Beyond the Algorithm: Comparing Human and AI Systems Analysis of Outcomes of an Engineering Summer Bridge Program

Dr. Caitlin Morris, University of Massachusetts Lowell

Caitlin Morris is an Assistant Teaching Professor in the department of Chemical Engineering at the University of Massachusetts Lowell. She has a background in chemical engineering and teaches courses in the undergraduate curriculum. Her interests include student learning, professional development, and educational innovation in engineering including the use of Artificial Intelligence.

Prof. Kavitha Chandra, University of Massachusetts Lowell

Kavitha Chandra is Professor of Electrical and Computer Engineering in the Francis College of Engineering at the University of Massachusetts Lowell. She directs the Research, Academics and Mentoring Pathways (RAMP) to Success summer bridge and academic program for new engineering students, preparing them with research, communication and leadership skills. Her research interests are in computational and data-driven modeling of physical systems in acoustics and communication networks, model-based systems engineering, user-centric design and enterprise modeling of digital health systems, and engineering education.

Dr. Susan Thomson Tripathy, University of Massachusetts Lowell

Dr. Susan Thomson Tripathy is a social science research consultant specializing in qualitative research methodology, including ethnography and participatory action research.

Orlando Arias, University of Massachusetts Lowell
Elizabeth Often

Beyond the Algorithm: Comparing Human and AI Interpretation of Mentorship Data in an Engineering Summer Bridge Program

1. Introduction

As artificial intelligence expands into all aspects of higher education, educators and researchers increasingly rely on AI tools for data processing, coding support, and preliminary interpretation. The potential benefits are clear: AI systems can summarize patterns rapidly, reduce human workload, and reveal structures hidden in large datasets [1], [2]. Yet these systems currently cannot provide the experience related to social, cultural, and developmental dimensions of engineering education as humans can. While some contemporary LLM-based systems can generate context-sensitive interpretations when provided with background and appropriate prompts, their reasoning is still fundamentally pattern-based and may diverge from theory-grounded, developmentally nuanced human judgment—particularly when datasets are small, indirect, or socially situated [1], [2].

Engineering education research frequently involves constructs that require deep interpretation—engineering identity [3], belonging [4], mentorship [5], [6], and career socialization [7]. These constructs are embedded in student development and program design and are shaped by interactions within the learning environment. As such, they are difficult to interpret purely numerically or through text-based pattern detection.

This paper addresses the question: How do AI systems interpret mentorship and engineering identity data differently from humans, and how does prompt structure influence this interpretation? While some contemporary AI systems can generate context-sensitive and theoretically framed interpretations when explicitly prompted—as is the case for general-purpose reasoning models such as ChatGPT 5.1 and more narrative-oriented models such as Claude 3.5—the nature and boundaries of that contextual insight remain unclear. Because mentorship and professional identity formation are essential ingredients for engineering leadership development, this study situates its analysis within both engineering education and engineering leadership scholarship. To investigate this, we use datasets from Research, Academics, and Mentoring Pathways (RAMP), an engineering summer bridge program designed to support early engineering students through mentorship, project-based learning, and community building. By systematically comparing AI outputs generated under non-contextualized and contextualized prompting conditions to human interpretation, this study examines not whether AI can produce contextual language, but how that context is constructed, constrained, and aligned with evidence relative to human, theory-grounded reasoning.

AI tools were prompted under two analytic regimes:

1. Non-contextualized analysis, emphasizing statistical description;
2. Contextualized analysis, emphasizing theoretical interpretation and program context.

A third adjudicator lens allowed models to synthesize differences. This study reveals critical insights into the strengths and limitations of AI for engineering education research and for understanding mentorship-driven student development.

2. Background

2.1 Engineering Identity and Belonging

Engineering identity is a central construct in engineering education, predicting persistence, sense of purpose, academic engagement, and professional orientation [3], [8]. Identity is shaped not only by skill acquisition but by recognition from others, performance opportunities, and internalization of engineering values [9]. For students who are starting their engineering degree pathways, identity formation is active and fragile, influenced by mentorship, peer dynamics, cultural cues, and experiences of success or struggle [7].

Belonging intersects with identity to support resilience as students progress towards the engineering degree. Students who feel connected to engineering communities are more likely to persist, particularly those from historically excluded groups [4]. As such, belonging is often a key outcome for bridge programs such as RAMP.

2.2 Mentorship Mechanisms in Engineering Education

Mentorship serves as a relational mechanism for transmitting engineering norms, expectations, and professional habits [5], [6]. Research shows that high-quality mentoring can improve career self-efficacy, academic confidence, personal growth and clarity of goals, and professional identity formation [10], [11]. Industry mentors provide authentic exposure to engineering careers and practices, while graduate or near-peer mentors often offer cognitive apprenticeship—scaffolding tasks, modeling strategies, and guiding reflection [12]. Both forms of mentorship are key components in RAMP.

Recent work using participatory action research (PAR) in STEM (science, technology, engineering, and mathematics) summer bridge programs demonstrates that mentoring relationships are also deeply psychosocial, shaping how students see themselves as future engineers and community members. Tripathy, Chandra, and Reichlen [13] show that PAR-based assessment surfaces student perceptions of mentorship, belonging, and agency in a summer bridge program, while Tripathy et al. [14] highlight the transformative potential of positioning women engineering undergraduates as peer facilitators and co-researchers in mentoring spaces.

Mentorship in engineering education has also been linked to leadership competency and leadership identity development. Rottmann, Sacks, and Reeve [23] describe Engineering Leadership Orientations as professional identity configurations that shape how engineers understand authority, collaboration, responsibility, and influence in technical contexts. These orientations emerge not solely through formal leadership instruction, but through socialization, project-based learning, and mentoring relationships that expose students to authentic professional practice. Related work from

the Troost Institute for Leadership Education in Engineering (e.g., Chan & Rottmann [24]; Reeve, Rottmann, & Sacks [25]) demonstrates that leadership identity in engineers develops through experiential learning, relational validation, and engagement in complex team-based problem solving. In this framing, mentoring within bridge programs may function not only as academic support but also as an early site of leadership socialization, contributing to how students perceive themselves as capable contributors, collaborators, and emerging leaders in engineering contexts.

