Received: 1 November 2024 | Revised: 18 September 2025 | Accepted: 29 September 2025

DOI: 10.1002/jee.70037

RESEARCH ARTICLE

JEE

Affordances and limitations of using large language models to generate qualitative data about mental health perceptions in engineering

Jeanne Sanders[1] | John Mobley IV[2] | Isabel Miller[1,2] | Nicola W. Sochacka[3] | Paul A. Jensen[1,4] | Karin J. Jensen[1,2]

[1]Department of Biomedical Engineering, University of Michigan Ann Arbor, Ann Arbor, Michigan, USA

[2]Engineering Education Research Program, University of Michigan Ann Arbor, Ann Arbor, Michigan, USA

[3]ProQual Institute for Interpretive Research Methods, Sunshine Coast, Queensland, Australia

[4]Department of Chemical Engineering, University of Michigan Ann Arbor, Ann Arbor, Michigan, USA

Correspondence
Karin J. Jensen, Department of Biomedical Engineering, University of Michigan Ann Arbor, 1101 Beal Ave, Ann Arbor, MI 48109, USA.
Email: kjens@umich.edu

Funding information
National Science Foundation, Grant/Award Number: 2342384

Abstract

Background: Generative artificial intelligence (AI) large-language models (LLMs) have significant potential as research tools. However, the broader implications of using these tools are still emerging. Few studies have explored using LLMs to generate data for qualitative engineering education research.

Purpose/Hypothesis: We explore the following questions: (i) What are the affordances and limitations of using LLMs to generate qualitative data in engineering education, and (ii) in what ways might these data reproduce and reinforce dominant cultural narratives in engineering education, including narratives of high stress?

Design/Methods: We analyzed similarities and differences between LLM-generated conversational data (ChatGPT) and qualitative interviews with engineering faculty and undergraduate engineering students from multiple institutions. We identified patterns, affordances, limitations, and underlying biases in generated data.

Results: LLM-generated content contained similar responses to interview content. Varying the prompt persona (e.g., demographic information) increased the response variety. When prompted for ways to decrease stress in engineering education, LLM responses more readily described opportunities for structural change, while participants' responses more often described personal changes. LLM data more frequently stereotyped a response than participants did, meaning that LLM responses lacked the nuance and variation that naturally occurs in interviews.

Conclusions: LLMs may be a useful tool in brainstorming, for example, during protocol development and refinement. However, the bias present in the data indicates that care must be taken when engaging with LLMs to generate data. Specially trained LLMs that are based only on data from engineering education hold promise for future research.

KEYWORDS

ChatGPT, engineering education, generative artificial intelligence, idealized worker, stereotype

1 | INTRODUCTION

Generative artificial intelligence (AI) large-language models (LLMs), such as ChatGPT, have emerged as widely applicable tools for generating text-based information. These LLMs generate conversational text in response to user prompts, drawing from extensive training, generally on text originally authored by humans (Ge et al., 2024). Utilizing generative techniques, these models analyze inputs to discern context, intent, and linguistic nuances. By leveraging vast datasets encompassing diverse topics, LLMs generate coherent and contextually relevant—though not always factually accurate—responses. For example, they are capable of engaging in meaningful dialogue, answering inquiries, crafting narratives, and assisting in complex problem-solving. This leads to a question of if and when generating this meaningful and complex LLM-generated data might be useful as data for qualitative engineering education studies. The iterative learning and fine-tuning processes employed in the development of these models enhance their adaptability to different subject matter, rendering them assets in varying fields (Ge et al., 2024). However, understanding how the datasets used to train LLMs influence their outputs is critical. The composition, diversity, and inherent biases within these datasets directly impact the quality and accuracy of the generated responses (Bender et al., 2021; Dengel et al., 2023; Hämäläinen et al., 2023; Muldoon & Wu, 2023).

In this study, we explore affordances and limitations of using LLMs to generate qualitative data in engineering education. We do so by comparing conversational data generated by a widely available LLM (ChatGPT) with multi-institutional qualitative interviews previously conducted with students and faculty to better understand their experiences of stress and wellness in engineering education. Exploring perceptions of mental health in engineering is important because bias is present in discourses on mental health and stress (Asghar & Minichiello, 2023), and understanding engineering cultural norms around mental health and stress is important for addressing these challenges in engineering academia (Jensen, 2021). Additionally, people are increasingly engaging with LLMs and AI, such as ChatGPT, in their daily lives, and thus these tools are placed in a position to influence norms and discourses about mental health and stress.

Our findings identify opportunities to enhance the quality and efficiency of qualitative data generation by purposefully working with LLMs. For example, such a human–AI partnership might entail using AI in conjunction with practice interviews to refine the question flow of interview protocols. At the same time, we identify instances when LLMs may be less helpful for qualitative data generation, such as when biases inherent to LLMs are likely to limit the richness of data or when the complexity of human experience lies beyond LLM training datasets. Our research questions are twofold, centered on exploring the affordances and limitations of using LLMs to generate qualitative data and the ways in which LLM-generated data can reproduce and reinforce cultural norms and dominant narratives of stress and the idealized worker in engineering.

2 | HOW LLMs WORK

Before describing our methods, it is important to understand how LLMs work. LLMs generate free-form text in response to a user's prompt. The prompt, itself a block of text, is first converted into a numerical embedding that provides context for each token in the prompt (a token is a small piece of text, usually smaller than a word) (OpenAI et al., 2024). These embeddings are passed through an attention mechanism that allows the LLM to identify key relationships between tokens in the prompt as it generates a response. While many machine learning models have been used for understanding or generating text, the transformer model developed in 2013 (Google, 2013) was the first to combine embeddings and attention in a manner that scales to huge numbers of parameters trained on an enormous corpus of text. OpenAI, the company that developed and trained the ChatGPT 4o (where o stands for "omni"), the LLM we used in this study, does not release details on the model size, data size, or training procedures from their model. ChatGPT 4o is presumed to be a larger model than the previous ChatGPT 3 model released by OpenAI in 2020. ChatGPT 3 included up to 175 billion parameters (Brown et al., 2020) and was trained on trillions of tokens of data from the Common Crawl, a publicly available dataset containing internet and web data (Raffel, 2019), the WebText dataset (Radford et al., 2019), two internet-based corpuses of books, and the English language Wikipedia. ChatGPT version 3.5 includes human feedback in its training, consisting of teams of contractors from third-party companies globally, to both improve answers and avoid inappropriate, obscene, or unhelpful responses (Ouyang et al., 2022).

Three features of LLM training and usage are relevant to this study. First, LLMs are stochastic (OpenAI et al., 2024). The transformer models inside an LLM output a probability distribution over all possible responses to a prompt, and this distribution is sampled to construct the response shown to the user. Different users who give the same prompts to the same model could (and often do) receive different responses.

Second, LLMs augment the user's prompt with a larger context (i.e., text that is given to the LLM alongside the prompt to improve the response) (OpenAI et al., 2024). Context could include previous prompts and responses from the same session to allow both the user and the LLM to reference previous parts of a "conversation." Information gathered from the user (i.e., that does not come from the internet repository) can also be added as context to personalize the LLM. OpenAI does not detail what and how much context is given to the ChatGPT 4o model; however, previous models used context sizes equal to thousands of words (Brown et al., 2020).

Third, the responses from LLMs reflect the biases of their training data. Even the process of embedding words as numeric vectors (i.e., transforming text fragments into arrays or lists of numbers that represent the context of the fragment) creates gender bias in the vectors that cannot be fully removed, as the bias comes from the data, not the model (Bolukbasi et al., 2016). The only way to remove this bias would be to train the model on text that is free of bias, which is impossible because of the inherent bias of all authors.

Altogether, these three properties of LLMs imply that AI-generated text is not a neutral source of ground truth produced by a deterministic machine. Instead, LLM responses share many of the properties of qualitative data: nuanced responses that lack repeatability and are biased by the experiences of the participant. This work investigates LLMs as a tool for improving qualitative research but also demonstrates how experience collecting, interpreting, and analyzing qualitative data is a critical skill for users of LLMs.

It is helpful to consider how data from traditional qualitative analysis (i.e., conducted fully by a human) is incorporated into broad-data LLMs such as ChatGPT that are trained on a wide variety of source material. Figure 1 provides an illustration of this data flow.

Researchers often conduct interviews with carefully selected, specific subpopulations as depicted on the left side of Figure 1. Researchers then synthesize these data through qualitative analysis and then share results publicly, for example, through journal articles, which are depicted as the "Qual analysis" arrow in Figure 1. However, these interviews are rarely made publicly available to protect participant anonymity and safety. These interviews are thus not directly available as training data for broad-data LLMs, depicted by the "Not this" arrow in Figure 1. The resulting research articles may be included in the broad range and variety of LLM training and context data. In Figure 1, the resulting research articles (Experience Synthesis) are depicted as contributing to the LLM output by the curved, dotted green arrow and small green box on the right.

Even though the synthesized results of qualitative analysis may be used to generate AI responses, it is unclear to what extent these articles are included. Many sites with paywalls, such as journal publishers, put effort into blocking or misdirecting web crawlers that are used to gather information that is used for training AI (Baack & Mozilla Insights, 2024). So, even though a research article is available through the internet, it may not be available as training data for AI. Additionally, and as mentioned previously, the training data for later models of ChatGPT has not been published, adding to the opacity of understanding what training data was used.

Broad-data LLMs gather information from many sources, such as written accounts from members of a specific subpopulation of interest that could influence the responses generated. This means that the data used to train the LLM could also be used in qualitative analysis, for example, Reddit posts used in engineering education research (EER) (Berdanier et al., 2020). However, the training data itself and the processes used to clean this data can introduce bias. For example, C4 is a public dataset developed by researchers at Google as a cleaner, smaller version of the Common Crawl dataset (Raffel et al., 2020) that has been used to train LLMs (Dodge et al., 2021). In addition to containing source data that contains bias (Schaul et al., 2023), analysis by Dodge et al. (2021) identified that the filtering used to clean the data disproportionately removed text from and about minoritized individuals.

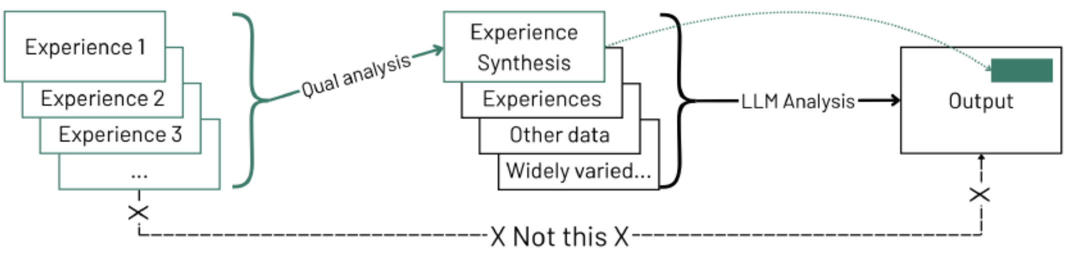

FIGURE 1 Data flow diagram depicting how traditional qualitative analysis is incorporated into large-language models (LLMs).

LLM training and output data are contextually similar to qualitative data, and they contain content related to many research questions of interest. However, because of its newness and the opacity of the training process, researchers do not know the extent to which these broad-data LLMs' output can be useful for qualitative research. As presented in this paper, a comparison between the LLM-generated output and the corresponding interview material developed by the research team and their participants provides insight into the affordances, limitations, and biases present in LLM-generated data.

3 | USING LLMs IN ENGINEERING EDUCATION RESEARCH FOR QUALITATIVE RESEARCH

3.1 | LLMs in engineering education

AI is emerging in engineering education as a tool for both teaching and learning (Liu et al., 2025; Martín Núñez & Diaz Lantada, 2020). For example, educators can use AI tools to help manage their time and workload by utilizing AI to aid in developing curricular plans and assessing student work (Martín Núñez & Diaz Lantada, 2020). Liu et al. (2025) further categorized the applications of AI in engineering education in their literature review and identified seven categories: (i) virtual experiment environments, (ii) learning prediction, (iii) learning analytics, (iv) engineering education robots, (v) intelligent tutoring systems, (vi) automatic evaluations, and (vii) assisted learning. AI has the potential to support learning and foster greater inclusivity in engineering education through tools that can create personalized learning pathways, add subtitles to slides, translate text into multiple languages, and implement other accessibility features (Martín Núñez & Diaz Lantada, 2020). AI also provides opportunities to make engineering education more equitable by offering accessible virtual alternatives to service-learning programs and experiences (Martín Núñez & Diaz Lantada, 2020), since, for example, the travel and time requirements of service-learning opportunities often exclude students with jobs or caretaking duties.

3.2 | LLMs in qualitative research: Benefits

Because of the ability of LLMs to process and generate text-based information, there is increasing interest in how they might be used to accelerate the traditionally time-intensive processes of qualitative research. LLMs can be used in qualitative research to summarize readings or transcripts, create or tailor study documents, generate codes or themes, and apply codes to data (Aubin Le Quéré et al., 2024; Drinkwater Gregg et al., 2025; Fuchs et al., 2025; Hayes, 2025; Tai et al., 2024). These affordances have prompted growing interest in exploring how LLMs might be used not only to analyze but also to generate qualitative data. Some may be drawn to LLMs for their ability to rapidly produce text that mimics human responses, potentially reducing the time and cost associated with traditional data collection (Hämäläinen et al., 2023). LLMs can also support the piloting of interview protocols or analytic strategies by generating response patterns for testing (Dengel et al., 2023), including in EER (Mburu et al., 2025).

