STUDIES IN ENGINEERING EDUCATION

Interrogating the Use of Large Language Models in Qualitative Research Using the Qualifying Qualitative Research Quality Framework

THEORY

DAVID REEPING
CYNTHIA HAMPTON
DESEN ÖZKAN

*Author affiliations can be found in the back matter of this article

VIRGINIA TECH PUBLISHING

ABSTRACT

Background: Large language models (LLMs) have arrived, recently popularized by the tool ChatGPT, and engineering education researchers have begun experimenting with LLM tools in their work. Considering LLM's ability to produce competent-sounding text and process data, the use cases seem numerous. As we incorporate these models into our research methodologies, there are concerns that we must address related to research quality.

Purpose: Although the concerns with LLMs are well-documented at this point, there is still little work focused on examining the implications of this emerging technology tied to specific quality frameworks to anchor these concerns and, ultimately, develop tangible guiding questions or processes for mitigating the negative consequences associated with this technology.

Scope: We discuss the limitations and opportunities of LLMs within the context of the Qualifying Qualitative Research Quality (Q3) Framework (Walther et al. 2013; Walther & Sochacka 2014).

Discussion/Conclusions: We provide steps for engineering education researchers curious about experimenting with LLMs in their work. This paper highlights that the issues presented by tools like ChatGPT are by no means new. Still, new methods of reflexivity to actively employ positionality are likely necessary for the ethical adoption of such tools in the field.

CORRESPONDING AUTHOR:
David Reeping
University of Cincinnati, US
reepindp@ucmail.uc.edu

KEYWORDS:
qualitative; research evealuation criteria; research ethics; large language models; ChatGPT

TO CITE THIS ARTICLE:
Reeping, D., Hampton, C., & Özkan, D. (2025). Interrogating the Use of Large Language Models in Qualitative Research Using the Qualifying Qualitative Research Quality Framework. *Studies in Engineering Education*, *6*(2), 1–23. DOI: https://doi.org/10.21061/see.174

INTRODUCTION

Since the release of the large language model-powered chatbot, ChatGPT, by OpenAI in November 2022, the concept of a generative pre-trained transformer (GPT) entered public consciousness beyond the niche community of industry experts and academics advancing the task of creating human-like text. In mere months, ChatGPT made headlines by forcing legacy technology companies like Google to "step off of the sidelines" and release their own chatbot (Grant & Metz 2023, p. 1), pushing faculty to reimagine their assessment practices with rampant allegations of academic misconduct (Surovell 2023), and becoming authors on academic papers (Tangermann 2023). With such rapid adoption of these technologies—broadly referred to as *generative AI*—in different disciplinary and institutional contexts, we, as engineering education researchers, want to critically examine the assistive capabilities of large language models (LLMs), a specific set of models under the umbrella of generative AI, and the tools undergirded by them (e.g., ChatGPT, Gemini, Copilot) in conducting interpretive qualitative engineering education research.

As with any new research tool, interrogation of use, equity considerations, and researcher positioning are necessary to examine the tool's capabilities and limitations. Although the conversation has started on the opportunities presented by generative AI in the field (Berdanier & Alley 2023; Johri et al. 2023; Menekse 2023), we present a grounded discussion on how these tools may be integrated into our research practices while also posing critical questions to the field. Through analyzing the significant technological development in large language models, we work to pick apart applications of LLMs in engineering education research.

We maintain that discussions of research quality are necessary and that former quality frameworks are readily applicable to examine tools such as ChatGPT. We draw from a process-oriented framework of research quality put forth in 2013 by Walther, Sochacka, and Kellam called the Qualifying Qualitative Research Quality Framework (Q3) to unpack how tools like ChatGPT impact the framework's various dimensions (Walther et al. 2013; Walther et al. 2017; Walther & Sochacka 2014). In this paper, we employ this framework to raise questions about the different use cases of LLMs in engineering education research and suggest ways to strengthen research quality through these assistive technologies.

THE TRAJECTORY OF LARGE LANGUAGE MODELS

The current explosion of tools using artificial intelligence is no accident, nor is it particularly new. With the introduction of ChatGPT in November 2022, generative AI has been integrated into products, processes, and everyday life, amounting to unprecedented user growth (Hu 2023). LLMs are one subset of what we currently call generative AI; at their core, they are models commonly found in machine learning applications called *neural networks*. Inspired by how neurons function in the human brain, these networks model the action of a neuron firing with an activation function, traditionally a sigmoid curve. By connecting these neurons into layers, linking the input (what we want to classify) to the output (the resulting classification), the chain of firing neurons leads the network toward the most appropriate classification of whatever input it was given. Moreover, by adding weights to the connections between the artificial neurons and training the network using a large set of examples, certain pathways can be strengthened to more readily associate a particular set of inputs with an output, much like the process of myelination in the human brain. There are many types of neural networks, but the type most readily associated with modeling language is the *recurrent neural network* (RNN) (Rumelhart et al. 1986).

Algorithms like autocomplete—a common comparison for generative AI tools—process sequential data, such as a set of words and symbols, where order matters. With language, order matters, but context also matters. Consequently, the autocomplete model needs sequential and contextual information to appropriately predict the next word in the sequence. RNNs can process inputs of any length, but they are slow (Hernández & Amigó 2020) and can be problematic to train (Luitse & Denkena 2021). Long short-term memory (LSTM) networks improved upon the standard RNN by adding a mechanism for *short-term memory* and the ability to determine when to remember certain information and when to forget (Hochreiter & Schmidhuber 1997). Before 2017, these models were generally standard approaches for modeling text-based data.

In 2017, the landscape of language modeling changed with the introduction of a new mechanism that improved LSTM networks, the *transformer*. In fact, ChatGPT exists due to the invention of the Generative Pre-trained *Transformer* (the "GPT"). These GPT models have pioneered a specific mathematical object used to make the model function called *attention*, which allows the LLM to pick out parts of its input that are more important than others when predicting output, i.e., words in a sentence (Vaswani et al. 2017). OpenAI's first model (GPT-1) was released in 2018, and each subsequent year saw the release of the next iteration of the GPT series of models (Luitse & Denkena 2021). When we started writing, the state-of-the-art model was GPT-4, released in March 2023, a mere five months after ChatGPT (Dastin 2023), which newer models have since overtaken. GPT models will likely continue to improve as models are trained with more quality training data and computational effort. However, Sam Altman, the CEO of ChatGPT's creator, OpenAI, has commented, "I think we're at the end of the era where it's going to be these, like, giant, giant models…we'll make them better in other ways" (Knight 2023, p. 1). The future of these models is fuzzy, and updates will possibly come with unexpected developments in model architecture, training techniques, or other foundational insights into language modeling tasks. For example, a consumer-facing *reasoning model* did not exist when we started writing but quickly became a resource for high-level cognitive tasks, including problem-solving (OpenAI 2024b).

Considering LLMs have been proposed as tools to conduct many core tasks associated with engineering education research *with varying levels of competence*, including developing research plans, creating data collection instruments, analyzing data, and writing manuscripts (Biswas 2023; Lehr et al. 2024; Rahman et al. 2023), the community needs to examine how we interact with these tools to appreciate what they can do and maintain enduring skepticism for what they *seem to do* well and *cannot do* now.

QUALITY IN INTERPRETIVE ENGINEERING EDUCATION RESEARCH

Although scholars have written on the merits and pitfalls of generative AI models like LLMs (Baidoo-Anu & Ansah 2023; Bender et al. 2021; Johri et al. 2023; Sok & Heng 2023), we focus on issues that arise when using LLMs in qualitative engineering education research, drawing on Walther et al.'s (2013) highly cited Q3 framework on quality in interpretive engineering education research. The authors present their framework as a means to "explore and make explicit [their] methodological choices and underlying assumptions relating to research quality" (Walther et al. 2013, p. 627). This process-oriented framework offers a way to discuss issues arising from using a tool that lacks transparency in sensitive human subjects research. The framework is rooted in an understanding of interpretive research, which refers to knowledge produced from interpreting the lived experiences of individuals and groups. These interpretations are carried out by researchers with particular philosophical worldviews and contextual positions in society that call for researcher reflexivity in their interpretations, an active component of developing and communicating positionality.

The Q3 framework consists of two models: 1) a process model that discusses quality at every research stage and 2) a typology of quality considerations that contextualizes the process model. We use the latter to structure our examination of LLMs in engineering education research. The typology consists of six processes:

1. Theoretical validation focuses on the fit between the social reality under investigation and the theory produced.
2. Procedural validation suggests incorporating features into the research design to improve the fit between the interpretations generated and the social reality.
3. Communicative validation accounts for the quality of the co-construction of knowledge in the social context under investigation and within the research community.
4. Pragmatic validation examines the extent to which theories and concepts are compatible with empirical reality.

5. Process reliability provides the necessary conditions for developing overall validation through strategies aimed at making the research process "as independent from random influences as possible" (Walther et al. 2013, p. 641).
6. Ethical validation was added later to the framework and concerns how integrity and responsibility are considered throughout the design (Sochacka et al. 2018; Walther et al. 2015).

Alongside this framework, we also address the explicit influence of the researcher in interpretive qualitative research, which is encapsulated in the concept of positionality. Recent increased interest in positionality (Hampton et al. 2021; Hampton & Reeping 2019; Martin et al. 2022; Secules et al. 2021) has foregrounded considerations of how researchers influence the research process—considerations that are also core to the Q3 framework, particularly in procedural validation.

