Doing Research with Help from ChatGPT: Promising Examples for Coding and Inter-Rater Reliability

Hanneke Theelen
Zuyd University of Applied Sciences, Netherlands

Joyce Vreuls
Zuyd University of Applied Sciences, Netherlands

Jim Rutten
All Pension Group (APG), Netherlands

To cite this article:

Doing Research with Help from ChatGPT: Promising Examples for Coding and Inter-Rater Reliability

Hanneke Theelen, Joyce Vreuls, Jim Rutten

Abstract

The rapid development of artificial intelligence and large language models (LLMs) has led to significant advancements in applying machine learning techniques across diverse disciplines, including educational science research. This study investigates the potential of LLMs like ChatGPT for qualitative data analysis, focusing on open, axial, selective coding, theme or pattern identification, and inter-rater reliability. Our findings indicate promising capabilities of ChatGPT in open coding, demonstrating accurate categorization of qualitative data. However, axial coding posed challenges due to the model's limited understanding, which was partially addressed by refining prompts based on ChatGPT's interpretation. ChatGPT also showed competence in selective coding and theme or pattern identification, providing additional insights. For inter-rater reliability, ChatGPT's performance varied across datasets, with improvements observed when providing contextual information. It is important to note the limitations and variability of LLMs such as ChatGPT, which is in public beta and subject to potential limitations in usage and reliability. Our study demonstrates ChatGPT's potential for coding and inter-rater reliability. Improved results are achieved with refined prompts and utilising ChatGPT's own definitions. The adoption of LLMs for qualitative analysis requires further exploration, including addressing algorithmic bias and the potential for inaccurate responses. Validation techniques are crucial in mitigating these risks.

Introduction

Qualitative analysis serves as a fundamental pillar in social science research (e.g., educational sciences), offering researchers valuable insights into human behaviour, social dynamics, and cultural phenomena. Traditionally, this analysis necessitated laborious manual coding and categorization of textual data, requiring significant time and resources. Recent advancements in artificial intelligence (AI) and large language models (LLMs), such as ChatGPT, have ushered in a new era of possibilities for automating and enhancing qualitative analysis, fundamentally transforming the research landscape.

Since the company OpenAI presented the generative AI tool ChatGPT for the world to play with late 2022, generative AI has taken a massive flight in popularity. Not only the availability in (free) AI tools such as ChatGPT

have surged, but the adoption of generative AI in professional domains such as marketing soared too riding the novelty wave of ChatGPT. Only few months later, generative AI tools are already widely being used to write catchy slogans, create logos, and generate realistic pictures using AI image generators such as Dall-E, Midjourney, and Stable Diffusion. Levi's for example experiments with AI generated models to display clothes on (Weatherbed, 2023), and Google offers merchants AI generated backgrounds for their product photo's (Malik, 2023). Marketing and brands already included the ability to use generative AI tools as a sought after skill set (Peres, Schreier, Schweidel, & Sorescu. 2023).

Several preliminary explorations have been conducted on the use of LLMs such as ChatGPT for research purposes in general. For instance, they have been utilised for text generation tasks (Dönmez, Idil, & Gulen, 2023) and idea generation in research (Rahman, Terano, Rahman, Salamzadeh, & Rahman, 2023). Additionally, their application for analysing interview data has also been investigated (Mesec, 2023). Numerous questions remain unanswered, such as whether LLMs like ChatGPT are capable of coding unstructured research data in an open, axial, and selective manner, as well as their suitability for theme or pattern identification and inter-rater reliability. The present study aims to fill this gap and address the following research question: *To what extent can the LLM ChatGPT be effectively utilised for coding (un)structured qualitative data and assessing inter-rater reliability?*

Theoretical Background

Qualitative Data Analysis: Coding and Pattern Identification in Educational Research

In the field of educational sciences, qualitative research methods are frequently employed alongside quantitative research methods. Qualitative research methods offer a more profound understanding by capturing individuals' experiences, perceptions, and opinions. This enables researchers to gain better insight into the motivations and behavioural patterns of individuals (Creswell, 2013; Seidman, 2013).

Common examples of qualitative research methods include interviews, focus groups, open-ended questions, and case studies. These methods generate large amounts of data, which take a lot of time and effort to process. Coding qualitative data requires careful attention to detail, accuracy, and systematic thinking. The objective is to generate meaningful patterns and insights that contribute to addressing research questions. Typically, patterns are identified through the process of coding the data. This coding process generally involves four main steps (Saldaña, 2015):

1. *Open coding:* During open coding, the researcher identifies concepts, ideas, and themes that emerge from the data. This is accomplished through careful reading and analysis of the data, followed by assigning codes to specific sections of text relevant to those concepts. Open coding is often iterative, with new codes being added as the analysis progresses.
2. *Axial coding:* Axial coding aids in organising and structuring the codes derived from open coding. It involves establishing relationships between codes and subcodes and creating (sub)categories to group the data based on overarching themes.
3. *Selective coding:* Selective coding entails selecting the most relevant, meaningful, and representative codes for further analysis. This involves prioritising codes that contribute significantly to understanding the research aim and research questions.

4. *Theme or pattern identification*: After coding and organising the data, researchers can identify overarching themes or patterns. This involves recognizing common elements, trends, contradictions, or other related aspects within the data. Themes can be identified by comparing different codes and seeking recurring ideas or concepts.

