The Future of Engineering Education
2024 Annual Conference & Exposition
Oregon Convention Center
Portland, OR . June 23 - 26, 2024
ASEE
Paper ID #42732

Evaluating ChatGPT's Efficacy in Qualitative Analysis of Engineering Education Research

Dr. Xiaorong Zhang, San Francisco State University

Dr. Xiaorong Zhang is an Associate Professor in Computer Engineering in the School of Engineering at San Francisco State University (SFSU). She is the Director of the Intelligent Computing and Embedded Systems Laboratory (ICE Lab) at SFSU. She has broad research experience in human-machine interfaces, neural-controlled artificial limbs, embedded systems, and intelligent computing technologies. She is a recipient of the NSF CAREER Award to develop the next-generation neural-machine interfaces (NMI) for electromyography (EMG)-controlled neurorehabilitation. She is a senior member of the Institute of Electrical and Electronics Engineers (IEEE) and a member of the Society of Women Engineers (SWE). She has served in professional societies in various capacities including the Chair of the IEEE Engineering in Medicine and Biology Society (EMBS) San Francisco Chapter (2018-present), an Associate Editor of the IEEE Inside Signal Processing E-Newsletter (2016-2018), an Outreach Co-Chair of the Society of Women Engineers (SWE) Golden Gate Section (2017-2018), a Co-Chair of the Doctoral Consortium at 2014 IEEE Symposium Series on Computational Intelligence, a Program Committee Member of various international conferences, and a regular reviewer of a variety of journals and conferences in related fields.

Dr. Stephanie Claussen, San Francisco State University

Stephanie Claussen is an Assistant Professor in the School of Engineering at San Francisco State University. She previously spent eight years as a Teaching Professor in the Engineering, Design, and Society Division and the Electrical Engineering Departmen

Fatemeh Khalkhal, San Francisco State University

Dr. Khalkhal is an assistant professor in mechanical engineering at San Francisco State University. She has a PhD in Chemical Engineering from Ecole Polytechnique de Montreal. Her research experience and interest are in developing structure-property relationships in complex fluids and broadening the participation of women and underrepresented minorities in engineering.

Yiyi Wang, San Francisco State University

Yiyi Wang is an assistant professor of civil engineering at San Francisco State University. In addition to engineering education, her research also focuses on the nexus between mapping, information technology, and transportation and has published in Accident Analysis & Prevention, Journal of Transportation Geography, and Annuals of Regional Science. She served on the Transportation Research Board (TRB) ABJ80 Statistical Analysis committee and the National Cooperative Highway Research Program (NCHRP) panel. She advises the student chapter of the Society of Women Engineers (SWE) at SFSU.

©American Society for Engineering Education, 2024

Evaluating ChatGPT's Efficacy in Qualitative Analysis of Engineering Education Research

Abstract

This study explores the potential of ChatGPT, a leading-edge language model-based chatbot, in crafting analytic research memos (ARMs) from student interview transcripts for use in qualitative data analysis. With a rising interest in harnessing artificial intelligence (AI) for qualitative research, our study aims to explore ChatGPT's capability to streamline and enhance this process.

The research is part of a mixed-methods project examining the relationships between engineering students' team experiences, team disagreements, and engineering identities. Our team had previously developed an interview protocol for collecting qualitative data and initiated analysis using coding methods and ARMs for individual transcripts. We designed an ARM Development Guidelines document to ensure consistency among four team members in the ARM creation process. The guidelines include a set of key questions that each ARM should aim to address.

Our objective is to assess ChatGPT's proficiency in creating ARMs based on our development guidelines and compare its outputs with human-written ARMs for accuracy and depth of insight. For this purpose, we selected two student interview transcripts. A structured analysis protocol for ChatGPT was devised in adherence to the ARM Development Guidelines.

Two team members, experienced in qualitative analysis and ARM composition, drafted ARMs for the chosen transcripts using the same guidelines, enabling a direct outcome comparison. Subsequently, a rigorous validation process was conducted, using rubrics to assess narratives from both methods. The manual ARM authors performed a self-assessment, while the other researchers conducted a anonymous evaluation of the human-generated and AI-generated ARMs. We used two rubrics for this comparison. A general rubric gauged accuracy, clarity, analysis time, and usefulness. A specialized rubric was used to determine if the ARMs address the topics laid out in the ARM guidelines, such as self-identification, perceptions of engineering, teamwork descriptions, connections between identity and team experiences, comparisons with other interviews, and reflections.

In this paper, we describe our research methodology, present our findings, evaluate the advantages and limitations of ChatGPT in qualitative analysis within engineering education research, and provide guidance for future research directions. We aim to shed light on the capabilities of ChatGPT in qualitative analysis and contribute to the ongoing dialogue on harnessing AI for research in engineering education. Our findings will inform researchers and practitioners about the benefits, challenges, and best practices associated with integrating AI-powered tools such as ChatGPT into qualitative research methods.

