenges encountered.

In our analysis, we randomly selected subsets from each data set to maintain a manageable volume and to ensure a diverse representation of content. Each subset consisted of 50 text pieces or rows, with an average word count of 568 (SD = 283) per row for the Reddit data and 549 (SD = 202) per row for the interview transcript. To adhere to the token limitations imposed by the GPT-4 API (8,000 tokens for prompt and answer) [5], we ensured that the word count did not exceed 6,000 total number of words for each subset.

6 Quantitative results

In this section, we report quantitative results we received after we conducted a statistical analysis of participant responses and log data. Qualitative results will be reported in the next section. In this section, we will break down results by each research question (refer to §1 for details).

6.1 RQ1: Usefulness of ThemeViz in assisting theme development

We hypothesized that ThemeViz would be more useful in assisting theme development compared to ChatGPT (H1), which we broke down into a set of sub-hypotheses.

6.1.1 Usefulness in theme development. First, we examined the number of different versions of themes each group made (H1.1). We quantified the number of iterations by tallying the various theme versions saved by users. Each saved theme was regarded as a unique iteration if it did not contain identical data to a prior save. A t-test revealed that ThemeViz users developed more versions of themes than ChatGPT users (see Table 2). This outcome supports H1.

Next, we examined whether ThemeViz enables users to view data from different perspectives (H1.2). We asked users to report the degree to which they think the system helped them to view data from different perspectives on a 7-point Likert scale. Employing a Mann-Whitney U test, we

	ThemeViz				ChatGPT				Test Results	
	M	Md	SD	95% CI	M	Md	SD	95% CI	t/U	Effect Size, 95% CI
H1.1	4.07	4.00	1.69	[3.10, 5.04]	1.86	1.00	1.29	[1.11, 2.60]	t(26) = 3.90, p < .001	d = 1.47 [0.60, 2.35]
H1.2	5.57	6.00	1.28	[4.83, 6.31]	3.36	3.00	1.65	[2.41, 4.31]	U = 167.0, p < .001	r = 0.60 [1.12, 3.31]
H1.3	4.69	4.86	1.19	[4.00, 5.38]	3.92	4.14	1.08	[3.30, 4.55]	U = 135.0, p = .046	r = 0.32 [-0.08, 1.62]
H1.4	5.36	6.00	1.21	[4.83, 5.89]	3.07	3.00	1.77	[2.30, 3.84]	U = 166.5, p < .001	r = 0.60 [1.18, 3.40]
H1.5	5.43	5.50	1.34	[4.86, 6.00]	3.64	3.50	1.91	[2.78, 4.50]	U = 150.0, p = .008	r = 0.46 [0.57, 3.01]
H1.6	5.57	6.00	1.28	[4.83, 6.31]	4.21	5.00	1.72	[3.22, 5.21]	U = 144.0, p = .015	r = 0.45 [0.29, 2.43]

Table 2. RQ1 Results. We observe general support for Hypothesis 1 concerning the usefulness of ThemeViz. For the test results, t denotes a t-test, while U represents the Mann–Whitney U test.

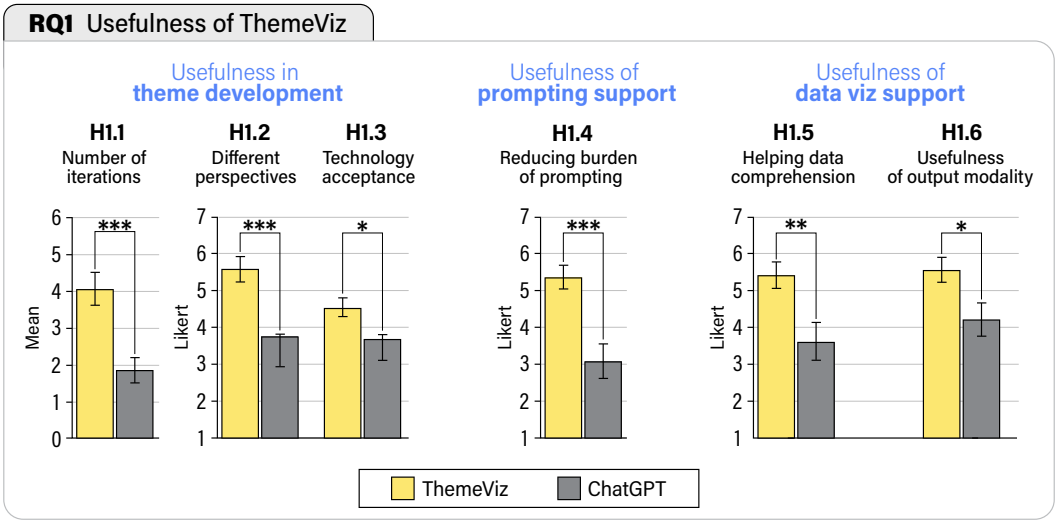

Fig. 8. RQ1 results. Refer to Table 2 for specific values.

found that ThemeViz users reported higher agreement, supporting this sub-hypothesis (see Table 2).

Finally, we also used a modified technology acceptance model (mTAM) [41] protocol in our survey (H1.3). Analyzing the results with a Mann-Whitney U test, we again found that ThemeViz users reported higher acceptance than that of control users (see Table 2).

6.1.2 Usefulness of prompting support. We hypothesized that users would find ThemeViz easier to prompt than ChatGPT (H1.4). We first examined how easy participants found the model prompting and response process to be using a 7-point Likert scale. A Mann-Whitney U test showed that ThemeViz users reported ThemeViz was easier to prompt compared to ChatGPT (see Table 2).