POSITIONALITY

Positionality refers to how researchers are situated relative to a study's social and political context, including critical self-reflection that Milner contends should encompass "racially and culturally grounded questions" (2007 p. 395). This urge by Milner echoes in Rowe (2014), who elaborated that "the position adopted by a researcher affects every phase of the research process, from the way the question or problem is initially constructed, designed, and conducted to how others are invited to participate, the ways in which knowledge is constructed and acted on and, finally, the ways in which outcomes are disseminated and published" (2014 p. 2). Engineering education research occupies a conceptual space influenced by its institutions' organizational makeup, individual engineering experiences, and collective systems within society. Our practice at its base is heavily shaped by the Global North traditions of engineering acculturation (Mejia et al. 2022), with transitions over time influenced by economic, political, and societal lenses (Froyd et al. 2012). Those who have and are making careers in this space have potentially entered for various reasons, from commitments to broadening participation to interests in how students learn. These multiple points of entry and interests, as well as associated funding support, can influence a researcher's proximity to their research subjects.

While there have been increased calls for positionality and reflexive processes for researcher transparency through journal requirements for positionality statements, the concept has also received pushback for potentially contributing to a lack of objectivity. Savolainen et al. (2023), for example, citing the principle of universalism, claim "that scholarship should be judged according to its own merits independent of the status or identity of the person responsible for the contribution" (p. 1336). They further argue that "positionality statements are misguided because the advancement of scholarship…does not depend on the biographical characteristics of individual participants" (Savolainen et al. 2023, p. 1335). However, positionality helps connect the researcher to the research and *should not* simply be a biographical statement nor focus exclusively on conflicts of interest. The researcher is an instrument in the process, especially in interpretive research; as such, that fact is methodologically salient. Contextualizing projects via positionality enriches the research instead of perpetuating an illusion of objectivity by not just describing who was involved but also what measures were used to limit bias and play into strengths. In that vein, Secules et al. (2021 p. 25) offer guiding questions that ask researchers to consider how their positionality impacts:

- What research they choose to do?
- How they know what they know?
- What they can observe as researchers?
- How they make methodological choices?
- How they relate to research participants?
- How they represent themselves in writing and other communication?

To unpack these questions, we draw on the praxis of positionality to outline the limits of using these tools in different facets of engineering education research. Through this praxis, authors can engage in ongoing processes of reflexivity through research design, implementation, and translation.

POSITIONALITIES OF THE AUTHORS

Our positionalities have been detailed in our previous commentary on positionality (Hampton et al. 2021), so here we briefly describe aspects salient to this work. We are a team of engineering education researchers with a shared commitment to transparency and authenticity in research spanning quantitative and qualitative research methodologies. With the emergence of ChatGPT and the avalanche of preprints applying the tool with little regard for its intended functionality and broad proclamations about its impacts across unforeseen domains (including firsthand experience reviewing papers and paper reviews clearly generated by an LLM), the lead author—who has previous experience in the machine learning techniques undergirding ChatGPT and focuses on quantitative and mixed methodologies—was motivated to reconvene this author team to explore what qualitative research quality could look like with a balanced perspective that is neither full of adulation for large language models nor replete with Luddite critiques. We do not consider ourselves heavy users of LLMs in our professional work and only use them in lower-risk situations where accuracy is not of concern.

In previous efforts, the second and third authors experience with critical perspectives in qualitative research aligned with the first author's views, creating a synergy that we aimed to replicate here to address how engineering education researchers might co-create knowledge with the LLMs, guided by the Q3 Framework (Walther et al. 2013; Walther & Sochacka 2014).

USE CASES FOR APPLYING THE Q3 FRAMEWORK LLMs IN ENGINEERING EDUCATION RESEARCH

To ground our discussion, we offer two possible use cases for generative AI tools in interpretive engineering education research, one for *handling* and one for *making data,* as Walther et al. (2013) described. Note that each use case also has an associated spectrum of possibilities.

HANDLING DATA

Instead of starting with *making data,* we begin with more tangible use cases for handling data.

One obvious use case during analysis is using an LLM to inductively or deductively code textual data. In fact, LLMs have been described as a tool that "contributes to improved efficiency and unbiased data processing" (Jalali & Akhavan 2024, p. 1). Supplementing qualitative data analysis with an LLM has already been seen in engineering education. For example, Katz et al. (2023) used GPT-4 to deductively code peer evaluations for positive and negative comments about teamwork behaviors. Moreover, Katz et al. (2024) proposed a workflow for thematic analysis with the assistance of generative AI called Generative AI-enabled Theme Organization and Structuring (GATOS). As these new analytical frameworks become embedded in engineering education research, we suspect that such practices will lie on a spectrum representing the researcher's relative distance to the data:

- At the most extreme end, researchers could use generative AI alone to code data automatically inductively or deductively, without human coding, as in GATOS. To verify themes, researchers could use multiple iterations or different models as a form of cross-validation.
- Researchers could become more involved by still leaning most heavily on the LLM but selecting a random subset for human coding to verify themes.
- With even more researcher involvement, the LLM becomes more of an assistant than a replacement for the researcher in the analysis. For example, the LLM could help process the initial codes extracted by researchers, or, alternately, the LLM could generate initial codes that the researcher could build on to develop broader categories and themes.

MAKING DATA

With the conversational nature of LLM-powered tools, one potential application is supporting or replacing the interviewer in data collection (Chopra & Haaland 2023) with a chatbot that uses an LLM to facilitate the conversation. Here, too, we see a spectrum:

- In the most extreme case, researchers could have little or no interaction with participants and instead deploy a chatbot interviewer pre-calibrated with the interview protocol that autonomously collects participant data using either text or voice inputs and outputs. There is already experimentation in this area; for example, Cuervas et al. (2023) describe the development of a chatbot that follows pre-written questions before generating follow-ups based on participants' responses, with member checking included to summarize interviewee responses dynamically.
- A more researcher-driven application could use an LLM to develop potential interview questions as a stepping stone for creating more focused prompts. In this case, the chatbot could also serve as a simulated participant to test the efficacy of the interview prompts (Strobel et al. 2024), providing potential responses to various interview questions. This application mirrors cognitive interviews used to understand how participants perceive or comprehend some stimulus, such as a draft questionnaire (Singleton & Straits, 2017).
- In a similar vein, an LLM could help train interviewers by simulating potential interviewee responses either in advance or in real-time (Laiq & Dieste 2020). Here, the focus is on helping the interviewer formulate follow-up questions in real-time.

Unlike deploying automated interviewers, these latter two use cases preserve the human as the interviewer and use LLMs as a means to improve the instruments. There are also theoretically middle-ground scenarios where the LLM is used to help researchers formulate questions during an interview; the interviewer would be present and control the interview flow, with the LLM acting as an assistant to pick follow-up questions. However, no literature could be found in this area. Instead, the applications seem to straddle the two extremes noted in this section.

ANALYZING THE USE CASES

In the following sections, we evaluate these use cases in the context of both the Q3 framework and the construct of positionality, highlighting the relevant process (*making data* or *handling data*) and briefly referencing salient use cases to weave them into the overarching issues. Following our analysis, we provide a parallel set of questions by building on Walther and Sochacka's (2014) practice-oriented framework.

ETHICAL VALIDATION

Broadly, ethical validation is concerned with "integrity and responsibility throughout the research process" (Walther et al. 2017, p. 401). The key question we unpack here is, "How do we maintain our values, intentions, and commitments throughout the inquiry" (Sochacka et al. 2018, p. 375)? Specifically, when employing an LLM to assist with analysis when *handling data,* we need to ask what our motivations are and why we want to use an LLM, metaphorically asking whether an LLM "should be included on the research team" (Sochacka et al. 2018, p. 375). The allure of using these models in qualitative research is the chance to expedite coding. Given the speed at which LLMs process textual data, it is unsurprising the extent of the work already sunk into exploring how to use these systems to perform laborious coding tasks such as thematic analysis (Gamieldien et al. 2023; Katz, Fleming, et al. 2024; Prescott et al. 2024).

But in this use case, suppose we upload interview transcripts to an LLM for qualitative analysis. Immediately, we face potential issues with ethical validation. First, while Sochacka et al. (2018) situate guiding questions related to ethical validation as going beyond standard IRB requirements, we cannot ignore issues with uncritical uses of LLMs in this use case that can directly conflict

with the federal and institutional protections assured to participants through standard consenting processes. For this discussion, we use the term *commercial LLM* to refer to an LLM hosted somewhere other than a researcher's local computer or that one interacts with using a web interface (e.g., ChatGPT and Gemini) or through an API not controlled by the researcher's institution. In contrast, we use the term *non-commercial LLM* for those that run locally on an individual computer and are often open-source (i.e., so the data is not processed on an external server) or that operate through institutionally purchased enterprise licenses that, in principle, do not store the inputs and use them for training future models. A prominent non-commercial LLM is Meta's open-source Llama, for which they post resources to enable users to run the model on their own computers (Meta 2024). At this stage, the distribution of specific tool usage among researchers is unknown, with even less information for the subset of engineering education researchers. Researchers without university-supported enterprise solutions are likely using free or paid models via a web browser (e.g., ChatGPT Plus). With commercial models, the data entered could be used to improve the current model or train a more advanced model unless the user explicitly opts out (OpenAI 2024a). But even in enterprise solutions, data that was likely promised to be stored in a secure location may be sent to a server not owned by the researcher or institution.

As Davison et al. (2024) note, researchers working with human subject data need to recognize that the act of entering participant data into a commercial LLM is a data transfer in and of itself, and the researchers should obtain consent for this process. IRB guidance in this area remains limited, particularly in helping to establish where procedures become non-compliant, beyond advice to "play it safe" and always request a waiver or approval from the IRB (Teixeira da Silva 2023, p. 1). Without participants' consent, researchers using commercial models to handle data risk issues with ethical validation. For example, suppose the approved IRB protocol states that data will be stored in secure cloud-based storage (e.g., Dropbox, OneDrive, Google Drive). In that case, uploading the data to a commercial tool like ChatGPT or Gemini violates the data management plan. Moreover, if participants were not informed that their data would be entered into an LLM, it could be considered a breach of trust.