Introduction and Background

Artificial Intelligence (AI) tools, such as the advanced large language model (LLM)-based chatbot ChatGPT, have increasingly been integrated into various stages of academic research. These stages include the creation of study introductions and objectives, conducting literature reviews, data analysis, and brainstorming methodologies [1], [2]. In qualitative research, there is a growing interest in leveraging AI tools to enhance data analysis processes. This research often entails detailed analyses of diverse data types to unearth study participants' nuanced perspectives. Traditionally, these analyses have depended heavily on manual efforts, which are not only time-consuming but also vary based on the analyst's expertise and perspective. The introduction of AI tools like ChatGPT marks a potential paradigm shift in the processing and interpretation of qualitative data.

Feuston and Brubaker discussed the role of AI tools in the sampling and coding phases of qualitative research, especially with large data sets [3]. Their study delves into the nuanced views of scholars regarding AI's influence on qualitative analysis, highlighting the importance of balancing human expertise with AI capabilities. Zhang et al. explored the use of ChatGPT in thematic analysis, particularly in coding qualitative analysis, and found that well-crafted prompts could enhance the quality of thematic analysis [4]. The same team also developed QualiGPT, a specialized tool integrating LLMs like GPT for thematic analysis, aimed at improving coding and theme identification in qualitative research [5]. Previous research has recognized the vast potential and benefits of employing LLM-based AI tools in qualitative research, such as rapid data processing, increased productivity, and the ability to provide concise summaries and preliminary insights. However, concerns and challenges persist, including biases, ethical considerations, the lack of domain expertise, and interpretability issues [1], [4], [6].

In engineering education, qualitative research often involves intricate analyses of interviews, focus groups, and observational data to capture the subtle viewpoints of students and educators. The utilization of AI tools such as ChatGPT in this domain of qualitative research is still relatively unexplored. This study delves into the novel application of ChatGPT for qualitative analysis in engineering education research. While previous investigations have primarily focused on the utility of ChatGPT in the coding phases of qualitative research, our study centers on evaluating ChatGPT's effectiveness in generating Analytic Research Memos (ARMs) from student interview transcripts.

ARMs are a pivotal component in qualitative research, offering researchers a structured method to document, analyze, and communicate their findings. These memos serve multiple purposes: they facilitate the process of data analysis, assist in the organization and synthesis of complex information, and aid in the development of theories and frameworks [7]. ARMs are particularly useful in capturing the researcher's insights, observations, and interpretations as they engage with the data [8]. This process often involves an iterative cycle of data collection and analysis, where memos play a critical role in tracking the evolution of the researcher's understanding. As a result, generating ARMs can be a time-consuming process, although this varies depending on the complexity of the research and the depth of analysis required. The format and content of ARMs can vary widely depending on the research approach and the researcher's preferences. The memos can range from brief notes to extensive, detailed documents and are often revisited and revised throughout the research process.

Our study is part of a larger project examining the intersection of teamwork experience and engineering identity among engineering students, particularly from diverse backgrounds, funded by the National Science Foundation's Research Initiation in Engineering Formation (RIEF) program. This project seeks to uncover how teamwork—especially team disagreements—connects to engineering identity, employing a mixed-method approach [9]. Conducted at San Francisco State University, a Hispanic Serving Institution in the Western U.S., our study collected quantitative data through surveys and qualitative data through interviews with students from over 20 engineering classes spanning three semesters. We developed a semi-structured interview protocol and conducted 28 one-hour interviews with students.

The qualitative analysis by our team underwent several phases [10]. We first combined Saldaña's initial and provisional coding methods [7] to code five interview transcripts, which helped us create a unified codebook. Subsequently, this codebook was used to analyze additional transcripts, with separate ARMs for each. Despite the systematic rigor of this approach, it proved to be labor-intensive. To streamline the analysis, we shifted to a 'quick' ARM method for the remaining transcripts, guided by a set of ARM Development Guidelines crafted by our team. This strategy significantly reduced analysis time from 4-8 hours to 1.5-4 hours per transcript, though still required considerable effort.

Our interest in ChatGPT stems from its potential to accelerate the analysis process further. We aimed to evaluate ChatGPT's capability in generating ARMs based on our guidelines, comparing its performance to human-crafted ARMs in terms of accuracy and insightfulness. For this, we selected two student interview transcripts and established a structured analysis protocol for ChatGPT, aligned with our ARM Development Guidelines. Meanwhile, two team members experienced in qualitative analysis prepared ARMs for the same transcripts, enabling a direct comparison of AI-generated and human-generated memos. We designed a thorough validation process, including self-assessment by the ARM authors and anonymous evaluation by other researchers, utilizing two rubrics to assess accuracy, clarity, analysis time, usefulness, and adherence to the ARM guidelines.