6.1.3 Usefulness of interactive data visualization support. We posited that interactive data visualization support in ThemeViz would help users to more effectively process AI-generated themes and conduct theme development. To capture the one part of this sub-hypothesis H1.5 we asked users to rate on a 7-point Likert scale how well the visualization helped them to understand the text by providing overviews (see Table 2). Comparing between ThemeViz users and ChatGPT users using a Mann-Whitney U test, we found that ThemeViz users generally reported higher ratings.

For another sub-hypothesis, H1.6, we asked users to rate the usefulness of the feedback that they received from the system (either visualizations or textual responses) on a 7-point Likert scale. As in the previous hypothesis, a Mann-Whitney U test showed that ThemeViz users reported higher feedback quality (see Table 2).

6.2 RQ2: User perception of ThemeViz's AI assistant as a collaborative partner

We investigated whether participants considered ThemeViz to be a collaborator more than in the ChatGPT condition (H2.1). A Mann-Whitney U test revealed ThemeViz's AI assistant was considered more to be a collaborator/co-worker than the control (see Table 3). The intuition here is that a more efficacious tool would be perceived as a collaborator rather than a blunt instrument.

To draw more insight on this point, we asked specifically about the tool's *ability* as a collaborator (H2.2). We employed a set of questions (see supplemental materials) concerning confidence and

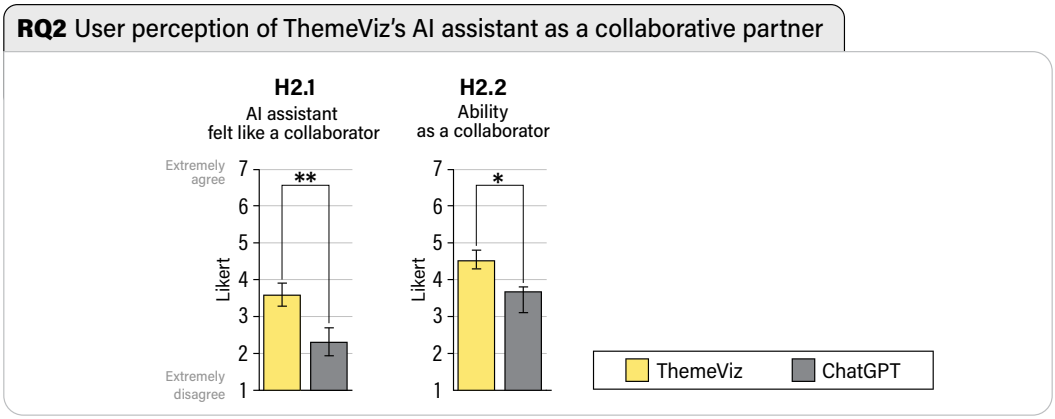

Fig. 9. RQ2 results. Refer to Table 3 for specific values.

perceived efficacy to triangulate on this point. A Mann-Whitney U test revealed that ThemeViz participants reported better ability as a collaborator than ChatGPT participants (see Table 3). We argue that this is a result of the specific workflow support ThemeViz provides (as opposed to burdensome textual prompting), and the intuitive embedding of the model in the interface.

7 Qualitative results

In this section, we present qualitative results to add nuance to the quantitative findings and address questions that remain unanswered by the quantitative data.

For qualitative data analysis, we used thematic analysis [7, 8] to analyze interview transcripts, following Braun and Clarke's six-phase approach. The three authors conducted and transcribed all interviews. Three authors performed open coding on the transcripts. Following the initial coding, we iteratively discussed and refined the codes to ensure they captured the essence of the data accurately. This process involved consolidating similar codes and addressing instances where our interpretations differed. We collaboratively developed a final set of codes that comprehensively represented the key concepts in the transcripts. To develop themes, we used a digital whiteboard tool, starting with a preliminary session to identify initial themes. Each author then independently refined these themes by cross-referencing the raw data and codes to ensure consistency and alignment with the data. After the first round of refinement, we convened multiple times to confirm that each theme clearly represented the data and was distinct from the others. Through these discussions, we finalized three primary themes, which we elaborate on in the following section.

	ThemeViz				ChatGPT				Test Results	
	M	Md	SD	95% CI	M	Md	SD	95% CI	U	Effect Size, 95% CI
H2.1	3.57	4.00	1.16	[2.90, 4.24]	2.29	2.0	1.44	[1.46, 3.12]	U = 148.5, p = .009	r = 0.44 [0.32, 2.25]
H2.2	4.55	4.83	0.95	[4.00, 5.10]	3.45	3.58	1.31	[2.70, 4.21]	U = 145.5, p = .015	r = 0.41 [0.25, 1.94]

Table 3. RQ2 Results. We find support for H2, as users perceive ThemeViz as more of a collaborator with better collaborative abilities. For the test results, U represents the Mann–Whitney U test.

7.1 ThemeViz expands perspectives, with visualizations aiding comprehension and prompts preventing AI hallucinations (RQ1)

Our quantitative results support the usefulness of ThemeViz in theme development (Section 6.1). In this section, we add depth to the understanding of the theme development support provided by ThemeViz. We present findings around three aspects: (1) the usefulness of AI assistance in theme development, (2) the usefulness of interactive visualization, and (3) the usefulness of prompting support.