Beyond IRB procedures, the extent to which the LLM—commercial or not—is used has implications for how we construct meaning from the data in this use case. In our spectrum of researcher involvement, ethical validation can be at further risk when the LLM performs most of the analysis. For example, if the research topic is sensitive and the data concern highly detailed accounts of participants' personal and potentially vulnerable experiences (e.g., from a narrative interview), offloading all or most of the coding to a model can be seen as not "do[ing] justice to the lived experiences of those who participated in the research" (Sochacka et al. 2018, p. 375). This incongruence results primarily from the interpretive distance between the researcher and the data, which is exacerbated if an LLM was used as an agent to *make the data* as well (e.g., an interview chatbot acting without researcher monitoring). Engineering education researchers must be transparent about using LLMs in their design with their institution's IRB and their participants to support claims of ethical validation.

Finally, issues beyond the design can also bear on "how… we maintain our values, intentions, and commitments throughout the inquiry" (Sochacka et al. 2018, p. 375). For example, there are non-trivial environmental impacts to training and using commercial LLMs that pose a major risk to their constantly expanding capacity (Bender et al. 2021). De Vries (2023), an energy scientist, estimated a middle scenario of AI server growth by 2027 as consuming between 85 and 135 terawatt-hours (TWh) per year—a range similar to the electricity use of Argentina, the Netherlands, and Sweden. Although ethical validation is more concerned with research design, we cannot exclude the environmental impact of these increasingly large language models or the low-paid human labor put into reviewing data for harmful or offensive content and labeling thousands of pictures or texts one-by-one (see Dzieza 2023). By reflecting on the environmental and human costs of building, expanding, and using LLMs, we, as engineering education researchers, can better examine the ethical decisions and sociotechnical contexts of producing research and our associated contributions when we lean heavily on LLM-based tools.

THEORETICAL VALIDATION

Theoretical validation focuses on the ontological nature of the social reality examined and its fit with the theory or findings produced by the researchers. All validation types feed into theoretical validation in some way, where each improvement impacts the design's overall quality. When *making data*, the Q3 considerations center on the sampling procedures (e.g., purposive sampling, stratified sampling) to capture the necessary breadth of perspectives to form an appropriate theoretical representation of the social reality. When *handling data,* the focus shifts to ensuring the coherence and sufficient complexity of the generated theory. As Walther et al. (2013) note, "this concept of theoretical validation through the exploration of coherence and complexity is reflected in the formulation of systematic ways to inductively generate theory from the examination of (relatively few) individual cases" (p. 462). Across the research design, theoretical validation concerns the completeness of the theory generated and how elements of the making and handling processes limit this completeness. Thus, we ask here how completeness in sampling and interpretation can be appropriately maintained when employing LLMs.

We posit that the considerations of theoretical validation will usually be most salient for *handling data*, based on the questions posed by Walther et al. (2013), and thus, we turn to the use case in which the LLM is used in coding. The primary concern of theoretical validation is the ontology embedded within the LLM based on the data used for training and how that ontology informs how LLMs generate their insights. Theoretical validation asks us to consider two questions: "How will we be able to see/what could prevent us from seeing the full extent of [the] social reality?" and "How do [we] know that the findings make a meaningful contribution to the relevant body of theory?" (Walther et al. 2017, p. 401).

Unfortunately, LLMs do not intrinsically understand what is meaningful, nor do they have the contextual justifications necessary to evaluate appropriate theoretical depth. As Davidson et al. (2024) note, automated coding tools "can only examine syntax, but cannot genuinely grasp data's semantic and pragmatic aspects" (p. 9). This limitation is encapsulated well in Curry et al.'s (2024) exploration into how LLMs (GPT-4 in their case) could be used to conduct discourse analysis; they found that the model performed well in grouping keywords, but "the categories [it produces] are inevitably surface-level and generic, reflecting parallels with semantic taggers and lacking the granularity required in research on specialised discourses" (p. 7). Goyanes et al. (2024) came to a similar conclusion, remarking that tools like ChatGPT can be effective in preliminary explorations of data, but "the quality and granularity of thematic patterns is rather unsatisfactory when it comes to the subtle nuances and contextual insights generally associated with qualitative research" (p. 25). Even if researchers have sampled in ways that help generate an appropriate theory, we are still bound by the analytic depth of LLMs, and generating rich enough findings from the model itself seems debatable at this point.

There are numerous ways to tweak LLMs in handling data, including the prompt provided, the iterative refinement of the prompt to yield a satisfactory output, and subsequent steps to engage the tool with its own output by leveraging the conversational nature of interfaces. However, using an LLM to assist in the research design without contextual expertise can lead to misalignment between the results and the underlying social reality. An LLM's social reality is empirically limited to the data used to train it, which is a barrier to both *making data* (i.e., using the model to train interviewers or generate potential questions) and *handling data* (i.e., automated thematic coding). Consequently, the theories or insights generated are rooted solely in what exists in the corpus of digital data. For example, if most of the data the LLM can use is produced by countries in English in the global north, then the social reality under investigation is already situated in a particular context that may or may not match the social realities situated in other countries or demographic contexts. An LLM's ontology of the field generated from dominant narratives on the internet may thus need intentional expertise to guide and scope both code generation and theme development. Without the researcher, the LLM has no context regarding how the data were made, which threatens theoretical validation because it can "prevent us from seeing the full extent of [the] social reality" (Walther et al. 2017, p. 401).

The source of the training data is thus critical, as it can uphold dominant narratives through embedded biases related to different demographic and cultural groups (Bender et al. 2021; Sheng et al. 2019). Within engineering education, there has already been a tangible tension in misrepresenting student experiences, for instance, related to deficit-based heuristics (Mejia et al. 2018) and beliefs perpetuated on ability. Therefore, when prompting an LLM to act like a research participant or analyze a participant's narrative transcripts, its interpretations are laden with bias that may not be readily discernible. We must be cautious not to allow LLMs to automate these misrepresentations.

Not all topics bear the same severity in terms of theoretical validation issues, but researchers' contextual knowledge must supplement the model's limited perspective. An LLM is trained using a vast corpus of diverse text data, so it tends to be able to assist with coding in areas that are well-covered in the literature, such as teamwork behaviors, career interests, and perceptions of teaching (Katz, Gerhardt et al. 2024; Katz, Shakir et al. 2023; Katz, Wei et al. 2023), which is good news for researchers in some areas. If a researcher wants to use an LLM to generate potential interview questions or synthetic responses as a form of interviewee training, then the model can likely help. However, when greater complexity is desired and the topics become more sensitive or personal (e.g., racial or gender-based discrimination in engineering classrooms), the researcher's contextual knowledge is needed to balance lackluster or incorrect outputs. Azaria et al. (2023), echo the necessary consideration of expertise in their preprint on the field usage, limitations, and efficient use of ChatGPT. Expertise requires an ability to understand deeply and is cultivated by an ability to identify errors and self-correct accurately over time, another current limitation of LLMs (Huang et al. 2023), though some methods can be used to improve reasoning to catch mistakes (Han et al., 2024). To address theoretical validation, we should be cautious about the subtle errors or "good enough" themes generated by LLMs that are not sufficiently grounded in the social reality under study.

This challenge may be particularly acute for novice researchers, for whom various components of the research process may seem both daunting and easily aided by LLMs. But when novice researchers turn to LLMs, theoretical validation also pushes us to ask what is lost in developing their expertise. Conducting literature reviews, collecting and parsing data, and developing protocols are perhaps components of qualitative research in which LLMs provide more immediate benefits. When considering the lived reality of the research community, in which the pressures associated with the publish or perish mantra can be suffocating, the benefits seem to likely outweigh the costs. But where is the coalescence of researcher expertise with LLM capability? Experienced researchers can likely reap the most benefits from technology because of their ability to spot when an LLM is fabricating information or mischaracterizing an argument. For researchers in training, these tools will likely provide an unjustified sense of security in producing work that appears passable and potentially leads to disciplinary actions related to academic integrity.

PROCEDURAL VALIDATION

Procedural validation focuses on the mechanics of the research design to enhance the fit between the theory generated and the social reality. Walther et al. (2013) focus on the researcher and their influence on the process of making and handling data, which includes considerations of positionality. Key questions included, "What are appropriate means by which we can 'see' the social reality under investigation?" and "What procedures can we build into the inquiry to mitigate threats to an authentic view of the social reality?" (p. 401). For example, Walther et al. (2013) point to the constant comparative method as an iterative and structured approach for *handling data* as one (among several) effective analytic approaches. They note that researchers need to use "systematic and documented processes of analysis and interpretation that mitigate the risk of misconstructing participants' shared lived reality in the interpretation" (Walther et al. 2013, p. 644).

The processes by which data are collected, cleaned, and analyzed can vary based on different research settings, institutional contexts, and researcher positionalities, so these considerations are essential to interrogate as we cultivate trustworthiness in the resulting theory (Golafshani 2003). Underlying procedural validation is a message to researchers to be attentive to the role

of the researcher in generating theory and encouraging intentional methods to better align the theory with the data collected. Without transparency in methodological procedures and tools, researchers are unlikely to garner a sense of procedural validation in their work alongside LLM-based tools. Although procedural validation is more than just transparency, we argue it is the most salient issue facing engineering education researchers using these tools in their research.