To facilitate the coding process, researchers often employ qualitative analysis software such as NVivo, ATLAS.ti, or MAXQDA. These software tools provide functionalities for organising, coding, and analysing data, allowing efficient management of large qualitative datasets. Nevertheless, coding remains a time-intensive process, often requiring substantial hours to complete.

Additionally, coding qualitative data is often a collaborative effort to reduce potential biases due to the perspective of individual researchers. To enhance the reliability of data analysis, inter-rater reliability is often employed. This involves having another researcher analyse a portion of the dataset and subsequently comparing their findings with those of the initial researcher until consensus is reached (Creswell, 2013; Saldaña, 2015).

The Role of a Large Language Model (LLM) in Qualitative Research

The time-consuming nature of coding large data sets raises the question if AI can help. AI is software simulating intelligence, where the software learns and adapts (i.e., getting smarter) the more data it processes. Thus, the system has a self-improving capability (Farrokhnia, Banihashem, Noroozi, & Wals, 2023). AI involves the development of computer systems that can perform tasks that normally require human intelligence, such as understanding natural language, decision-making, problem-solving, and learning (Russell & Norvig, 2016). Among the various types of AI, the Large Language Model (LLM) stands out. LLMs are software specifically trained to analyse substantial volumes of text and establish connections within it (Radford et al., 2019).

LLMs possess the ability to generate text on demand, whether it involves producing new texts, summarising existing ones, or categorising them (Radford et al., 2019). ChatGPT is a LLM that has also been trained to engage in natural conversations with its users, to generate text that is perceived as written by human (Susnjak, 2022) and to perform tasks such as:
1. Summarising a news item into a concise four-line post suitable for LinkedIn.
2. Identifying the sentiment expressed in responses to open-ended questions.
3. Extracting the most significant keywords from a given text or generating a summary encapsulated in a single label.

Recent developments in generative AI, combined with the emergence of LLMs, offer significant opportunities for teaching and research. These developments feature, among other things, the ability to iteratively improve their own performance (Mann, 2023), offer improved access to a wealth of information (Cascella et al., 2023) - often tailored to individual preferences (Haque et al., 2022) - and reduce workload (Farrokhnia, et al., 2023). However, several concerns about the application of generative AI technologies have arisen from this proliferation. Concerns include the potential shallowness of AI deep understanding, the inherent dependence of generative models on the

quality of their training data, inherent limitations in AI's ability to understand nuanced contexts and complex situations, and complicated privacy and security issues (Baidoo-Anu & Ansah, 2023).

Despite these concerns, the use of LLMs in qualitative research can potentially bring relief by automating and speeding up certain aspects of the research process. The integration of AI techniques into qualitative analysis may hold several potential advantages. Firstly, it might enable researchers to analyse vast quantities of textual data in a fraction of the time typically required for manual coding. This newfound efficiency empowers researchers to explore larger datasets and derive comprehensive insights. Secondly, AI can assist in data exploration by identifying themes, patterns, and connections within the text, aiding researchers in uncovering key findings. Thirdly, AI can potentially contribute to enhancing inter-rater reliability by enabling researchers to validate the reliability of their own coding. Moreover, AI's versatility in processing diverse data sources, including interviews, focus group transcripts, and surveys positions it as a flexible tool applicable to various domains within the realm of social science.

Preliminary research has already been done on the application of AI for research purposes in social sciences, including educational sciences. For instance, Dönmez, Idil, and Gulen (2023) examined the potential use of ChatGPT in writing research articles. The findings indicate that while AI technologies offer researchers opportunities in terms of efficiency, creativity, and diverse perspectives, there are concerns regarding content reliability, as well as ethical and intellectual property issues. ChatGPT was found to be incapable of independently producing a complete research article. However, it can provide tips and support during the writing process. This lack of deep understanding was also identified as a weakness in the study by Farrokhnia and colleagues (2023).

Another example is an introductory literature review on the use of ChatGPT for research purposes, which revealed its potential as an interesting tool for generating new research ideas. Nevertheless, challenges may arise when utilising ChatGPT for literature synthesis, citation generation, problem statement formulation, identifying research gaps, and conducting data analysis. Therefore, researchers should exercise caution when incorporating ChatGPT into academic research (Rahman, Terano, Rahman, Salamzadeh, & Rahman, 2023). Regarding qualitative data-analysis Mesec (2023) conducted an initial promising exploration into the identification of themes from interview data, comparing the performance of ChatGPT to that of a human researcher. It should be noted that the text used in the study was already moderately structured. Consequently, an important question arises regarding the performance of ChatGPT when confronted with more unstructured text.

Research Aim

In summary, the emergence of AI and LLMs, with ChatGPT serving as a notable example, opens a promising frontier for qualitative analysis in educational science research. It is important to consider the strengths and limitations of LLMs in this context. Further exploration and empirical validation are needed to determine the extent to which LLMs can improve the qualitative research process. This is a first attempt to investigate whether the LLM ChatGPT can be used for coding (un)structured, qualitative data and assessing inter-rater reliability. Through our exploration, we aim to provide researchers with valuable insights to inform their qualitative analysis

endeavours when incorporating AI techniques.