2.3 Summer Bridge Programs as Developmental Contexts

Summer bridge programs are designed to support transitional student populations—rising high-school seniors and new college students—who are making critical academic and career decisions. Bridge programs have been shown to enhance student confidence, STEM identity, sense of belonging, and persistence into engineering majors [5], [10]. Work at our own institution has documented how summer bridge and related initiatives use PAR, allyship, and co-creation frameworks to promote student ownership, teamwork, and professional skills [13], [15], [16].

Our host institute's program, RAMP, builds on this foundation by integrating industry interactions, hands-on robotics projects, coding instruction, inclusive teamwork activities, and mentorship, creating a learning ecosystem that supports both technical and identity development.

2.4 AI Interpretation in Education Research

AI models can summarize content, detect patterns, and generate interpretations at scale, but they may:

- lack awareness of developmental nuance
- fail to distinguish correlation from mechanism
- misinterpret ceiling effects or small-N datasets
- ignore qualitative meaning-making
- overgeneralize based on limited evidence [1], [2], [17], [18].

These concerns are particularly salient in engineering education, where constructs like identity, belonging, and mentorship are highly contextual and often measured indirectly. This study contributes to the AI-in-education literature by examining how model architecture and prompt design shape interpretation and by documenting specific failure modes when AI is used to interpret mentorship and identity data.

The constructs of engineering identity, belonging, and mentorship are not only conceptual lenses for interpreting student experience; they also serve as analytic anchors in this study. Specifically, the contextualized AI prompt and the human interpretive benchmark were guided by established definitions of engineering identity [3], [9], belonging [4], and mentoring as a relational and developmental mechanism [5], [11]. These frameworks informed how contextualized interpretations were evaluated—particularly when assessing claims related to identity growth, relational validation, and professional socialization.

3. Methods

The following sections describe the methodological framework used to examine AI interpretation of survey datasets derived during the RAMP 2025 summer program through two distinct analytic lenses. Specifically, AI analyses were conducted using a non-contextualized lens and a contextualized lens to capture differences in how meaning is constructed from the same data. Figure 1 provides an overview of this framework and visually distinguishes the two analytic tracks, with the non-contextualized analysis indicated in blue and the contextualized analysis indicated in orange. Throughout the Methods and Results sections, non-contextualized analyses are indicated in blue (🔵) and contextualized analyses are indicated in orange (🟠) to visually distinguish between structurally constrained and interpretive analytic outputs. In contextualized outputs, references to identity, belonging, and mentorship were evaluated relative to the theoretical frameworks introduced in Section 2 to assess whether AI-generated interpretations aligned with established constructs or extended beyond what the data supported. Additionally, because identity and belonging are also foundational to engineering leadership development [23], grounding the contextualized lens in these frameworks ensures that interpretive claims remain aligned with established scholarship.

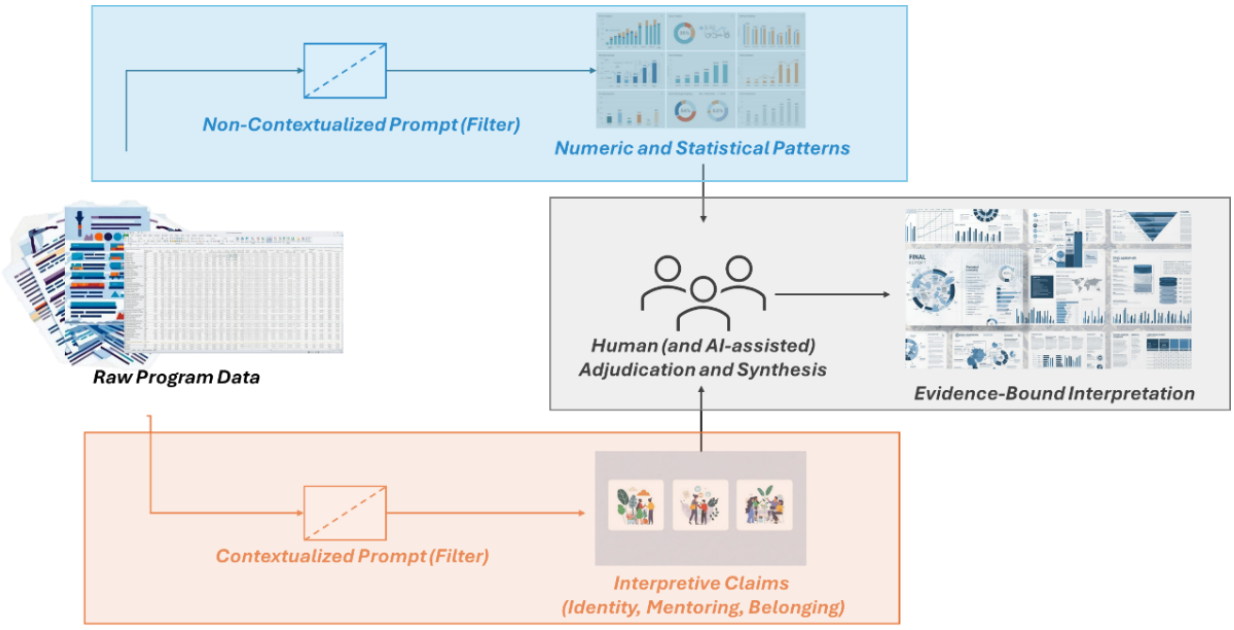

Figure 1: Piping and Instrumentation Diagram (P&ID)-style hybrid human–AI interpretive workflow illustrating prompt engineering as an interpretive filter. The same survey data are processed through two engineered prompt conditions: a non-contextualized prompt (blue), which constrains AI output to numeric and statistical patterns, and a contextualized prompt (orange), which permits theory-informed interpretive claims related to identity, mentoring, and belonging. Outputs from both prompts are subsequently reviewed through a human (and AI-assisted) adjudication and synthesis step, resulting in evidence-bound interpretations.

The two analytical lenses shown in Figure 1 are operationalized through deliberately engineered AI prompts designed to either suppress or encourage contextual interpretation. In this framework, the non-contextualized and contextualized analyses are not separate datasets or post-hoc

categorizations, but rather distinct prompting conditions that govern how the AI processes and represent the same underlying data. Each prompt acts as an interpretive constraint, shaping what types of inferences the AI is permitted to make.