7.1.1 ***Usefulness of AI assistance in theme development.*** Participants reported that ThemeViz's AI assistant helped them expand perspectives, refine questions, and provided inspiration, especially when AI-generated themes differed from their own. For instance, participants in the ThemeViz condition said that ThemeViz helped them develop "different perspectives" (P10) of data, "refine(d) research questions" (P13), and "expand[ed] themes" (P8). Participants also suggested that ThemeViz's AI assistant helped them "re-think the data [they] clustered... by looking at different themes" (P9) that the AI generated. When the themes did not align with their ideas, they served as a "source of inspiration, prompting deeper thinking into the research" (P13). P2 gives an example:

> *... [when I was working with Reddit data, at first] I just had a rough idea that most comments were about AI art, with some focused on legal issues, but that's all I had. When I asked the AI [ThemeViz] to develop more themes, it offered more perspectives. Some of these included cultural and social implications. It definitely provided a broader view than I could have achieved by developing perspectives through reading alone. (P2)*

Even participants who were initially hesitant about using AI in theme development developed a more positive view of human-AI collaboration after experiencing how AI-generated theme variations could serve as a source of inspiration. Before using ThemeViz, many participants "did not realize the efficiency AI could bring" (P12). For instance, P11 thought it was a "stupid idea to ask AI to help with thematic analysis," but once she used ThemeViz, she saw "potential." Many participants described ThemeViz's AI feature as "extremely useful" (P6, P14), "efficient" (P2), "effective" (P12, P13), "saves time and effort" (P12), and "less time-consuming" (P6). P6 elaborates:

> *The goal of qualitative data analysis is to explore all perspectives and engage deeply with the data, reviewing it repeatedly. Maybe ThemeViz could help streamline this process. Instead of taking weeks, it could reduce the time to just a few days for iterative theme development. (P13)*

7.1.2 ***Usefulness of interactive visualization.*** Regarding the usefulness of interactive visualization, participants believed the visualization helped them more clearly understand the "overall structure" (P2) and "relationship between data" (P10) in "less time" (P2). P14 reports, "[the bubble chart] simplifies complex data. The structure becomes easier to understand, allowing me to focus on specific parts. It reduces the cognitive load compared to reading extensive text." P6 also mentioned that the visual overview of themes is useful for "seeing if one particular has too many or too few codes" and deciding whether he needs to break it down. We believe that the interactive visualization of themes assisted the sensemaking of textual data by displaying high-level patterns while still allowing users to focus on a granular level.

7.1.3 ***Usefulness of prompting support.*** Participants identified prompting support as another useful feature. While ChatGPT participants frequently reported instances of AI 'hallucination,' where themes were generated without grounding in the data, ThemeViz users did not report such issues; instead, they only noted cases where the AI misclassified codes under incorrect themes. Also, they were able to detect such mistakes by examining classified code using interactive visualization.

For example, participants who participated in the ChatGPT condition found a theme that appeared to be interesting but had no real supporting data. P17 expressed feeling unable to "judge whether what GPT told [her] is correct or not," leading to a "loss of control." Concerns grew in the ChatGPT condition as participants noticed it "randomly makeup like new things. (P17)" P24 also pointed out synthetic quotes "combined from different parts of different quotations" that "did not really make sense."

ThemeViz participants faced no issues with themes not being grounded in data, as ThemeViz consistently embeds raw data and related metadata within prompts to ensure AI-generated themes remain data-driven. Participants only reported occasional misclassification of codes; for example, P5 noted instances where AI "misplaced codes in other themes," and P3 found cases of incorrect theme categorization, though they did not observe themes generated from non-existing codes or non-existing raw data.

7.2 AI falls short as a collaborative analyst (RQ2)

We revealed in our quantitative results section (Section 6.2) that participants viewed the AI assistant in ThemeViz as more of a collaborator than ChatGPT, perceiving it as having a greater capacity for collaboration. We believe ThemeViz's design, which supports user autonomy with data visualization and prompting features, contributed to the AI assistant receiving higher evaluation scores. However, our qualitative results indicated that, despite ThemeViz scoring higher than ChatGPT, researchers still did not fully regard the AI assistant as a collaborator. In our interviews, we prompted participants to reflect on the role that the AI model played in their analysis session. Was it a tool they used, a collaborator helping them, or something in between? In general, most participants (27 out of 28 participants) thought of the AI assistance they received from the LLM embedded system as a "tool," not as a collaborator, regardless of condition. Participants explained this perspective was due to two primary factors: (1) a lack of agency and (2) a communication style that did not facilitate an in-depth understanding of the dataset.

7.2.1 ***Beyond pattern recognition: why AI falls short of the essential role of agency in qualitative analysis collaboration.*** Participants who participated in the ThemeViz condition believed AI assistant was a tool because of lack of agency. They believed LLM does not have the agency to take responsibility for the analysis results, and participants did not trust that "AI could conduct theme development alone without human supervision (P7)". P10 said they believe the AI assistant is a tool because it does not have its own "mindset" while human "collaborators have their own." P8 believed current LLM-based AI agents do not have the capability to understand "nuances [of within dataset] and [conduct] philosophical reflections that are expected from a collaborator." P24 from ChatGPT condition also agreed with P8, P24 thought GPT "doesn't really understand what exactly would be an interpretation that generates knowledge because just producing themes shouldn't be about summarizing, it should be about meaning." P24 added more explanation of why they expect an ability to "think" in a collaborator:

> *...GPT is very good at explaining very generic themes that I think any untrained undergrad can come up with, but if I'm working with an RA, I would expect them to challenge themselves and their ability to think. Is this a meaningful theme? I think that part is missing [in AI]. It's really hard for AI to be considered a collaborator in that sense. (P24)*

This illustrates that researchers do not consider AI assistants as equal collaborators if they do not have human-level agency and reflection capability. However, P19 from the ChatGPT condition raised a question of whether AI should take an equal collaborator role to be a collaborator by mentioning the "power dynamic" within "every [human-human] collaboration." P19 said, "[when you are writing a paper,] you have a first author and second author. There may be differences in

experiences, such as one person being more senior. To collaborate successfully with AI, I think the human needs to take the role of the first author and lead it [AI] accordingly."