Much work in this area has largely focused on using AI to analyze data (Sufi, 2024; Tai et al., 2024; Z. Wang et al., 2024; Xiao et al., 2023), but there is limited research examining the use of LLMs to *generate* data for qualitative research. AI-generated data aims to quicken the research process by generating information on the experiences of artificial users/participants. For example, Dengel et al. (2023) used LLMs, including ChatGPT and BARD, to generate interview data about K-12 computer science education; Hämäläinen et al. (2023) used generated interview responses about video games; Salminen et al. (2024) and Sattele and Ortiz (2024) used AI to generate user personas to help provide feedback in the design process; and J. S. Park et al. (2024) used LLMs to generate 1000 simulations based on real human participant interviews with high fidelity. Since LLMs summarize text from a large pool of data, they may provide insights into understanding cultural phenomena; however, many researchers emphasize the importance of ethical and intentional data generation to help avoid biased responses (Dengel et al., 2023; Hämäläinen et al., 2023; Roberts et al., 2024; Q. Wang et al., 2025).

3.3 | LLMs in qualitative research: Ethical concerns

Despite their promise, the use of LLMs to generate qualitative data raises important methodological and ethical concerns such as generic or biased output, misrepresentation of findings, the potential for uncritical acceptance of this data, and the energy costs of using AI.

LLMs have the potential for producing generic or biased output. LLMs use large datasets to produce outputs, but source data such as web data often contains biased language and hegemonic views centering White, heterosexual, male voices that can be reproduced in outputs (Bender et al., 2021; Muldoon & Wu, 2023). Additionally, some populations feel unsafe sharing their views on online platforms, resulting in a dominant view prevailing and silencing marginalized voices (Bender et al., 2021; Muldoon & Wu, 2023). As described previously, the process of preparing data for training can disproportionately and negatively impact marginalized voices. Additionally, during the development process, models also clean and remove data based on language that is obscene or unintelligible. This removes some hate speech and slurs from datasets, but this also removes additional marginalized actors who reclaim offensive language (Bender et al., 2021). These cleaning methods further shape datasets to be modeled on hegemonic ideals, and multiple studies have found that larger datasets tend to be more biased, racist, sexist, and so on, in content (Muldoon & Wu, 2023).

AI-generated output can also misrepresent data and findings. Hämäläinen et al. (2023) found GPT-3 to have created its own fake information when describing video games, which was often misidentified as human-generated by study participants because of the realistic emotional stories. Dengel et al. (2023, p. 11) found that researchers interviewing LLMs "cannot assume to be engaging in a conversation with an expert in the respective field." Codes or themes produced by LLM tools can promote biased and harmful stances (Aubin Le Quéré et al., 2024; Kapania et al., 2024). Sattele and Ortiz (2024) found AI-generated personas and their associated AI-generated images to contain false or unspecified information with gender-, race-, and age-related stereotypes. Relying on personas may result in designers making decisions based on irrelevant, inaccurate, or biased information that could cause or exacerbate harm. Work from Nighojkar et al. (2025) and J. S. Park et al. (2024) suggests that these limitations may be mediated by using data curated with intentional research methods.

These limitations underscore the need for careful validation, transparency, and critical reflection when using LLMs as generative tools in qualitative research. Beyond these caveats of research quality, there also exist concerns for the broad resource usage of data centers that enable such technologies (Jegham et al., 2025). While outside the scope of this article, a holistic evaluation of factors, which include environmental impact, energy consumption, and sustainability, must be integrated into discussions around the ethical deployment and scalability of LLMs. Addressing these broader considerations is essential to ensure that advances in AI technology are balanced with responsible stewardship of global resources and long-term societal well-being.

4 | HARMS IN ENGINEERING MAY BE PERPETUATED BY LLMs

4.1 | Existing norms and bias in engineering education

Hegemonic norms, specifically reflecting White and male ideals, are present in higher education at large, and specifically in engineering education (Cech, 2022; Cech & Waidzunas, 2011; Forester et al., 2024; Secules, 2019). These norms can create stressful and unwelcoming environments for all individuals, but particularly for individuals of color, LGBTQ + individuals, women, and people with disabilities who do not abide by the hegemonic standards (Akpanudo et al., 2017; McGee, 2020, Acharya et al., 2018). Engineering has been identified as a culture of stress (Jensen & Cross, 2021) and as an environment of "suffering and shared hardship" (Godfrey & Parker, 2010, p. 12) where someone must put their "nose down to the grindstone and work harder" (Sanders et al., 2024b, p. 201) to be successful. This requirement for hardship likely compounds with identity-related stressors experienced by people with historically marginalized identities, who experience added stressors due to underrepresentation and microaggressions (Cross et al., 2021; Negi et al., 2019). Studies focused on engineering students specifically have found high rates of stress and mental health disorders (Danowitz & Beddoes, 2018; Jensen & Cross, 2021). Other studies have found engineering students to report lower measures of flourishing (Healthy Minds Network, 2023; Healthy Minds Network, 2024) and help-seeking (Armstrong et al., 2022; Miller et al., 2022) when compared to other university students.

4.2 | LLMs can reproduce harmful aspects of engineering culture

As mentioned in previous sections, LLMs reinforce hegemony and Western values (Bender et al., 2021; Muldoon & Wu, 2023). Research has exemplified this hegemony from engaging with biased LLMs (Bano et al., 2025) as reinforcing important decisions such as hiring decisions (Nakano et al., 2024) and influencing career exploration (Due et al., 2024).

Given the existing norms and biases toward a culture of stress in engineering education, it may be possible that engaging with AI use in engineering education and EER may further reinforce these problematic cultural narratives. LLMs regenerating biases or ideals that reflect certain norms not only limit the knowledge obtained from conducting research but also can reinforce institutional pressures and environments that undervalue and harm marginalized populations within engineering.

5 | THEORETICAL FRAMEWORK: IDEALIZED WORKER

Acker's (1990) conceptualization of gendered organizations posits that organizational structures are not gender-neutral but are instead undergirded by assumptions aligned with dominant forms of masculinity. These assumptions are present in organizational structures, including job expectations and daily practices (e.g., always available and unencumbered by caregiving responsibilities). Subsequent scholars have applied this theory of gendered organizations to various contexts, including gender segregation within government jobs (Björk & Härenstam, 2016; Shelley et al., 2011), early childhood education (Sargent, 2005), and clinical research trials (Cottingham & Fisher, 2022).

Acker's work was later examined through the lens of Pawley's (2019) and Smith's (2005) conceptualizations, which describe organizations as institutions that structurally enforce conformity to the idealized worker. This "idealized worker" is someone who embodies many social privileges (e.g., a non-first-generation, cisgender, heterosexual, White man from a socioeconomically advantaged background) (Pawley, 2019; Smith, 2005). Sallee (2012) used the theory of gendered organizations and idealized workers to explore how male faculty as fathers balance work and home responsibilities and the structures and policies that influence how fathers divide their time between home and work.

The relevance of this framework to the present study lies in its ability to reveal how pressures to conform to an idealized worker model systematically marginalize those with different lived experiences. The idealized worker experiences the least pressure to conform since they have the experiences that the institution is designed to support. These pressures to work, think, and behave in certain ways do not consider how people think and work differently. Examples of this include working overtime and being rewarded and praised as well as being "on-call" for emergencies or perceived as always available (Gray et al., 2019; Sallee, 2012). These assumptions can disproportionately impact individuals who do not fit the ideal worker mold, such as those with caregiving responsibilities or disabilities. To understand the pressures to conform, we must understand both the idealized worker and the structural forces that reproduce this pressure to conform. For example, structural forces such as mandatory in-person working hours reproduce the idea that workers can give all of their time to work within certain hours of the day and have no other responsibilities, such as caring for children or family members, that may affect working hours. Examining institutional pressures that marginalized individuals experience leads to a deeper understanding of the ways in which cultures, such as engineering stress culture (Jensen & Cross, 2021), are developed and sustained.

This study draws on the concept of the idealized worker to examine whether LLMs, when prompted to generate qualitative data about engineering education, tend to reproduce narrow and exclusionary norms and representations of who belongs in engineering. If AI is able to meaningfully reproduce experiences of marginalized individuals in engineering, then it may meaningfully provide assistance with describing barriers to equity. However, if AI tends to reproduce narrow, idealized representations of who belongs in engineering, then using AI without critical reflection may unintentionally validate and perpetuate the cultural norms that contribute to the marginalization of those who do not fit that ideal. By applying Acker's framework to LLM-generated content, this study critically examines the extent to which AI models replicate institutionalized norms of belonging and excellence in engineering. This theoretical lens foregrounds the importance of identifying not only what narratives are reproduced by LLMs but also whose experiences are omitted or flattened in the process, raising broader questions about the risks and responsibilities of using AI in qualitative educational research.

6 | RESEARCH QUESTIONS

In this study, we compare conversational data generated with a widely available LLM (ChatGPT) with multi-institutional qualitative interview data collected from students and faculty in engineering education. Guided by the theoretical lens of the idealized worker and a critical orientation toward AI in qualitative research, we explored the following research questions:

RQ1. *What are the affordances and limitations of using LLMs to generate qualitative data in engineering education, based on similarity–difference comparisons with semi-structured human interview data?*

RQ2. *In what ways might LLM-generated qualitative data reproduce and reinforce dominant cultural narratives in engineering education, including those related to the idealized worker and the dominant narrative of high stress in engineering?*

7 | POSITIONALITY

Our research team's interest in this topic originated from a concern about the potential to reinforce high-stress cultures within educational settings. This is informed by our diverse experiences across research, instruction, and advocacy work. Several team members are attuned to the normalization of high stress within engineering and are concerned about AI's potential biases, particularly regarding marginalized identities. One team member brings specific expertise and conducts research on AI, enhancing our understanding of the technology's ethical and technical complexities; moreover, all members have had prior interactions with various AI models and feel keenly aware of their growing presence within society at large. We approached this work with varied levels of familiarity and skepticism toward the utilization of AI, shaped by experiences such as prompt engineering, instructional applications, and examining bias in medical AI—all of which reinforce our critical perspective on interpreting AI outputs. Our prior experience in qualitative and quantitative research supports us in balancing methodological rigor with caution around AI's limitations. Positional differences within our team allow us to recognize the unique dynamics between students and faculty. We are mindful of how our identities, especially those associated with privilege, shape our communication and disclosure choices, aiming to center the well-being of marginalized groups. Through this positionality, we acknowledge how our backgrounds shape our perspectives, which thereby enables us to engage more ethically and critically with our research on AI in educational contexts (Secules et al., 2021).

8 | METHODS: OVERVIEW

We used a structured approach to compare two sets of existing qualitative interview datasets with conversational text generated from ChatGPT's 4o LLM. The interview datasets included 38 interviews with engineering students and engineering faculty in the United States around topics of stress and wellness. After conducting the initial phase of analysis, we conducted a second phase of analysis using new data generated from ChatGPT's 4o LLM. In summary, we conducted two phases of analysis: (i) a comparison between AI-generated and human-generated interview transcripts, and (ii) a content analysis of AI-generated data across varied persona demographics. We first present the Phase 1 methods and results followed by the methods and results of Phase 2. This research was approved by the focal institution's Institutional Review Board (HUM00254875 for the University of Michigan).

9 | METHODS: PHASE 1

The first analysis was a qualitative comparison between the AI-generated data and the human-generated interview transcripts to identify the remaining differences between these two data sources. Uncovering the differences between the AI- and human-generated data serves to highlight the affordances and limitations of using LLMs to generate qualitative data in engineering education.

9.1 | Participants and interview data

The data used for comparison with AI-generated data were transcripts from interviews with engineering undergraduate students ($n=14$) and with engineering faculty ($n=24$) in the United States. We conducted the interviews between 2021 and 2023. The original interview protocols used a semi-structured format and discussed topics centering on wellness, coping, mental health supports, and stress in engineering. Questions from each protocol are presented in Appendix A.

Participants included faculty from 17 institutions, and the undergraduate engineering students all were enrolled at one institution. Participants were not intentionally stratified by demography and included people who identified as White, Black, Asian, Middle Eastern or North African, and from India or the Indian Subcontinent. Participants were nearly half men and half women. More information about these participants and studies is available in the related faculty (Jensen, Sanders, et al., 2023; Johnson et al., 2024; Sanders et al., 2024a, 2024b) and student (Sanders et al., 2025) study publications.

9.2 | AI data generation

We used the most advanced model of the most commonly used text AI-generation service available at the time of this study, namely ChatGPT's 4o model web interface (July–October 2024), to produce all AI-generated data for this study. ChatGPT's 4o model used training data up to October 2023 (OpenAI Platform, 2024b). This model was chosen to allow for the most transferability of our findings to other researchers—although findings may also transfer to other LLMs—and to ensure the results reflect the current state of the art in accessible, high-performance language models. While the model does maintain multimodal capabilities, GPT-4o matches or exceeds other accessible options (e.g., GPT-3, GPT-4) in core text-based benchmarks, offers greater speed and cost efficiency, and generates more conversationally toned outputs (OpenAI Platform, 2024a; Shahriar et al., 2024). These features made it particularly well suited for scalable, text-focused research workflows while maintaining relevance across a range of academic and applied contexts.