Methods

Providing Datasets and Tasks to ChatGPT to Code Qualitative Data

An experimental approach was employed to investigate the coding capabilities of ChatGPT. The primary objective was to assess whether ChatGPT can perform open, axial, and selective coding and to evaluate its suitability for inter-rater reliability. We aimed to determine if the codes generated by ChatGPT align with those produced by human researchers and whether ChatGPT can contribute to improving inter-rater reliability.

Experiment Setup

We signed up for a free account from OpenAI to use ChatGPT. ChatGPT was approached using both the chat interface (chat.openai.com) using natural conversation techniques to request certain data and via an API connection (Application Programming Interface, so that two systems can communicate with each other) between a Google Spreadsheet and ChatGPT, using this plugin:

> https://workspace.google.com/marketplace/app/gpt_for_sheets_and_docs/677318054654

In this latter setup, research data is organised in rows and columns and questioned using programmatic 'prompts'. Questioning AI is typically referred to as prompting, where a prompt can be either a natural sentence (i.e., "Please name five types of birds") or more programmatic (GPT_LABEL, C1:C20,5,0.3; In order: invoking a predefined function of ChatGPT, cell-range, number of labels to generate, level of creativity) used in the Google Spreadsheet.

In this experiment we have incorporated several datasets, consulted coding literature to define the prompts, and utilised datasets previously coded by two independent researchers with established inter-rater reliability. The datasets were selected through convenience sampling, taking into account a diverse range of datasets to ensure a well-rounded perspective. This variability extended to the level of dataset structure, where some datasets had pre-existing coding categories and others did not. Moreover, data sensitivity was a crucial factor in selection, with priority given to datasets with sufficient anonymisation and non-sensitive subjects. Given the uncertainty about the ultimate use of this data within ChatGPT, these precautions were taken carefully.

For this experiment, we employed (un)structured data consisting of three datasets:
- Dataset 1: An unstructured dataset consisting of two interview transcripts with curriculum developers, exploring the theme of responsive curriculum development (Vreuls, Koeslag-Kreunen, van der Klink, Nieuwenhuis, & Boshuizen, 2022). This study investigated how curriculum developers respond to changes in professional practice during the process of developing educational programs (curricula) for higher professional education and training. It revealed the characteristics of this complex development process. This dataset was previously coded and inter-rater reliability has already been established by researchers.
- Dataset 2: A structured dataset about students tagging and scoring video clips based on four predefined

sensitising categories in a study focused on observing teachers' interpersonal behaviour (Theelen, van den Beemt & den Brok, 2019). Again, this dataset was also previously coded and inter-rater reliability has already been established by researchers. We verified if ChatGPT was familiar with the four in literature defined categories prior to the study.

- Dataset 3: A self-constructed dataset by ChatGPT about listening levels. This dataset consists of a transcript of a group meeting in which team members deliberated and sought consensus on the future direction of a company. We consulted ChatGPT for insights on the theory of levels of listening (Covey & Walsmit, 2011) in conversation and requested a transcript of the meeting. The first two authors then coded the transcript themselves to first establish 'human' inter-rater reliability.

The Coding Process

To guide ChatGPT's coding process, we provided prompts. For coding dataset 1 we used the chat interface. ChatGPT received specific selections of the transcript and the following prompt for open coding: "Label the following text using as many labels as necessary and indicate which sentences/ quotes the labels correspond to". Subsequently, we compared the labels assigned by ChatGPT with the open-coded text generated by human researchers. Next, we provided the following prompt for axial coding: "Label the following text using the labels: XX, XX, XX", and the prompt for selective coding: "Select the most relevant, meaningful, and representative codes for further analysis. Prioritise codes that contribute significantly to understanding the research aim and research questions (i.e., how do curriculum developers define and give substance to responsive curriculum development)."

Finally, regarding the identification of themes or patterns, the chat interface, ChatGPT, was provided with the following prompt after applying the previously mentioned coding strategies. These themes were derived from a large segment of a single interview. "Can you perform a Theme or pattern identification with this output? This involves recognizing common elements, trends, contradictions, or other related aspects within the data. Themes can be identified by comparing different codes and seeking recurring ideas or concepts."

For this dataset, we also used the API connection between Google spreadsheets and ChatGPT using the following three syntax: (1) =GPT("Think of one word that describes the core of the text",A3), (2) =GPT("Think of one label that describes the core of the text",A3), and (3) =GPT_TAG(A6,,A$9:B$14). The latter prompt asks that ChatGPT invents as many labels as it deems useful and gives five examples of data and corresponding labels.

Dataset 2 was used to evaluate the suitability of ChatGPT for inter-rater reliability when coding the textual information in accordance with the given set of four categories. For this, we used the API connection between Google spreadsheets and ChatGPT. We 'trained' ChatGPT by giving some examples of correctly coded tags. We used prompts like this: =GPT_TAG(A24,N$13:O$16,A$25:B$40,1).

Dataset 3 was also used to evaluate the suitability of ChatGPT for inter-rater reliability when coding the textual information in accordance with the given set of five categories. For this, we used the chat interface. In the prompt,

we gave ChatGPT its own definitions of levels of listening. We then entered the transcript of the fictitious meeting. Finally, we gave the following instruction: "Specify which fragments of the text the labels apply to".