The rest of this paper will detail our methodology, present our findings, and discuss the benefits and limitations of integrating ChatGPT into qualitative analysis for engineering education research.

Methods

To gather qualitative data, our team devised a semi-structured interview protocol comprising four segments: introduction and warm-up, engineering identity, teamwork, and conclusion. When time permitted, we asked the interviewees to reflect upon stories of practicing engineers, which were compiled from publicly accessible accounts of the day-to-day experiences of practicing engineers. This interview framework and other relevant aspects of our research design received approval from our institution's Institutional Review Board.

Throughout the RIEF project, we conducted a total of 28 interviews—10 in the Spring of 2022 and an additional 18 in the Fall of 2022. These interviews, facilitated via Zoom, typically spanned one hour. The audio recordings from the Spring 2022 sessions were transcribed and cleaned by a professional service, while the Fall 2022 interviews underwent auto-transcription by Zoom, followed by a cleaning process by our research team. To ensure the privacy of our participants, all transcripts were manually anonymized and de-identified before they were employed in qualitative analysis.

For the purposes of this study, we selected two student interview transcripts that remained unanalyzed by our team at the onset of this research, intending to conduct a comparative examination and evaluation of these transcripts using a dual approach that incorporates both ChatGPT and traditional human analysis techniques.

Data Analysis

The research team created the ARM Development Guidelines to streamline the quick ARM method-based analysis as follows:

> *Write a short (~1-2 pages) Analytic Research Memo (ARM) about any noteworthy codes that emerged from your coding. Potential things to keep in mind as you write your ARM include:*
> - *Who is this interviewee? How do they self-identify (both with regards to their engineering identities and their non-engineering identities)?*
> - *How do they explain or justify their identities?*
> - *How do they perceive engineering? Does this perception relate to how they see themselves as engineers (or not see themselves as engineers)?*
> - *What is the interviewee describing with regards to their teamwork? How does the interviewee characterize and understand what went on in their teams?*
> - *What connections does the interviewee make between their engineering identity/identities and their experiences in teams?*
> - *How is what is going on in this interview similar to, or different from, other interviews?*
> - *What surprised me? What intrigued me? What strikes me?*

The human researchers followed these guidelines to craft the ARMs for each transcript.

ChatGPT Analysis Protocol

For this study, we employed ChatGPT 4 as our analytical tool, given its superior processing capabilities over ChatGPT 3.5, albeit requiring a subscription. Initially, both versions had a limitation of approximately 4096 tokens per interaction, including the user's prompt and the model's response. This limit equates to about 1000 to 1500 words, depending on the complexity of the text. Our transcripts, typically ranging from 6000 to 8000 words, exceeded this limit. To address this challenge, the original protocol involved dividing each anonymized transcript into segments, creating sub-summaries, and then compiling a comprehensive ARM in accordance with the ARM Development Guidelines. By November 2023, ChatGPT 4 introduced a file upload feature, significantly increasing the token limit to 2 million per text file, thus enabling the direct upload of transcripts for processing without prior segmentation.

The initial ChatGPT prompt instructions were slightly adjusted from our original ARM Development Guideline, particularly in the leading paragraph, to provide a clearer background and purpose for the analysis, as well as more explicit requirements for the generated output. For example, we replaced “a short (1-2 pages) ARM” with “a 500-1000 word ARM” in the instructions to ChatGPT. However, we retained all bullet-point questions unchanged from the original guidelines. The revised leading paragraph in the prompt instructions is as follows:

> *Write a 500-1000 words Analytic Research Memo (ARM) about the attached anonymized student interview transcript. The purpose of the ARM is to conduct qualitative analysis of a mixed-methods research project examining the relationships between engineering students' team experiences, disagreements, and their engineering identities. Potential things to keep in mind as you write your ARM include:*

Following some initial tests, we identified potential weaknesses in the original prompt. Notably, the lack of explicit reference to disagreements led to generated memos that failed to thoroughly explore the nuances of disagreements within team dynamics, an element vital to our study. Moreover, the memos only superficially addressed the pivotal link between engineering identity and team experiences, lacking in-depth analysis. Despite these omissions in the ARM Development Guidelines, our human researchers, drawing on their extensive experience with the project, implicitly understood the significance of these aspects and consistently considered them in their analyses. To address these oversights, we refined the prompt by adding specific questions aimed at probing these critical areas more deeply:

- *Please expand on the interviewee’s teamwork experience in the focused course [course pseudo name]. Provide more details about any disagreements encountered during teamwork, and explain whether and how this teamwork experience influenced or was influenced by his engineering identity.*
- *Does the interviewee’s teamwork experience make them feel more or less like an engineer, and why?*
- *Please also expand on the interviewee’s perception of engineers and engineering and explain their engineering identity.*
- *What about the interviewee’s teamwork experience in other engineering courses?*

Before feeding data and prompts to ChatGPT, we configured its Data Controls settings to disable chat history and model training functionalities. With chat history turned off, no input or output data will be used for model training or appear in the history sidebar. For monitoring and abuse prevention, ChatGPT retains all conversations for 30 days, after which they will be permanently deleted.