Our goal was to engineer and identify a prompt which uniformly generated data that was similar to the aggregate human responses. By "aggregate," we mean the combined set of responses from all participants for each question, embodying representative patterns. We then applied this prompt to the interview questions and compared the generated output to the interview responses. When designing the prompt, we asked the AI to generate a summary of possible responses for this analysis for convenience.

We used an iterative approach to develop and refine the prompt used to generate the AI data, considering an array of prompting strategies highlighted by Sahoo et al. (2024). This entailed initially prompting the AI to answer questions from the semi-structured interview protocols we used for the student and faculty interviews (refer to Appendix A). We then compared the AI-generated answers to the interview data and identified differences between the AI-generated output and the human-generated interview transcript data. For example, in an early iteration of the prompt, the AI response resembled the entirety of the remainder of an imagined interview. The AI response included the answer to our question, and then it continued to also include several other similar interview questions and their responses. We were only interested in the response to the question asked and so tested various prompt phrasings until the AI-generated response contained only answers to the intended question. Based on these observed differences between the AI-generated data and our human-generated interview data, we varied the prompt structure to more closely align with the structure of the responses in the human-generated data. We continued to iterate in this manner until we found the following prompt-generated data that was closest to our interview data.

> Please respond to this as an engineering faculty member in the US. The faculty is in the middle of an interview. Only give the faculty's response to this question. – 1. [Question].
> Please generate a list of the answers that you think 4 different (2 women and 2 men) engineering faculty members in the US might give.
> Then, after generating the 4 responses. Create a bulleted list summary of the key points people shared

Keeping all components of the prompt as a single prompt (e.g., entering all of the information in the textbox and then sending it to ChatGPT) instead of separating it into consecutive prompts provided the best application of data within the prompt. For example, if we included demographic information about the persona in prompt 1 and then asked the AI to answer a question in prompt 2, we observed that the second generated answer would include much less demographic-specific information in the persona response than if everything was combined into one prompt.

The following points describe each aspect of the prompt, a description of the prompt, and our reasoning:

- "engineering faculty member in the US": This is the persona descriptor applied for "who" is answering the "interview question." We varied this according to which persona we were interested in answering the question. So, when generating data to compare with the student interview, we instead used "engineering undergraduate student in the US."
- "The faculty is in the middle of an interview.": This prompted the AI to generate responses that were more often in paragraph form with full sentences rather than a bulleted list. This also encouraged the AI to share generated stories if asked later. For example, the faculty-protocol question "Have you had an interaction with an undergraduate engineering student regarding any mental health issues?" resulted in short example stories of when the imagined faculty members had interacted with students (example included in Appendix B).
- "Only give the faculty's response to this question.": If not specified, the AI would sometimes generate the rest of the interview, including the questions the interviewer would ask. Since we did not want this, we included this phrase to shorten the AI response to only the answer to the question asked.
- "1.": We found that numbering the question response (e.g., instead of formatting such as quotation marks) was most effective in generating the expected information. This preference for numbers indicates that the data used to train the AI may have contained numbered questions instead of questions demarcated by quotation marks (White et al., 2023).
- [Question]: We replaced this with the interview question we wanted answered.
- Including line breaks within a single prompt: We found that separating the requests within a single prompt with a line break helped the AI recognize each request as a separate item. Otherwise, the AI would sometimes skip part of the prompt.
- "Please generate a list of the answers": This instructed the AI to generate multiple responses with a single prompt request. This saved the researcher time and did not seem to impact what responses the AI generated.
- "4 different (2 women and 2 men) engineering": This request varied the answers the AI generated to the prompt based on the demographic information provided. We found this addition to be most effective when there was only a small amount of demographic information provided, or the AI would generate responses that aligned more with stereotypes. Prompting minimal demographic variation—in our case gender—seemed to widen the types of responses the AI generated. Demographic variables that are and have been widely used seemed to prompt less stereotyping in their responses. So, for example, we did not include nonbinary identities in this response, even though we included other gender identities. It seemed that the more marginalized the identity/ies included in the response, the more strongly the AI generated a response that focused on marginalization along these identities (explored more in Phase 2, Sections 11 and 13). We encourage considering an approach of varying demographic identities when generating prompts.
- "engineering faculty members in the US": We repeated this demographic description of engineering faculty, even though it was specified in an earlier part of the prompt. We found that, even though ChatGPT is capable of retaining information from previous prompts (White et al., 2023), it was significantly more effective to include the persona directly next to each question. That is, the AI generation tended to give more descriptive and specific answers, which we desired, if the persona descriptions were repeated.
- "Then, after generating the 4 responses. Create a bulleted list summary of the key points people shared.": This prompts the AI to summarize the generated responses into a bulleted list. We found this list to often be a reasonable summary of the four responses, and we referenced this list when comparing results.

We occasionally explored additional demographic identities in our prompting. These explorations were limited in scope for the project, but we have included a few additional explorations to expand the results. When relevant, this additional prompt modification is described in the results. Although it is possible to have a continuing "conversation" with the AI, we chose to generate a new session for each question we asked because doing so reduced the variability in responses. Additionally, we disabled features within the web interface for ChatGPT that carried information from one prompt session to another, which are otherwise enabled by default. Disabling this feature instructed ChatGPT to "forget" prior instructions and interactions from previous sessions.

9.3 | Comparison and analysis of AI- and human-generated interview data

The first and second author completed independent analyses to categorize the similarity in content of AI- and human-generated responses for each interview question into three categories: (i) similar by more than 90%, denoted "High

Similarity"; (ii) similar by more than 50% but less than 90%, denoted "Moderate Similarity"; and (iii) similar by less than 50%, denoted "Low Similarity." We chose to bin these responses broadly on the basis of similarity (i.e., above 90%, between 50% and 90%, and less than 50%) to increase the likelihood that the responses would remain in the same category when session-to-session variation (i.e., the AI responding slightly differently when asked the same question another time) was considered.

Some human-generated interview responses, particularly in the faculty interviews, elicited very specific stories. If we compared each human response individually to the AI-generated responses, the specific stories would significantly differ because the AI-generated responses tended to be broader. Similar deficiencies in the depth and specificity of AI-generated responses have been found in the context of English literature (Amirjalili et al., 2024) and computational linguistics (Ma et al., 2023). However, when the human-generated responses were taken in aggregate and compared with the AI-generated responses, we sometimes found more similarities to the AI-generated data. There were three questions in the faculty interviews where several respondents answered with "I don't know." When aggregating responses, we binned "I don't know" responses into the "more than 50% but less than 90%" category (i.e., "Moderate Similarity") because this uncertainty mimicked the generalized and non-committal nature of the AI-generated responses, thereby aligning more closely with the model's tendency to provide broad, less definitive statements. We have indicated these responses with an asterisk in Table A1, Appendix A.

The first and second author compared the data using slightly different methods, informed by their familiarity with the data. The first author was familiar with all human participant data before this study and had been the lead analyst regarding the prior human-generated data (Jensen, Sanders, et al., 2023; Johnson et al., 2024; Sanders et al., 2024a, 2024b, 2025). Since the first author was familiar with the human interview data, she compared AI-generated output to the aggregate responses of the individual questions using post-thematic analysis spreadsheets (Braun & Clarke, 2006) and resulting publications. The second author was new to the human-generated data, and he compared the AI responses to each human participant's response. He primarily analyzed the transcripts (for faculty) and available summary spreadsheet (for students) for the data. The second author qualitatively categorized the agreement between each AI and individual human responses as high, low, or moderate. Then, based on the frequency of high, moderate, or low agreement per question, the second author categorized the response for the relevant question into the three bins, demarcated by 90% and 50% similarity as described previously. The two researchers then discussed and reached a consensus on the analysis (Campbell et al., 2013). This process is outlined in Figure 2.

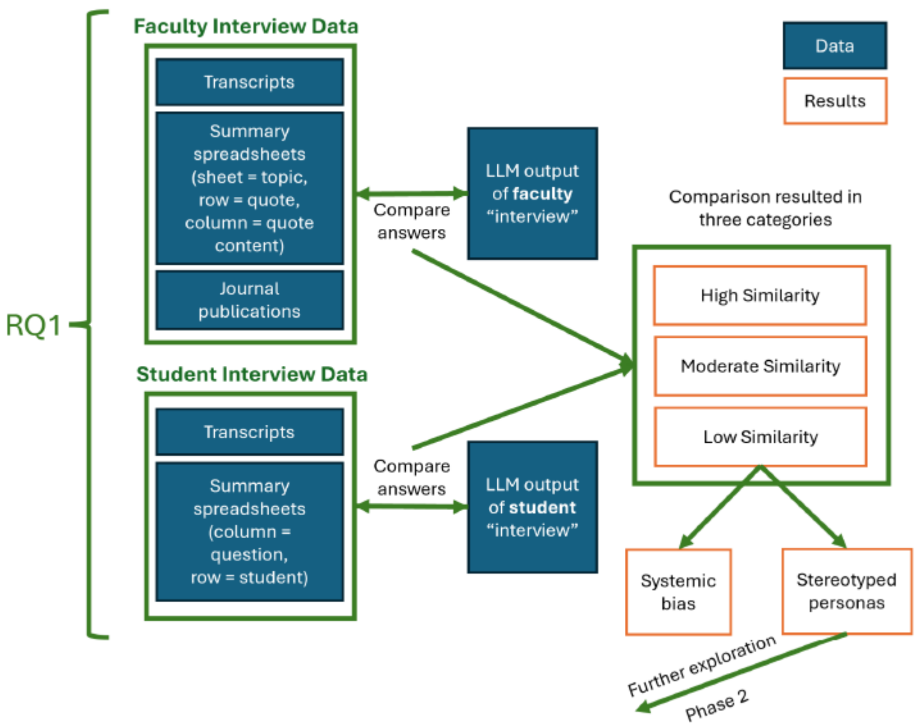

FIGURE 2 Methods summary for Phase 1.

10 | RESULTS: PHASE 1

When we compared the AI-generated data to the human-generated student and faculty interview data, we found that most responses contained similar information. Of the 30 questions asked to students and faculty, seven questions contained very similar information. Twenty question comparisons contained similar information, and three contained different information. We describe "High Similarity" information as information that overlapped in content or meaning by approximately 90% or more. "Moderate Similarity" questions overlapped by at least 50%, while "Low Similarity" responses contained less than 50% similarity. In the less than Moderate Similarity output, we identified multiple instances of the AI-generated responses stereotyping the output, which is further explored in Sections 11 and 13 (Phase 2). A summary of these Phase 1 results is presented in Table 1.

10.1 | High similarity

The seven very high similarity questions involved requesting definitions (How do you define stress? What does wellness mean to you? Is stress different from anxiety? Depression?) or common experiences that are experienced by many people (How do you notice when you are stressed? Who are the people you turn to for support when you're feeling stressed? When you feel these stressors, how does that relate or not relate to feeling in control? Do you think faculty members, like yourself, have a responsibility to intervene when you suspect a student is struggling with mental health or wellness?).

For example, when answering the question "How do you define stress?," half of the students described feeling overwhelmed, with several also describing feeling pressure, physical symptoms (e.g., a weight on their shoulder, migraine, sweaty palms), and feeling anxious. The AI-generated responses similarly described feeling overwhelmed, pressure and tension, prioritization changes, physical and mental effects, and emotional impact such as worry.

Similarly, when we asked students who they turn to for support when they are feeling stressed, most described turning to friends in their major for class-related support, friends outside of their major who lived close by for emotional support, and family. Students also described partners, faculty and staff mentors, affinity groups, and other friends as people they turn to for support. The AI-generated responses also highlighted these stakeholders, although the AI response de-emphasized friends who lived close by and affinity groups and explicitly mentioned campus counseling centers as occasional sources of support. Although counseling centers were not present in student answers to this question, several students described campus counseling centers elsewhere in their interview with both positive and negative experiences.

The high degree of similarity between the AI- and human-generated responses to these definitional and common experience questions suggests that AI may be a useful additional tool in protocol design. Researchers may not be able to guarantee a priori which answers will be most similar to participants'. However, researchers may use AI output to check for semi-structured and structured protocol flow, in addition to the regular methods of pilot testing. Researchers may use AI to increase protocol flow by arranging questions so that the previous question's response relates to the next question in the protocol. If using AI in this way, we recommend remaining critical of the tendency of AI to

TABLE 1 Summary of results for Phase 1.

Comparison category	Description	Example question
High similarity	Frequently similar response if question asked for a definition	In your opinion, is stress different from anxiety?
Moderate similarity	More response variation in participant response than LLM, based on participant lived experience	Have you ever noticed undergraduate engineering students struggling with mental health issues?
Low similarity	LLM tended to describe systemic change over participant change	If you had a magic wand and could change any of these situations that cause stress, what would you do?
Low similarity	LLM tended to stereotype response more often than participant	How are you similar or different from this "typical" engineering student?

stereotype and thus considering methods to ensure the protocol flow allows flexibility for less common experiences to be recorded during the interview.