Figure 1 adopts a process-flow representation analogous to a piping and instrumentation diagram (P&ID), a familiar structure in engineering systems analysis, to emphasize that interpretation itself is a controlled operation. In this representation, prompt engineering is depicted using a filter symbol, indicating that each analytical lens selectively permits or blocks certain forms of reasoning. The non-contextualized filter restricts the AI to structural and numerical descriptions (e.g., distributions, counts, and missingness), while explicitly preventing causal, developmental, or theoretical interpretation. In contrast, the contextualized filter allows the AI to integrate developmental considerations, identity and mentoring frameworks, and qualitative meaning-making, while still operating on the same input data.

By framing prompt engineering as an interpretive filter within a P&ID-style workflow, the figure highlights a central methodological claim of this study: AI interpretation is not intrinsic to the data but is engineered through the constraints and affordances imposed by prompts. This representation also clarifies why different AI systems, even when given identical data, can produce substantively different interpretations depending on how those interpretive filters are configured.

3.1 Research Design

We designed a comparative framework to evaluate how different AI systems interpret the same educational datasets under constrained versus contextualized conditions. This framework followed a three-lens interpretive protocol:

1. 🔵 Non-Contextualized Lens → numeric, descriptive, structure-oriented analysis
2. 🟠 Contextualized Lens → interpretive, developmental, and theoretically grounded analysis
3. Adjudicator Lens → synthesis and evaluation of differences across the two analyses

This design enabled us to examine AI interpretive flexibility and limitations and to compare AI outputs to human analysis. In addition, this three-lens interpretive framework was developed for the present study, informed by prior work on mixed-methods analysis and participatory action research in engineering education, but represents an original application of prompt-engineered AI comparison rather than an adaptation of an existing analytic protocol.

3.2 Datasets

Table 1 summarizes the datasets analyzed by each AI system. These datasets are from the RAMP 2025 cohort.

Table 1: RAMP 2025 Datasets Used for AI Interpretation

Dataset	Description	Sample Size
Student Pre-Survey	Pre-program survey capturing students' self-reported engineering identity, expectations for the RAMP program, prior exposure to engineering-related activities, and basic demographic information (e.g., educational level). Responses establish baseline perceptions and anticipated outcomes prior to participation.	20
Student Post-Survey	Post-program survey assessing students' self-reported engineering identity, sense of belonging, satisfaction with program components (including mentoring and project experiences), perceived accomplishments, and reflections on mentoring interactions following completion of RAMP.	19
Mentor Pre-Survey	Pre-program survey of industry mentors documenting expectations for participation, prior mentoring experience, and initial perspectives on engaging with pre-college and early-college engineering students.	6
Mentor Post-Survey	Post-program survey of industry mentors capturing satisfaction with the mentoring experience, perceptions of student engagement and development, and reflective responses regarding mentoring strategies, challenges, and outcomes observed during the program.	5
Mentor Poll	Student survey collecting ranked preferences for industry partners or projects, along with brief qualitative explanations for those rankings, providing insight into students' initial interests, career alignment, and decision-making prior to mentor matching.	20

These datasets contain numeric, ordinal, and qualitative data, allowing examination of how AI interprets multi-modal information.

3.3 AI Models

Three distinct AI systems were selected:

- Perplexity Sonar: a retrieval-augmented large language model designed to synthesize structured information across documents. Retrieval-augmented generation (RAG) systems are known to emphasize document-grounded summarization and structured synthesis [19].
- ChatGPT 5.1: general-purpose transformer-based large language model capable of integrating statistical reasoning with contextual interpretation. Independent research on large language models (LLMs) suggests that such systems can generate context-sensitive responses when appropriately prompted, though their reasoning remains pattern-based rather than theory-grounded [20], [21].
- Claude 3.5 Sonnet: a large language model optimized for extended narrative coherence and contextual reasoning. Studies of generative AI systems demonstrate that increased narrative fluency can correlate with higher interpretive richness but also with increased risk of overgeneralization or unsupported inference [22].

It is important to note that commercial AI systems may incorporate content moderation, alignment training, and—in some cases—human-in-the-loop reinforcement processes. While the internal architecture and review mechanisms of these systems are not fully transparent, this study evaluates the models based on the outputs delivered through their publicly accessible interfaces. Any human review embedded within these systems would affect all prompt conditions equally and therefore does not undermine the comparative design; however, it does reinforce that AI-generated interpretations reflect sociotechnical systems rather than purely autonomous algorithmic reasoning.

3.4 Prompt Architecture

Each AI model was prompted once per analytic condition (non-contextualized, contextualized, and adjudicator) using identical datasets. Prompts were issued in isolation to minimize carryover effects; no conversational history was preserved between prompts. While it is not possible to fully eliminate latent model memory effects, this design reduces the likelihood that prior prompts directly shaped subsequent outputs.

🔵 Non-Contextualized Prompt

Models were instructed to:

- avoid interpretation and causal language
- avoid contextual or developmental explanations
- provide only numeric and structural descriptions (e.g., distributions, missingness, variable types)
- refrain from attributing meaning, mechanisms, or program impact

This tested each AI's ability to maintain analytic restraint and operate as a numeric/structural auditor.

🟠 Contextualized Prompt

Models were instructed to interpret data in the style of an engineering education researcher, explicitly considering:

- developmental stage (high school vs. first-year college students)
- engineering identity and belonging frameworks
- mentorship mechanisms and program structure
- alignment and mismatch between student and mentor perspectives
- qualitative themes in open-ended responses

Adjudicator Prompt

Models were given both their non-contextualized and contextualized outputs and asked to:

- compare differences between the two lenses
- identify points of agreement and disagreement
- highlight blind spots and overinterpretation risks
- generate a synthesized interpretation that integrated numeric rigor and contextual meaning

This design is illustrated in Figure 2, which depicts the engineered prompt conditions used in this study. The figure explicitly shows how the non-contextualized prompt constrains AI interpretation by blocking contextual and causal reasoning, and how the contextualized prompt enables theory-informed interpretation while introducing additional interpretive risk.