Results

Open Coding

To investigate ChatGPT's potential for open coding, dataset 1 about responsive curriculum development was employed. The labels generated by ChatGPT for dataset 1, which consisted of interview data, exhibited a high level of accuracy, as indicated in Table 1. On the left side of the table, the labels given by the researchers of the original dataset are presented. The right side shows the labels assigned by ChatGPT in the current study. Although there were occasional instances where ChatGPT utilised different terminology than what the researcher had used, the underlying content-based labels remained accurate. It is important to note that Table 1 presents a sample, but the complete interview was consistently coded accurately by ChatGPT.

Table 1: Original Labels Versus Labels ChatGPT of Dataset 1 (sample)

Original label	Label ChatGPT
Vision	Outdated vision document
Anticipation	Anticipating the future
Learning to learn	Development of learning strategies/ learning to learn
Curriculum development	Curriculum development
Co-creation	Co-creation

Axial and Selective Coding

To explore ChatGPT's proficiency in axial and selective coding, dataset 1 was utilised once again. During the process of axial coding, a sufficient level of similarity was observed between the findings generated by ChatGPT and those of the original research, despite the use of different terminology by ChatGPT and the researchers. Table 2 showcases the categories identified by the researchers alongside the labels generated by ChatGPT.

Table 2. Original Categories Versus Categories ChatGPT of Dataset 1 (sample)

Original categories researchers	Categories ChatGPT
Quality curriculum developers	Difficulty for teachers
Responsive curriculum	Future of education
Authentic curriculum	Learning environment instead of school
Social-political process educational institution	Role of an educational institution

During the process of selective coding, ChatGPT identified and prioritised several relevant, meaningful, and representative codes that are pertinent to the research objective of understanding how curriculum developers give substance to responsive curriculum development. The following output provides examples of the selective codes mentioned by ChatGPT:

"1. Responsive Pedagogy: Codes related to the integration of responsive pedagogical approaches and instructional strategies within the curriculum, considering the needs and interests of the learners.

2. Authentic Assessment: Codes related to the use of authentic and meaningful assessment methods that align with the goals and objectives of the responsive curriculum.

3. Collaboration and Partnerships: Codes related to the collaboration and partnerships with various stakeholders, such as teachers, students, families, community organisations, and experts, to co-develop and implement the curriculum.

4. Continuous Professional Development: Codes related to providing ongoing professional development opportunities for curriculum developers and teachers to enhance their knowledge and skills in implementing responsive curriculum development.

5. Flexibility and Adaptability: Codes related to designing the curriculum with flexibility and adaptability in mind, allowing for modifications and adjustments based on evolving needs and contexts."

All these codes were addressed in the original research, deemed relevant to the research objective, and in alignment with the existing literature on responsive curriculum development (e.g., Vreuls, Koeslag-Kreunen, van der Klink, Nieuwenhuis, & Boshuizen, 2022). This congruence establishes a robust groundwork for conducting further analysis and gaining a deeper understanding of how curriculum developers effectively gave substance to responsive curriculum development.

Theme or Pattern Identification

In order to explore ChatGPT's capability for theme or pattern identification, dataset 1 was utilised. Once again, a notable convergence was observed between the findings generated by ChatGPT and the themes elucidated in the original research. To elucidate these findings, the subsequent Table 3 showcases pertinent quotes that exemplify the identified themes and patterns as delineated in the original research article.

Table 3. Original Themes Versus Themes ChatGPT of Dataset 1 (sample)

Fragments about themes in the original research article	Themes ChatGPT
An open flexible and authentic curriculum Curriculum developers mentioned how they implemented communities where learning was centred around authentic tasks. These communities consisted of students, professionals, and teachers. In such communities, the learning and working	*Shift in teaching approach* These codes suggest a shift in the traditional teaching approach towards a more experiential and practical learning environment. The focus seems to be on providing relevant and contextual information, prioritising real-world experiences, and

environments were interconnected and highly interdependent, which aligns with notions of a hybrid learning environment.

distinguishing between knowledge and mere information.

The challenging context, roles, and responsibilities of curriculum developers at the school level
Curriculum developers mentioned that the curriculum development process was a complex and time-consuming task. They indicated that qualification frameworks and classifications, (institutional) rules and regulations, exam regulations and examination boards were particularly restrictive in pursuing a responsive curriculum development process.

Challenges faced by teachers
These codes indicate the challenges and changes experienced by teachers. It suggests that teachers face difficulties in adapting to changing educational practices and may require support in adjusting their roles and teaching methods.

An open, flexible, and authentic curriculum
Curriculum developers considered the openness, flexibility and authenticity of the curriculum to be important characteristics of a responsive curriculum. In their view, a responsive curriculum is a kind of (partially) open curriculum, with the form of a "curriculum vitae", rather than a finite (four-year) programme.

Future of education
These codes highlight the evolving nature of education in response to the changing needs of the future and the labour market. There is an emphasis on creating a learning environment that goes beyond traditional classroom settings and aligns with the demands and interests of students.

Coding for Inter-Rater Reliability

To investigate ChatGPT's ability to encode texts using predefined categories provided by human researchers, all three datasets were used. With Dataset 1, ChatGPT encountered minimal difficulties as seen below in table 4. Table 4 displays meaningful segments of interview transcripts (excerpts) from dataset 1. These excerpts are presented in the left column. The original labels assigned by the researchers of dataset 1 are shown in the middle column, in comparison to the labels provided by ChatGPT in the right column.