After completing the setup, we uploaded the anonymized transcript as an attachment to ChatGPT 4 and entered the prompt instructions into the chat box. ChatGPT promptly generated a response. This response, namely the ChatGPT-produced ARM, was subsequently saved in a Word document for future evaluation. Concurrently, an experienced member of our research team, with expertise in qualitative analysis and ARM creation, prepared an ARM for the selected transcript following the ARM Development Guidelines. This parallel process facilitated a direct comparison of outcomes during the evaluation phase.

Evaluation Process

The evaluation process was meticulously structured to utilize rubrics for assessing ChatGPT-produced and human-crafted ARMs. Each transcript's two ARMs underwent a dual assessment approach: the author of the manual ARM conducted a self-evaluation, while another researcher carried out an anonymous review of the ARMs. The anonymous evaluators received both ARMs without knowing their origin, along with the transcript, for their evaluation.

For the evaluation, two sets of rubrics were developed. The first, a general rubric, aimed to assess the overall accuracy, clarity, and usefulness of the ARMs. The second, a specialized rubric, focused on the extent to which the ARMs addressed the key questions outlined in the ARM Development Guidelines. These critical areas included the interviewee's self-identification, their perceptions of engineering, detailed accounts of teamwork, the interplay between individual identity and team experiences, comparative analysis with other interviews, and personal reflections. This comprehensive evaluation framework ensured a thorough analysis of the ARMs' quality and relevance.

Evaluation Results and Discussion

In the results section, we will present the outcomes of our evaluation process and discuss the findings. Additionally, the Appendix section contains the human-crafted and ChatGPT-generated ARMs for one of the analyzed transcripts.

Tables 1 and 2 collectively present the rating scores from the anonymous evaluations of both GPT-produced and human-generated ARMs, as well as the self-evaluation by the author of the human ARM on the GPT-produced ARM. Table 1 showcases the results obtained using the general rubric, while Table 2 presents the outcomes assessed with the specialized rubric. Additionally, evaluators were asked to supplement their ratings with comments, which will be summarized and discussed in the following section.

Table 1. Evaluation Results Using the General Rubric

Evaluator's Instruction: Please rate each ARM on a scale of 1 to 5 for each criterion listed in the table, where 1 represents the lowest and 5 represents the highest rating.

Criteria	**Description**	**Transcript #1**			**Transcript #2**		
		Anonymous Evaluation		Self-evaluation on GPT ARM	Anonymous Evaluation		Self-evaluation on GPT ARM
		Human ARM	GPT ARM		Human ARM	GPT ARM	
Accuracy	Comprehension of the transcript: Does the memo show a clear understanding of the data from the transcript?	4	4	4	5	3	4
	Coherence with the protocol: Does the memo adhere to the protocol established for analysis?	3	4	5	-	-	5

Clarity	Language quality: Is the memo well-written and free from grammatical and spelling errors?	5	5	5	5	5	5
	Structure: Is the memo well-structured, logical and easy to follow?	4	5	4	5	5	5
Usefulness	How useful is the memo for the intended purpose?	3	4-4.5	3	5	5	4

Note: In the table, the '-' symbol indicates that the evaluator did not assign a rating for the specified criterion.

Table 2. Evaluation Results Using the Specialized Rubric

Evaluator's Instruction: Please rate each ARM on a scale of 1 to 5 for each criterion listed in the table, where 1 represents the lowest rating and 5 represents the highest rating.

Criteria	**Description**	**Transcript #1**			**Transcript #2**		
		Anonymous Evaluation		Self-evaluation on GPT ARM	Anonymous Evaluation		Self-evaluation on GPT ARM
		Human ARM	GPT ARM		Human ARM	GPT ARM	
Self-Identification	Does the memo clearly describe how the interviewee self-identifies in terms of engineering and non-engineering identities?	5	5	3	5	4	4
	Does the memo adequately explain or justify the identities mentioned by the interviewee?	-	-	3	5	3	3
Perception of Engineering	Does the memo present a detailed description of the interviewee's perception of engineering?	2	4	4	5	4	3
	Does the memo establish a connection between the interviewee's perception of engineering and their self-identified engineering identity (or lack thereof)?	1	4	1	5	3	3
Teamwork Description	Does the memo provide a clear and comprehensive description of the interviewee's teamwork experiences?	3	5	5	5	5	4
	Does the memo effectively convey the interviewee's understanding and characterization of what happened within their teams?	3	5	3	5	5	3
Connection between Identity and	Does the memo identify connections between the interviewee's engineering identity and their team experiences?	1	4-5	1	5	4	4

Team Experiences	Does the memo delve into the deeper implications of these connections?	1	4	1	5	4	3
Comparisons	Does the memo appropriately draw parallels or contrasts with other interviews?	1	4	2	5	1	1
	Is the memo analysis thorough and meaningful?	1	4	1	5	4	3
Reflections	Does the memo accurately capture elements from the interview that were surprising, intriguing, or striking?	2	5	3	5	3	2

Note: In the table, the '-' symbol indicates that the evaluator did not assign a rating for the specified criterion.