While P19's remark shows that collaboration is not always between collaborators with equal ability, P23 argued that AI still should not be regarded as a collaborator, as it lacks the agency to take responsibility. P23 elaborates:

> *.. GPT [may] provide some incorrect information. If that [GPT] is a tool, then it's solely my responsibility. However, being a collaborator implies that it has agency and can share responsibility for those [mistakes or incorrect information]. However, I don't think it [GPT] has that responsibility, and I also don't think researchers should delegate their roles to GPT. (P23)*

Our results illustrate that, although previous research on human-AI collaboration in qualitative data analysis suggested that large language models (LLMs) could take on a collaborative role by offering new perspectives in data interpretation, researchers' expectations for AI in interpretative tasks remain high for it to be considered a true collaborator. Our findings suggest that, beyond the ability to suggest new themes, researchers expect collaborators to possess agency, reflection capabilities, and the ability to take responsibility for their actions or mistakes.

7.2.2 ***Beyond compliance: why AI's communication style fails to meet qualitative research needs.*** Another reason researchers view AI assistants as tools rather than collaborators is that interactions with LLMs follow an instruction-response style, resulting in a one-way conversation. The LLMs compliantly and faithfully respond to user requests, whereas researchers expect a bi-directional dialogue where both parties challenge each other during theme development.

The reason why they wanted a "mutual relationship rather than a one-sided instructional one" (P13) with AI is because it is common for researchers to "discuss (P1)" different perspectives with other researchers and help each other "think deeper" (P13) in real-life collaboration. For instance, our participants wanted to know more about the data in depth instead of only getting potential themes generated by AI. P4 shared her experience when she was developing themes using ThemeViz's AI assistance. P4 noted:

> *... there was a moment [the interviewee] talked about, 'Oh, my father sent me to school, but it didn't matter.' As a researcher, I was like, 'Why? Why she didn't think this matter, you know?', but then [AI] clustered it as 'family background' [instead of delving in-depth into the meaning]. (P4)*

Instead of simply receiving potential themes, researchers wanted the AI to highlight interesting parts of the dataset to help them explore the data's deeper meaning. They believed that being "challenged" (P24) would be an effective way to enhance their understanding of the data. As P24 noted, "Good collaboration is like two people who either argue with each other or challenge each other to think, to create something new." P10 echoed this sentiment, adding, "The AI should be able to ask questions, prompting the user to think deeper."

In addition to the need for discussion and questioning, some participants felt that for true collaboration, it should not only be the AI assisting them in developing themes; human researchers should also contribute to the AI, as collaboration is inherently bi-directional and mutually beneficial. As P28, who participated in the ChatGPT condition, noted:

> *...[If] you're collaborating with me. You're getting something from me, and I'm getting something from you. So even though this computer [AI] is getting my data and doing whatever with the data, I don't feel that it's benefiting the computer. Maybe it [AI] is learning from my behavior so it can help other people, but it's not like a direct thing [benefit to AI], so it's hard to see that as a collaborator. (P28)*

Our qualitative results indicate that, although our quantitative findings (Section 6.2) showed participants viewed the AI assistant in ThemeViz as more collaborative than ChatGPT, perceiving it as having a greater capacity for partnership, it still falls short of the level of collaboration researchers expect. Researchers anticipate meaningful, two-way conversations in which the AI challenges ideas, probes for deeper understanding, and mutually benefits from the interaction rather than simply complying with user instructions.

7.3 Limitations in addressing ethical concerns (RQ3)

Our interview results revealed that ThemeViz's AI assistance has limitations in addressing researchers' ethical concerns. Here, we report their concerns regarding issues of privacy and bias.

7.3.1 ***Privacy concerns: risk of conflicting with ethical duties to participants in qualitative studies.*** ThemeViz condition participants had broad concerns about "data leakage" (P8) of sensitive information, and interviewees "feared that [AI] platforms might use their personal information" (P13). They felt that analyzing data using an AI may break the "trust that the researcher built [with the interviewees]" (P11). This concern resonated with participants who participated in the ChatGPT condition. They believed this issue was more critical when the data was about "vulnerable populations" (P28). They also cited risks related to intellectual property (P17) and IRB data protection requirements (P23). They believe if they need to use AI-embedded tools for the analysis, they need to make sure data is "protected" (P23) and want to know if data they put into systems becomes "intellectual property" (P17) of an AI company or whether they "should be concerned" (P17) about these.

7.3.2 ***AI bias concerns: how black-box model poses risks to essential principles of cultural context and researcher positionality in qualitative research.*** Another concern was about bias in AI models. As a salient topic for the population we engaged for our study, participants expressed broad concerns about the need for independence in qualitative research and the ways that bias may creep into the process. Participants noted specific concerns about "where training data [were] coming from" (P10). P5 made an interesting observation about their work where "[GPT-4] might have biased opinions to evaluate data from non-western settings, or may misinterpret cultural contexts present in the data from those countries."

Participants believed that the black-box nature of LLMs goes against the qualitative data analysis practice, where the positionality of the researcher is important. P18 said they always "include a position statement the researcher takes towards the given subject" to take "responsibility and accountability in the perspectives." P24 also believed that positionality is important because the "unique perspective" researchers can bring based on the subjectiveness of the researcher is the "beauty" of qualitative research. P24 noted that representing what "everyone thinks" is not the value of "original scholarship." Thus, they questioned whether AI bias could lead to the misrepresentation of data or whether AI bias lacks explicit positionality, creating misalignment with qualitative research principles. These results suggest that it is essential to consider researchers' ethical concerns around privacy and bias when designing data analysis tools with embedded LLMs.