Table 4. Excerpts, Original Labels Versus Labels ChatGPT (sample)

Excerpt	Original label	Label ChatGPT
Yes, uh, I think, think we are doing very well, a number of things I think are very good. Uhm. Particularly the didactic design is very good, uh also uh, the choice for uh, working in an exemplary way, i.e., not putting everything in, but rather choosing a	Didactic design, Curriculum development, Authentic learning situation	Didactic design, Curriculum development, exemplary work

number of uh well-considered elements and elaborating them systematically, so that students also learn to recognise the coherence and the underlying principles and can apply them. And, uh, and in which we have also really tried and occasionally made very good successes

1: and that is the collaboration with professional practice and also involving uhm other stakeholders like the students and uh.... 2: uh are those examples of which could be better? 1: yes that could be done better, yes we have made really good successes in that, hey with that uh, we started with that uh really design groups where all parties had a role uh in, and uh we really kept that up for a long time, we are uh, we have chosen uh critical professional situations, as a starting point that uh yes that were the basis for the elaboration of the further modules, and yes uh for example the patients and students really played an important role in those choices, but at a certain point there was such pressure to deliver things as well, and yes with all uhm yes, knowing that, how important it is to involve those parties, hey yes. it does delay. And at a certain point it all goes faster if you uh, if you do it with a team of teachers uh, so yes then there was also a bit, it was somewhat at the expense of involving external stakeholders.	Co-creation/stakeholder engagement	Collaboration with professional practice and stakeholders

It is important to emphasise that Table 4 provides only a condensed representation of the entire dataset. The labels generated by ChatGPT maintained a one-to-one correspondence with the labels provided by the researchers, resembling the process of open coding. Whether employing a single word, a single label, or multiple meaningful labels, ChatGPT consistently matched the original text.

In dataset 2, a lower inter-rater reliability was observed, as evidenced by the results presented in Table 5. The basis for compiling this dataset was formed by tags given by preservice teachers after watching video excerpts of key classroom events. These tags included insights into teachers' interpersonal behaviour. The tags were evaluated using a framework with four predefined categories:

1. Descriptions: meaning they only contain information about observable/ overt classroom events.

2. Positive or negative evaluations: meaning they contain an appraisal of what is seen in the video.
3. Analytic chunks: these are tags that contain information about underlying principles to learning and teaching.
4. Prescriptions for teacher action: which contain information about alternatives for teacher action.

The 'original score' reflects the assessment assigned by human researchers to the tags, accompanied by the corresponding score generated by ChatGPT. In the initial attempt the inter-rater reliability yielded a score of 34 percent. However, upon providing ChatGPT with supplementary information pertaining to the underlying theoretical model via syntax, the inter-rater reliability score increased to 43 percent. Again, the table is only a condensed representation of the entire dataset.

Table 5. Tags, Original Scores, and Scores ChatGPT of Dataset 2 (sample)

Tag	Original score	Score ChatGPT
No patience	2	1
Very strict	2	2
Unclear	2	2
Correcting	3	3
Unsatisfied	3	4

In order to expound upon the plausible reasons contributing to the observed low inter-rater reliability, dataset 3 was utilised, encompassing a transcript of a team meeting. This transcript was evaluated using five predefined listening levels (Covey & Walsmit, 2011):

- Listening Level 0: Ignoring the other person. There is clearly no listening involved here.
- Listening Level 1: Pretending to listen. You say 'yes, yes' and perhaps nod your head, but your thoughts are elsewhere.
- Listening Level 2: Selective listening. Based on our own frame of reference, we make an unconscious selection of the information we receive. While other information passes us by, we remember what fits into our frame of reference.
- Listening Level 3: Attentive listening. When we listen attentively, we are fully engaged with another's story. This level of listening is often experienced as very uplifting because someone is really there for you.
- Listening Level 4: Empathic listening. Empathic listening goes one step beyond attentive listening. Besides listening to what someone says, you listen to how they tell their story. You pay attention to their non-verbal cues, put yourself in their emotional world, and ask how they are feeling. By offering emotional reflections, you show them that you empathise with them and acknowledge how they're feeling.

Before using ChatGPT for inter-rater reliability of dataset 3, it was verified whether ChatGPT comprehended the fundamental theoretical constructs. For this dataset, an inter-rater reliability of 70 percent was established, as depicted in Table 6. Table 6 presents meaningful excerpts from the transcript of dataset 3 in the left column. The original scores assigned by the researchers of this study are displayed in the middle column, in comparison to the scores provided by ChatGPT in the right column.

Table 6. Excerpts, Original Scores, and Scores ChatGPT of Dataset 3 (sample)

Excerpt	Original score	Score ChatGPT
I think we need to strike a balance between marketing and product development. Both are important for the success of our project	3	3
Let us not forget that the financial side is also important. We need to make sure we stay within our budget	2	2
I understand we have different perspectives, but I think we should focus on customer satisfaction and feedback.	4	3

Discussion and Conclusion

Exploring the Potential of ChatGPT for Qualitative Data Analysis: Findings and Future Directions

We conducted an initial study to determine whether language models such as ChatGPT can serve as valuable tools for qualitative data analysis and, in particular, for open, axial and selective coding, theme or pattern identification, and inter-rater reliability. Our findings show promising potential, although the effectiveness of the results vary depending on the specific requirements and the quality of the prompt.