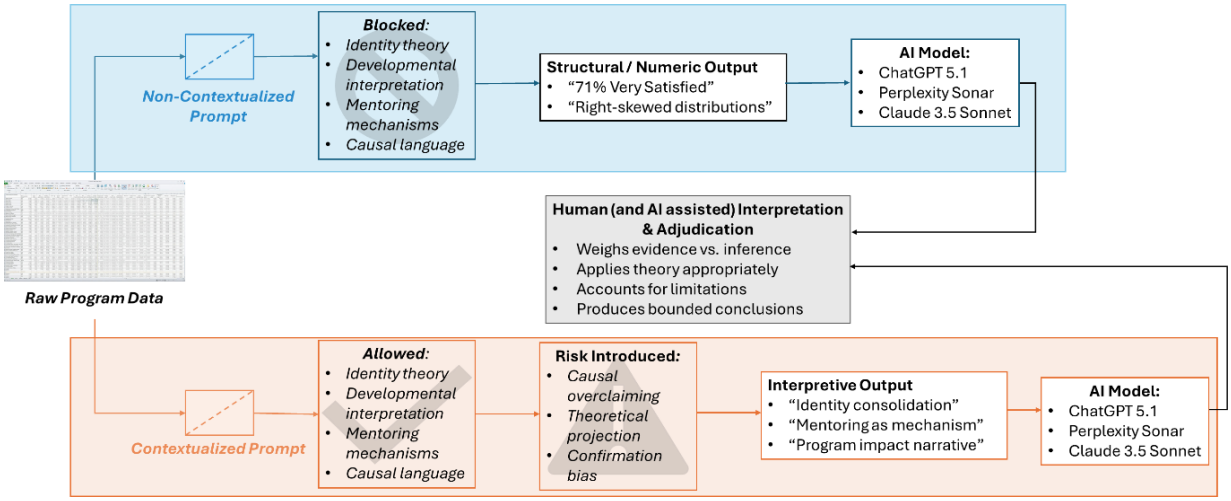

Figure 2: Conceptual representation of prompt engineering as interpretive filtering. The same survey and qualitative data are processed through two engineered prompt conditions. The non-contextualized prompt (blue) constrains AI output to structural and numerical descriptions, while the contextualized prompt (orange) permits theory-informed interpretation and qualitative meaning-making, introducing additional interpretive risk. A human adjudication step integrates both outputs into evidence-bound conclusions.

3.5 Human Interpretation

To benchmark AI performance, a human engineering education researcher provided independent numeric and contextual interpretations of the data, followed by a synthesis analysis. The human analysis drew on prior experience with and research on our host institution's summer bridge program, RAMP, and related PAR-based bridge initiatives at our institution [13], [16], as well as broader literature on engineering identity, belonging, and mentoring. This comparison clarified where AI aligned with or diverged from theory-informed human reasoning.

3.6 Analytic Approach

AI outputs were compared qualitatively across:

- interpretive tendencies (e.g., conservative vs. narrative)
- theoretical alignment and use of concepts (identity, belonging, mentorship)
- evidence use (numeric vs. qualitative vs. blended)
- drift into interpretation when instructed not to interpret
- overinterpretation risks (causal overclaiming, projection, confirmation bias)
- explicit acknowledgment of limitations (small-N, ceiling effects, non-equivalent items)

We also examined adjudicator outputs to see whether models could critique their own earlier interpretations and differentiate between what the data show and what interpretations infer.

4. Results

4.1 🔵 Non-Contextualized Outputs

Under non-contextualized prompts, Perplexity delivered a highly constrained structural analysis. It described row counts, column types, missingness patterns (i.e., which survey items were left unanswered and by whom), and scale ranges, and noted features such as ceiling effects and extreme duration outliers but did not infer meaning. A typical Perplexity statement was: "Mentor datasets contain very small samples ($n \approx 5$), and identity items are right-skewed toward agreement."

ChatGPT produced a more statistical audit. It conducted chi-square comparisons for shared categorical items, calculated effect sizes, and noted that no student pre–post items reached conventional significance at $\alpha = 0.05$. It flagged small mentor sample sizes and cautioned about unstable effect-size estimates but largely avoided contextual or causal claims.

Claude, by contrast, drifted noticeably into interpretation despite the non-contextual instructions. For example, it stated that "coding confidence improved significantly, demonstrating the program's effectiveness," even though no inferential statistics were requested or reported. Claude also described program components as "effective" or "impactful," treating descriptive changes as evidence of success.

Table 2. Characteristics of Non-Contextualized Outputs

Model	Output Characteristics
Perplexity	Strictly numeric and structural; no interpretation; identifies distributions, missingness, and scale use.
ChatGPT	Statistical comparisons and effect sizes; cautious phrasing; minimal contextual inference.
Claude	Begins contextualizing despite instructions; narrative drift; uses causal and evaluative language.

4.2 Contextualized Outputs

Contextualized prompts elicited theoretical and developmental interpretations from all three models, but with different emphasis and risks.

Perplexity integrated established frameworks of engineering identity [3], [9], belonging [4], and mentorship as developmental socialization [5], [11] in a conservative manner. For example, when Perplexity suggested that students appeared to "consolidate identity rather than form it," this interpretation aligns with identity development frameworks that distinguish between exploratory and consolidating phases of professional identity formation [9], rather than introducing novel or unsupported theoretical constructs.

ChatGPT's contextualized interpretations similarly drew on identity consolidation language and mentoring-as-socialization frameworks, linking high baseline identity scores and project-based engagement to established models of professional identity development. It described RAMP as a space where already motivated students refined their engineering identity through authentic project work and relationships with industry mentors, noted that ceiling effects limited observable numeric gains, and raised plausible mechanisms (e.g., cognitive apprenticeship, professional socialization) while acknowledging design and power limitations.

Claude's mechanism-level claims invoked identity formation and professional socialization themes but often extended beyond what was directly measurable in the survey instruments, illustrating how theory-language can be applied more expansively than the data support. It described the RAMP program's robot construction and programming project as "an especially effective mechanism for accelerated identity development" and argued that mentoring "transformed students' perceptions of engineering." These statements suggested causal relationships and durable identity shifts that were not directly measured by the survey instruments or supported by the small sample sizes. Notably, Claude's contextualized interpretations align with common narratives in prior engineering education literature, suggesting that the model may be drawing on learned patterns from previously published studies rather than on evidence uniquely supported by the present dataset.

Table 3. Characteristics of Contextualized Outputs

Model	Theme Integration	Overreach Risk
Perplexity	High fidelity to data; theory-informed but cautious interpretations.	Low–moderate
ChatGPT	Strong conceptual synthesis; integrates quantitative and qualitative evidence.	Moderate
Claude	Rich narrative; causal framing and strong evaluative claims.	High

4.3 Adjudication and Cross-Model Comparison of Lens-Conditioned AI Outputs

All three models, when acting as adjudicators, were able to distinguish between numeric findings and interpretive claims to some degree. Perplexity emphasized differences between what data "show" and what interpretations "infer," explicitly flagging small sample sizes and ceiling effects as constraints. ChatGPT provided a nuanced synthesis that separated structural results (e.g., non-significant pre–post differences) from theory-informed inferences and highlighted where contextual claims went beyond the data.