8 Discussion and design implications

Our findings indicate that ThemeViz was effective in helping qualitative researchers conduct theme development by enabling them to go through more iterations and understand data from different perspectives. ThemeViz's support for prompting and visualizing AI responses benefited users as well. However, contrary to expectations set by prior work [64] — which suggested that LLMs could act as collaborators in interpretative tasks like theme development — qualitative researchers did not view the ThemeViz AI assistant (or ChatGPT) as a collaborator. Our qualitative results reveal

that researchers did not view AI assistants as collaborators due to (1) the lack of agency in AI and (2) the absence of bi-directional discussion and mutual benefit. Finally, we found that ThemeViz's AI assistance has limitations in addressing concerns related to data privacy and AI bias. These results indicate that, although ThemeViz's design was effective to some extent, there is room for improvement in future designs of LLM-embedded qualitative data analysis tools and in enhancing human-AI collaboration for theme development, as discussed in the following section.

8.1 Rethinking human-AI collaboration in theme development

8.1.1 Challenge me: designing proactive, questioning AI agents to foster critical thinking in theme development. Previous research on human-AI collaboration in thematic analysis examined whether LLMs could generate themes of quality comparable to those created by humans, with the idea that if LLMs produced sufficiently high-quality themes, they could take on a collaborative role [38, 49, 65]. However, our findings reveal that although researchers found the themes suggested by LLMs helpful, they did not view LLMs as collaborators due to their conversational style. In Section 7.2, we reported that the current AI conversational style, which adopts a passive role by faithfully responding to users' inquiries rather than challenging researchers, asking questions, or engaging in discussions of disagreement, does not align well with theme development tasks.

In theme development, deepening the understanding of data is essential and often facilitated by dynamic, in-depth discussions with collaborators. Future research could explore designing AI agents that focus on asking questions rather than providing answers [1] and investigate their impact on theme development. Additionally, future designs might consider qualitative data analysis tools that foster critical thinking through methods such as Socratic dialogue [50]. We are encouraged by a new line of research examining the potential of LLMs to support critical thinking [31]. Furthermore, recent studies have explored the personalization of LLM agents [42, 46], suggesting the feasibility of assigning them proactive personalities.

8.1.2 Supporting researchers' autonomy in human-AI collaboration in theme development. Assigning AI the role of proactively asking challenging questions may not be sufficient for effective human-AI collaboration in theme development. As participants noted in Section 7.2, many researchers regard AI as lacking the agency that human collaborators possess and are thus hesitant to consider AI as a full collaborator. Consequently, some researchers may prefer to treat AI strictly as a tool rather than as a collaborator. In such cases, providing an option to access a more subservient model with a limited role could enhance researchers' sense of autonomy and control over their work.

This debate about AI's agency remains ongoing, as some scholars argue that AI lacks intentionality, making it incapable of bearing moral responsibility [35]. Others contend, however, that since AI actions can have social consequences and resemble human behaviors in certain respects, AI should share some responsibility [24, 56].

A recent study on AI in collaborative ethical decision-making further underscores the complexity of assigning responsibility. It suggests that while humans tend to trust AI for its specialized capabilities, they still rely on human experts for moral judgment. When AI accountability is limited, human experts may be blamed for negative AI-driven outcomes, especially in situations where their autonomy is restricted [59]. Therefore, supporting researchers' autonomy in their interactions with AI is crucial, as it enables them to maintain control over their work.

8.2 Trustworthiness and transparency

In ThemeViz, we focused on providing a useful LLM-embedded system to support theme development. While ThemeViz received positive feedback for its usefulness in helping to develop themes, it

fell short in addressing users' privacy and bias concerns. A key issue for participants in the qualitative analysis space was understanding how privacy would be maintained, especially when models involve the ingestion of transmitted data or reserve certain rights in their policy documentation. One possible way forward is to consider adopting open-source large language models such as Llama [60] and indicate to users how data will be analyzed and whether it will be used for training and stored. Interest is increasing in safeguarding sensitive data and privacy within large language models and systems [40, 55].

Lastly, participants expressed concerns about biases within AI models, a widespread issue across the AI field that has become an area of intensive research focus. In thematic analysis, the subjectivity and biases of human researchers are often integral to the interpretive process [7, 13]. However, unlike AI models, human researchers can explicitly address their biases through positionality statements in their papers, acknowledging how their personal perspectives may shape their work. Making one's positionality intentional and transparent with respect to the research topic can enhance the validity of findings by situating knowledge within its social, cultural, and political contexts [12].

In contrast, large language models often operate as "black boxes," concealing the sources and nature of their biases. Nevertheless, there is an opportunity to enhance transparency by exploring the "positionality" of these models. Recent studies have begun to quantify the positionality of NLP datasets and models [54], offering insights into the types of data informing AI outputs. Incorporating model positionality into future human-AI collaboration systems for thematic analysis could provide researchers with a clearer understanding of model biases, enabling more informed and transparent engagement with AI-generated insights.

8.3 Additional use cases of ThemeViz

While ThemeViz was designed to support inductive thematic analysis (see Section 3 - DG7), its utility may vary depending on the type of analysis and who is conducting the analysis. For example, in deductive thematic analysis, where researchers apply predefined theories to guide theme development [8], users could provide a theoretical framework as a prompt, so ThemeViz's AI assistant can map codes to established theoretical constructs. Another noteworthy use case is the variation in ThemeViz's utility based on researchers' experience levels, even though this was not explicitly considered during its design.