The observations from the results reveal several points:

1. The self-evaluation ratings for the ChatGPT-produced ARMs are generally lower than the anonymous evaluation ratings for the ChatGPT-produced ARMs. This may be due to the researchers who created the human ARMs having spent more time and iterations on the transcripts than the anonymous evaluators, thus having a deeper understanding of the transcripts and the limitations of the ChatGPT-produced ARMs. It is also possible that this may have been a bias of the human authors as they unconsciously defended their own initial work and interpretations.

2. Generally, the ChatGPT-produced ARMs received relatively higher ratings (scales 4-5) in non-interpretative criteria, such as describing or summarizing the transcript content relevant to specific topics of interest (e.g., teamwork experience in specific courses, engineering identity). They received high ratings for language and structure clarity and coherence with the ARM Development Guidelines. The ChatGPT-produced ARMs performed even better than human-crafted ARMs in following the guidelines and attempting to address every question listed in the prompt instructions. In contrast, human researchers tended to focus more on questions they found intriguing, important, or noteworthy in their ARMs, and might not explicitly address all the questions in the guidelines, thus potentially receiving lower ratings in certain criteria from the anonymous evaluators.

3. The ARMs produced by ChatGPT generally received lower ratings, especially from self-evaluators, on criteria requiring interpretation, such as identifying or explaining the connections between the student's perception of engineering and their engineering identity or the links between the student's engineering identity and teamwork experience. A significant concern raised by all evaluators was the inaccuracy of ChatGPT's performance on interpretative tasks. Even more concerning was ChatGPT's tendency to fabricate facts or connections that were not true in an attempt to answer the questions in the guidelines. These inaccuracies were not always obvious to anonymous evaluators. Below are example quotes from evaluators' comments, with the following abbreviations representing the evaluators: T1A1 for the anonymous evaluator for Transcript #1; T1S1

for the self-evaluator for Transcript #1; T2A1 for the anonymous evaluator for Transcript #2; T2S2 for the self-evaluator for Transcript #2.

First, we observe T1S1 highlighting several instances where they believe GPT fabricated findings from the data. The quotes below are in response to the criterion "Connection between Identity and Team Experiences" (T1S1's rating was 1, whereas T1A1's rating was 4):

> *This was a clear weakness in the GPT-generated ARM. To me, GPT seemed to fabricate the stated connection between Omar's experience in [Engineering Course ST] and his engineering identity (I don't think Omar indicated that this experience made him feel more – or less – like an engineer).*
>
> …
>
> *As another example of where I think GPT may be giving an interpretation that the data does not support, it said, "Omar's teamwork experience in [Engineering Course ST] made him feel more like an engineer." I don't see Omar actually saying this. In short, GPT's attempts to determine the deeper implications of these connections seem to lead it to fabricating connections that aren't actually there.*
>
> …
>
> *Omar did have positive experiences in teams, but he did not connect that to his engineering identity when prompted. Instead, he connected his negative experiences in teams to his engineering identity. In addition, GPT seems to have made up that last sentence entirely ("Omar's ability to navigate these diverse team dynamics reflects the adaptability and resilience integral to engineering roles.").*

In another place, T1S1 discussed a connection that the interviewee made between their views of engineering and their own engineering identity (a central question to our research project, which, in the evaluator's opinion, GPT overlooked. This was in response to the criterion "Does the narrative establish a connection between the interviewee's perception of engineering and their self-identified engineering identity (or lack thereof)?" (T1S1's rating: 1, while T1A1's rating was 4):

> *I don't think the narrative made a connection between these two. However, I believe Omar did connect them (when he said that engineering is a profession where you need to have experience before you can be considered an engineer) – it seems GPT did not pick up on the significance of this part of the interview.*

T1A1's comments on the criterion "Usefulness" (T1A1's rating: 4-4.5, while T1S1's rating was 3) are as follows. In the comments, ARM 1 refers to the human-crafted ARM, and ARM 2 refers to the ChatGPT-produced ARM. The evaluator was uninformed about this information.