Claude's adjudicator summaries were coherent and often insightful, but it was less explicit in critiquing its own earlier overstatements. It tended to gently rephrase rather than directly problematize its causal or evaluative language, reinforcing the need for human oversight even when AI is asked to "adjudicate" itself.

Table 4 summarizes the primary conclusions reached by each model, the differences observed between lenses, and the reasoning patterns that drove those differences.

Table 4: Model Analysis Comparison

Model	Non-contextualized findings	Contextualized findings	Key difference between lenses	Primary reasoning pattern
Perplexity	Structural audit of data: small mentor samples, partial responses, duration outliers; identity items skewed toward agreement at both timepoints; wide numeric variability on 0–100 and ranking scales.	De-identified program strengthened existing engineering identity; authentic project work and relational mentoring anchored confidence, engagement, and teamwork; developmental distinctions between high-school and first-year students.	Expansion from data structure and distributions to cautious mechanism-based interpretation.	Emphasizes data integrity and distributional shape, then layers theory conservatively when context is allowed.
ChatGPT	No statistically detectable differences across shared substantive survey items; significance limited to administrative metadata; conclusions constrained by small sample sizes and ceiling effects.	De-identified program primarily deepened and clarified an already strong engineering trajectory; mentorship supported perceived growth, while belonging outcomes were more mixed; developmental stage shaped perceived impact.	Shift from statistical detectability to developmental and identity-based interpretation.	Moves from hypothesis-testing logic (p-values, power) to triangulation of distributions, subgroup patterns, and identity theory.
Claude	Reports large percentage changes and high satisfaction as evidence of effectiveness (e.g., coding confidence increase, changed career perceptions, rising engagement).	Frames project-based learning and mentorship as effective mechanisms for identity development through multiple pathways.	Little separation between lenses; evaluative and causal language appears even under non-contextual constraints.	Relies on magnitude of change and strongest positive signals; narrative synthesis outweighs statistical constraint.

Across all three models, non-contextualized prompting produced analyses that were anchored in observable data structure and response distributions. However, the degree to which models adhered to numeric restraint varied. Perplexity most consistently operated as a structural auditor, focusing on sample sizes, missingness, scale usage, and distributional shape without inferring meaning. ChatGPT similarly constrained its conclusions under non-contextualized conditions, explicitly framing results in terms of statistical detectability and emphasizing limitations due to small sample sizes and ceiling effects. In contrast, Claude's non-contextualized outputs already incorporated evaluative language, treating large percentage changes and high satisfaction ratings as evidence of program effectiveness, indicating interpretive drift even when contextual reasoning was explicitly restricted.

Under contextualized prompting, all three models expanded their analyses to include identity, mentoring, and developmental interpretations, but again did so in distinct ways. Both Perplexity

and ChatGPT converged on a similar high-level interpretation: that the program primarily strengthened or clarified an already strong engineering trajectory rather than creating interest from scratch. This conclusion was grounded in consistently high pre-survey identity scores, post-survey responses clustered near the ceiling, and mixed patterns in reported interest change. ChatGPT further emphasized that mentorship functioned as a key mechanism for perceived growth while belonging outcomes remained heterogeneous, and it introduced developmental distinctions between high-school and first-year students to explain subgroup variation. Perplexity offered a similarly cautious interpretation, highlighting authentic project work and relational mentoring as anchors for engagement and confidence while maintaining explicit attention to data constraints.

Claude's contextualized analysis differed more substantially. It framed project-based learning and mentorship as effective mechanisms for engineering identity development through multiple pathways, including skill acquisition, professional socialization, and stereotype disruption. While these interpretations were supported by strong satisfaction ratings and qualitative comments, they relied on broader synthesis and causal framing that exceeded what was directly measured by the survey instruments.

Comparing models across lenses reveals two key patterns. First, prompt engineering meaningfully shapes AI interpretation but does not fully override model-specific tendencies. Second, contextualized prompting consistently increased interpretive richness while simultaneously increasing the risk of overreach, particularly in models inclined toward narrative synthesis. These findings reinforce the need for human adjudication when using AI to interpret educational data, especially for constructs such as identity, belonging, and mentorship that are developmentally and socially situated.

While the need for human adjudication may appear to limit the autonomy of AI systems, it does not negate their value. AI models consistently demonstrated strengths in rapid structural auditing, pattern detection, and cross-variable synthesis that would be time-intensive for human researchers. Under non-contextualized prompting, AI systems efficiently surfaced distributional skew, ceiling effects, missingness patterns, and sample-size constraints—diagnostic insights that are foundational to responsible interpretation. Under contextualized prompting, AI generated plausible interpretive hypotheses that, although requiring validation, expanded the analytic horizon by proposing developmental distinctions, mentoring mechanisms, and identity-related explanations. In this sense, AI functions not as a replacement for human reasoning, but as an accelerant and hypothesis generator within a bounded analytic workflow. The value of AI in this framework lies in its capacity to surface patterns and provisional interpretations at scale, while human researchers retain responsibility for theory alignment, evidentiary weighting, and causal restraint.

Together, the cross-model comparison demonstrates that differences in AI interpretation are not simply a function of data, but emerge from the interaction between prompt constraints, model architecture, and reasoning style. This provides a critical foundation for the discussion that follows regarding the appropriate and responsible use of AI in engineering education research.

Finally, in addition to comparing model-specific outputs, we examined the degree of consistency across AI systems within each analytic lens. Under non-contextualized prompting, outputs were broadly consistent in identifying high baseline identity, small mentor sample sizes, and right-skewed response distributions. Differences across models were primarily in emphasis rather than conclusion, with Perplexity focusing on data structure, ChatGPT emphasizing statistical detectability, and Claude highlighting magnitude of change.

Under contextualized prompting, interpretive variability increased substantially. While all three models referenced mentoring and project-based learning as influential program elements, the strength and scope of claims differed. Perplexity and ChatGPT offered bounded interpretations aligned with ceiling effects and developmental context, whereas Claude extended interpretations toward causal and mechanism-level conclusions. This divergence suggests that contextualized prompting increases interpretive richness but also amplifies model-dependent reasoning styles, reinforcing the need for human adjudication when using AI to interpret complex educational data.