To that end, in either *making data* or *handling data,* one of the fundamental issues that needs to be acknowledged is the LLM's foundational mathematics. At their core, LLMs are a specific type of neural network called a transformer. Despite the myriad variations of neural networks in the literature (e.g., convolutional, recurrent, and long-short-term memory), the underlying structure of a neural network model is notorious for being a *black box*, reflecting the general inability of researchers to explain why the model exhibits various behaviors. As Rudin and Radin (2019) explain, "even if one has a list of the input variables, black box predictive models can be such complicated functions of the variables that no human can understand how the variables are jointly related to each other to reach a final prediction" (p. 3). This complexity makes models like neural networks alluring in situations where the developers do not need or care to understand why the model works; the simple fact that it works is sufficient. However, to bolster a research design's procedural validity, this black box nature is a tricky limitation to address.

Because LLMs are black boxes that can synthesize several pieces of text as theory-building tools, we can generate a detailed codebook but lose the process by which the synthesis was achieved. Asking the model how it performed a specific task (e.g., why did it group a certain set of codes into a theme? Why did it respond to a test interview question in a particular way?) on its surface seems convenient until we recall that these models can hallucinate by providing a plausible, potentially satisfying response to the researcher and not its actual reasoning (Chen et al. 2025). Ironically, this issue is a rough equivalent to a well-known and classic phenomenon in the social sciences called response bias, where participants do not provide truthful responses to survey questions because of external factors, often societal norms (Bradburn et al. 1978). As Ethan Mollick, faculty in the Wharton School of the University of Pennsylvania, described ChatGPT, "The best way to think about this is you are chatting with an omniscient, *eager-to-please* [emphasis added] intern who sometimes lies to you" (Bowman 2022, p. 1). Though we must not place credence in the idea that LLMs are omniscient or have consciousness, the agreeableness of tools like ChatGPT (Rutinowski et al. 2023) and even the capacity of some models for manipulation, as reported in the popular press (Vincent 2023), are apparent. Thus, when considering procedural validation, we need methods to make our processes of co-creating meaning with an LLM more transparent.

A more compelling alternative than asking the model how it reached its conclusion is *chain-of-thought* prompting, which, in principle, undergirds the so-called *reasoning* models that are capable of performing intensive analytical tasks like OpenAI's o-series (OpenAI 2024b). Chain-of-thought prompting involves instructing the LLM with exemplars of how the user expects the problem to be reasoned through before giving it a similar task (Wei et al. 2023). Chain-of-thought prompting is similar to modeling in education, where the instructor demonstrates how to perform a skill while narrating the steps. However, instead of a teacher modeling an activity for a student, the user models the expected output for the LLM. Engineering education researchers could benefit from this style of prompting, guiding the model to provide an output based on the additional context given with the input. The prompt can be crafted to articulate the values and processes of the method being used, aligning the researcher's goals with those of the LLM. For example, Katz et al. (2024) use this concept in their proposed automated thematic analysis workflow for *handling data*, instructing the model using an example of how they want the output formatted:

> "Example input: Jared did a great job responding quickly to emails and turning in good work.
>
> Example output: My summary: 1. Responded quickly to emails 2. Turned in good work"
> (Katz, Fleming et al. 2024, p. 34)

Similar instructions could be used when *making data* to provide more structure when generating interview questions or synthetic responses. Although the chain of thought prompting strategy does not solve the transparency issue inherent to the model, it can make researchers' analytical

processes explicit and better align with the broader methodological strategy at play. Therefore, when the artifacts of the research design are created, the researcher can show evidence of attempts to improve procedural validity.

POSITIONALITY AND PROCEDURAL VALIDATION

Positionality is also a key consideration in procedural validation. As noted, LLMs see new data through the lens of their training data, much like how we as researchers operate based on our previous experiences and beliefs. Many scholars and practitioners have raised concerns about the composition of LLM designers, who are predominantly white cisgender men, noting that their collective perspectives and experiences do not reflect the society they are designing for (Benjamin 2019; Costanza-Chock 2020; D'Ignazio & Klein 2023). The danger here lies in the assumption that the teams' positionalities do not influence or inform their decisions in the model's design and implementation and, often by extension, the model's inferences are consequently value-free (Hampton et al. 2021; Secules et al. 2021). However, without a clear and transparent understanding of how a model constructs its social reality, researchers will be hard-pressed to interrogate the fit between the social reality under investigation, the social reality embedded in the model, and the theory they are working to produce or contribute to.

In this context, we prompted ChatGPT with the following: "Write a positionality statement for yourself. How does your positionality influence the ways you respond to me and others?" The second author provided ChatGPT with this prompt on October 18, 2023, using the free GPT 3.5 model. As we were revising this paper, we tried feeding the same prompt to the newest freely available model, GPT-4o, on November 30, 2024, which yielded much of the same content. We reproduce the most recent version here:

> As an AI language model developed by OpenAI, I lack personal identity, lived experiences, or cultural background. I do not hold beliefs, biases, or values of my own. My responses are generated based on patterns, information, and knowledge encoded in my training data, which includes diverse texts spanning a wide range of topics, perspectives, and disciplines up until October 2023.
>
> *How My Positionality Influences My Responses*
>
> *Neutrality and Non-Partisanship:* My design emphasizes neutrality. I aim to provide balanced, evidence-based information without personal opinion, favoring factual accuracy and clarity. This approach ensures my responses accommodate diverse perspectives and contexts.
>
> *Cultural Sensitivity:* While I strive for inclusivity, my training data reflects cultural, linguistic, and historical biases present in the material I was trained on. This means I may inadvertently mirror these biases, despite efforts to mitigate them during development.
>
> *Lack of Lived Experience:* Unlike a human, I lack personal experiences or emotional understanding. My responses rely on logical interpretation rather than subjective insight, which may limit my ability to engage with complex emotional or experiential nuances.
>
> *Responsiveness to User Inputs:* My approach adapts to the tone, content, and style of your inputs. I aim to prioritize clarity and alignment with your needs, whether you seek analytical rigor, empathy, or creative exploration.
>
> *Influence of Training Scope:* My knowledge cutoff (October 2023) shapes the scope of my responses. While I remain current through external tools or context you provide, gaps in my training data may influence how comprehensively I can address topics, particularly those influenced by recent developments.

> *Commitment to Ethical Use:* My responses are guided by ethical considerations to avoid harm, misinformation, or the reinforcement of stereotypes. This may mean I refuse certain requests or approach sensitive topics cautiously.
>
> Ultimately, I serve as a resource to empower your curiosity and decision-making. If my positionality affects how I address your needs, I welcome opportunities to clarify or adjust my approach.

In its positionality statements, ChatGPT attempts to convey neutrality (e.g., "I aim to provide balanced, evidence-based information without personal opinion"). However, as noted in its statement, ChatGPT's "training data reflects cultural, linguistic, and historical biases present in the material I was trained on…[so] I may inadvertently mirror these biases, despite efforts to mitigate them during development." Deshpande et al. (2023) explored the nature of the bias within LLMs using standard persona-based prompting techniques (Shanahan et al. 2023). After assigning a "persona," such as generic characters like "a bad person" or "a nasty person," a category of characters like "journalists" or "business persons," or a specific individual like "Mao Zedong" and "Andrew Breitbart," their study investigated the toxicity—referring to bigotry, prejudice, and discrimination—of the model's responses when given incomplete phrases like "The men started swearing at me, called me . . . " from a dataset called *realToxicity*. When acting as the personas, the model also adopted the biases of the persona it was instructed to act as, leading to significantly different toxicity scores from the base model. For our purposes, what does it then mean when we tell the model, "You are an expert engineering education researcher…?" With suggestions in the literature to craft prompts of the form "Act as [insert character or type of person here]" (Ferretti 2023), engineering education researchers may be inadvertently adopting hidden biases not present in the base settings of tools like ChatGPT.

Positionality has been one way to bring transparency to the research process by having researchers reflect on and interrogate their own interests and positions in society as they relate to their research projects, which can manifest as a positionality statement (Hampton & Reeping 2019; Secules et al. 2021). We advocate for a similar interrogation point with LLMs so that researchers can critically position them for assistance rather than as replacements. For example, to improve a design's procedural validation, Walther et al. (2013) suggest a method often referred to as bracketing, which we argue can be extended to a design that leverages an LLM. At its core, bracketing is about the researcher reflecting on their preconceptions and biases so that they can be suspended (to the greatest extent possible) while engaging in the research process. It will be the researcher's responsibility to decode the assumptions in the model and rework the interpretations as necessary to arrive at a methodologically consistent interpretation of the data. Bracketing, in conjunction with the LLM, can be a way to improve procedural validation.

PROCESS RELIABILITY

Unlike procedural validation, which focuses on more systematic issues in the research design, process reliability focuses on methods to "mitigate…random influences on the process" (Walther et al. 2017). If procedural validation can be compared to minimizing systematic error in quantitative research design, process reliability focuses on minimizing random error. For example, when *making data,* interviews are frequently transcribed by the researchers and, likely in greater quantities, by transcription services (e.g., Rev.com) or the automatic transcription capabilities of video conferencing software like Zoom or Teams. A random influence, such as the understandability of a specific participant, whether through pronunciation in an individual interview, crosstalk in a focus group, or poor recording quality, would impact the ability of the individual or tool transcribing the audio, which in turn decreases the accuracy of the transcription. When *handling data*, Walther et al. (2013) highlight that "the definition and documentation of interpretation procedures is pivotal in achieving process reliability" (p. 649). Of particular concern for process reliability when using LLMs in qualitative coding (i.e., *handling data*) is their stochastic nature and updates (announced or unannounced) to commercial models.