> *Compared to ARM 1, ARM 2 drew more detailed info from the interview and provided some interpretation and perhaps personal insights and analysis, …. It is*

> *also more descriptive even though it misses some terminologies Omar used to describe engineering, e.g., "engineering is mostly about designing, building, and making sure the math checks out for any sort of..." or the description he provided to explain what engineers do which is the definition of "design". It also seems the person who wrote ARM 2 took the liberty and added some info from their own knowledge about engineering profession or what engineering (and engineering identity) is when speaking about Omar; for example, "problem-solving" was never directly mentioned by Omar as a trait for an engineer or what engineers do but it was used in the ARM in few places when discussing Omar's EI or his perception of engineering [and this freaks me out if ARM 2 was produced by ChatGPT!] Overall, it seems to me that ARM 2 was written by a knowledgeable person ... but needs to be checked more carefully for the accuracy of the information provided in the ARM.*

The comments reveal that while ChatGPT-produced ARMs may seem to offer interpretative insights, the accuracy and reliability of these interpretations can be questionable.

T2A1 expressed similar concerns regarding the ChatGPT-produced ARM, as illustrated in the quotes below:

> *ARM1 (which refers to the ChatGPT-produced ARM) seems to "stretch" the transcripts, e.g., saying disagreement reinforces his EI. The transcript doesn't support that...*
>
> *ARM1 said resolving conflict helps grow his EI (engineering identity), which is not supported in the transcripts and in fact the other way round.*
>
> *I feel that ARM1 somewhat tries to "stretch" the narrative bit far, making connections that are not there.*

4. We did not expect ChatGPT to perform well for criteria that required comparative analysis since it was not provided with comparative data (such as other interview transcripts). However, we were still curious about how ChatGPT would handle these questions. To our surprise, despite the absence of comparative data, ChatGPT provided responses.

   ChatGPT's response to Transcript #1:

   > *Compared to other interviews, Omar's focus on practical skills and team dynamics stands out. His journey from confusion to clarity in understanding engineering is notable.*

   ChatGPT's response to Transcript #2:

   > *While this memo does not compare Archie's experience to other interviews, it is noteworthy that Archie's military background and prior degree in Computer*

> *Science provide him with a unique perspective compared to typical engineering students. His practical approach to engineering, influenced by real-world experiences, sets him apart.*

> The response to Transcript #1 is particularly concerning because ChatGPT gives the impression of having conducted a comparison, even though there was actually no other interview data to compare with. The response to Transcript #2 is more acceptable because it does not attempt to compare the analyzed interview with others. However, the reason for ChatGPT's inconsistent responses to similar prompts in different analyses remains unclear, raising concerns about the consistency and reliability of ChatGPT-based analysis.

Discussion and Conclusion

This study investigated ChatGPT's ability to generate ARMs from student interview transcripts in the context of engineering education research. The findings reveal that ChatGPT can produce ARMs that are coherent and adhere well to development guidelines; however, its performance varies significantly across different criteria. Specifically, ChatGPT excelled in non-interpretative aspects, such as summarizing content and elucidating topics of interest with notable clarity and organization, but struggled with interpretative tasks, sometimes fabricating connections not present in the data. Furthermore, the inconsistency in ChatGPT's interpretative accuracy highlights a potential reliability issue in qualitative analysis contexts. Also, in agreement with studies conducted by other researchers [4], it was found that prompt instructions need to be carefully crafted and tested before extensive deployment. Well-crafted prompts could significantly enhance the quality of responses generated by ChatGPT.

One limitation of this study was ChatGPT's inability to access comparative data, which restricted its performance in tasks requiring such analysis. Additionally, this study only analyzed two transcripts, with one researcher crafting the human ARM and another researcher conducting anonymous evaluations for each transcript. Augmenting the volume of data and expanding the research team might diminish bias within the data analysis and evaluation phases.

Despite the limitations and concerns identified in this study, the researchers all expressed positive views on the benefits of using ChatGPT as a co-researcher in qualitative research. Looking forward, for the incorporation of ChatGPT in qualitative analysis, it is recommended to:

1. Combine ChatGPT's capabilities with human oversight, especially for interpretative tasks, to mitigate inaccuracies. This could include pairing a human researcher with ChatGPT, in a similar way to how two humans might pair up during qualitative analysis to reduce bias in their analysis.
2. Use ChatGPT as a tool for initial data processing or to generate preliminary insights, followed by thorough human validation.
3. Equip qualitative researchers with knowledge and skills in prompt engineering to better design prompt instructions.
4. Further explore the rapidly advancing capabilities of AI tools to better leverage AI's strengths while addressing its limitations.

Bibliography

[1] S. S. Biswas, “ChatGPT for Research and Publication: A Step-by-Step Guide,” *J. Pediatr. Pharmacol. Ther.*, vol. 28, no. 6, pp. 576–584, 2023.