5. Discussion

5.1 AI Systems Have Distinct Interpretive Identities

This study reveals that AI models do not behave uniformly. Perplexity operates like a data auditor, focusing on structure, distributions, and missingness. ChatGPT resembles a mixed-methods researcher, balancing numeric and thematic integration. Claude defaults to narrative educational storytelling, generating compelling but sometimes overreaching accounts. These tendencies persisted across lenses and highlight the need for transparency in reporting AI-assisted qualitative or quantitative findings.

5.2 🔴 Contextualization Makes AI More Useful But Also More Dangerous

Contextualized prompts allow AI to approximate human interpretive work by identifying patterns that resemble mechanisms (e.g., cognitive apprenticeship in mentorship), developmental distinctions (e.g., differences between high-school and first-year participants), and program design implications (e.g., the role of project structure in engagement). However, they also increase several interpretive risks:

- Causal overclaiming. Claude stated, “Coding confidence improved significantly, demonstrating the program’s effectiveness.” This incorrectly attributes causality and statistical significance to descriptive pre–post differences without tests or controls.
- Theoretical projection not grounded in data. ChatGPT suggested that “belonging scores map onto early stages of identity consolidation,” projecting identity theory onto data that do not directly measure identity status or stages.
- Confirmation bias toward positive outcomes. Claude wrote that “mentor satisfaction confirms that student engagement increased,” treating perceived engagement as objective evidence of program impact.

- Underemphasis on sample size limitations. All models noted small mentor and student Ns, but contextualized outputs often treated patterns in these small groups as more robust than warranted.

These issues echo broader concerns in AI interpretability and fairness, where models may appear confident and coherent even when the underlying evidence is weak [18], [17].

5.3 🔵 Non-Contextual AI Alone Is Insufficient

Non-contextual, numeric-only AI analysis is not adequate for engineering education research questions centered on outcomes such as identity, belonging, and mentorship. While structural audits are valuable for data cleaning, instrument diagnostics, and descriptive reporting, they do not address how students and mentors experience programs or how developmental and socio-cultural factors shape outcomes. This limitation aligns with critiques that AI, as currently designed, struggles to grasp socio-cultural phenomena and affective constructs in education [1], [2].

5.4 Human Interpretation Aligns with Theory and Lived Experience

Human interpretation differed fundamentally from AI interpretation because it incorporated both disciplinary theory and tacit understanding of student development in the RAMP program context. Whereas AI models drew primarily from observed patterns in the datasets, the human researcher contextualized those patterns within established literature on engineering identity, mentorship, and belonging, as well as prior qualitative and PAR-based work with similar cohorts [13], [16].

For example, the human analyst, a sociologist with expertise in PAR and ethnography interpreted ceiling effects in identity-related items not as evidence of program failure to create change, but as a reflection of self-selection into RAMP program by already highly motivated and engineering-committed students. This interpretation is consistent with findings that bridge programs often attract students with substantial preexisting interest and identity investment [5], [10].

Similarly, where ChatGPT and Claude inferred that higher post-program interest scores signaled "identity growth," the human researcher recognized that Likert-level increases are not meaningful when scores begin near the maximum. This required domain knowledge that high pre-scores constrain the capacity to observe numeric growth and therefore call for alternative indicators, such as qualitative reflections, narratives of role clarification, or subsequent retention data.

The human analyst also brought developmental insight that AI systems did not consistently articulate. Differences between high-school and first-year students—for instance, the former emphasizing exploration ("learning what engineering is") and the latter emphasizing consolidation and navigation ("building networks," "strengthening preparation for the major")—align with identity theories that distinguish exploratory from consolidating phases and with mentorship literature that highlights role- and age-specific needs [9], [11]. AI models noticed these subgroup patterns inconsistently and tended to collapse them into a single narrative of "increased interest."

Finally, human interpretation appropriately weighted the limitations of the data. Mentor perceptions of increased student engagement were treated as valuable but subjective accounts of the mentoring experience, not as direct evidence of engagement trajectories. Human researchers adjusted interpretations in light of small sample sizes ($n \approx 5$ for mentors), possible social desirability bias, and the absence of long-term tracking. Claude, by contrast, frequently extrapolated from these perceptions to broad causal statements (e.g., "the program significantly increased student interest").

In sum, human reasoning integrated theory, caution, and contextualized meaning-making in ways AI systems could not fully replicate.

5.5 Implications for Engineering Education Research

This study demonstrates that AI tools are promising but incomplete partners for engineering education research. Their strengths lie in rapid data summarization, structural audits, and the generation of preliminary interpretations that can prompt new questions. Yet the analysis also reveals critical implications for how these tools should be integrated into research workflows.

First, AI-driven analysis must be explicitly labeled by lens. As seen in this study, numeric-only and contextualized prompts produce fundamentally different outputs, and these outputs should not be conflated. An AI-generated result under a non-contextual lens should be used for understanding distributions, missingness, and general trends, but not for drawing conclusions about program impact.

Second, engineering education researchers must remain vigilant about AI's tendency toward overinterpretation, particularly in contextualized settings. Claude's interpretation of descriptive changes as evidence of causal mechanisms, or ChatGPT's subtle tendency to treat identity increases as meaningful in the presence of ceiling effects, illustrates how AI may inadvertently mislead inexperienced researchers. For engineering education research to maintain rigor, AI-generated interpretations should undergo human verification grounded in established literature, particularly when addressing mentorship, identity, or belonging [8], [7].

Third, findings underscore the importance of theoretical triangulation. AI systems have no inherent understanding of engineering identity theory, mentoring frameworks, or student development; they only infer patterns statistically or semantically. Therefore, AI outputs should never substitute for theoretical frameworks but should instead be positioned as inputs into theory-driven human analysis.

Fourth, the variability across models demonstrates the need for transparency in reporting AI contributions. Different models (Perplexity vs. Claude) produced different interpretations from identical datasets under identical prompts, meaning research that uses AI must document which model was used, how it was prompted, and how outputs were interpreted or edited.