Open Coding

Initially, we aimed to assess ChatGPT's ability to *open label* qualitative data accurately. In our experiment, ChatGPT demonstrated considerable competence. We provided several texts for labelling, and the assigned labels were consistently correct. The labels generated by ChatGPT for dataset 1 turned out to be very accurate. Occasionally, ChatGPT employed different terminology than what we had used, but the content-based labels remained accurate. This indicates that the model successfully identified and labelled the various themes present in the interviews without any errors. Achieving such a high level of accuracy demonstrates ChatGPT's effectiveness in open coding tasks and highlights its potential as a valuable tool in the analysis of qualitative data.

We did encounter some challenges in formulating the appropriate prompts to elicit the desired results. Initially, we encountered difficulty in associating texts with their respective labels, as we were only shown the labels

without clear indications of the corresponding texts. However, using the interactive chat interface, we quickly refined our prompts until we achieved the desired outcomes.

Axial and Selective Coding, and Theme or Pattern Identification

Thus far, our experience with ChatGPT has been quite positive. Open coding, which is often the most time-consuming aspect of data analysis, was facilitated satisfactory by ChatGPT's capabilities. However, we encountered greater challenges when attempting axial coding. ChatGPT appeared to have a lesser understanding of this coding technique, which also seemed to be attributed to the inadequacy of our prompts in guiding it through these specific tasks. Upon inquiring about the meaning of axial coding to ChatGPT, it provided us with a definition that we utilised to modify the prompt according to its interpretation of axial coding. With this refined prompt, ChatGPT appeared capable of successfully executing the task. Although ChatGPT used different terminology than our own, it successfully grasped the fundamental concepts, ensuring the accuracy of the results. The model's proficiency in open and axial coding tasks emphasises its capabilities and suggests substantial potential in various stages of the coding process. Moreover, an interesting observation emerges from the analysis: at times, the researcher provided more nuanced details, while in other instances, ChatGPT demonstrated its ability to comprehend nuanced details within the data, showcasing its capacity to capture complex information during qualitative analysis. Furthermore, ChatGPT demonstrated competence in selective coding as well as theme or pattern identification, as it was able to generate additional information that complemented the researcher's insights.

Coding for Inter-Rater Reliability

Additionally, we were interested in examining whether ChatGPT is capable of coding texts using predetermined categories or labels that we provided. This aspect is particularly relevant for establishing inter-rater reliability in our own coding process. In order to ensure objectivity and accuracy in the analysis, it is common practice for a second researcher to independently code a portion of the dataset, allowing for the calculation of agreement levels between different coders. We conducted experiments using all three datasets.

When using ChatGPT to code texts based on predetermined categories, we observed variations in its performance, depending on the dataset in question. With Dataset 1, ChatGPT encountered minimal difficulties. The labels corresponded one-to-one with the labels provided by the researchers, similar to the process of open coding. Whether using a single word, a single label, or multiple meaningful labels, they consistently matched the original text. This can be attributed to the fact that the labels are often directly present in the text itself, which explains the high level of agreement.

With dataset 2, ChatGPT encountered significant difficulties. The tags primarily comprised single words lacking contextual information; however, they were associated with an underlying educational model known as the Interpersonal Circle Teacher (Horowitz & Strack, 2011). Within the framework of this theoretical model, tags incorporating words such as 'strict,' 'understanding,' and 'friendly' presented challenges in assigning accurate scores due to the limited visibility of their connection to the underlying theory in the absence of additional context.

Consequently, ChatGPT faced challenges in effectively coding based on these tags. Nevertheless, ChatGPT was well-acquainted with the theoretical model (learning to notice model; Sherin & van Es, 2005) that formed the basis for the encodings. During the initial attempt, we achieved an inter-rater reliability of 34 percent. ChatGPT exhibited noticeable challenges primarily in dealing with code 2, which encompassed both positive and negative evaluations of teacher behaviour. The complexity arose from the inherent ambiguity involved in accurately interpreting and categorising such evaluations. The lack of explicit contextual cues made it difficult for ChatGPT to accurately distinguish the intended sentiment of ratings from the teacher's behaviour as described in the tags. As a result, the model struggled to consistently code these instances. It became apparent that further refinements were necessary to enhance ChatGPT's ability to accurately capture the nuances of teacher evaluations, encompassing both positive and negative dimensions. However, by providing ChatGPT with additional information about the underlying theoretical model through syntax, we observed an increase in the score to 43 percent. Although this score still remains relatively low, it emphasises the importance of enhancing the prompts to improve the accuracy of ChatGPT's coding.

Due to the complexity and high contextual dependency of Dataset 2, we decided to generate a third dataset using ChatGPT. The aim was to investigate whether a dataset that did not explicitly contain labels (as in Dataset 1) but still provided more contextual information, thus being less complex than Dataset 2, would yield improved results for ChatGPT. In this case, ChatGPT autonomously generated the labels based on existing literature, implying a level of familiarity with them. For this analysis we utilised the Chat interface. As anticipated, ChatGPT demonstrated significantly improved coding performance on this dataset. Upon initial attempt the inter-rater reliability reached 70 percent. ChatGPT seemed to have the most difficulty in correctly assessing feeling reflections in this particular context.