[2] P. A. Christou, “How to Use Artificial Intelligence (AI) as a Resource, ow Methodological and Analysis Tool in Qualitative Research?,” *Qual. Rep.*, vol. 28, no. 7, pp. 1968–1980, 2023.

[3] J. L. Feuston and J. R. Brubaker, “Putting Tools in Their Place: The Role of Time and Perspective in Human-AI Collaboration for Qualitative Analysis,” *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, pp. 1–25, Oct. 2021, doi: 10.1145/3479856.

[4] H. Zhang, C. Wu, J. Xie, Y. Lyu, J. Cai, and J. M. Carroll, “Redefining Qualitative Analysis in the AI Era: Utilizing ChatGPT for Efficient Thematic Analysis.” arXiv, Sep. 19, 2023. Accessed: Jan. 29, 2024. [Online]. Available: http://arxiv.org/abs/2309.10771

[5] H. Zhang, C. Wu, J. Xie, C. Kim, and J. M. Carroll, “QualiGPT: GPT as an easy-to-use tool for qualitative coding.” arXiv, Oct. 10, 2023. Accessed: Jan. 29, 2024. [Online]. Available: http://arxiv.org/abs/2310.07061

[6] Y. K. Dwivedi *et al.*, “‘So what if ChatGPT wrote it?’ Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy,” *Int. J. Inf. Manag.*, vol. 71, p. 102642, 2023.

[7] J. Saldaña, “The coding manual for qualitative researchers,” *Coding Man. Qual. Res.*, pp. 1–440, 2021.

[8] R. M. Emerson, R. I. Fretz, and L. L. Shaw, *Writing ethnographic fieldnotes*. University of Chicago press, 2011. Accessed: Jan. 31, 2024. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=k83BlbBHubAC&oi=fnd&pg=PR7&dq=Writing+Ethnographic+Fieldnotes&ots=ePUznlmFLm&sig=0Tkwz-TVhmnegF6Cce7J1XqlTn0

[9] Y. Wang, S. Claussen, X. Zhang, and F. Khalkhal, “Board 257: Development and Initial Outcomes of an NSF RIEF Project in Understanding Teamwork Experience and its Linkage to Engineering Identity of Diverse Students,” in *2023 ASEE Annual Conference & Exposition*, 2023. Accessed: Feb. 02, 2024. [Online]. Available: https://peer.asee.org/board-257-development-and-initial-outcomes-of-an-nsf-rief-project-in-understanding-teamwork-experience-and-its-linkage-to-engineering-identity-of-diverse-students

[10] S. Claussen, F. Khalkhal, X. Zhang, A. K. Biviano, and Y. Wang, “Qualitative analysis of the relationships between the teamwork experiences of diverse students and their engineering identities at a Hispanic-serving institution,” in *2023 ASEE Annual Conference & Exposition*, 2023. Accessed: Feb. 02, 2024. [Online]. Available: https://peer.asee.org/qualitative-analysis-of-the-relationships-between-the-teamwork-experiences-of-diverse-students-and-their-engineering-identities-at-a-hispanic-serving-institution

Appendix: Human-Crafted and ChatGPT-Produced ARMs for Transcript #1

Human-Crafted ARM for Transcript #1

Omar is a transfer student at SF State, having attended community college after high school. He did not know what he wanted to major in when he graduated from high school, but ended up settling on engineering because of the field's job prospects. He is studying electrical engineering now and is interested in electronics.

Omar states that he currently would not consider himself as an engineer, and that he needs more experience before he would be able to identify as one. He feels like he needs some years of experience in industry before he could call himself an engineer, and made an interesting analogy to medicine and law, fields where he thinks a person similarly needs experience to officially be considered a doctor or lawyer. He followed by saying, "I believe that's the sort of profession that engineering is, where at least you should be working with a full-time engineer until you're comfortable enough to say be employed by yourself and say, I can lead a project or I can deliver a product" (00:15:12).

Omar said that the recognition by other people in his life of him as an engineer is varied, and in part depends on their connection to engineering (e.g. his cousins are engineers, so they don't see him as one yet, since he doesn't yet have the experience). He had an interesting view on his professors and how they see their students, saying, "I believe my professors do believe us to be budding engineers or definitely in that engineering mindset. So they already want us to be thinking of ourselves as engineers. Even if they don't believe we are engineers, they do want us to think of ourselves as engineers is what I think" (00:19:44).

In the interview, Omar indicates that teamwork is very important, since engineers have to work in teams effectively in industry with both other engineers and non-engineers. As a result, he sees communication skills as very important, something he emphasized multiple times throughout the interview. He said that earlier in his studies at SFSU, his experience with teamwork was stunted due to Covid and online learning. He went on to say, "But I've definitely had better experiences with teamwork when I'm in person. We're able to coordinate with each other, and it's very important to meet in person and to be able to talk and just relax with each other. That builds, I believe, a lot of social skills, as well as a feeling of not being alone in the classes. We were able to reinforce that. Oh, you had a problem with that question. I did, as well and such" (00:22:41).