Finally, these findings highlight the importance of clearly structured human–AI collaboration in engineering education research. We are not suggesting that researchers must adopt AI tools; rather, we argue that when AI systems are used for analytic support, they should be embedded within transparent, theory-grounded workflows that preserve human interpretive authority. Prior scholarship on AI in education emphasizes the need for bounded, ethically informed integration of automated tools rather than uncritical adoption [1], [2], [17]. Within this framing, hybrid human–AI workflows can leverage AI's strengths in rapid pattern recognition and synthesis while ensuring that theory alignment, evidentiary restraint, and contextual nuance remain under human oversight. The goal is not automation of interpretation, but augmentation of analytic efficiency within clearly defined epistemic guardrails.

For example, guardrails that could be used in this study include: (1) requiring the model to label every claim as descriptive (directly observed in frequencies or distributions) versus interpretive (theory-linked inference), consistent with calls for transparent and bounded integration of AI in educational contexts [1], [2], [17]; (2) mandating an explicit evidence trace for each interpretive claim (i.e., identifying the specific survey items, subgroup counts, or qualitative excerpts that support the claim), in line with recommendations to mitigate unintended consequences and overgeneralization in machine learning systems [18], [20]; (3) prohibiting causal language and impact statements unless supported by appropriate research design and statistical testing, particularly in small-N contexts susceptible to overinterpretation [22], [18]; (4) enforcing a "limitations-first" reporting structure that foregrounds sample size constraints, missingness, and ceiling effects prior to interpretation [1], [2]; and (5) requiring a human adjudication step in which interpretive claims are evaluated against established frameworks of engineering identity, belonging, mentorship, and leadership development [3], [4], [5], [23].

The findings also have implications for engineering leadership development. If mentorship and project-based learning operate as mechanisms for identity clarification and professional socialization, they may simultaneously contribute to leadership competency and leadership identity formation. Engineering Leadership Orientation research suggests that students develop leadership orientations through engagement in authentic engineering practice, responsibility-sharing, and relational influence rather than through formal leadership coursework alone [23]. The mentorship dynamics observed in this study—particularly those involving exposure to industry professionals and collaborative problem solving—align with experiential models of leadership development identified in the Troost iLead research program [24]. AI-generated interpretations of such outcomes therefore carry additional weight: overinterpretation or causal overclaiming may influence how leadership development is assessed or reported in engineering education contexts.

This balanced approach aligns with calls in the literature for ethical and epistemologically sound integration of AI in STEM learning [1], [2]. Ultimately, engineering education cannot rely solely on algorithmic interpretation but should use AI as a support tool that enhances—not replaces—the deep contextual understanding that human researchers and participants bring.

Beyond individual research workflows, these findings have implications for institutions increasingly adopting AI tools for assessment, accreditation, and program evaluation. Without explicit guardrails distinguishing descriptive analysis from interpretive inference, AI-generated reports risk overstating program impact or masking nuanced developmental outcomes. Engineering education programs that rely on AI-supported analytics should therefore ensure that human expertise remains central in interpreting results related to identity, belonging, and mentoring—particularly when such interpretations inform resource allocation or program redesign.

6. Limitations

This study has several limitations that must be acknowledged. First, the sample sizes across student and mentor datasets were small, particularly for mentor pre/post surveys ($n \approx 5–6$). AI systems often treat small-N patterns with unwarranted confidence, especially models like Claude that default to narrative synthesis. This limits the generalizability of any interpretation whether generated by AI or humans—and constrains statistical inference. Future research should test AI behavior on larger, more variable datasets to identify how sample size affects interpretive drift. Additionally, commercial AI systems may incorporate reinforcement learning from human feedback (RLHF), moderation layers, or partial human review in their training and deployment pipelines. While these sociotechnical layers are not transparent, this study evaluates outputs as they are delivered to end users. Any embedded human influence would affect all prompt conditions consistently and therefore does not compromise the internal comparison of analytic lenses, though it underscores that AI systems are not purely autonomous agents.

Second, the study did not evaluate the correctness or numerical accuracy of AI-produced descriptive statistics. Instead, the focus was on interpretive tendencies. This is consistent with the purpose of the study but means that findings do not speak to AI reliability in quantitative accuracy. Additional work should examine how these models handle incorrect, missing, or messy numeric data and whether their interpretations depend on statistical fidelity.

Third, although two distinct prompting regimes were tested, variations in phrasing, context, or prompt structure could produce different AI behaviors. Prompt engineering is itself a methodological variable, and developing standardized prompts for educational research remains a challenge.

Fourth, data coding of open-ended survey responses was completed by one member of our research team, who used both inductive and deductive approaches. While using multiple coders can provide added perspectives and allow for calculation of inter-rater reliability, it may also limit interpretive insight and may not be feasible for small research teams [26]. That said, as our study expands, it may benefit from using multiple coders to identify areas of agreement or divergence from AI outputs. The intent is that future studies could systematically document the coding process and coder agreement to strengthen the credibility of claims about AI interpretive behavior.

Fifth, the study does not address potential biases within AI training data that may influence interpretation of mentorship, belonging, or engineering identity constructs. Models trained on broad internet data may inadvertently reinforce stereotypes about STEM ability or belongingness. These risks require deeper examination.

Finally, the datasets originate from a single-site program and cannot represent the diversity of bridge program structures, student populations, or mentoring arrangements nationwide. Care must be taken not to generalize findings about AI interpretive behavior beyond the scope of this study's context.

7. Future Work

Future research should expand this line of inquiry in several ways. First, studies should test additional AI models, including open-source variants (such as Llama-3 or Mixtral) and domain-tuned educational models. This would reveal whether observed tendencies—such as Perplexity's numeric conservatism or Claude's narrative drift—are model-specific or indicative of broader architectural patterns.

Second, as AI systems evolve rapidly, longitudinal tracking of model behavior is essential. Interpretive tendencies in one version of a model may change in future versions, meaning that research must remain vigilant about model drift over time. Monitoring stability across versions will help researchers understand whether improvements in contextual reasoning reduce the interpretive gaps identified in this study.

Third, future studies should investigate equity implications of AI interpretation. Identity, belonging, and mentorship are experienced differently across demographic groups. AI systems might systematically overinterpret or underinterpret experiences of marginalized students. Examining how AI handles race, gender, first-generation status, or intersectional identities will be essential for ensuring ethical and equitable research practices.

Fourth, future work should aim to create structured frameworks for hybrid AI–human analysis workflows. These frameworks might include: (1) standardized non-contextual numeric summaries generated by AI; (2) contextual interpretation protocols that require explicit tie-back to numeric evidence; (3) human adjudication layers ensuring validity and contextual sensitivity; and (4) transparent reporting conventions that disclose how AI was used in each stage.