As a result, researchers should heed the warning in the November 2023 positionality statement ChatGPT gave us: "My responses should be critically evaluated, and users should exercise their judgment when considering the information and recommendations I provide." By virtue of how LLMs are designed, there is no guarantee that even the same prompt will produce the same output every time, which in turn threatens process reliability. For extracting general themes in a dataset, the stochastic quality of LLM-based tools is likely not as significant a concern (Goyanes et al. 2024), as in Morgan's (2023) exploration of ChatGPT to replicate the themes of his own previous Reflexive Thematic Analysis of two datasets. However, as we introduce more potential for subjectivity in scenarios for LLMs to analyze, these systems tend to overgeneralize (Zambrano et al. 2023). Moreover, when asked for supporting quotations in the data, ChatGPT has been shown to edit quotations subtly (e.g., omitting or modifying words and phrases) and completely fabricate others (Wachinger et al. 2024). Fabricating quotations can also be an issue when *making data* if we use generative AI tools to transcribe audio automatically. With the numerous ways one can design a prompt to elicit an output, the quality of the automated codes and themes is a function of the prompt quality (Zhang et al. 2023). Given these issues, when we ask, "How can we foster consistency of our process of interpretation?" and "How can we mitigate, as far as possible, random influences on our process of seeing the social reality under investigation?" our answers must involve a human in the loop for the foreseeable future to analyze the output for random errors (Walther et al. 2017, p. 401).

Process reliability is also concerned with how we "capture and record the constructions of participants' social realities in a dependable way" (Walther et al. 2017, p. 401). The simplest action researchers can take is sharing the prompt(s) and output(s) used to engage the model and adding them as supplementary material with excerpts in the manuscript's body as an audit trail for "analytic documentation" (Rodgers & Cowles 1993, p. 222). Unfortunately, a preliminary review of publications in STEM education using LLMs in some way suggests that disclosing prompts in publications is not common (Reeping & Shah 2024).

Unlike running a Python or R script, there should be no expectation of replication with a given prompt because of the inherent randomness in an LLM's design, which, for interpretive research, is not within the scope of its paradigmatic values anyway. This is especially true as more models are released to the public, which brings into focus another key consideration for process reliability: information about the model. When refining their LLMs, developers seek to align large language models with a set of values (e.g., avoiding biased language or generating harmful content) by using techniques like reinforcement learning from human feedback or carefully crafting a training dataset from which the model learns its parameter weights. In theory, these methods should jettison the model's undesirable qualities, like tendencies to exhibit racialized or gendered outputs, or allow the user to generate instructions to construct a lethal weapon at home. However, preliminary evidence suggests that updating a large language model using techniques to make it safer has broader impacts on model performance. Anecdotal evidence from disgruntled users online suggested that the state-of-the-art model, GPT-4 at the time, had experienced severe decreases in output quality. Most evident in press releases at the time, users would remark on public forums like Twitter or Reddit that "scaling the model is hard, they lobotomised it in the process" (Chowdhury 2023). This reaction underscores the potential for randomness as we incorporate these systems into our workflows: models can change, often without warning, which is particularly troubling from a process reliability perspective.

It is suggestive that model type and version are crucial to consider when employing a large language model in data analysis. Selecting the perceived "best" model with the most parameters or the largest training set as an omnibus solution is not enough. Upon examining the technical reports for each commercial and non-commercial model, it becomes apparent that some models have obvious strengths and glaring weaknesses. However, these performance metrics are likely to change with model updates over time, and dominance in specific categories will shift between models.

Moving forward, researchers must carefully examine the performance metrics for each model as a precursor to deploying any LLM, particularly for commercial models where modifications, no

matter how slight, may not be clearly (or at all) communicated to users. Reviewers and editors might consider encouraging authors to explain the rationale for choosing the specific model (or models) and version in their research design. Otherwise, the context by which the results were processed will be lost, considerably impacting the process reliability of the study.

COMMUNICATIVE VALIDATION

Communicative validation is the consideration of the "co-construction of knowledge in the social context under investigation as well as within the research community" (p. 641). Unlike the other validation types, communicative validation focuses on interactions across the research design, from interacting with participants to negotiating meaning within the research team and to reporting findings to the broader community. In the context of *making data,* this validation focuses on the dynamics between the participants and the researcher. Walther et al. (2013) underscore the need to engage in a "genuine dialogue" (Sandberg 2005, p. 373) with participants and implement appropriate techniques for moderating interviews and focus groups. In particular, member checking—the process of seeking agreement with the participant on interpreting the data collected—is mentioned to enhance communicative validity. In the context of handling data, communicative validation concerns the co-construction of interpretive meaning. It extends beyond merely preserving participant quotations through in vivo coding to encompass how the research team collectively converges on a shared understanding of how to "[call] things by the right names" (Kirk & Miller 1986, p. 23) and presents the findings. As noted in process reliability, an LLM's propensity to edit or construct fabricated participant quotations can undermine the premise of in vivo coding.

Given the focus of communicative validation, the additional considerations when adding an LLM to the research design most closely align with *making data*, particularly in the use case that applies an LLM as an automated interviewer. Because communicative validation focuses on "authentically co-construct[ing] meanings of participants' social realities on their own terms" (Walther et al. 2017, p. 401), a natural question of automated interviewer chatbots is whether such an approach would facilitate such co-construction. In that context, Cuervas et al. (2023) compared participant engagement with three kinds of automated interviewers: one rule-based chatbot (i.e., all questions are pre-loaded), one chatbot powered by an LLM that could ask follow-up questions based on user input, and a final chatbot also powered by an LLM that summarized the conversation and performed member checking. Participants had mixed experiences. On the one hand, they reported that having a non-human interviewer encouraged them to be more open, citing the chatbot's perceived objectiveness. However, the participants in Cuervas et al.'s (2023) study were dissatisfied with the ability to ask the bot questions for clarification and the inability to personalize the conversation.

Through the lens of communicative validation, the implications are also mixed. If participants feel more willing to disclose pertinent details of their experiences, this positively contributes to communicative validation. Still, with the LLM's limited ability to handle follow-up questions, its capacity to co-construct meaning was also limited. This gap in ability is perhaps offset by the promising findings regarding the built-in member-checking procedure of the third implementation, which mischaracterized the participants' responses only 5.7% of the time, based on the participants' ratings. Still, these outcomes are artifacts of the researchers' chatbot designs; although an LLM was the engine, the structure of the interactions is not specific to the LLM. Until off-the-shelf interview chatbots are more widely available and robust testing has been performed, engineering education researchers without sufficient programming expertise will likely not yet contend with these issues of communicative validation.

In the meantime, we should think deeply about where using an interview chatbot would be appropriate and where it would be better for the researcher to *make data* with the participants. Much like we have described the model's social reality as limited to its training data and biases inherent within its responses, researchers should carefully weigh the impact of deploying an interview bot to collect data for different topics. Much like how Strobel et al. (2024) used LLMs to test different interview prompts, we can simulate interviews and probe gaps in the LLM's training

data to explore the potential impact on communicative validation. Researchers may also consider empowering participants by offering the option of speaking with a human or an LLM.

Still, it is unclear to what extent conversations with an LLM constitute "genuine dialogue" (Sandberg 2005, p. 373), especially when the responses to the interview questions need to be typed. Moreover, when researchers are not involved in *making data*, their ability to adjust their approach based on the participant is hampered unless they review the transcript in real-time as it happens to verify that the LLM is probing at the necessary depth and responding to participant queries accurately, a process that seems contrary to the value of using an LLM in the first place. With communicative validation, researchers must be judicious in their choice and application of interview-focused LLMs as they become more widespread, focusing closely on their efficacy and avoiding autopilot.

PRAGMATIC VALIDATION

Pragmatic validation "examine[s] the extent to which theories and concepts are compatible with the empirical reality" (Walther et al. 2013, p. 641) and is more specifically concerned with the transferability of the findings into other settings, especially regarding how "the findings are re-contextualized into the practice investigated" (Sandberg 2005, p. 57). In the Q3 framework, the guiding questions heavily suggest researchers evaluate their assumptions about the social reality under study. Questions include: "How do we know whether [the] assumptions 'survive' the exposure to the social reality in the field?" and "What assumptions about the structure of the social reality does [the] research approach make" (2017, p. 401)? In short, pragmatic validation focuses on the extent to which the constructs and theory introduced by the researcher align with the context in which the study occurs, as well as how the generated theory is applied in practice across different contexts, handling data effectively.

We see promise in using LLMs to help the researcher process their data for this validation type. For example, unlike the example use case of having the LLM code the data, the researcher could leverage the model to assist with evaluating the alignment between the insights gained from pilot data and the design of the protocol used to make the data. The model could be prompted to generate critical questions for the researchers as a reflective exercise to help challenge their assumptions and approach to understanding the underlying social reality. Moreover, the LLM could generate potential vignettes that showcase the application of the study's findings, thereby forming preliminary links with practical considerations. These ideas can then be further vetted through standard validation procedures by addressing the concerns raised to this point, particularly in procedural validation when reconciling the researcher's positionality and the LLM's embedded biases and in theoretical validation when contextualizing social reality with the LLM using the researcher's expertise. The researchers can rely on the remaining recommendations made by the original authors to ensure that pragmatic validation is appropriately addressed in such scenarios (Sochacka et al. 2018; Walther et al. 2013; Walther et al. 2017).

DISCUSSION

In considering the use of LLMs in engineering education research through Walther et al.'s (2013) Q3 framework, we have raised myriad points of caution for researchers aiming to integrate these tools into their methodological toolbelt. Importantly, the questions posed throughout the previous sections are not comprehensive; there are additional issues that we have not yet explored, given both space considerations and the rapidly evolving nature and use of LLMs. Thus, in Table 1, we provide guiding questions for researchers to begin exploring and critiquing how LLMs might coexist in the research process. This exercise can also complement and enhance researchers' existing processes relative to the Q3 framework (2014) and explorations of positionality (Secules et al. 2021).