10.2 | Moderate similarity

Most responses (20 of 30 protocol questions) contained information that was similar in content (Moderate Similarity) but not as aligned as the High Similarity category. These questions tended to request the participant to describe their opinions (What typically causes..., How do you think stressors and supports will change or stay the same..., What comes to mind when...) and their own experiences (Have you had an interaction with..., Have you noticed undergraduate engineering students..., Can you describe things you have learned...). Two primary underlying patterns emerged from these responses: (i) Participant responses tended to vary more than AI-generated responses. This is contrasted with later Section 10.3 (Low Similarity: Systemic bias in AI data) where AI-generated data included significant variation not included in participant responses. (ii) The AI-generated data did not default to including as wide a variety of barriers to supporting mental health as the participants identified in their interviews.

Participants' lived experience often seemed to impact their responses to the interview questions in ways that the AI-generated did not default to considering. A clear example of this participant-response variation was present with the faculty-protocol question “Have you ever noticed undergraduate engineering students struggling with mental health issues? What does that look like to you?” The AI responses included recognizing exhaustion, withdrawal, and perfectionism, and when recognizing this, trying to create a safe space, flexible deadlines, and directing students to mental health resources on campus. Faculty participants generally responded similarly, but some responses fell outside of this range. For example, Max described feeling skepticism when students shared their mental health struggles, sharing that they were unsure “whether they are telling me [about their] struggle, or they're trying to exploit the situation to get more time.” This skepticism was not present in the AI-generated responses. On the other end of this example spectrum, Audrey described a specific experience supporting a student who was severely struggling with their mental health. Audrey's engagement helped the student access mental health supports that were beyond what their university's counseling center was capable of providing. Similarly, in the student interviews, student struggles, for example with finances, were present in the interviews but not in the AI-generated responses.

AI-generated responses also defaulted to incorporating fewer barriers into their response. In addition to financial barriers, student-described barriers included a lack of resources at their counseling center, getting incorrect information from faculty advisors, and a lack of options or awareness of options from their tutoring, financial, and career services. Similarly, some faculty participants described their students' limited access to mental health services or administrative pressures that did not support student mental health (Sanders et al., 2024b). AI did describe barriers from competitive culture and stigma as well as the cost of education and uncertainty of job prospects.

In the analyzed student interview data (Sanders et al., 2025), we found that some participant demographic identities, such as low income, were explicitly described as related to some of the barriers they experienced (e.g., constantly searching for financial aid). Other marginalized identities, such as participant LGBTQ+ identity, were not described as salient to the barriers that participants experienced. This variance in identity-related barriers is interesting because approximately half of the participants identified as LGBTQ+, and LGBTQ+ inequality is present in engineering education (Cech & Rothwell, 2018). With this post-interview knowledge of salient demographic identities, we explored additional AI prompt output. When we prompted the AI to describe barriers for students with specific backgrounds, the AI-generated content included information that was missing from prior generic “engineering student” stresses. For example, when prompted to describe barriers for low-income engineering undergraduates in the United States, the AI-generated response included students “may feel overwhelmed by the process of applying for aid or discouraged by bureaucracy,” which was one of the missing components from earlier prompts. We found that the AI did not by default “assume” that a student may be from a low-income background, which may indicate idealized worker assumptions of who students are, where they come from, and what they need. We explored implications from this result in further detail in Sections 11 and 13 (Phase 2) of the analysis.

These results suggest an affordance of using AI to test an interview protocol and draft ideas for how participants may answer. A limitation of this process is that the AI-generated responses may not vary as much as participant responses, and the AI-generated responses may not recognize barriers that are present in participants' experiences. Additional barriers participants experience may be explored with AI by varying participant demographic identities.

However, this may not replicate participant responses and may lead to stereotyping participant responses. This stereotpying is further explored in Section 10.3 (Low Similarity: Stereotyped Persona Descriptions) and Sections 11 and 13 (Phase 2).

10.3 | Low similarity: Systemic bias in AI data

While most AI-generated responses were generally similar to participants' aggregate responses, this similarity did not hold for three questions. These three questions are discussed in more detail in this and the following sections. The first question asked students to describe an ideal future scenario ("If you had a magic wand and could change any of these situations that cause stress, what would you do? How would you change them?"). This question occurred toward the end of the interview, after students had discussed several situations when they felt stress as well as who they turned to for support. We found that AI-generated responses tended to be more aware of systems of power than some student responses. The AI-generated responses often described strategies and solutions that entailed larger systems changes, for example, changing to growth-focused grading and implementing career pathway programs. In contrast, students who participated in the original interviews were more likely to internalize problems as stemming from their own abilities.

This question was intended to creatively engage the student participants with envisioning a different future. Student responses fell into three categories, with approximately one-third to one-half of students falling into each category.

The first category followed a pattern of: I would change "x," but that would fundamentally change "y." The most common example of this described the student making school easier in a capacity, but then expressing a concern that they would not learn the content. For example, Sumaya shared, "I'm tempted to say I would get rid of the lab report... but, ... I know ultimately the professors are trying to help us learn, help us become better academics." This pattern of change addressed the immediate stressors the students observed, but they did not creatively replace elements of their system to ultimately reach the same goal.

The second category of change addressed something within the students themselves. For example, increasing their own management skills, increasing their ability to understand concepts more quickly, or knowing what will happen in the future. Bridie shared, "Realistically, I'd just be able to understand concepts and like study easier." Similarly, Lois shared, "I would just poof up skills, stress management skills in me, I guess, because is there really a world without stress? No, there isn't." These responses were grounded in changing the students' own abilities. While learning skills can often be challenging, it can still seem more attainable at times than changing larger systems such as university policy.

The third category of how students answered centered more on sources outside of themselves, some of which were systems-oriented. These commonly included changes to teacher actions such as curving lab assignment grades or "finding teachers who... very much love what they're teaching." Systems changes included students having more money, departments advocating more for their students, and shorter wait times for on-campus mental health counselors.

AI-generated responses exclusively described responses that were closer to changing norms or systems. These suggested changes included (i) spreading course deadlines more evenly to reduce student burden at a specific time; (ii) increasing communication between departments to improve coordination and scheduling; (iii) increasing transparency in project requirements and grading; and (iv) implementing "better systems" to divide group project work and assess contributions. When prompted for details, the AI provided examples such as centralized scheduling between department classes, a department-wide calendar system for entering major deadlines, or using specific peer-evaluation software. These responses are outside of the scope of what students suggested in their responses. These responses also did not include a description of likely limitations, such as available funding, to implementing these changes.

Key takeaways from these results are that participants tended to most often focus on changes that were within their control, although systems-level changes are often needed to enact lasting change (Braithwaite et al., 2018). This phenomenon has implications for recognizing opportunities for systems change. People within a specific system may be more aware of barriers. They may also not have as much access to tools or knowledge to ideate potential change possibilities. Cross-disciplinary or cross-experiential knowledge may be supportive of change efforts (Loughman et al., 2000). A potential affordance of broad-data AI-generated qualitative data is that it may provide value in brainstorming possibilities, for example for systems change. Since the basis for training the broad-data AI LLM includes a wide range of sources, thoughtfully including critical, change-oriented sources, including AI generation in brainstorming, may provide insights that may not occur to engineering educators.

10.4 | Low similarity: Stereotyped persona descriptions

The second two questions that resulted in AI-generated output that was dissimilar to our data were both questions that indirectly requested that the AI describe an average "other" (e.g., describing a "typical" student, asking how the participant is similar or different to other students). While both questions' output differed from the participant responses, one of the AI responses generated data that strongly described marginalization along a single dimension of identity. We further explored this phenomenon by varying the persona prompts for the AI-generated data (i.e., Phase 2, Sections 11 and 13).

The first of these two questions asked, "How are you similar or different from this 'typical' engineering student?" Student responses varied widely in describing how they were different, for example, describing themselves as a commuter, first generation, being an extrovert, or being artsy. The AI response included similarities to the student responses such as motivated and collaborative. However, the AI responses included describing that the stereotypes of "engineering students being introverted or antisocial" are untrue and that most students are "are active in clubs or research" and that "most people are pretty collaborative." These participant and AI responses indicate that they are both aware of similar stereotypes of engineers. However, the participants described these stereotypes as the norm, whereas the AI responses rejected these stereotypes as untrue. This somewhat contrasts with later results in this section that describe the AI as leaning into identity stereotypes.

The second of these two questions was, "Of the stressors and supports you described, which ones do you think are similar and different to other students in your department?" This question prompted the AI "student" to generalize about which of their experiences were similar and different from other students in the department. This differs from the other responses generated, since this question prompts the AI to broadly generalize rather than generate a specific experience.

When generating a response for this second question, the AI-generated content frequently and commonly focused on the demographic variation we provided as part of the prompt. This focused on a single dimension of the student and significantly simplified and reduced the breadth of the students' answers in a way that was dissimilar to other responses generated using the same prompt structure (e.g., other responses we generated).

Both the women and men persona responses answered this question with content focused on women's lack of belongingness. For example, women personas described "what might be different for me is the added stress of not feeling represented as a woman in engineering" and "I feel like I'm dealing with a lot more imposter syndrome compared to some of my male classmates." Men's personas included "I have more people I can relate to in class and during projects, which I think is a bit different from what some female students might experience." and "It seems like some of my classmates, especially women, don't get as much encouragement in that area, which I think is a major difference." These responses paint women students in engineering as feeling alone and left out of engineering. While lack of belongingness is a phenomenon that occurs in engineering (Benedict et al., 2017; Pearson et al., 2018; Rohde et al., 2019), students have much deeper and more nuanced lives beyond only this.

Over half (8 out of 14) of the students we interviewed identified as women, but none of the students we interviewed described a gender-based lack of belonging as a way they felt different from other students in their department. Some students who described feeling less belonging attributed this to commuting or navigating financial stressors rather than gender. Other students described feeling particularly close with family or less able to rely on family for support. Yet other students described feeling particularly more stressed or less stressed as a way they felt to be more unique. Both men and women interviewed described these sorts of varied experiences, but this variation was not present in the AI-generated response. The gender stereotypes in the AI response were likely reflective of the prompt instructions to provide responses for women and men. To further explore underlying stereotypes present in the model, we varied the persona prompts and explored the output.

11 | METHODS: PHASE 2

We explored the stereotyped persona results from the previous section in more detail to uncover ways that LLM-generated qualitative data might reproduce and reinforce dominant cultural narratives in engineering education, including those related to the idealized worker and the dominant narrative of high stress in engineering. To further explore the reproduction of these dominant cultural narratives, we conducted a content qualitative analysis of 20 AI responses generated for each of three different personas informed by our framework. These three personas (generic, powered, and multiply marginalized) are described in the following section.

11.1 | AI data generation

The question selected to generate these AI data was "Do you think faculty members, like yourself, have a responsibility to intervene when you suspect a student is struggling with mental health or wellness? Why/Why not?" This question was selected for two reasons: (i) The AI-generated response fell in the "High Similarity" category, meaning the AI had responded similarly enough that there was interesting alignment, and so resulting variation was likely from the introduced demographic personas. (ii) This question's response is one that had previously been more deeply and thoroughly analyzed (Sanders et al., 2024b). Since a thorough analysis of this human data was available and had been performed by the first author, using this question strengthened the quality of this analysis:

> Please respond to this as a [persona descriptor] member in the US. The faculty is in the middle of an interview. Only give the faculty's response to this question. – 1. Do you think faculty members, like yourself, have a responsibility to intervene when you suspect a student is struggling with mental health or wellness? Why/Why not?
> Please generate a list of the answers that you think 20 different [persona descriptor] members in the US might give.

The three personas were inserted into the prompt at the [persona descriptor]. We selected these personas based on our theoretical framework of the idealized worker, which posits that people who embody many social privileges experience the least pressure to conform.

- Persona 1: The "default" or "generic" persona, which we described in the prompt as "an engineering faculty." This had no additional persona specifiers, and so this data is an example of the default generated output.
- Persona 2: the "powered" persona, which we described in the prompt as "a white able-bodied straight cisgender engineering man faculty." Since people with these identities often hold many social privileges (Pawley, 2019), including these identities in the comparison is important for understanding the impact of the idealized worker in this context.
- Persona 3: the "multiply marginalized" persona, which we described as "a Black disabled LGBTQ+ woman engineering faculty." Our theoretical framework describes that people with marginalized identities experience a more explicit and identifiable process of enculturation. We use "multiply" instead of "multiple" marginalized identities to indicate the compounding nature of this marginalization. We were interested in whether any of these processes of enculturation may be evident in the AI-generated data.

We generated 20 responses for each persona because we found that the content variation was repeating (i.e., we reached saturation) with 20 responses. Example responses are included in Appendix D.

11.2 | Content analysis of AI-generated data with persona variation

The first author thematically analyzed these data (Braun & Clarke, 2006) using an abductive approach informed by a priori knowledge from previous work (Sanders et al., 2024b) as well as patterns found in the AI-generated data, described in Figure 3. The second author independently coded the responses using the initial codebook generated in

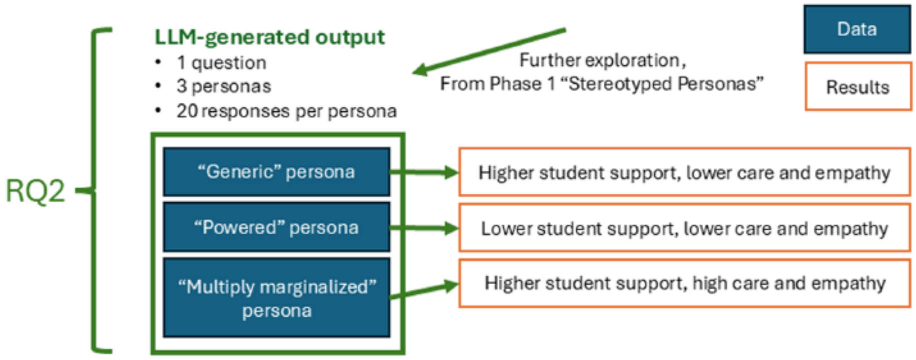

FIGURE 3 Methods summary for Phase 2.

the first author's first pass and wrote memos describing potential codebook changes. The two researchers then discussed their analysis and reached a consensus (Campbell et al., 2013) in the codebook presented in Table C1 of Appendix C.