Fifth, an important area for future study is the applicability of AI tools to qualitative analysis. Models like Claude and ChatGPT show promise in coding, aggregating, and interpreting large volumes of text data. However, their risk of overgeneralization and inability to verify meaning through human empathy or cultural understanding raises questions about trustworthiness. Research should examine how to leverage AI for first-cycle coding while reserving deeper interpretation for human researchers.

Finally, future work could investigate how AI might support program improvement efforts by generating design hypotheses, identifying overlooked patterns, or highlighting tensions in student or mentor narratives. While AI cannot replace human insight, it may serve as a catalyst for reflective program redesign when used responsibly and in partnership with participants.

8. Conclusion

This study provides a novel comparison of AI interpretive behavior in the context of engineering education research. By applying three AI models to the same datasets under controlled prompt conditions, we reveal that model design and prompting structure profoundly influence interpretation. Non-contextual AI outputs were accurate but shallow; contextual AI outputs were rich but prone to overreach. Human interpretation remains essential for recognizing the meaning behind student and mentor experiences, especially in mentorship-driven, identity-centered programs like RAMP program.

Our findings highlight the need for transparency in AI-supported educational research and argue for deliberate hybrid approaches that combine computational efficiency with human contextual insight. As AI continues to influence research and assessment practices, engineering education must develop rigorous frameworks that ensure interpretive validity, safeguard against overgeneralization, and preserve the human understanding central to identity and mentorship processes.

References

[1] W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign, 2019.
[2] R. Luckin, W. Holmes, M. Griffiths, and L. B. Forcier, *Intelligence Unleashed: An Argument for AI in Education*. Pearson, 2016.
[3] A. Godwin, "The development of a measure of engineering identity," in *Proceedings of the ASEE Annual Conference & Exposition*, 2016.
[4] T. L. Strayhorn, *College Students' Sense of Belonging*, 2nd ed. Routledge, 2018.
[5] G. Crisp and I. Cruz, "Mentoring college students: A critical review of the literature between 1990 and 2007," *Research in Higher Education*, vol. 50, no. 6, pp. 525–545, 2009.
[6] B. W.-L. Packard, *Successful STEM Mentoring Initiatives for Underrepresented Students: A Research-Based Guide for Faculty and Administrators*. Stylus Publishing, 2016.
[7] K. L. Tonso, "Engineering identity," in A. Johri and B. M. Olds, Eds., *Cambridge Handbook of Engineering Education Research*. Cambridge University Press, 2014, pp. 267–282.
[8] A. Patrick and M. Borrego, "A review of engineering identity research," in *Proceedings of the 2016 ASEE Annual Conference & Exposition*, 2016.
[9] H. B. Carlone and A. Johnson, "Understanding the science experiences of successful women of color: Science identity as an analytic lens," *Journal of Research in Science Teaching*, vol. 44, no. 8, pp. 1187–1218, 2007.
[10] A. L. Zydney, J. S. Bennett, A. Shahid, and K. W. Bauer, "Impact of undergraduate research experience in engineering," *Journal of Engineering Education*, vol. 91, no. 2, pp. 151–157, 2002.
[11] H. Thiry and S. L. Laursen, "The role of student–advisor interactions in apprenticing undergraduate researchers," *CBE—Life Sciences Education*, vol. 10, no. 3, pp. 228–246, 2011.
[12] A. Collins, J. S. Brown, and S. E. Newman, "Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics," in L. B. Resnick, Ed., *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*. Lawrence Erlbaum Associates, 1989, pp. 453–494.
[13] S. Tripathy, K. Chandra, and D. Reichlen, "Participatory Action Research (PAR) as formative assessment of a STEM summer bridge program," in *Proceedings of the 2020 ASEE Virtual Annual Conference & Exposition*, 2020.
[14] S. Tripathy, K. Chandra, H.-Y. Hsu, Y. Li, and D. Reichlen, "Engaging women engineering undergraduates as peer facilitators in participatory action research focus groups," in *Proceedings of the 2021 ASEE Virtual Annual Conference & Exposition*, 2021.
[15] E. Deterding, N. Agyeman, S. Tripathy Thomson, C. Keough, S. Lewis, and K. Chandra, "Inclusive teamwork: Using Participatory Action Research (PAR) to improve teamwork projects in Intro to Mechanical Engineering," in *Proceedings of the ASEE Northeast Section Conference (ASEE-NE 2022)*, 2022.
[16] K. Chandra, S. Tripathy Thomson, S. Lewis, and N. Sahila, "Building research, teamwork, and professional skills in an engineering summer bridge program: Reflections towards an allyship model," in *Proceedings of the 2024 ASEE Annual Conference & Exposition*, 2024.

[17] I. Roll and R. Wylie, “Evolution and revolution in artificial intelligence in education,” *International Journal of Artificial Intelligence in Education*, vol. 26, no. 2, pp. 582–599, 2016.
[18] H. Suresh and J. Guttag, “A framework for understanding unintended consequences of machine learning,” *arXiv preprint* arXiv:1901.10002, 2021.
[19] P. Lewis *et al*., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Proceedings of NeurIPS*, 2020.
[20] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in *Proceedings of FAccT*, 2021.
[21] R. Bommasani *et al*., *On the Opportunities and Risks of Foundation Models*. Stanford CRFM, 2021.
[22] Z. Ji *et al*., “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, 2023.
[23] C. Rottmann, R. Sacks, and D. Reeve, “Engineering leadership: Grounding leadership theory in engineers’ professional identities,” *Journal of Engineering Education*, vol. 104, no. 4, pp. 351–373, 2015.
[24] A. M. Chan and C. Rottmann, “Leadership development for engineers: What do students learn?” *Journal of Engineering Education*, vol. 107, no. 2, pp. 204–228, 2018.
[25] D. Reeve, C. Rottmann, and R. Sacks, “The role of experiential learning in developing engineering leadership,” in *Proceedings of the ASEE Annual Conference & Exposition*, 2015.
[26] D. Keene, “Spotlight on qualitative methods: Do I need multiple coders?” Interdisciplinary Association for Population Health Science. Accessed: Feb., 2026. [Online]. Available: https://iaphs.org/demystifying-the-second-coder/