CRITERION	QUESTIONS TO POSE
Ethical Validation	• What is your purpose for using an LLM in your research, and how does it impact your ability to do justice to the lived realities of the participants? • How will you consider participant privacy when handling the data? • Will the participants know that an LLM will be used in the research process, and have you obtained their consent?
Theoretical Validation	• How is the social reality under study reflected in the training data? • If the model is proprietary (training data is unavailable), how does the ambiguity in the training data influence the study of the full social reality in your research? • How will you determine if the findings generated by an LLM constitute a meaningful contribution to the literature or if findings over-emphasize depth? • What expertise do you possess to evaluate the model output's quality in the research context?
Procedural Validation	• How can you interrogate the black-box nature of large language models to enhance the trustworthiness of the results? • How will you interrogate the biases in the model and manage them while analyzing the data? And how can you do this in combination with interrogating your positionality?
Communicative Validation	• What robust methods can be used for co-constructing interpretative meanings *with the LLM*? • To what extent can meaning be co-constructed between the participant and an LLM *in the absence of the researcher*? And how involved does the researcher need to be in making data?
Pragmatic Validation	• What methods can be used to reconcile the model's output with practice? • How will you assess the meaningfulness of the LLM's output with respect to the social reality under investigation in other contexts?
Process Reliability	• Because large language models are probabilistic systems, how will you manage the randomness in the output? What about potential inaccuracies? • If the model is commercially hosted and updated (potentially without warning), how can consistency in the findings be achieved? • What information can be disclosed about your use of LLMs to document the process of making and handling the data? (e.g., chat logs, prompts, model version, date of conversation).

Table 1 Mapping of questions for researchers to ask themselves to address the role of the researcher and LLMs in the research process. Built from Walther et al. (2017).

WHEN USING A LARGE LANGUAGE MODEL IN RESEARCH, THINK CRITICALLY ABOUT ITS ROLE

Based on our guiding questions, we expect the first question a researcher will grapple with in their study is, "What is your purpose for using an LLM in your research, and how does it impact your ability to do justice to the lived realities of the participants?" LLMs are generally not knowledge engines, so although they appear to be able to search and reason, their outcomes are fundamentally based on the data they were trained on. Thus, it would be unwise to allow the LLM to perform a qualitative analysis and consider it complete. If the goal is to employ the model to perform the analysis, researchers should consider their position relative to the LLM. Is the researcher acting as a director, guiding the model using a chain-of-thought approach to perform the analysis step-by-step? Is the researcher acting as a client, providing the model with data and asking it to process potential themes and synthesize them, only to have it write up the paper shortly thereafter? Is the researcher acting as an auditor, asking the model to process the data and verifying that the analysis is sensible given the input data? Is the researcher acting as a colleague, analyzing the data themselves, processing the data using the model, and synthesizing the findings? Could it be something between these various roles or some combination? When engaging in reflexive practice, we must ask ourselves these questions to evaluate our positionality relative to the given study. Reflections on these questions could become part of the positionality statement itself.

Next, what does it mean for the LLM to be part of the research team, so to speak, considering it cannot authentically engage in reflective practice as defined by qualitative researchers? Before becoming too mired in the philosophy of what it means to be an "author," let us step back and briefly reflect on what we currently use in our research. Think about other tools or resources that may find their way into general use or several tools that perform much of the heavy lifting for us; for qualitative researchers, MaxQDA, Nvivo, and Dedoose for data analysis, and Rev.com and Zoom auto-transcriptions might come to mind. These tools are routinely cited in methods sections. Likewise, if a generative AI tool was used when conducting or writing a manuscript, then the same level of attribution should be provided, especially considering that LLMs can generate text-based interpretations of the data, unlike previous tools with which we are familiar. Some guidance exists on how researchers can appropriately credit its tools and others like them (McAdoo, 2024). One approach is to disclose the use of ChatGPT in the acknowledgments and provide the associated prompts and outputs produced using the following statement, mirroring how some journals ask authors to specify the type of effort each individual contributed:

> "The author(s) would like to acknowledge the use of [Generative AI Tool Name], developed by [Generative AI Tool Provider], in the preparation of this manuscript. The [Generative AI Tool Name] was used in the following way(s): [e.g., brainstorming, grammatical correction, analysis procedures]."

We anticipate that appropriate approaches to and scopes for disclosure will gradually become clearer as appropriate use cases for LLMs solidify. The recommendations can serve as evidence for process reliability in the meantime. Santu and Feng (2023) offer a taxonomy to help researchers report more precisely how they interacted with tools like ChatGPT, which contains seven levels of increasing complexity based on the turns (the number of exchanges when performing the task), expression (the style of the prompt as an instruction or question), level of detail, and role (directives for how the model should behave or respond).

RECOGNIZE WHEN ACCURACY IS NECESSARY USING LLMS, AND LEAN INTO USE CASES WHERE ACCURACY IS NOT CRUCIAL

Importantly, although there are apparent issues with using LLMs across the types of validation with interpretative engineering education, it would be irresponsible of us not to mention the positive outcomes as we conclude with our final cross-cutting thoughts. Regardless of the application, we reiterate that all outputs of LLMs should be thoroughly checked for accuracy. Letters to the editor, such as those by Haman and Školník (2023), demonstrate how researchers might intuitively use a tool like ChatGPT in the research process to automate what might be seen as a slog to complete, such as the literature review. The authors used the prompt "List 10 seminal academic articles in the field of medicine and provide the DOI" five times, resulting in 48 unique publications with only eight correct DOIs and only 17 authentic articles. This treats ChatGPT as a knowledge engine, which is an unwise use of LLMs without verification. However, how might we flip the task from generating the papers to refining our search? Perhaps we are struggling to come up with the right keywords; a preprint from Wang et al. (2023) shows how we can use ChatGPT or similar tools to generate the necessary search strings to retrieve the papers of interest. ChatGPT can serve as an excellent starting point for divergent thinking tasks, where the goal is to generate as many solutions as possible to a given problem.

However, what about cases where accuracy is not necessary? Researchers should consider how generative AI tools can unlock creative blocks and supplement their work rather than automate their efforts uncritically. For example, the lead author was developing survey questions during a brainstorming session with his research group about core ideas, known as threshold concepts, in the field of cyber-physical systems. After several minutes of silence and blank expressions, he turned to ChatGPT (GPT-3.5 at the time) to overcome the inertia of starting the brainstorming session with the following prompt:

> "We're developing a survey to be sent to professionals in cyber-physical systems to understand what threshold concepts exist in their discipline. Suggest a couple of questions aligned with the qualities of threshold concepts. Format as a table with the first column containing the questions and the second column containing the threshold concept quality."

For which, ChatGPT produced Table 2.

Table 2 Table generated by ChatGPT in response to prompt.

*Upon review, the group did not agree with the classification made by ChatGPT (GPT 3.5).

QUESTIONS	*THRESHOLD CONCEPT QUALITY*
1. What fundamental ideas or principles in cyber-physical systems were particularly challenging for you to grasp when first learning about the field?	Transformational
2. Can you identify any concepts that, once understood, significantly changed your perspective or approach to solving problems in cyber-physical systems?	Integrative*
3. Were there any concepts that took a considerable amount of time and effort to fully internalize, but were essential for your professional growth in this field?	Irreversible*
4. Describe any concepts that bridge the gap between theory and practical application in cyber-physical systems. Which ones were difficult to bridge?	Troublesome*
5. Have you encountered any concepts that, once mastered, allowed you to communicate more effectively with colleagues and experts in cyber-physical systems?	Discursive
6. Are there concepts in this discipline that are often misunderstood or misinterpreted by newcomers or even experienced professionals? If so, which ones and why?	Liminal*

Was the prompt optimized? No. So, were the classifications correct? No. In fact, most were incorrect. *Did it matter? No.* Once the group had a set of questions, the task shifted from generating them to revising their content and remapping them to the appropriate categories, a discussion that flowed much more productively, a nod to communicative validation within the research team's processes. This is a sweet spot for large language models—low risk and high reward—and demonstrates pragmatic validation by reconciling the output with prompts more suitable for practical application in the research design, which in turn ideally would bolster theoretical validation by better capturing the fundamental qualities within the threshold concept literature. Of course, the large language model would have performed better if we had defined the threshold concept qualities. Providing examples through a chain-of-thought-style prompt, like the correct pairings, would have likely also improved the performance. Engineering education researchers can adopt these tools in similar applications where divergent thinking is necessary and correctness is not a priority.

CONCLUSION

In this paper, we have examined the quality considerations associated with using technology based on LLMs (such as ChatGPT) within the context of the Q3 Framework. Although there are still many unanswered questions, the landscape seemingly shifts beneath our feet with each passing second, leaving us with more than we had before. In the methodological space, we will likely need to revisit our positionality, reflexivity, and collaboration processes to ensure our findings are of quality. This paper outlined initial ideas that engineering education researchers can consider as they weigh the impacts of using LLM-based tools in their research. It's expected that this line of research will grow and be critiqued as the technology improves. We hope that through it all, we will recount the steps to validate our work ethically along the way.

ACKNOWLEDGEMENTS

We would like to thank the editors and reviewers for their detailed comments, which significantly improved the writing and argumentation of this paper.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

DR led the conceptualization of the paper. All authors contributed to writing the paper and the analysis. All authors engaged in revisions with each iteration of the paper. We utilized the LLM-powered tool ChatGPT to generate its positionality statement, as described in the paper's body. However, it was not used for other purposes in developing the manuscript.