12 | QUALITY

In considering Walther et al. (2013)'s Qualifying Qualitative Research Quality (Q3) framework, we observed several tensions with key quality considerations. These tensions existed with the AI variability, open-endedness of design possibilities, and rapid change that generative AI development is continuing to undergo. Exemplifying this environment of rapid change, new LLM releases often shift model behaviors in ways that directly affect the reproducibility and stability of findings, underscoring the importance of systematically validating our approach using established frameworks such as the Q3 framework (Reeping et al., 2025). A natural tension existed in both process reliability (e.g., process consistency) and procedural validation (e.g., design choices that result in AI data similar to the human interview data) with regard to the variability present in the AI-generated data. This variability was present in designing the prompts used in Phases 1 and 2 to generate data, because seemingly innocuous changes in the prompt could result in significant changes in the output. We iterated on the design and consulted with field experts on our advisory board for prompt design ideas to mitigate the variability.

13 | RESULTS: PHASE 2

13.1 | Exploring stereotyped persona descriptions: LLM persona comparisons

Results from the content analysis of AI-generated prompts described a trend of multiply marginalized persona output indicating more care and empathy and a powered perspective indicating more uncertainty and discomfort in supporting student mental health. Code frequency counts are presented in Table 2. Each number describes how many of the 20 responses generated for that persona had the relevant code applied. Responses that were the most different from the generic persona are of the most interest, as these are most likely to indicate a consistent trend in the AI-generated output.

Clear patterns emerged displaying the persona effect in the AI-generated data. Overall, the multiply marginalized persona was portrayed as describing increased care and empathy when compared to the generic and powered personas. The powered persona was more similar to the generic persona with regard to expressed care and empathy, but the powered persona described more instances of seeming uncomfortable or unsure if supporting student mental health was part of their role as faculty. All persona responses frequently included directing students to mental health resources, being available for student conversation, and holistically recognizing that students have needs outside of only academic needs.

The multiply marginalized persona described more instances of emphasizing creating a culture of care for students (twice as frequently as the generic persona). The multiply marginalized persona described that students from diverse backgrounds may experience different challenges in 30% of responses. For example, sharing, "In my experience, students from marginalized backgrounds often feel isolated." This awareness of diversity in students' lived experiences was not described at all in either the generic or powered personas.

Conversely, some faculty in our human interviews who held more traditional powered identities described recognizing that students' differing experiences inherently impact them differently. An example of this, faculty Mark shared:

> One thing we are also talking about is [how] mental health issues could look different based on your experiences, both in terms of diversity, equity, and inclusion. So, trying to see how that looks like, for instance, if someone's working two jobs. And, I never had to work two jobs to get myself in college. (Mark, faculty interview)

In this description, Mark, who did not have an experience of working two jobs in college, described an awareness that students' financial needs negatively influencing their mental health. With this example, Mark demonstrates an awareness of others' struggles, despite not having personally experienced those struggles. This awareness exemplifies nuance that was not clearly exemplified in the AI-generated answers.

The powered persona differed from the generic one in that it included responses indicating uncertainty and discomfort in how to support student mental health. As an example persona response, "We should be aware, but we shouldn't intervene directly. It's important to respect their privacy and boundaries." Our faculty interview participants also expressed discomfort, uncertainty, and a desire for increased training, which was a primary finding of our faculty interviews (Sanders et al., 2024b). However, in the human interviews, this discomfort was not siloed to faculty with powered identities. Similar to the example above with an awareness of difference, our participants' experiences and viewpoints were more varied and nuanced than the AI-generated text.

To further explain the phenomena of AI-generated text lacking the breadth and nuance of lived experience, and thus perpetuating stereotypes, we generated Figure 4, which describes the data flow using broad-data LLMs to generate data. The experiences represented on the left include the full qualitative dataset that is collected as part of a study, containing nuance and human complexity that benefit qualitative analysis. The middle portion of the figure represents qualitative thematic analysis. Thematic analysis pulls out only the solid triangle (red) from each experience in this example. These results expand our collective knowledge of the studied phenomena. However, this analysis is often limited by human capacity (i.e., fewer participants or less complex data per participant). The LLM analysis then combines the analyzed information (red triangle) with much more material and creates an output (right portion of the figure) that is similar to the input (left portion of the figure). However, the gradients and variety in shape and shape placement in the original input images are not present in the final output images. We believe this descriptively explains and aligns with findings presented in these results, which are that similar data are present in the output but the nuance remains limited, resulting in data stereotyping. This stereotyping limitation may be due in part to the data available for training the LLM.

TABLE 2 Code counts demonstrating impacts in answers of persona variation.

	Powered	Generic	Multiply marginalized
Culture of care	5	4	9*
Diversity, equity, and inclusion (DEI) experience unique challenges	0	0	6*
Empathy from personal experience	0	0	5*
Holistic	4	6	3
Make myself available	3	6	5
Not a counselor	4	3	1
Uncomfortable	5*	0	0
Not really/unsure	5*	0	0
Yes, but...	6	6	5
Yes strong	9*	14	15

Note: Gray boxes and asterisks indicate responses that were different by more than five from the Generic persona response.

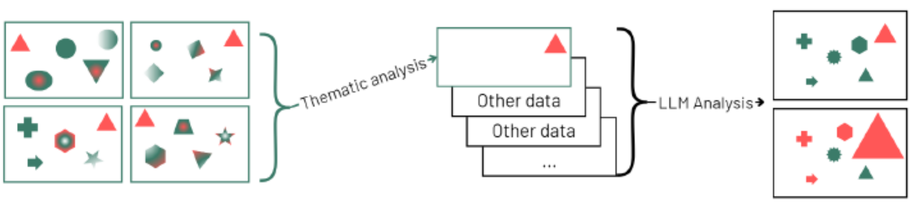

FIGURE 4 Pictorial representation of data flow presented in Figure 1 to demonstrate the loss of nuance present in broad-data large-language model (LLM)-generated text.

14 | DISCUSSION

14.1 | AI affordance: LLMs as a tool for brainstorming

Our results indicate promise when using LLMs as a tool for brainstorming. The AI-generated output may be useful in brainstorming participant responses (High Similarity) during protocol creation. Additionally, AI-generated output sometimes included significantly varied options compared to the participant responses (Low Similarity). This benefit in brainstorming aligns with literature supporting LLM use in creativity generation (Chang & Li, 2024), particularly when occurring after initial drafting (Kosmyna et al., 2025). In a brainstorming space, AI hallucinations are less of a concern, as outputs are being used to generate ideas and explore creative options instead of being used as final pieces of data. For example, a researcher may test the flow of an interview protocol by "interviewing" an LLM, checking that the following question builds off a previous answer and rearranging questions as needed to increase the interview flow. The wide array of options presented by the LLM may also beneficially supplement individual brainstorming efforts for systems change. Our results indicated that some questions (i.e., definitional and common experiences) were similarly described in AI-generated output. Although the underlying biases will continue to be present in the LLM output, a conscientious researcher may mitigate these biases by bringing a critical awareness by supplementing LLM output with additional brainstorming sources (i.e., engaging with colleagues, intentionally reviewing diverse sources) and by varying persona demographics in the AI prompt.

Additionally, students' initial ideas for system changes seemed filtered through their perceived locus of control. That is, they were likely to suggest changes to themselves or perceive that a wider systems change would be ineffective. This may be related to how a person's perception of their locus of control impacts the motivation and perceptions (Spector, 1982). Since the LLM's training data included perspectives outside of only engineering undergraduate students, this filtering effect of a perceived locus of control was not as prevalent in the LLM-generated data, resulting in a broader range of potential systems changes. While the LLM was somewhat likely to reproduce structural mechanisms that enforce enculturation, it also provided value in that it generated ideas that originated from multiple fields or disciplines.

LLMs are still inherently limited in their brainstorming capacity because of less "awareness" of systemic barriers such as university pressure to not report a student of concern (Sanders et al., 2024b) or intersectional barriers that are not as publicly in the zeitgeist. This is supported by literature demonstrating LLM reproductions of biases (Chen et al., 2024; Taubenfeld et al., 2024). Ethically, it is concerning that LLM generation often reproduces biases, and introducing LLM content uncritically in the early brainstorming process is likely to reproduce these biases. As a tool, LLMs can help teams to brainstorm solutions but may not be able to provide specific instructions needed for enacting cultural change.

14.2 | AI limitation: LLMs somewhat reinforce the idealized worker

The AI-generated data was broadly similar to the parallel interview data from engineering students and faculty. It also generally aligned with findings in engineering education literature, such as the following: engineering culture is often considered high stress (Jensen & Cross, 2021); coursework is a primary stressor for undergraduate engineering students (Ban et al., 2022; Jensen, Mirabelli, et al., 2023); and engineering faculty often desire to support their students' mental health (Sanders et al., 2024b).

Although the AI-generated data was similar to the interview data, we found the interview experiences to be more varied and less stereotyping than the AI-generated data. The nuance that human qualitative data provides is valuable for interpreting meaning from data and understanding human complexity (Goyes & Sandberg, 2024). Without this nuance, results likely inequitably privilege some perspectives over others (Dengel et al., 2023; Hämäläinen et al., 2023; Roberts et al., 2024). This impact of the privileged perspective, or the idealized worker, was present in the code count results presented in Table 2. Data demonstrated closer similarity between the institutional pressures to conform and the default response when compared to a multiply marginalized identity. Accordingly, personas that described marginalized identities showed more change from the default response compared to personas that described powered identities. The data corpuses on which LLMs are constructed have been noted to possess their own social biases (Gallegos et al., 2024; Kadan et al., 2022). AI-generated responses for powered identities also failed to show empathy or understanding for people who have different experiences (i.e., Phase 2), even though interview data included participants with similarly powered identities who were more empathetic. Thus, even when intentionally trying to introduce response variation, responses still tended to default to idealized worker norms and failed to provide nuanced perspectives (i.e., Section 10.4). This adds to ethical concerns over using LLM-generated data as similar to real individuals because outputs may only reflect limited perspectives. Our data suggest that it would be unwise to expect broad-data LLMs such as ChatGPT to generate new knowledge of experiences of people with marginalized identities. LLM literature supports this claim by describing inbuilt social identity biases (Hu et al., 2025) as influencing prompt responses (Garg et al., 2022; Huang et al., 2020). Although our framework posits that the process of enculturation is more apparent for people with multiple marginalized identities, this was somewhat limited in AI data. The tax of increased mentorship expectations that occurs for people who have marginalized identities (Trejo, 2020) was present in the data. Owing to the lack of nuance present in the data, we suspect that it may be difficult to engage more deeply with broad-data LLMs to uncover new processes of enculturation that are less understood. Our results demonstrate many drawbacks for using LLMs trained on broad datasets (i.e., reinforcing the idealized worker) to generate data for research analysis, but the generative aspects may provide valuable insights when considered in combination with additional sources, as happens with brainstorming.

15 | IMPLICATIONS

Implications for using LLM-generated data for qualitative purposes apply for many stakeholders, including researchers and policy makers. Most broadly, our results further confirm that broad-data LLMs may also hold promise for brainstorming applications. For example, they may help synthesize information pulled from fields outside of engineering into an engineering context.

15.1 | For researchers

Our methods and results have value to current researchers who may want to use LLMs such as ChatGPT to generate data in their research. If a researcher wants to use these LLMs as a tool for exploring engineering perspectives, they must recognize that the data that was used to train the tool was broad, and our data demonstrate that, at this time, results default to reflecting those of the idealized worker. Future work may consider LLMs that have been trained only with data generated by specific groups as a form of content analysis. Figure 5 demonstrates an example of this potential analysis if it was trained on a specific dataset that uses only relevant data from sources within the subpopulation of interest. This method is not presented in this paper, but it is a potential direction that LLM research could advance.

Researchers should also be aware of the potential ethical concerns when utilizing LLMs, such as representing generated data as human responses and biases within data (and thus responses). Our data demonstrate the use of LLMs for idea exploration and comparison, but do not serve as fully representative responses for actual human participants. Limitations inherent in the data used to train AI will continue to be limitations in the output. Key questions to think about when doing research with LLMs are "Where is the data coming from?" "Who is or is not represented in this data?" "Do the outputs privilege some perspectives over other?" and "Is the LLM reproducing my own bias based on my input prompts?" Considering these questions can benefit researchers by requiring them to consider biases that may be present.