This variability of results in terms of inter-rater reliability has been previously established in a study conducted by Khademi (2023). Our experiment underscores the significance of fine-tuning the prompts. Moreover, we observed that utilising ChatGPT via the API connection linking Google Spreadsheets and ChatGPT and feeding it examples, appear to enhance the level of inter-rater reliability. We believe this is due to programmatic prompts being well-defined and always the same when prompting sequentially, whereas the Chat interface is more a conversational model that is trained to receive and produce on conversational input – never giving the same answer to avoid corresponding like a computer. These findings highlight the potential benefits of feeding the LLM relevant theoretical information before requesting output and to utilise additional interfaces to improve the performance of ChatGPT in terms of coding accuracy.

Furthermore, it is important to keep in mind that LLMs such as ChatGPT are currently in public beta, with potential plans for transitioning to a paid only service in the future. The quality of these models varies, and their reliability raises some concerns. The free version of the model imposes limitations on the number of queries per minute, which is particularly relevant when integrating with Spreadsheets or other time-sensitive applications.

Above all, we see that getting valuable results from LLMs such as ChatGPT for research purposes only succeeds if the user has a good understanding of the research topic to create solid prompts and to validate the output.

Limitations

One limitation, however, was that the free version of ChatGPT can only process a small segment of text at a time. Consequently, we had to divide larger texts, such as interview transcriptions, and input them in a staggered manner. Nonetheless, this did not significantly impact the quality of the results obtained. This indicates that the paid version of ChatGPT which can process larger texts would still return similar qualitative results.

The adoption and suitability of ChatGPT for qualitative analysis remains relatively uncharted territory. Moreover, associated challenges and considerations, such as mitigating algorithmic bias, should be further explored, as ChatGPT possesses the potential to generate inaccurate or misleading responses (Mesec, 2023). Employing appropriate validation techniques becomes imperative in mitigating such risks.

Another limitation relates to the use of datasets obtained for this study through convenience sampling. This approach poses challenges to the generalisability of our research findings, as the datasets may not be representative for all types of qualitative research. There are also limitations to the use of three datasets. Qualitative data manifest themselves in different ways and at different levels. It is plausible that alternative datasets might have yielded different results.

In qualitative research, collecting 'sensitive' data is common practice. But even when data are not sensitive, they are consistently anonymised to avoid traceability to respondents. Moreover, in the context of academic research, much attention is paid to the secure storage of research data. The potential implications of introducing such data into ChatGPT remain uncertain. The forthcoming implications of this action are difficult to assess at this time. In the context of the current study, the selection of datasets was determined by these considerations, albeit limiting the choices available.

Improved Input Leads to Better Output

The primary finding of our study is the promising potential of ChatGPT for various coding facets (open, axial, selective, and theme or pattern identification) and assessing inter-rater reliability. The most crucial lesson learned is that improved input leads to better output. Notably, the utilisation of ChatGPT's own definitions appears to be particularly advantageous in this regard. Therefore, it is essential for researchers and other users intending to employ ChatGPT for coding datasets to receive adequate training in utilising ChatGPT's functionalities and creating the most accurate prompts. We, too, acknowledge that we have yet to fully explore and comprehend all the capabilities and functionalities of ChatGPT. As this investigation represents an initial exploration, there are still numerous opportunities for further exploration and advancement.

Implications of ChatGPT in Social Science Research: Advancing Qualitative Data analysis and Acknowledging Limits

The rapid evolution of AI and LLMs has led to significant advances in the integration of machine learning methods

in a wide range of fields, including social science research. The implications of our study's findings extend into the realm of social science research and resonate within the broader discourse on the integration of AI tools into academic endeavours. AI technologies are constantly evolving and their integration into research practices is gaining importance. Exploring ChatGPT's potential in qualitative data analysis is in line with the academic community's growing interest in using AI for research purposes. The positive results in open coding and insights into thematic patterns offer a glimpse of the potential efficiency gains AI can offer social science researchers when analysing extensive qualitative datasets.

Moreover, acknowledging the limitations and challenges of AI tools places itself in the ongoing debate on ethical and methodological considerations in the application of AI. While academia struggles with concerns about algorithmic bias, interpretability, and reliability of AI results, this study's recognition of ChatGPT's limitations contributes to a nuanced understanding of the role of AI in qualitative research.

In essence, this study contributes to the discourse on AI tools in academia by providing empirical insights into the usefulness and limitations of ChatGPT for qualitative data analysis. The implications of the study include both potential benefits and cautions and enrich the ongoing dialogue on the integration of AI tools within the methodological landscape of social science research.

Future Research: AI as Part of Researchers' Future

The world of generative AI is still evolving, with ChatGPT recently releasing version 4 (we used version 3.5 in our research), competitor Bard from Google entering the stage in Europe, and image generators such as Dall-E and Midjourney creating increasingly more realistic images. As this current research is a first exploration of how a LLM can contribute to social science, many opportunities to further investigate lie ahead. Following up on our research, it is interesting to explore how ChatGPT handles different types of datasets, or how different LLM's interpret the same datasets. And, on another level, could LLM's trained to converse in natural language assist with creating interview formats, or even take the interviews. As the world is still figuring out how generative AI is going to be part of all of our lives, as researchers, we too have a mission to explore the benefits for science of this great leap in technology.

Getting Started Coding in ChatGPT Yourself?