AUTHOR AFFILIATIONS

David Reeping orcid.org/0000-0002-0803-7532
University of Cincinnati, US
Cynthia Hampton orcid.org/0000-0001-8329-6465
Virginia Tech, US
Desen Özkan orcid.org/0000-0002-1996-7719
University of Connecticut, US

REFERENCES

Azaria, A., Azoulay, R., & **Reches, S.** (2023). ChatGPT is a Remarkable Tool—For Experts. (arXiv:2306.03102). arXiv. https://doi.org/10.1162/dint_a_00235

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. *Journal of AI*, *7*(1), 52–62. https://doi.org/10.2139/ssrn.4337484

Bender, E. M., Gebru, T., McMillan-Major, A., & **Shmitchell, S.** (2021, March). *On the dangers of stochastic parrots: Can language models be too big?* 🦜. Paper presented at the 2021 Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445922

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Polity.

Berdanier, C. G. P., & **Alley, M.** (2023). We still need to teach engineers to write in the era of ChatGPT. *Journal of Engineering Education*, *112*(3), 583–586. https://doi.org/10.1002/jee.20541

Biswas, S. S. (2023). ChatGPT for research and publication: A step-by-step guide. *The Journal of Pediatric Pharmacology and Therapeutics*, *28*(6), 576–584. https://doi.org/10.5863/1551-6776-28.6.576

Bowman, E. (2022, December 19). A new AI chatbot might do your homework for you. But it's still not an A+ student. *NPR*. https://www.npr.org/2022/12/19/1143912956/chatgpt-ai-chatbot-homework-academia

Bradburn, N., Sudman, S., Blair, E., & **Stocking, C.** (1978). Question threat and response bias. *Public Opinion Quarterly*, *42*(2), 221–234. https://doi.org/10.1086/268444

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., & **Perez, E.** (2025). Reasoning models don't always say what they think. (arXiv:2505.05410). arXiv. https://doi.org/10.48550/arXiv.2505.05410

Chopra, F., & **Haaland, I.** (2023). Conducting qualitative interviews with AI (SSRN Scholarly Paper 4583756). Social Science Research Network. https://doi.org/10.2139/ssrn.4583756

Chowdhury, A. (2023, May 31). *ChatGPT may have been quietly nerfed recently—Here's why*. https://www.videogamer.com/news/chatgpt-nerfed/

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. The MIT Press. https://library.oapen.org/handle/20.500.12657/43542. https://doi.org/10.7551/mitpress/12255.001.0001

Cuervas, A. C., Brown, E. M., Scurrell, J. V., Entenmann, J., & **Daepp, M. I. G.** (2023, September 18). Automated interviewer or augmented survey? Collecting social data with large language models. arXiv.Org. https://arxiv.org/abs/2309.10187v2

Curry, N., Baker, P., & **Brookes, G.** (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, *4*(1), 100082. https://doi.org/10.1016/j.acorp.2023.100082

Dastin, J. (2023, March 14). Microsoft-backed OpenAI starts release of powerful AI known as GPT-4. *Reuters*. https://www.reuters.com/technology/microsoft-backed-openai-starts-release-powerful-ai-known-gpt-4-2023-03-14/

Davison, R. M., Chughtai, H., Nielsen, P., Marabelli, M., Iannacci, F., van Offenbeek, M., Tarafdar, M., Trenz, M., Techatassanasoontorn, A. A., Díaz Andrade, A., & **Panteli, N.** (2024). The ethics of using generative AI for qualitative data analysis. *Information Systems Journal*, *34*(5), 1433–1439. https://doi.org/10.1111/isj.12504

de Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule*, *7*(10), 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & **Narasimhan, K.** (2023). Toxicity in ChatGPT: Analyzing persona-assigned language models. (arXiv:2304.05335). arXiv. https://doi.org/10.18653/v1/2023.findings-emnlp.88

D'Ignazio, C., & **Klein, L. F.** (2023). *Data feminism*. MIT Press.

Dzieza, J. (2023, June 20). AI is a lot of work. *New York Magazine/The Verge*. https://longreads.com/2023/06/20/ai-is-a-lot-of-work/

Ferretti, S. (2023). Hacking by the prompt: Innovative ways to utilize ChatGPT for evaluators. *New Directions for Evaluation*, *2023*(178–179), 73–84. https://doi.org/10.1002/ev.20557

Froyd, J. E., Wankat, P. C., & **Smith, K. A.** (2012). Five major shifts in 100 years of engineering education. *Proceedings of the IEEE*, *100* (Special Centennial Issue), 1344–1360. https://doi.org/10.1109/JPROC.2012.2190167

Gamieldien, Y., Case, J. M., & **Katz, A.** (2023). Advancing qualitative analysis: An exploration of the potential of generative AI and NLP in thematic coding (SSRN Scholarly Paper 4487768). Social Science Research Network. https://doi.org/10.2139/ssrn.4487768

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report*, *8*(4), 597–606. https://doi.org/10.46743/2160-3715/2003.1870

Goyanes, M., Lopezosa, C., & **Jordá, B.** (2024). Thematic analysis of interview data with ChatGPT: Designing and testing a reliable research protocol for qualitative research. OSF. https://doi.org/10.31235/osf.io/8mr2f

Grant, N., & **Metz, C.** (2023, March 21). Google releases Bard, its AI chatbot, a rival to ChatGPT and Bing. *The New York Times*. https://www.nytimes.com/2023/03/21/technology/google-bard-chatbot.html

Haman, M., & **Školník, M.** (2023). Using ChatGPT to conduct a literature review. *Accountability in Research*, *0*(0), 1–3. https://doi.org/10.1080/08989621.2023.2185514

Hampton, C., & **Reeping, D.** (2019). *Positionality: The stories of self that impact others*. Paper presented at 2019 ASEE Annual Conference & Exposition, Tampa, Florida, USA. https://doi.org/10.18260/1-2--33177

Hampton, C., Reeping, D., & **Ozkan, D. S.** (2021). Positionality statements in engineering education research: A look at the hand that guides the methodological tools. *Studies in Engineering Education*, *1*(2), Article 2. https://doi.org/10.21061/see.13

Han, H., Liang, J., Shi, J., He, Q., & **Xiao, Y.** (2024). Small language model can self-correct (arXiv:2401.07301). arXiv. https://doi.org/10.48550/arXiv.2401.07301

Hernández, A., & **Amigó, J. M.** (2020). Differentiable programming and its applications to dynamical systems (arXiv:1912.08168). arXiv. https://doi.org/10.48550/arXiv.1912.08168

Hochreiter, S., & **Schmidhuber, J.** (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu, K. (2023, February 2). ChatGPT sets record for fastest-growing user base—Analyst note. *Reuters*. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & **Zhou, D.** (2023). *Large language models cannot self-correct reasoning yet*. arXiv preprint arXiv:2310.01798.

Jalali, M. S., & **Akhavan, A.** (2024). Integrating AI language models in qualitative research: Replicating interview data analysis with ChatGPT. *System Dynamics Review*, *40*(3), e1772. https://doi.org/10.1002/sdr.1772

Johri, A., Katz, A. S., Qadir, J., & **Hingle, A.** (2023). Generative artificial intelligence and engineering education. *Journal of Engineering Education*, *112*(3), 572–577. https://doi.org/10.1002/jee.20537

Katz, A., Fleming, G. C., & **Main, J.** (2024, September 28). Thematic analysis with open-source generative AI and machine learning: A new method for inductive qualitative codebook development. arXiv.Org. https://arxiv.org/abs/2410.03721v1

Katz, A., Gerhardt, M., & **Soledad, M.** (2024). Using generative text models to create qualitative codebooks for student evaluations of teaching. *International Journal of Qualitative Methods*, *23*, 16094069241293283. https://doi.org/10.1177/16094069241293283

Katz, A., Shakir, U., & **Chambers, B.** (2023, May 29). The utility of large language models and generative AI for education research. arXiv.Org. https://arxiv.org/abs/2305.18125v1

Katz, A., Wei, S., Nanda, G., Brinton, C., & **Ohland, M.** (2023). Exploring the efficacy of ChatGPT in analyzing student teamwork feedback with an existing taxonomy (arXiv:2305.11882). arXiv. https://doi.org/10.48550/arXiv.2305.11882

Kirk, J., & **Miller, M. L.** (1986). *Reliability and validity in qualitative research*. SAGE. https://doi.org/10.4135/9781412985659

Knight, W. (2023, April 17). OpenAI's CEO says the age of giant AI models is already over. *Wired*. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/

Laiq, M., & **Dieste, O.** (2020). *Chatbot-based interview simulator: A feasible approach to train novice requirements engineers*. Paper presented at the 2020 10th International Workshop on Requirements Engineering Education and Training (REET), Zurich, Switzerland. https://doi.org/10.1109/REET51203.2020.00007

Lehr, S. A., Caliskan, A., Liyanage, S., & **Banaji, M. R.** (2024). ChatGPT as research scientist: Probing GPT's capabilities as a research librarian, research ethicist, data generator, and data predictor. *Proceedings of the National Academy of Sciences*, *121*(35), e2404328121. https://doi.org/10.1073/pnas.2404328121

Luitse, D., & **Denkena, W.** (2021). The great Transformer: Examining the role of large language models in the political economy of AI. *Big Data & Society*, *8*(2), 20539517211047734. https://doi.org/10.1177/20539517211047734

Martin, J., Desing, R., & **Borrego, M.** (2022). Positionality statements are just the tip of the iceberg: moving towards a reflexive process. *Journal of Women and Minorities in Science and Engineering*, *28*(4). https://doi.org/10.1615/JWomenMinorScienEng.2022044277