Novice qualitative researchers, who may be unfamiliar with thematic analysis, could struggle with its inherent ambiguity [8, 48], especially when developing coherent themes from unstructured textual data. In such cases, ThemeViz can provide initial theme suggestions and support the creation of multiple theme versions, which users can then critically evaluate and refine. This process may also help them build essential analytical thinking skills. For more experienced qualitative researchers who may have more established routine for analysis, ThemeViz may serve less as a training or support tool and more as an efficiency enhancer. These researchers might use ThemeViz more selectively such as when seeking alternative perspectives on their data or when working with particularly complex or large datasets that benefit from computational support.

Future qualitative data analysis tools ought to consider incorporating features to support diverse user needs based on type of the analysis and experience levels. For example, the system could benefit from adjustable levels of user-expertise assistance, for novice users, system could provide more transparent reasoning behind suggested themes to help them learn about the thematic analysis as a method, and for expert researchers, the system could provide customizable workflows that align with established research practices. However, when supporting novice researchers they may be more vulnerable to negative AI influence. Therefore, it is important to support researchers' autonomy to ensure they still take charge of the analysis instead of solely relying on AI's results.

9 Limitations

In our research, each study session lasted one hour per condition to minimize cognitive load and fatigue. Given the extensive time required for theme development, an hour may not provide a realistic time scale for analysis, even if we reduced the data size to compensate. Future studies could explore the effectiveness of LLM-embedded tools for qualitative data analysis (QDA) over extended sessions or real-world deployments. These could also offer a fuller picture of their long-term effects and usability. In addition, our research was limited by the token count of GPT-4, which prevented the use of very long texts. Assessing the impact of AI assistance across various text lengths could enhance our understanding of its role. Additionally, most participants in our study were relatively young researchers, and our snowball sampling resulted in participants mainly from HCI backgrounds, which limits representing diverse qualitative researchers. Future research might benefit from examining the responses of more seasoned researchers from diverse disciplines to AI support in thematic analysis. Beyond that, while we focused on one-human-one-AI collaboration, there are opportunities to support more diverse collaboration scenarios such as collaborative theme development among multiple researchers. Future work can explore these additional collaborative situations and their dynamics.

Finally, regarding the system design, we designed our system to provide AI assistance exclusively during the theme development phase. This focused approach was intentional to understand the clear impact of AI assistance on theme development while preserving user autonomy by offering support for manual coding and manual theme development. Future work could explore the effects of integrating AI assistance throughout the entire process, investigating ways to balance user autonomy when AI support is applied more broadly.

10 Conclusion

In this paper, we investigated whether LLMs could take a collaborator's role in interpretive tasks like theme development by providing alternative interpretation of data. For this purpose, we designed, built, and studied 'ThemeViz,' an interactive visual system that utilizes GPT-4 to support AI-assisted theme development. We investigated three research questions centered around 1) the usefulness of ThemeViz, 2) users' perception of AI as a collaborative partner, and 3) the limitations of AI support in ThemeViz. Through experiments with 28 qualitative researchers, our results suggest that ThemeViz is generally perceived as having high utility in terms of assisting in developing multiple perspectives of data by helping theme iterations. While ThemeViz scored higher than ChatGPT in user perceptions of AI assistance as a collaborator, qualitative researchers did not perceive it as a true collaborator due to its lack of agency and passive communication style. Our study also highlighted ethical concerns related to privacy and bias. These findings provide empirical evidence of the extent to which a graphical interface specifically designed for theme development support can outperform a chat interface, revealing the values researchers prioritize in collaboration, as well as the current limitations of LLMs in theme development. Ultimately, we share design implications for future research in designing more effective human-AI collaboration within intelligent systems for qualitative data analysis.

References

[1] Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (Merida, Mexico) *(WSDM '24)*. Association for Computing Machinery, New York, NY, USA, 8–17. doi:10.1145/3616855.3635856

[2] Atlas. 2023. ATLAS.ti in 45 minutes. https://atlasti.com/trainings/atlas-ti-in-45-minutes.