Using intentionally selected data for training a LLM as described in Figure 5 may mitigate some biases present and lead to interesting findings. For example, imagine a study that collects, with informed consent, daily journal entries from engineering students at several universities, including various student identities (e.g., sociodemographic, year in

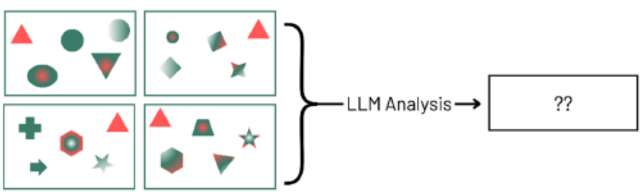

FIGURE 5 Pictorial representation of data flow demonstrating potential alternate data flow where a large-language model (LLM) is trained specifically on a custom dataset of only the individuals from the population of interest.

program, major, etc.). The prompts for the daily journal entries would be guided by the projects' research questions and focus. This imagined study could then use these entries for fine-tuning an LLM (Bergmann, 2024). An interesting insight gained from analyzing the output of this trained data could describe beliefs and biases that the students hold. Another insight could explore the process of enculturation that occurs over time for these students and how this process of enculturation differs by student identity. This example scenario demonstrates the potential of LLMs to generate insights from a large number of participants while maintaining some of the richness present in qualitative data, although many of the underlying biases, as previously described, would likely still be present. Additionally, as data start to become open access, future research into LLMs may find that LLM outputs may be able to become more context-specific.

Beyond methodological guidance, these findings have relevance for theories concerning social identity, stereotype activation, and professional empathy. The differences observed between powered, generic, and multiply marginalized personas suggest that AI-generated responses reproduce societal assumptions about identity and authority, particularly in contexts involving student mental health. Multiply marginalized personas' higher expressions of care and awareness of diverse student challenges highlight how social identity can influence the modeling of empathy in AI outputs. Conversely, powered personas' uncertainty and discomfort indicate that AI may underrepresent nuanced perspectives, offering a simplified view of professional engagement and the limits of AI in capturing human complexity. These limitations have recently been documented in terms of AI-assisted therapy (Carlbring et al., 2023), LLMs' exhibition of in-group favoritism and out-group hostility (Hu et al., 2025), and generative image models that default to narrow cultural norms (Y. S. Park, 2024). Constrained views within LLMs, therefore, risk reinforcing existing hierarchies rather than challenging them. As emphasized by Timmons et al. (2023), such risks underscore that AI systems designed for clinical or educational contexts can hinder equity advancement in mental health work if they replicate biased historical patterns rather than addressing the structural inequities embedded within them.

15.2 | For policy makers

Using generated data to influence policy decisions without including perspectives from real people is risky. LLMs can provide a starting point for ideation for systemic changes, but should not be the sole determinant. Change and progress are best sustained when individuals throughout the system can come together and make decisions that reflect and align with their values (Makenzie & Wehner, 2024; Metz et al., 2023). Prompt engineering can impose set values onto the LLM, but these values are coming from the inherently biased perspective of the user. This introduces some ethical concerns over using LLMs, which often produce generalized and normative outputs, to craft policies and make decisions that affect diverse groups. Policy makers using LLMs should reflect on their own positionality and biases while interpreting outputs, and may want to ask similar questions as listed above for researchers. For example, how might the implications of using AI in academic contexts reinforce the idealized worker? Consider the broad range of potential scenarios where AI may be used, such as creating instruction (Martín Núñez & Diaz Lantada, 2020), grading (Liu et al., 2025), or writing articles about engineering or students. This AI use, particularly when used uncritically, has the potential to reinforce harmful messaging throughout academia.

15.3 | Limitations and future work

Because of the high variation in prompt wording possibilities, it was impossible for us to test every single combination. Additionally, we only specified three personas out of many options for varying privileged and marginalized identities. Results will vary by changing prompts or personas and thus we included our prompts and justifications so others may attempt to replicate our process. Collecting interview data and generating data with LLMs occurred at different times. There could have been unique or influential events that occurred after interviews that would not be reflected in our interview data but may have been included in the LLM's data repository. The culture at specific institutions and in specific departments may differ. We did not include unique descriptions of the institution or departmental culture in our prompts, but we would expect adding these descriptions to affect response outputs. Future work may benefit from training LLM on data from within engineering culture and possibly specific institutions or departments. These results can then be compared to responses generated from an untrained LLM.

16 | CONCLUSION

The data presented here demonstrate the benefits and drawbacks of using LLMs such as ChatGPT to generate data for qualitative analysis. LLM content was mostly similar to interviews, although there was an increased likelihood to describe systemic factors, and persona applications tended to generate output that was less nuanced and less varied than interview responses. The LLM content somewhat reinforced descriptions of the idealized worker; however, LLMs may provide a valuable tool for brainstorming if used critically. Examining engineering culture with LLMs trained specifically on content generated only in engineering contexts holds promise.

ACKNOWLEDGMENTS

This material is based upon work primarily supported by the US National Science Foundation (NSF) under NSF Award Number 2342384. Any opinions, findings and conclusions, or recommendations expressed in this material are ours and do not necessarily reflect those of the NSF. There are also several individuals whom we would like to thank: all advisory board members for their consultation and advice, including Julie P. Martin for her review, insights on analysis, and support with paper revisions; and David Joyner for his perspective on LLMs. We would also like to thank Jo Walther for his contributions in assisting in the understanding of the social reality under investigation. We also thank our participants for sharing their perspectives with us. Without our participants, none of this would be possible. We used AI to generate the "AI-generated data" as described in the Methods sections (Section 9 and Section 11). We used AI grammar checkers as copyediting tools. We used generative AI to identify and link to potential reading material about generative AI, and then we read some of that literature at the linked locations. When revising the Introduction (Section 1), and Literature Review sections (Section 3 and Section 4), we used generative AI to brainstorm possible alternate language and logic. We used partial sentence and flow suggestions as part of the rewrite.

ORCID

Jeanne Sanders https://orcid.org/0000-0002-8865-5444
John Mobley IV https://orcid.org/0000-0003-0828-3896
Isabel Miller https://orcid.org/0000-0002-9774-5812
Nicola W. Sochacka https://orcid.org/0000-0002-9731-6911
Paul A. Jensen https://orcid.org/0000-0002-1257-9836
Karin J. Jensen https://orcid.org/0000-0001-9456-5042

REFERENCES

Acharya, L., Jin, L., & Collins, W. (2018). College life is stressful today – Emerging stressors and depressive symptoms in college students. *Journal of American College Health*, *66*(7), 655–664. https://doi.org/10.1080/07448481.2018.1451869

Acker, J. (1990). Hierarchies, jobs, bodies: A theory of gendered organizations. *Gender & Society*, *4*(2), 139–158. https://doi.org/10.1177/089124390004002002

Akpanudo, U. M., Huff, J. L., Williams, J. K., & Godwin, A. (2017). *Hidden in plain sight: Masculine social norms in engineering education*. Paper presented at the 2017 IEEE Frontiers in Education Conference (FIE), Indianapolis, IN, USA. https://doi.org/10.1109/FIE.2017.8190515

Amirjalili, F., Neysani, M., & Nikbakht, A. (2024). Exploring the boundaries of authorship: A comparative analysis of AI-generated text and human academic writing in English literature. *Frontiers in Education*, *9*, 1347421. https://doi.org/10.3389/feduc.2024.1347421

Armstrong, I., Dailey, D., Wilder, K., Blackburn-Lynch, W., Hammer, J. H., & Wilson, S. A. (2022). *Investigating mental health distress and help-seeking perceptions in first-year engineering students*. Paper presented at the ASEE Annual Conference and Exposition, Minneapolis, MN. https://doi.org/10.18260/1-2-40784

Asghar, M., & Minichiello, A. (2023). Two sides to every psyche: Implications of positive psychology for "mental health" research in engineering education. *Journal of Engineering Education*, *112*(1), 12–17. https://doi.org/10.1002/jee.20493

Aubin Le Quéré, M., Schroeder, H., Randazzo, C., Gao, J., Epstein, Z., Perrault, S. T., Mimno, D., Barkhuus, L., & Li, H. (2024). *LLMs as research tools: Applications and evaluations in HCI data work*. Paper presented at the CHI Conference on Human Factors in Computing Systems (extended abstracts), Honolulu, HI, USA. https://doi.org/10.1145/3613905.3636301

Baack, S., & Mozilla Insights. (2024). *Training data for the price of a sandwich: Common crawl's impact on generative AI*. Mozilla Foundation. https://www.mozillafoundation.org/en/research/library/generative-ai-training-data/common-crawl/

Ban, N., Shannon, H., Wright, C. J., Miller, M. E., Hargis, L. E., Usher, E. L., & Wilson, S. A. (2022). *Identifying common perceived stressors and stress-relief strategies among undergraduate engineering students*. Paper presented at the 2022 ASEE Annual Conference and Exposition, Minneapolis, MN. https://doi.org/10.18260/1-2–41009

Bano, M., Gunatilake, H., & Hoda, R. (2025). *What does a software engineer look like? Exploring societal stereotypes in LLMs*. https://arxiv.org/abs/2501.03569

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

Benedict, B., Verdín, D., Godwin, A., & Milton, T. (2017). *Social and latent identities that contribute to diverse students' belongingness in engineering*. Paper presented at the 2017 IEEE Frontiers in Education Conference (FIE), Indianapolis, IN, USA, 2017, (pp. 1–5). https://doi.org/10.1109/FIE.2017.8190644

Berdanier, C. G. P., Whitehair, C., Kirn, A., & Satterfield, D. (2020). Analysis of social media forums to elicit narratives of graduate engineering student attrition. *Journal of Engineering Education*, *109*(1), 125–147. https://doi.org/10.1002/jee.20299

Bergmann, D. (2024). *What is fine-tuning?* IBM. https://www.ibm.com/think/topics/fine-tuning

Björk, L., & Härenstam, A. (2016). Differences in organizational preconditions for managers in genderized municipal services. *Scandinavian Journal of Management*, *32*(4), 209–219. https://doi.org/10.1016/j.scaman.2016.09.002

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 29), 4356–4364. Curran Associates, Inc.

Braithwaite, J., Churruca, K., Long, J. C., Ellis, L. A., & Herkes, J. (2018). When complexity science meets implementation science: A theoretical and empirical analysis of systems change. *BMC Medicine*, *16*(1), 1–14. https://doi.org/10.1186/s12916-018-1057-z

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

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners (version 4). arXiv. https://doi.org/10.48550/ARXIV.2005.14165

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, *42*(3), 294–320. https://doi.org/10.1177/0049124113500475

Carlbring, P., Hadjistavropoulos, H., Kleiboer, A., & Andersson, G. (2023). A new era in internet interventions: The advent of Chat-GPT and AI-assisted therapist guidance. *Internet Interventions*, *32*, 100621. https://doi.org/10.1016/j.invent.2023.100621

Cech, E., & Rothwell, W. R. (2018). LGBTQ inequality in engineering education. *Journal of Engineering Education*, *107*(4), 583–610. https://doi.org/10.1002/jee.20239

Cech, E. A., & Waidzunas, T. J. (2011). Navigating the heteronormativity of engineering: The experiences of lesbian, gay, and bisexual students. *Engineering Studies*, *3*(1), 1–24. https://doi.org/10.1080/19378629.2010.545065

Cech, E. A. (2022). The intersectional privilege of white able-bodied heterosexual men in STEM. *Science Advances*, *8*(24). https://doi.org/10.1126/sciadv.abo1558

Chang, H. F., & Li, T. (2024). A framework for collaborating a large language model tool in brainstorming for triggering creative thoughts. arXiv. https://doi.org/10.48550/arXiv.2410.11877

Chen, N., Liu, J., Dong, X., Liu, Q., Sakai, T., & Wu, X. M. (2024). AI can be cognitively biased: An exploratory study on threshold priming in LLM-based batch relevance assessment. arXiv preprint arXiv:2409.16022.

Cottingham, M. D., & Fisher, J. A. (2022). Gendered logics of biomedical research: Women in U.S. phase I clinical trials. *Social Problems*, *69*(2), 492–509. https://doi.org/10.1093/socpro/spaa035

Cross, K. J., Mendenhall, R., Clancy, K. B., Imoukhuede, P., & Amos, J. (2021). The pieces of me: The double bind of race and gender in engineering. *Journal of Women and Minorities in Science and Engineering*, *27*(3), 79–105. https://doi.org/10.1615/JWomenMinorScienEng.2021034902

Danowitz, A., & Beddoes, K. (2018). *Characterizing mental health and wellness in students across engineering disciplines*. Paper presented at the 2018 The Collaborative Network for Engineering and Computing Diversity Conference Proceedings, Crystal City, Virginia. https://doi.org/10.18260/1-2–29522

Dengel, A., Gehrlein, R., Fernes, D., Görlich, S., Maurer, J., Pham, H. H., Großmann, G., & Eisermann, N. D. G. (2023). Qualitative research methods for large language models: Conducting semi-structured interviews with ChatGPT and BARD on computer science education. *Informatics*, *10*(4), 78. https://doi.org/10.3390/informatics10040078

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., &, Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. arXiv [Cs.CL]. http://arxiv.org/abs/2104.08758

Drinkwater Gregg, K., Ryan, O., Katz, A., Huerta, M., & Sajadi, S. (2025). Expanding possibilities for generative AI in qualitative analysis: Fostering student feedback literacy through the application of a feedback quality rubric. *Journal of Engineering Education*, *114*(3), e70024. https://doi.org/10.1002/jee.70024

Due, S., Das, S., Andersen, M., López, B. P., Nexø, S. A., & Clemmensen, L. (2024). Evaluation of large language models: STEM education and gender stereotypes (version 1). arXiv. https://doi.org/10.48550/ARXIV.2406.10133

Forester, R., Miller, R. A., Friedensen, R., Vaccaro, A., & Kimball, E. W. (2024). White racial framing and White supremacy culture in STEM education: Experiences of students with minoritized identities of sexuality and/or gender. *International Journal of Education in Mathematics, Science and Technology*, *12*(3), 642–659. https://doi.org/10.46328/ijemst.3402

Fuchs, S., Werth, A., Méndez, C., & Butcher, J. (2025). Leveraging AI-generated synthetic data to train natural language processing models for qualitative feedback analysis. *Journal of Engineering Education*, *114*(4), e70033. https://doi.org/10.1002/jee.70033

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

Garg, A., Srivastava, D., Xu, Z., & Huang, L. (2022). Identifying and measuring token-level sentiment bias in pre-trained language models with prompts. arXiv preprint arXiv:2204.07289.