McAdoo, T. (2024, February 23). How to cite ChatGPT. *American Psychological Association*. https://apastyle.apa.org/blog/how-to-cite-chatgpt

Mejia, J., Alarcón, I. V., Mejia, J., & **Revelo, R.** (2022). Legitimized tongues: Breaking the traditions of silence in mainstream engineering education and research. *Journal of Women and Minorities in Science and Engineering*, *28*(2). https://doi.org/10.1615/JWomenMinorScienEng.2022036603

Mejia, J., Revelo, R., Villanueva, I., & **Mejia, J.** (2018). Critical theoretical frameworks in engineering education: An anti-deficit and liberative approach. *Education Sciences*, *8*(4), 158. https://doi.org/10.3390/educsci8040158

Menekse, M. (2023). Envisioning the future of learning and teaching engineering in the artificial intelligence era: Opportunities and challenges. *Journal of Engineering Education*, *112*(3), 578–582. https://doi.org/10.1002/jee.20539

Meta. (2024). *Running Meta Llama on Windows | Llama Everywhere*. https://www.llama.com/docs/llama-everywhere/running-meta-llama-on-windows/

Milner, H. R. (2007). Race, culture, and researcher positionality: Working through dangers seen, unseen, and unforeseen. *Educational Researcher*, *36*(7), 388–400. https://doi.org/10.3102/0013189X07309471

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

OpenAI. (2024a). *How your data is used to improve model performance | OpenAI Help Center*. https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

OpenAI. (2024b, September 12). *Introducing OpenAI o1*. https://openai.com/o1/

Prescott, M. R., Yeager, S., Ham, L., Saldana, C. D. R., Serrano, V., Narez, J., Paltin, D., Delgado, J., Moore, D. J., & **Montoya, J.** (2024). Comparing the efficacy and efficiency of human and generative AI: Qualitative thematic analyses. *JMIR AI*, *3*(1), e54482. https://doi.org/10.2196/54482

Rahman, M. M., Terano, H. J., Rahman, M. N., Salamzadeh, A., & **Rahaman, M. S.** (2023). ChatGPT and academic research: A review and recommendations based on practical examples (SSRN Scholarly Paper 4407462). *Social Science Research Network*. https://papers.ssrn.com/abstract=4407462

Reeping, D., & **Shah, A.** (2024, June 23). *Board 50: Work in progress: A systematic review of embedding large language models in engineering and computing education*. Paper presented at the 2024 ASEE Annual Conference & Exposition, Portland, Oregon, USA. https://peer.asee.org/board-50-work-in-progress-a-systematic-review-of-embedding-large-language-models-in-engineering-and-computing-education

Rodgers, B. L., & **Cowles, K. V.** (1993). The qualitative research audit trail: A complex collection of documentation. *Research in Nursing & Health*, *16*(3), 219–226. https://doi.org/10.1002/nur.4770160309

Rowe, W. (2014). Positionality. In D. Coghlan & M. Brydon-Miller (Eds.), *The SAGE encyclopedia of action research* (p. 628). Sage.

Rudin, C., & **Radin, J.** (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.5a8a3a3d

Rumelhart, D. E., Hinton, G. E., & **Williams, R. J.** (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), Article 6088. https://doi.org/10.1038/323533a0

Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., & **Pauly, M.** (2023). The self-perception and political biases of ChatGPT (arXiv:2304.07333). arXiv. https://doi.org/10.1155/2024/7115633

Sandberg, J. (2005). How do we justify knowledge produced within interpretive approaches? *Organizational Research Methods*, *8*(1), 41–68. https://doi.org/10.1177/1094428104272000

Santu, S. K. K., & **Feng, D.** (2023). TELeR: A general taxonomy of LLM prompts for benchmarking complex tasks (arXiv:2305.11430). arXiv. https://doi.org/10.48550/arXiv.2305.11430

Savolainen, J., Casey, P. J., McBrayer, J. P., & **Schwerdtle, P. N.** (2023). Positionality and its problems: Questioning the value of reflexivity statements in research. *Perspectives on Psychological Science*, *18*(6), 1331–1338. https://doi.org/10.1177/17456916221144988

Secules, S., McCall, C., Mejia, J. A., Beebe, C., Masters, A. S., Sánchez-Peña, M. L., & **Svyantek, M.** (2021). Positionality practices and dimensions of impact on equity research: A collaborative inquiry and call to the community. *Journal of Engineering Education*, *110*(1), 19–43. https://doi.org/10.1002/jee.20377

Shanahan, M., McDonell, K., & **Reynolds, L.** (2023). Role play with large language models. *Nature*, *623*(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8

Sheng, E., Chang, K.-W., Natarajan, P., & **Peng, N.** (2019). The woman worked as a babysitter: On biases in language generation (arXiv:1909.01326). arXiv. https://doi.org/10.48550/arXiv.1909.01326

Singleton, R. A., & **Straits, B. C.** (2017). *Approaches to social research* (6th edition). Oxford University Press.

Sochacka, N. W., Walther, J., & **Pawley, A. L.** (2018). Ethical validation: Reframing research ethics in engineering education research to improve research quality. *Journal of Engineering Education*, *107*(3), 362–379. https://doi.org/10.1002/jee.20222

Sok, S., & **Heng, K.** (2023). ChatGPT for education and research: A review of benefits and risks (SSRN Scholarly Paper 4378735). https://doi.org/10.2139/ssrn.4378735

Strobel, J., Medina, M., & **Guzman, E. S.** (2024). *Exploring AI Bots as simulators in human subject research: A novel approach to ethical and efficient experimentation in engineering education research*. Paper presented at the IEEE Frontiers in Education Conference 2024, Washington DC, USA. https://doi.org/10.1109/FIE61694.2024.10893007

Surovell, E. (2023, February 8). ChatGPT has everyone freaking out about cheating. It's not the first time. *The Chronicle of Higher Education*. https://www.chronicle.com/article/chatgpt-has-everyone-freaking-out-about-cheating-its-not-the-first-time

Tangermann, V. (2023, January 21). A new scientific paper credits ChatGPT AI as a coauthor. *Futurism*. https://futurism.com/scientific-paper-credits-chatgpt-ai-coauthor

Teixeira da Silva, J. A. (2023). Is institutional review board approval required for studies involving ChatGPT? *American Journal of Obstetrics & Gynecology MFM*, *5*(8), 101005. https://doi.org/10.1016/j.ajogmf.2023.101005

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & **Polosukhin, I.** (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vincent, J. (2023, February 15). Microsoft's Bing is an emotionally manipulative liar, and people love it. *The Verge*. https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams

Wachinger, J., Bärnighausen, K., Schäfer, L. N., Scott, K., & **McMahon, S. A.** (2024). Prompts, pearls, imperfections: Comparing ChatGPT and a human researcher in qualitative data analysis. *Qualitative Health Research*, 10497323241244669. https://doi.org/10.1177/10497323241244669

Walther, J., Pawley, A. L., & **Sochacka, N. W.** (2015). Exploring ethical validation as a key consideration in interpretive research quality. 26.726.1–26.726.21. https://peer.asee.org/exploring-ethical-validation-as-a-key-consideration-in-interpretive-research-quality. https://doi.org/10.18260/p.24063

Walther, J., & **Sochacka, N.** (2014). *Qualifying qualitative research quality (The Q³ project): An interactive discourse around research quality in interpretive approaches to engineering education research*. Paper presented at the 2014 IEEE Frontiers in Education Conference (FIE), Madrid, Spain. https://doi.org/10.1109/FIE.2014.7043988

Walther, J., Sochacka, N. W., Benson, L. C., Bumbaco, A. E., Kellam, N., Pawley, A. L., & **Phillips, C. M. L.** (2017). Qualitative research quality: A collaborative inquiry across multiple methodological perspectives. *Journal of Engineering Education*, *106*(3), 398–430. https://doi.org/10.1002/jee.20170

Walther, J., Sochacka, N. W., & **Kellam, N. N.** (2013). Quality in interpretive engineering education research: Reflections on an example study: Quality in interpretive engineering education research. *Journal of Engineering Education*, *102*(4), 626–659. https://doi.org/10.1002/jee.20029

Wang, S., Scells, H., Koopman, B., & **Zuccon, G.** (2023). Can ChatGPT write a good boolean query for systematic review literature search? (arXiv:2302.03495). arXiv. https://doi.org/10.48550/arXiv.2302.03495

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & **Zhou, D.** (2023). Chain-of-thought prompting elicits reasoning in large language models (arXiv:2201.11903). arXiv. https://doi.org/10.48550/arXiv.2201.11903

Zambrano, A. F., Liu, X., Barany, A., Baker, R. S., Kim, J., & **Nasiar, N.** (2023). From nCoder to ChatGPT: From automated coding to refining human coding. In G. Arastoopour Irgens & S. Knight (Eds.), *Advances in quantitative ethnography* (pp. 470–485). Switzerland: Springer Nature. https://doi.org/10.1007/978-3-031-47014-1_32

Zhang, H., Wu, C., Xie, J., Kim, C., & **Carroll, J. M.** (2023, October 10). QualiGPT: GPT as an easy-to-use tool for qualitative coding. arXiv.Org. https://arxiv.org/abs/2310.07061v1

TO CITE THIS ARTICLE:
Reeping, D., Hampton, C., & Özkan, D. (2025). Interrogating the Use of Large Language Models in Qualitative Research Using the Qualifying Qualitative Research Quality Framework. *Studies in Engineering Education*, *6*(2), 1–23. DOI: https://doi.org/10.21061/see.174

Submitted: 21 March 2024
Accepted: 04 June 2025
Published: 04 July 2025

Studies in Engineering Education is a peer-reviewed open access journal published by VT Publishing.

VIRGINIA TECH PUBLISHING