[3] Benoit A. Aubert and Barbara L. Kelsey. 2003. Further Understanding of Trust and Performance in Virtual Teams. *Small Group Research* 34, 5 (2003), 575–618. doi:10.1177/1046496403256011
[4] Tehmina Basit. 2003. Manual or electronic? The role of coding in qualitative data analysis. *Educational Research* 45 (08 2003), 143–154. doi:10.1080/0013188032000133548
[5] Sebastian Becker. 2023. Maximum Token length in GPT-4. https://community.openai.com/t/maximum-token-length-in-gpt-4/385914. Accessed: 2025-07-01.
[6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa
[7] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association. doi:10.1037/13620-004
[8] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. doi:10.1080/2159676X.2019.1628806
[9] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* 18, 3 (2021), 328–352. doi:10.1080/14780887.2020.1769238
[10] Virginia Braun and Victoria Clarke. 2022. Conceptual and design thinking for thematic analysis. *Qualitative psychology* 9, 1 (2022), 3. doi:10.1037/qup0000196
[11] David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & quantity* 56, 3 (2022), 1391–1412. doi:10.1007/s11135-021-01182-y
[12] Scott Allen Cambo and Darren Gergle. 2022. Model Positionality and Computational Reflexivity: Promoting Reflexivity in Data Science. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 572, 19 pages. doi:10.1145/3491102.3501998
[13] Karen A Campbell, Elizabeth Orr, Pamela Durepos, Linda Nguyen, Lin Li, Carly Whitmore, Paige Gehrke, Leslie Graham, and Susan M Jack. 2021. Reflexive thematic analysis for applied qualitative health research. *The Qualitative Report* 26, 6 (2021), 2011–2028.
[14] Nan Cao and Weiwei Cui. 2016. *Introduction to text visualization.* Vol. 1. Springer.
[15] Ashley Castleberry and Amanda Nolen. 2018. Thematic analysis of qualitative research data: Is it as easy as it sounds? *Currents in Pharmacy Teaching and Learning* 10, 6 (2018), 807–815. doi:10.1016/j.cptl.2018.03.019
[16] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. *arXiv preprint arXiv:2306.14924* (2023). doi:10.48550/arXiv.2306.14924
[17] Prokopis A. Christou. 2023. How to Use Artificial Intelligence (AI) as a Resource, Methodological and Analysis Tool in Qualitative Research? *The Qualitative Report* (2023).
[18] Laura Ann Chubb. 2023. Me and the Machines: Possibilities and Pitfalls of Using Artificial Intelligence for Qualitative Data Analysis. *International Journal of Qualitative Methods* (2023). doi:10.1177/16094069231193593
[19] Victoria Clarke and Virginia Braun. 2013. *Successful qualitative research: A practical guide for beginners.* Sage publications ltd. http://digital.casalini.it/9781446281024
[20] Dovetail. 2023. Analyze data in minutes, not hours. https://dovetail.com/analysis/.
[21] Zackary Okun Dunivin. 2024. Scalable Qualitative Coding with LLMs: Chain-of-Thought Reasoning Matches Human Performance in Some Hermeneutic Tasks. *arXiv preprint arXiv:2401.15170* (2024). doi:10.48550/arXiv.2401.15170
[22] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 473–482. doi:10.1145/2207676.2207741
[23] Jessica L. Feuston and Jed R. Brubaker. 2021. Putting Tools in Their Place: The Role of Time and Perspective in Human-AI Collaboration for Qualitative Analysis. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 469 (oct 2021), 25 pages. doi:10.1145/3479856
[24] Luciano Floridi. 2016. Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 2083 (2016), 20160112. doi:10.1098/rsta.2016.0112
[25] Rebecca Freeth and Guido Caniglia. 2020. Learning to collaborate while collaborating: advancing interdisciplinary sustainability research. *Sustainability science* 15, 1 (2020), 247–261. doi:10.1007/s11625-019-00701-z
[26] Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. 2023. CoAIcoder: Examining the Effectiveness of AI-assisted Human-to-Human Collaboration in Qualitative Analysis. *ACM Trans. Comput.-Hum. Interact.* 31, 1, Article 6 (Nov. 2023), 38 pages. doi:10.1145/3617362
[27] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. CollabCoder: A Lower-barrier, Rigorous Workflow for Inductive Collaborative Qualitative Analysis with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 11, 29 pages. doi:10.1145/3613904.3642002

[28] Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L. Glassman, and Toby Jia-Jun Li. 2023. PaTAT: Human-AI Collaborative Qualitative Coding with Explainable Interactive Rule Synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 362, 19 pages. doi:10.1145/3544548.3581352

[29] Dongyu Gong, Xingchen Wan, and Dingmin Wang. 2024. Working Memory Capacity of ChatGPT: An Empirical Study. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 9 (Mar. 2024), 10048–10056. doi:10.1609/aaai.v38i9.28868

[30] Leo A Goodman. 1961. Snowball sampling. *The annals of mathematical statistics* (1961), 148–170. http://www.jstor.org/stable/2237615

[31] Ying Guo and Daniel Lee. 2023. Leveraging chatgpt for enhancing critical thinking skills. *Journal of Chemical Education* 100, 12 (2023), 4876–4883. doi:10.1021/acs.jchemed.3c00505

[32] Leah Hamilton, Desha Elliott, Aaron Quick, Simone Smith, and Victoria Choplin. 2023. Exploring the Use of AI in Qualitative Analysis: A Comparative Study of Guaranteed Income Data. *International Journal of Qualitative Methods* (2023). doi:10.1177/16094069231201504

[33] Susan Havre, Beth Hetzler, and Lucy Nowell. 2000. ThemeRiver: visualizing theme changes over time. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings.* 115–123. doi:10.1109/INFVIS.2000.885098

[34] Adam S Hayes. 2025. "Conversing" with qualitative data: Enhancing qualitative research through large language models (LLMs). *International Journal of Qualitative Methods* 24 (2025), 16094069251322346. doi:10.1177/16094069251322346

[35] Kenneth Einar Himma. 2009. Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology* 11 (2009), 19–29. doi:10.1007/s10676-008-9167-5

[36] Sirkka L. Jarvenpaa, Kathleen Knoll, and Dorothy E. Leidner. 1998. Is Anybody out There? Antecedents of Trust in Global Virtual Teams. *Journal of Management Information Systems* 14, 4 (1998), 29–64. doi:10.1080/07421222.1998.11518185

[37] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R. Brubaker. 2021. Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 94 (apr 2021), 23 pages. doi:10.1145/3449168

[38] Awais Hameed Khan, Hiruni Kegalle, Rhea D'Silva, Ned Watt, Daniel Whelan-Shamy, Lida Ghahremanlou, and Liam Magee. 2024. Automating Thematic Analysis: How LLMs Analyse Controversial Topics. *arXiv preprint arXiv:2405.06919* (2024). doi:10.48550/arXiv.2405.06919

[39] Elisabeth Kirsten, Annalina Buckmann, Abraham Mhaidli, and Steffen Becker. 2024. Decoding Complexity: Exploring Human-AI Concordance in Qualitative Coding. *arXiv preprint arXiv:2403.06607* (2024). doi:10.48550/arXiv.2403.06607