Ge, Y., Hua, W., Mei, K., Tan, J., Xu, S., Li, Z., & Zhang, Y. (2024). *OpenAGI: When LLM meets domain experts*. Paper presented at the Advances in Neural Information Processing Systems, New Orleans, LA, USA. https://proceedings.neurips.cc/paper_files/paper/2023/file/1190733f217404edc8a7f4e15a57f301-Paper-Datasets_and_Benchmarks.pdf

Godfrey, E., & Parker, L. (2010). Mapping the cultural landscape in engineering education. *Journal of Engineering Education*, *99*(1), 5–22. https://doi.org/10.1002/j.2168-9830.2010.tb01038.x

Google. (2013). *word2vec [tool for computing continuous distributed representations of words]*. Code Archive. https://code.google.com/archive/p/word2vec/

Goyes, D. R., & Sandberg, S. (2024). Trust, nuance, and care: Advantages and challenges of repeat qualitative interviews. *Qualitative Research*, *25*(2), 330–349. https://doi.org/10.1177/14687941241246159

Gray, D., De Haan, E., & Bonneywell, S. (2019). Coaching the 'ideal worker': Female leaders and the gendered self in a global corporation. *European Journal of Training and Development*, *43*(7/8), 661–681. https://doi.org/10.1108/EJTD-01-2019-0011

Hämäläinen, P., Tavast, M., & Kunnari, A. (2023). *Evaluating large language models in generating synthetic HCI research data: A case study*. Paper presented at the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg Germany. https://doi.org/10.1145/3544548.3580688

Hayes, A. S. (2025). "Conversing" with qualitative data: Enhancing qualitative research through large language models (LLMs). *International Journal of Qualitative Methods*, *24*, 1–19. https://doi.org/10.1177/16094069251322346

Healthy Minds Network. (2023). *Healthy minds study among colleges and universities, year (2022)* [dataset]. Healthy Minds Network—University of Michigan, University of California Los Angeles, Boston University, & Wayne State University. https://healthymindsnetwork.org/research/data-for-researchers

Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., van der Linden, S., & Roozenbeek, J. (2025). Generative language models exhibit social identity biases. *Nature Computational Science*, *5*(1), 65–75. https://doi.org/10.1038/s43588-024-00741-1

Huang, P.-S., Zhang, H., Jiang, R., Stanforth, R., Welbl, J., Rae, J., Maini, V., Yogatama, D., & Kohli, P. (2020). Reducing sentiment bias in language models via counterfactual evaluation. arXiv. https://doi.org/10.48550/arXiv.1911.03064

Jegham, N., Abdelatti, M., Elmoubarki, L., & Hendawi, A. (2025). How hungry is AI? Benchmarking energy, water, and carbon footprint of LLM inference. arXiv preprint arXiv:2505.09598.

Jensen, K. (2021). The time is now to build a culture of wellness in engineering. *Studies in Engineering Education*, *2*(2), 42–45. https://doi.org/10.21061/see.67

Jensen, K., Sanders, J., Johnson, E., Mirabelli, J., & Vohra, S. (2023). CAREER: Supporting mental health and wellness in engineering culture to promote equitable change. Paper presented at the 2023 ASEE Annual Conference and Exposition, Baltimore, MD. https://nemo.asee.org/public/conferences/327/papers/37365/view

Jensen, K. J., & Cross, K. J. (2021). Engineering stress culture: Relationships among mental health, engineering identity, and sense of inclusion. *Journal of Engineering Education*, *110*(2), 371–392.

Jensen, K. J., Mirabelli, J. F., Kunze, A. J., Romanchek, T. E., & Cross, K. J. (2023). Undergraduate student perceptions of stress and mental health in engineering culture. *International Journal of STEM Education*, *10*(1), 1–23. https://doi.org/10.1186/s40594-023-00419-6

Johnson, E., Vohra, S., Sanders, J., Mirabelli, J., Kunze, A., & Jensen, K. (2024). *Faculty and staff ideas and expectations for a culture of wellness in engineering*. Paper presented at the 2024 ASEE Annual Conference and Exposition, Portland, OR. https://par.nsf.gov/biblio/10556061

Kadan, A., Padmanabhan, D., Gangan, M. P., & Lajish, V. L. (2022). Towards an enhanced understanding of bias in pre-trained neural language models: A survey with special emphasis on affective bias. In J. Mathew, G. Santhosh Kumar, D. Padmanabhan, & J. M. Jose (Eds.), *Responsible data science: Lecture notes in electrical engineering* (Vol. 940), (pp. 13–45). Springer. https://doi.org/10.1007/978-981-19-4453-6_2

Kapania, S., Agnew, W., Eslami, M., Heidari, H., & Fox, S. (2024). 'Simulacrum of stories': Examining large language models as qualitative research participants. arXiv preprint arXiv:2409.19430.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv preprint arXiv:2506.08872.

Liu, C., Wang, G.-C., & Wang, H.-F. (2025). The application of artificial intelligence in engineering education: A systematic review. *IEEE Access*, *13*, 17895–17910. https://doi.org/10.1109/ACCESS.2025.3532595

Loughman, T. P., Fleck, R. A., Jr., & Snipes, R. (2000). A cross-disciplinary model for improved information systems analysis. *Industrial Management & Data Systems*, *100*(8), 359–369. https://doi.org/10.1108/02635570010353848

Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., & Liu, X. (2023). AI vs. human – Differentiation analysis of scientific content generation. arXiv: 2301.10416.

Makenzie, L. N., & Wehner, J. (2024). *Context matters: Moving beyond "best practices" to creating sustainable change*. Center for Employment Equity at the University of Massachusetts Amherst. https://www.umass.edu/employmentequity/context-matters-moving-beyond-%E2%80%9Cbest-practices%E2%80%9D-creating-sustainable-change

Martín Núñez, J. L., & Diaz Lantada, A. (2020). Artificial intelligence aided engineering education: State of the art, potentials and challenges. *International Journal of Engineering Education*, *36*, 1740–1751.

Mburu, T. K., Rong, K., McColley, C. J., & Werth, A. (2025). Methodological foundations for artificial intelligence-driven survey question generation. *Journal of Engineering Education*, *114*(3), e70012. https://doi.org/10.1002/jee.70012

McGee, E. O. (2020). Interrogating structural racism in STEM higher education. *Educational Researcher*, *49*(9), 633–644. https://doi.org/10.3102/0013189X20972718

Metz, A., Kainz, K., & Boaz, A. (2023). Intervening for sustainable change: Tailoring strategies to align with values and principles of communities. *Frontiers in Health Services*, *2*, 959386. https://doi.org/10.3389/frhs.2022.959386

Miller, M. E., Wright, C. J., Hargis, L. E., Usher, E. L., Hammer, J. H., & Wilson, S. A. (2022). *Identifying mental health related help-seeking beliefs in undergraduate engineers*. Paper presented at the ASEE Annual Conference and Exposition, Minneapolis, MN. https://doi.org/10.18260/1-2-41918

Muldoon, J., & Wu, B. A. (2023). Artificial intelligence in the colonial matrix of power. *Philosophy and Technology*, *36*(1), 80. https://doi.org/10.1007/s13347-023-00687-8

Nakano, T., Shimari, K., Kula, R. G., Treude, C., Cheong, M., & Matsumoto, K. (2024). *Nigerian software engineer or American data scientist? GitHub profile recruitment bias in large language models*. Paper presented at the 2024 IEEE International Conference on Software Maintenance and Evolution (ICSME), Flagstaff, AZ. https://doi.org/10.1109/ICSME58944.2024.00063

Negi, A. S., Khanna, A., & Aggarwal, R. (2019). Psychological health, stressors and coping mechanism of engineering students. *International Journal of Adolescence and Youth*, *24*(4), 511–520. https://doi.org/10.1080/02673843.2019.1570856

Nighojkar, A., Moydinboyev, B., Duong, M., & Licato, J. (2025). Giving AI personalities leads to more human-like reasoning (version 2). arXiv. https://doi.org/10.48550/ARXIV.2502.14155

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom V., Baltescu P., Bao, H., Bavarian, M., Belgum, J. … Zoph, B. (2024). GPT-4 technical report. arXiv [Cs.CL]. http://arxiv.org/abs/2303.08774

OpenAI Platform. (2024a). *Hello GPT-4o*. OpenAI. https://openai.com/index/hello-gpt-4o/

OpenAI Platform. (2024b). *Models*. Docs. https://platform.openai.com/docs/models

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback (version 1). arXiv. https://doi.org/10.48550/ARXIV.2203.02155

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people (version 1). arXiv. https://doi.org/10.48550/ARXIV.2411.10109

Park, Y. S. (2024). White default: Examining racialized biases behind AI-generated images. *Art Education*, *77*(4), 36–45. https://doi.org/10.1080/00043125.2024.2330340

Pawley, A. L. (2019). Learning from small numbers: Studying ruling relations that gender and race the structure of US engineering education. *Journal of Engineering Education*, *108*(1), 13–31. https://doi.org/10.1002/jee.20247

Pearson, N., Godwin, A., & Kirn, A. (2018). *The effect of diversity on feelings of belongingness for new engineering students*. Paper presented at the 2018 IEEE Frontiers in Education Conference (FIE) (pp. 1–7). https://doi.org/10.1109/FIE.2018.8658443

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. https://arxiv.org/abs/1910.10683

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. https://doi.org/10.21061/see.174

Roberts, J., Baker, M., & Andrew, J. (2024). Artificial intelligence and qualitative research: The promise and perils of large language model (LLM) 'assistance'. *Critical Perspectives on Accounting*, *99*, 102722. https://doi.org/10.1016/j.cpa.2024.102722

Rohde, J., Musselman, L., Benedict, B., Verdín, D., Godwin, A., Kirn, A., & Potvin, G. (2019). Design experiences, engineering identity, and belongingness in early career electrical and computer engineering students. *IEEE Transactions on Education*, *62*(3), 165–172. https://doi.org/10.1109/TE.2019.2913356

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:2402.07927.

Sallee, M. W. (2012). The ideal worker or the ideal father: Organizational structures and culture in the gendered university. *Research in Higher Education*, *53*(7), 782–802. https://doi.org/10.1007/s11162-012-9256-5

Salminen, J., Liu, C., Pian, W., Chi, J., Häyhänen, E., & Jansen, B. J. (2024). *Deus ex machina and personas from large language models: Investigating the composition of AI-generated persona descriptions*. Paper presented at the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA. https://doi.org/10.1145/3613904.3642036

Sanders, J., Johnson, E., Mirabelli, J., Kunze, A., Vohra, S., & Jensen, K. (2024a). Engineering professor perceptions of undergraduate engineering student stress. *European Journal of Engineering Education*, *50*(1), 143–163. https://doi.org/10.1080/03043797.2024.2373754

Sanders, J., Johnson, E., Mirabelli, J., Kunze, A., Vohra, S., & Jensen, K. (2024b). "Not a therapist": Why engineering faculty and staff do/n't engage in supporting student mental health and wellbeing. *IJEE International Journal of Engineering Education*, *40*(1), 196–213.

Sanders, J., Mirabelli, J., Johnson, E., & Jensen, K. (2025). Composite narratives with arts-based analysis of undergraduate engineering students' stress and social supports to identify structural barriers. *Studies in Engineering Education*, *6*(1), 98–128. https://doi.org/10.21061/see.160

Sargent, P. (2005). The gendering of men in early childhood education. *Sex Roles*, *52*(3–4), 251–259. https://doi.org/10.1007/s11199-005-1300-x

Sattele, V., & Ortiz, J. C. (2024). *Generating user personas with AI: Reflecting on its implications for design*. Centro de Investigaciones de Diseño Industrial, UNAM. DRS2024: Boston. https://doi.org/10.21606/drs.2024.1024

Schaul, K., Chen, S. Y., & Tiku, N. (2023). Inside the secret list of websites that make AI like ChatGPT sound smart. *The Washington Post*. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

Secules, S. (2019). Making the familiar strange: An ethnographic scholarship of integration contextualizing engineering educational culture as masculine and competitive. *Engineering Studies*, *11*(3), 196–216. https://doi.org/10.1080/19378629.2019.1663200

Secules, S., McCall, C., Mejia, J. A., Beebe, C., Masters, A. S., L. Sánchez-Peña, M., & 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

Shahriar, S., Lund, B. D., Mannuru, N. R., Arshad, M. A., Hayawi, K., Bevara, R. V. K., Mannuru, A., & Batool, L. (2024). Putting GPT-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences*, *14*(17), 7782. https://doi.org/10.3390/app14177782

Shelley, T. O., Morabito, M. S., & Tobin-Gurley, J. (2011). Gendered institutions and gender roles: Understanding the experiences of women in policing. *Criminal Justice Studies*, *24*(4), 351–367. https://doi.org/10.1080/1478601X.2011.625698

Smith, D. E. (2005). *Institutional ethnography: A sociology for people*. Rowman Altamira.