Engaging with ChatGPT: As of May 2023, an OpenAI-account (https://openai.com/) for ChatGPT is currently available free of charge. Users have the opportunity to access and utilise the model for various tasks.

Integration with Google Spreadsheets (plugin + instruction; https://gptforwork.com/): ChatGPT can be seamlessly integrated with Google Spreadsheets, which is Google's equivalent of Microsoft Excel. This integration is facilitated through a dedicated plugin that enables users to leverage ChatGPT's capabilities within the spreadsheet environment. Detailed instructions are provided to guide users on how to install and utilise the plugin effectively.

Upcoming Availability of Google's Bard in the Netherlands: Google's Bard (https://bard.google.com/?hl=en), a natural language processing system similar to ChatGPT, is anticipated to become accessible in the Netherlands in the near future. This development will provide Dutch users with the opportunity to engage with Bard's language generation capabilities.

References

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative Artificial Intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *SSRN*. Retrieved on 22nd August, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4337484.

Cascella, M., Montomoli, J., Bellini, V., & Bignami, E. (2023). Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. *Journal of Medical Systems, 47*(1), 1–5. https://doi.org/10.1007/s10916-023-01925-4

Covey, S. R., & Walsmit, V. (2011). *De zeven eigenschappen van effectief leiderschap* [The seven qualities of effective leadership]. Business Contact.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications.

Dönmez, I., Idil, S., & Gulen, S. (2023). Conducting academic research with the AI interface ChatGPT: Challenges and opportunities. *Journal of Science, Technology, Engineering, Mathematics and Art Education, 2*(6), 101-118. https://doi.org/10.55290/steam.1263404.

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023): A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, DOI:10.1080/14703297.2023.2195846.

Haque, M. U., Dharmadasa, I., Sworna, Z. T., Rajapakse, R. N., & Ahmad, H. (2022). "I think this is the most disruptive technology": Exploring sentiments of ChatGPT early adopters using Twitter data. arXiv. https://doi.org/10.48550/arXiv.2303.03836.

Horowitz, M. L., and S. Strack. 2011. *Handbook of Interpersonal Psychology*. New York: Wiley.

Khademi, A. (2023). Can ChatGPT and Bard generate aligned assessment items? A reliability analysis against human performance. C*omputer Science,* https://doi.org/10.48550/arXiv.2304.05372.

Malik, A. (2023). Google introduces Product Studio, a tool that lets merchants create product imagery using generative AI. *TechCrunch*. Retrieved on 22nd August, from https://techcrunch.com/2023/05/23/google-product-studio-tool-lets-merchants-create-product-imagery-using-generative-ai/?guccounter=1.

Mann, D. L. (2023). Artificial Intelligence discusses the role of artificial intelligence in translational medicine. *JACC: Basic to Translational Science, 8*(2), 221–223. https://doi.org/10.1016/j.jacbts.2023.01.001

Mesec, B. (2023). The language model of artificial intelligence chatGPT - a tool of qualitative analysis of texts. https://doi.org/10.22541/au.168182047.70243364/v1

Peres, R., Schreier, M., Schweidel, D., & Sorescu, A. (2023). On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *International Journal of Research in Marketing, 40*(2), 269-275.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised*

multitask learners. OpenAI.

Rahman, M., Terano, H. J. R., Rahman, N., Salamzadeh, A., Rahaman, S. (2023). ChatGPT and academic research: A review and recommendations based on practical examples. *Journal of Education, Management and Development Studies. 3*(1). 1-12. doi: 10.52631/jemds.v3i1.175

Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach (3rd ed.).* Pearson.

Saldaña, J. (2015). *The coding manual for qualitative researchers.* Sage Publications.

Seidman, I. (2013). *Interviewing as qualitative research: A guide for researchers in education and the social sciences.* Teachers College Press.

Sherin, M. G., & Van Es, E. A. (2005). Using video to support teachers' ability to notice classroom interactions. *Journal of Technology and Teacher Education, 13*(3), 475–491.

Susnjak, T. (2022). ChatGPT: The end of online exam integrity? arXiv. https://doi.org/10.48550/arXiv.2212.09292.

Theelen, H., van den Beemt, A. & den Brok, P. (2019). Using 360-degree videos in teacher education to improve preservice teachers' professional interpersonal vision. *Journal of Computer Assisted Learning, 35*, 582–594. https://doi.org/10.1111/jcal.12361

Vreuls, J., Koeslag-Kreunen, M., van der Klink, M., Nieuwenhuis, L., & Boshuizen, H. (2022). Responsive curriculum development for professional education: Different teams, different tales. *The Curriculum Journal, 33*(4), 636-659. https://doi.org/10.1002/curj.155

Weatherbed, J. (2023). Levi's will test AI-generated clothing models to 'increase diversity'. *The Verge.* Retrieved on 22nd August, from https://www.theverge.com/2023/3/27/23658385/levis-ai-generated-clothing-model-diversity-denim.

Author Information

Hanneke Theelen

iD https://orcid.org/0000-0003-0169-7922

Zuyd University of Applied Sciences

Netherlands

Contact e-mail: *hanneke.theelen@zuyd.nl*

Joyce Vreuls

iD https://orcid.org/0000-0001-5761-4773

Zuyd University of Applied Sciences

Netherlands

Jim Rutten

iD https://orcid.org/0009-0009-6071-0358

All Pension Group (APG)

Netherlands