[40] Ashutosh Kumar, Sagarika Singh, Shiv Vignesh Murty, and Swathy Ragupathy. 2024. The Ethics of Interaction: Mitigating Security Threats in LLMs. *arXiv preprint arXiv:2401.12273* (2024). doi:10.48550/arXiv.2401.12273

[41] Urška Lah, James R. Lewis, and Boštjan Šumak. 2020. Perceived Usability and the Modified Technology Acceptance Model. *International Journal of Human–Computer Interaction* 36, 13 (2020), 1216–1230. doi:10.1080/10447318.2020.1727262

[42] Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459* (2024). doi:10.48550/arXiv.2401.05459

[43] Shixia Liu, Xiting Wang, Christopher Collins, Wenwen Dou, Fangxin Ouyang, Mennatallah El-Assady, Liu Jiang, and Daniel A. Keim. 2019. Bridging Text Visualization and Mining: A Task-Driven Survey. *IEEE Transactions on Visualization and Computer Graphics* 25, 7 (2019), 2482–2504. doi:10.1109/TVCG.2018.2834341

[44] Megh Marathe and Kentaro Toyama. 2018. Semi-Automated Coding for Qualitative Research: A User-Centered Inquiry and Initial Prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173922

[45] Matthew B. Miles. 1979. Qualitative Data as an Attractive Nuisance: The Problem of Analysis. *Administrative Science Quarterly* 24, 4 (1979), 590–601. http://www.jstor.org/stable/2392365

[46] Vinod Muthusamy, Yara Rizk, Kiran Kate, Praveen Venkateswaran, Vatche Isahagian, Ashu Gulati, and Parijat Dube. 2023. Towards large language model-based personal agents in the enterprise: Current trends and open problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6909–6921. doi:10.18653/v1/2023.findings-emnlp.461

[47] Oliver Müller, Iris Junglas, Jan vom Brocke, and Stefan Debortoli. 2016. Utilizing big data analytics for information systems research: challenges, promises and guidelines. *European Journal of Information Systems* 25, 4 (2016), 289–302. doi:10.1057/ejis.2016.2

[48] Lorelli S Nowell, Jill M Norris, Deborah E White, and Nancy J Moules. 2017. Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods* 16, 1 (2017), 1609406917733847. doi:10.1177/1609406917733847

[49] Stefano De Paoli. 2023. Can Large Language Models emulate an inductive Thematic Analysis of semi-structured interviews? An exploration and provocation on the limits of the approach and the model. *arXiv:2305.13014* (2023). doi:10.48550/arXiv.2305.13014

[50] Richard Paul and Linda Elder. 2007. Critical thinking: The art of Socratic questioning. *Journal of developmental education* 31, 1 (2007), 36. https://www.proquest.com/scholarly-journals/critical-thinking-art-socratic-questioning/docview/228487383/se-2

[51] Devi Yulia Rahmi and Nurul Indarti. 2019. Examining the relationships among cognitive diversity, knowledge sharing and team climate in team innovation. *Team Performance Management: An International Journal* 25, 5/6 (2019), 299–317. doi:10.1108/TPM-11-2018-0070

[52] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 314, 7 pages. doi:10.1145/3411763.3451760

[53] Tim Rietz and Alexander Maedche. 2021. Cody: An AI-Based System to Semi-Automate Coding for Qualitative Research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 394, 14 pages. doi:10.1145/3411764.3445591

[54] Sebastin Santy, Jenny T Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. *arXiv preprint arXiv:2306.01943* (2023). doi:10.48550/arXiv.2306.01943

[55] Glorin Sebastian. 2023. Privacy and data protection in chatgpt and other ai chatbots: Strategies for securing user information. *International Journal of Security and Privacy in Pervasive Computing (IJSPPC)* 15, 1 (2023), 1–14. doi:10.4018/IJSPPC.325475

[56] Bernd Carsten Stahl. 2006. Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and information technology* 8 (2006), 205–213. doi:10.1007/s10676-006-9112-4

[57] Krishna Subramanian, Johannes Maas, Jan Borchers, and James Hollan. 2021. From Detectables to Inspectables: Understanding Qualitative Analysis of Audiovisual Data. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 518, 10 pages. doi:10.1145/3411764.3445458

[58] Paul Thompson. 2019. Pioneers of Social Research, 1996-2018. [data collection]. doi:10.5255/UKDA-SN-6226-6

[59] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 160, 17 pages. doi:10.1145/3491102.3517732

[60] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023). doi:10.48550/arXiv.2302.13971

[61] Yakun Wang and Fuxiang Yu. 2023. Visualization and Analysis of Mapping Knowledge Domains for AI Education System Studies. In *2023 11th International Conference on Information and Education Technology (ICIET)*. 475–479. doi:10.1109/ICIET56899.2023.10111115

[62] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 75–78. doi:10.1145/3581754.3584136

[63] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. doi:10.1145/3544548.3581388

[64] He Zhang, Chuhao Wu, Jingyi Xie, ChanMin Kim, and John M Carroll. 2023. QualiGPT: GPT as an easy-to-use tool for qualitative coding. *arXiv preprint arXiv:2310.07061* (2023). doi:10.48550/arXiv.2310.07061

[65] He Zhang, Chuhao Wu, Jingyi Xie, Yao Lyu, Jie Cai, and John M Carroll. 2023. Redefining Qualitative Analysis in the AI Era: Utilizing ChatGPT for Efficient Thematic Analysis. *arXiv preprint arXiv:2309.10771* (2023). doi:10.48550/arXiv.2309.10771

Received October 2024; revised April 2025; accepted August 2025