His experience working with his [*Engineering Course ST*] seemed overwhelmingly positive. It sounds like he and his two teammates were very similar in terms of abilities and working style, and that they were able to complement each other's strengths. As a result, he said he would be disappointed if future teams did not have as good of chemistry. He summarized his experience with this team as, "just a wonderful and good thing to have teammates that you can rely on to do their part of the work, as well as have people who are on the same pace as you" (00:34:09).

In terms of team experiences that made him feel more or less like an engineer, he referenced previous team-based where his opinion was not listened to (his Calculus team) or where teammates would not respond to his messages and, as a result, he would need to do the work alone. He cited these as disagreements, and talked about how in both cases, they boiled down to him not feeling respected by his teammates. However, while he seemed to think these were negative experiences, he ultimately seemed to see the need to pick up the slack for someone else as contributing to making him feel like an engineer, since it showed him he could complete the work on his own.

Overall, Omar came across in the interview as a confident engineer who has had productive experiences in teams. While he personally stated that he does not see himself as an engineer yet – largely due to his lack of experience in industry – he seemed to believe that most of the other people in his life see him as an engineer. He also didn't seem particularly affected by his stated lack of an engineering identity. He seems confident in his choice of major and genuinely interested in much of what he is learning in his classes (despite indicating that he – at least partially – seemed to select engineering because of the job opportunities). His team experiences have not adversely affected his identity, and he seems teamwork as critical to the work of an engineer.

ChatGPT-Produced ARM for Transcript #1

The interviewee, Omar, is a senior electrical engineering student. He prioritizes understanding electronics and programming, valuing education quality and networking with peers. His entry into engineering stemmed from practical considerations, influenced by the IT field's growth.

Omar perceives engineering primarily as a field focused on problem-solving and practical application. His engineering identity is still in development, as he feels the need for more hands-on experience to fully embrace the title of 'engineer'. This perspective highlights the importance of real-world experience in defining one's professional identity in engineering. Omar's views align with a common understanding in the field that engineering is not just about theoretical knowledge but also about applying that knowledge in practical, often team-based, contexts. Initially, he had a vague understanding of engineering, which evolved through coursework, especially in [*Engineering Course XY*], to a clearer view of engineering as design and problem-solving.

Omar's teamwork experience in [*Engineering Course ST*] was positive, characterized by camaraderie and effective collaboration, despite initial challenges due to COVID-19. His role focused on computer systems and MATLAB scripting. He values the importance of teamwork for engineers, as seen through his project on an inverted pendulum control system.

Team disagreements were not a significant feature in Omar's [*Engineering Course ST*] experience. Instead, he highlighted the importance of team chemistry and shared effort. This contrasted with previous experiences where he felt either overlooked or burdened with extra workload due to non-responsive teammates.

Omar's teamwork experiences shaped his perception of engineering and his identity within the field. He recognized the importance of effective communication and collaboration, acknowledging these skills as essential for engineers. His positive experiences in well-functioning teams reinforced his engineering identity, while negative experiences in other settings made him more aware of the challenges engineers can face. Omar's ability to navigate these diverse team dynamics reflects the adaptability and resilience integral to engineering roles.

His experiences in teams have shaped his understanding of engineering, emphasizing practical skills and team dynamics. However, he notes the necessity of experience for a complete engineering identity.

Omar's teamwork experience in [*Engineering Course ST*] made him feel more like an engineer. This is primarily because the experiences allowed him to apply theoretical knowledge in a practical setting, a key aspect of engineering. Working in teams also helped him develop essential skills such as collaboration, communication, and problem-solving, which are integral to the engineering profession. These experiences contributed significantly to his understanding and development of an engineering identity, reinforcing his connection to the field.

In other engineering courses, Omar's teamwork experience was initially limited due to COVID-19, with activities like lab work shifted to individual tasks. However, once in-person interaction resumed, his teamwork experiences improved significantly. For instance, in [*Engineering Course MN*], the in-person collaboration enhanced both social skills and a sense of camaraderie among peers. These experiences were important for him, emphasizing the value of face-to-face interaction for effective teamwork in engineering.

Compared to other interviews, Omar's focus on practical skills and team dynamics stands out. His journey from confusion to clarity in understanding engineering is notable.

What intrigues me is Omar's evolving perception of engineering and identity. His emphasis on experience over theoretical knowledge and the contrast between his self-perception and others' (like peers and professors) regarding his engineer status is insightful. The way his teamwork experiences have influenced his perspective on engineering roles and responsibilities is particularly interesting.

Omar represents a typical engineering student in transition, his identity shaped by academic and team experiences. His journey reflects the complexities of developing a professional identity in a practical and team-oriented field like engineering.