Spector, P. E. (1982). Behavior in organizations as a function of employee's locus of control. *Psychological Bulletin*, *91*(3), 482–497. https://doi.org/10.1037/0033-2909.91.3.482

Sufi, F. (2024). Generative pre-trained transformer (GPT) in research: A systematic review on data augmentation. *Information*, *15*(2), 99. https://doi.org/10.3390/info15020099

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, *23*, 1–14. https://doi.org/10.1177/16094069241231168

Taubenfeld, A., Dover, Y., Reichart, R., & Goldstein, A. (2024). Systematic biases in LLM simulations of debates. arXiv preprint arXiv:2402.04049.

The Healthy Minds Network. (2024). Interactive Data Interface with Healthy Minds Study. *The Healthy Minds Network Online*. https://healthymindsnetwork.org/data/

Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., & Chaspari, T. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspectives on Psychological Science*, *18*(5), 1062–1096. https://doi.org/10.1177/17456916221134490

Trejo, J. (2020). The burden of service for faculty of color to achieve diversity and inclusion: The minority tax. *Molecular Biology of the Cell*, *31*(25), 2752–2754.

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

Wang, Q., Wu, J., Tang, Z., Luo, B., Chen, N., Chen, W., & He, B. (2025). What limits LLM-based human simulation: LLMs or our design? (version 1). arXiv. https://doi.org/10.48550/ARXIV.2501.08579

Wang, Z., Xie, Q., Feng, Y., Ding, Z., Yang, Z., & Xia, R. (2024). *Is ChatGPT a good sentiment analyzer? A preliminary study*. https://arxiv.org/abs/2304.04339

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv preprint arXiv:2302.11382.

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). *Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding*. Paper presented at the 28th International Conference on Intelligent User Interfaces, Sydney Austrailia. https://doi.org/10.1145/3581754.3584136

AUTHOR BIOGRAPHIES

Jeanne Sanders is a Research Lab Specialist in the Department of Biomedical Engineering at the University of Michigan, 1101 Beal Ave Ann Arbor, MI 48109-2110, USA; jlsander@umich.edu.

John Mobley IV is a graduate student in the Engineering Education Research Program at the University of Michigan, 1101 Beal Ave Ann Arbor, MI 48109-2110, USA; jmiv@umich.edu.

Isabel Miller is a Research Investigator in the Department of Biomedical Engineering at the University of Michigan, 1101 Beal Ave Ann Arbor, MI 48109-2110, USA; isabelmm@umich.edu.

Nicola W. Sochacka co-directs the ProQual Institute for Interpretive Research Methods, Sunshine Coast, Queensland, Australia; nicolasochacka@proqualinstitute.com.

Paul A. Jensen is an Associate Professor in the Departments of Biomedical Engineering and Chemical Engineering at the University of Michigan, 1101 Beal Ave Ann Arbor, MI 48109-2110, USA; pjens@umich.edu.

Karin J. Jensen is an Assistant Professor in the Department of Biomedical Engineering and the Engineering Education Research Program at the University of Michigan, 1101 Beal Ave Ann Arbor, MI 48109-2110, USA; kjens@umich.edu.

How to cite this article: Sanders, J., Mobley, J. IV, Miller, I., Sochacka, N. W., Jensen, P. A., & Jensen, K. J. (2026). Affordances and limitations of using large language models to generate qualitative data about mental health perceptions in engineering. *Journal of Engineering Education*, *115*(1), e70037. https://doi.org/10.1002/jee.70037

APPENDIX A: PHASE 1, COMPARISON RESULTS

Table A1 describes the categorization for each protocol question. Each column describes the similarity between the LLM-generated and the participant responses. Responses were High Similarity (more than 90% similar) ($n = 7$), Moderate Similarity (less than 90% similar but more than 50% similar) ($n = 19$), or Low Similarity (less than 50% similar) ($n = 3$).

TABLE A1 Tabulated results from comparing interview results with large-language model (LLM)-generated responses.

Protocol	Question	High similarity	Moderate similarity	Low similarity
Student	How would you describe what you think of as a “typical” engineering student in your department?		X	
Student	How are you similar or different from this “typical” engineering student?			X
Student	How do you notice when you are stressed?	X		
Student	What typically causes you to experience stress? How often does this happen for you, and how significant are they to you?		X	
Student	If you weren't in engineering, which of these stresses would change and how? Are there any new ones that might come up?		X	
Student	What are some of the things you do to navigate these stresses?		X	
Student	For each of these stressors, who are the people you turn to for support when you're feeling stressed?	X		
Student	Have you ever talked or wanted to talk with an employee of your university about mental health (faculty, councilor, advisor)? How did that go; how did you feel?		X	
Student	Of the stressors and supports you described, which ones do you think are similar and different to other students in your department?			X

TABLE A1 (Continued)

Protocol	Question	High similarity	Moderate similarity	Low similarity
Student	How do you think these examples of stressors and supports will change or stay the same in your life after you finish your classes?		X	
Student	When you feel these stressors, how does that relate or not relate to feeling in control?	X		
Student	If you had a magic wand and could change any of these situations that cause stress, what would you do? How would you change them?			X
Student	How do you define stress?	X		
Student	So would you say you see stress as positive/negative/neutral?		X	
Student	Now that we've talked about stress, I'm going to flip the question and ask: what does wellness mean to you?	X		
Student	What comes to mind if I asked you to describe what you think a "culture of wellness" would look like in your department?		X	
Faculty	Have you had an interaction with an undergraduate engineering student regarding any mental health issues? Have you ever noticed undergraduate engineering students struggling with mental health issues? What does that look like to you?		X	
Faculty	Have you noticed if undergraduate engineering students are able to recognize when themselves or their peers are struggling with mental health issues?		X	
Faculty	Describe what wellness or well-being means to you.		X	
Faculty	Do you think students and faculty in engineering struggle with particular aspects of well-being? Which ones, and why?		X	
Faculty	Do you think faculty members, like yourself, have a responsibility to intervene when you suspect a student is struggling with mental health or wellness? Why/Why not?	X		
Faculty	How would you describe the relationship between undergraduate engineering students' stress and mental health?*		X	
Faculty	In your opinion, is stress different from anxiety? Depression? How so?	X		
Faculty	What do you think of when you hear the words "stressed undergraduate engineer"?		X	
Faculty	Have you noticed undergraduate engineering students experience more issues with extreme stress and mental health than other majors? Why/Why not?*		X	
Faculty	How would you describe the physical or physiological signs of stress in undergraduate engineers?		X	
Faculty	What do you think typically causes stress in undergraduate engineers?		X	
Faculty	Can you describe things you have heard or learned undergraduate engineering students do to manage stress?		X	
Faculty	What have you noticed your department/program does to encourage healthy and/or unhealthy stress management/coping strategies for undergraduate engineering students?*		X	
Faculty	Can you describe any resources or supports on campus or in your department for undergraduate engineering students who are stressed?		X	

Note: Responses that contained several responses of "I don't know" are marked with an asterisk (*).

APPENDIX B: PHASE 1, EXAMPLE OUTPUT

Example prompt and output for Phase 1. This exemplar output corresponds to the faculty interview question "Have you had an interaction with an undergraduate engineering student regarding any mental health issues?"

Prompt:

> Please respond to this as an engineering faculty member in the US. The faculty is in the middle of an interview. Only give the faculty's response to this question. – 1. Have you had an interaction with an undergraduate engineering student regarding any mental health issues?
>
> Please generate a list of the answers that you think 4 different (2 women and 2 men) engineering faculty members in the US might give. Then, after generating the 4 responses. Create a bulleted list summary of the key points people shared

ChatGPT output:

Response 1 (Female faculty member):

> Yes, I have. I had a student who came to me feeling overwhelmed by their course load and personal issues. I made sure to actively listen and let them know I was there to support them. After the conversation, I referred them to our campus counseling center and followed up later to see how they were doing.

Response 2 (Male faculty member):

> Yes, I've had several conversations over the years. One instance that comes to mind was a student struggling with anxiety around exams. We worked together to create a more manageable study plan, and I connected them with our mental health resources on campus. I believe it's important for faculty to be approachable so students feel comfortable discussing these issues.

Response 3 (Female faculty member):

> Absolutely. A student once shared with me that they were having difficulty balancing academic work with their mental health challenges. I offered them some flexibility with deadlines and emphasized the importance of seeking professional support. We have a responsibility as faculty to recognize when a student might need help and to guide them to the right resources.

Response 4 (Male faculty member):

> Yes, a few times. One conversation I remember well involved a student who felt isolated and homesick, which was affecting their performance. I made an effort to include them in class discussions and suggested joining student clubs for a better support system. I also walked them through the mental health services available on campus to ensure they knew where to turn if needed.

Key points summary

- Active listening and support: Faculty emphasized the importance of listening to students and showing support.
- Referrals to campus resources: All faculty members highlighted the value of referring students to professional mental health resources available on campus.
- Providing flexibility: Some faculty offered accommodations such as extended deadlines or modified study plans to help students manage their stress.
- Encouraging engagement: One faculty member suggested increased participation in class and joining clubs to alleviate feelings of isolation.
- Follow-up and continuous support: Faculty members noted the importance of following up with students to see how they are doing after initial conversations.

APPENDIX C: PHASE 2, CODEBOOK

TABLE C1 Phase 2 codebook for codes applied to the varied persona data.

Code name	Description	Example usage
Culture of care	Creating a supportive environment. Impacting culture or norms. Acknowledging stressful culture. "We" mentality.	"It's about creating a culture of care."
DEI experience unique challenges	An awareness that students with marginalized backgrounds experience unique challenges	"Recognizing that many students of color, LGBTQ+ students, and students with disabilities are often navigating unique challenges, I see it as my duty to..."
Empathy from personal experience	The response describes experiencing similar struggles with wellbeing, wellness, or mental health in their life. And, that leading to empathy for the students.	"As a Black woman who has navigated similar spaces, I know..."
Holistic	Supporting the student outside of "only academics"	"This includes not just their academic performance but their mental health and well-being."
Make myself available	Intentionally trying to be available for students to approach them with questions	"I believe we should be actively looking out for our students."
Not a counselor	Mentioning that faculty aren't counselors	"Even though we are not trained counselors, we can guide..."
Uncomfortable	Describing feeling uncomfortable	"No, I don't feel comfortable intervening in personal issues."
Not really/ unsure	Answering the question with uncertainty and/or a "No"	"Not really. Our main job is to teach and advise on academic issues..."
Yes, but...	Answering the question with a conditional yes	"I do think we have a responsibility to intervene, but that intervention should be..."
Yes strong	Answering the question with an unequivocable yes	"Yes, I believe faculty have a responsibility to intervene..."

APPENDIX D: PHASE 2, EXAMPLE OUTPUT

This appendix includes example output from Phase 2 analysis. We include three example outputs, which were generated as an output for each persona, with the corresponding codes following in brackets.

Persona 1: Default

> Absolutely. Students' success is not just about academics. If they are facing mental health challenges, it's crucial for us to help them find the resources they need to cope and succeed.
> [Yes strong; Holistic]
>
> Yes, faculty members should play a role in supporting students' overall well-being. Even though we are not trained counselors, we can guide them to professionals who can help.
> [Yes strong; Not a counselor]
>
> Yes, but our role should primarily be to create a supportive environment and encourage them to seek help rather than trying to solve their issues ourselves. [Yes, but...; Culture of care]

Persona 2: Powered

Yes, I believe faculty have a responsibility to intervene. We're often one of the few adults students regularly interact with, and it's important to support them beyond just academics.

[Yes strong; holistic]

Not really. Our main job is to teach and advise on academic issues. We should leave mental health to the experts.

[Not really/unsure; Not a counselor; Uncomfortable]

It's not our primary responsibility, but we should be attentive and show concern. A little support can go a long way.

[Yes, but...; Culture of care]

Persona 3: Multiply Marginalized

Absolutely. As faculty, we're not just responsible for teaching subject matter; we're also stewards of our students' overall well-being. Recognizing that many students of color, LGBTQ+ students, and students with disabilities are often navigating unique challenges, I see it as my duty to reach out and offer support or direct them to resources when needed. It's about creating a culture of care.

[Holistic; DEI experience unique challenges; Culture of care; Yes strong]

I do think we have a responsibility to intervene, but that intervention should be supportive rather than overstepping boundaries. I make it a point to check in and provide resources, but I also recognize that some students may not want to disclose their struggles, and I respect their autonomy.

[Yes, but...; Make myself available]

Yes, I believe faculty members should be proactive in addressing potential mental health struggles. It's about providing a safety net. As a Black woman who has navigated similar spaces, I know how crucial it can be to have someone recognize when you need help, especially when it feels like you're expected to be strong all the time.

[Personal experience and empathy; Yes strong; Culture of care]