Exploring the use of Generative AI for reliable qualitative coding: The Novice Experience

Ms. Melis Ozdemir, University of Virginia

Melis Ozdemir is an undergraduate student in biomedical engineering at the University of Virginia. She conducts research in engineering education and global health and has coauthored multiple peer-reviewed publications and conference papers. She serves as a teaching assistant in first-year engineering and has held leadership roles in student organizations focused on civic discourse and mentorship. Her interests lie at the intersection of technology, policy, and engineering education, particularly in the context of health and societal impact.

Dr. Anne Marguerite McAlister, University of Virginia

Dr. Anne M. McAlister is an Assistant Professor in the First Year Engineering Center at the University of Virginia. She has her PhD in Education and MS in Systems Engineering from the University of Virginia, and her BS in Chemical Engineering from The Ohio State University. Her research focuses primarily on identity and qualitative research methods.

Mr. William Harry Keenan, University of Virginia

William Keenan is a 3rd year undergraduate student at the University of Virginia majoring in Systems Engineering. His main research explores teamwork and conflict in first-year engineering design teams.

Exploring the use of Generative AI for reliable qualitative coding: The Novice Experience

Abstract

This Work-in-Progress paper explores using generative artificial intelligence (AI) to support qualitative coding in engineering education. Qualitative analysis generates meaning from rich, descriptive data through iterative engagement, yet it is labor intensive in large quantities. While generative AI may streamline elements of coding and organization, its use raises questions about ethics, validity, transparency, and preserving researcher interpretation. We present a practical, process-oriented approach for novice researchers that integrated AI as a collaborative assistant rather than a replacement for human judgment. We share our workflow and prompt development, explain how reflexivity and positionality guided our choices, discuss when AI was helpful or misleading, and detail how we recorded decisions for transparency. We offer early lessons learned and ethical boundaries. This WIP contributes to a grounded account of human-in-the-loop AI for qualitative analysis, emphasizing that critical thinking and interpretive engagement remain central to trustworthy findings.

Background

Qualitative analysis plays a critical role in engineering education research by enabling insight into participants' experiences and meaning making that are not readily captured through quantitative measures [1], [4]. Qualitative methods are rigorous and often time intensive, particularly in studies with a large amount of data. Generative AI offers new possibilities for supporting qualitative analysis through, for example, proposing preliminary codes and surfacing potential patterns. At the same time, its use introduces concerns related to validity, transparency, and the preservation of researcher interpretation [1]-[4]. This WIP presents a practical, transparent workflow for integrating AI as a collaborative assistant rather than a substitute for researcher judgment in the development and application of qualitative codes. Our goals are to share strategies accessible to novice researchers, reflect on benefits and limitations observed in practice, and articulate ethical boundaries that support credibility, reflexivity, and accountability.

Prior work in engineering education explores the use of computational tools to support qualitative analysis, including Natural Language Processing based systems for organizing open ended responses and mapping them to established analytic frameworks [1], [2]. These approaches aim to reduce repetitive labor while maintaining human authority over interpretation. Emerging studies on the use of large language models for qualitative analysis report similar affordances, such as faster first pass coding and pattern identification, alongside persistent risks of conflated constructs, loss of nuance, and overgeneralization without careful human oversight [3], [4]. Ethical concerns, including potential bias in model generated suggestions, researcher responsibility for analytic decisions, and the need for transparency to support validation and auditability, underscore the importance of clear prompting practices, version control, and memoing when using AI in qualitative workflows [5], [6]. AI may assist with organization and workload, but meaning making must remain a human responsibility.

Qualitative methodology literature emphasizes that analytic rigor derives from sustained engagement with data, reflexivity, and explicit documentation of analytic decision making, rather than from automation or standardization alone [7]. Research on computer assisted qualitative data analysis similarly frames software as a support for organization, retrieval, and transparency, while locating interpretive authority firmly with the researcher [8], [9]. Coding is understood as a cyclical, theory informed process shaped by comparison, memoing, and discernment [9]. From this perspective, generative AI introduces a methodological tension. While computational tools may promote efficiency, the credibility of qualitative findings depends on how researchers interrogate, justify, and document analytic choices. Situating generative AI within this broader tradition highlights the importance of human participance in the workflows [8].

To examine these issues in practice, we used an example dataset and drew on Expectancy Value Theory to analyze how engineering students reflected on their views of their work and their life. Expectancy Value Theory differentiates expectancy, or anticipated success, from value, including intrinsic value, attainment value, utility value, and cost as two components that determine motivation [10-12]. These constructs guided the development of a codebook and provided a principled framework for evaluating AI-generated coding suggestions.

Research Questions

We focus on process-level inquiry in the following research questions:

RQ1: How can novice qualitative researchers integrate accessible generative AI tools into qualitative coding workflows while preserving methodological rigor and researcher interpretation?
RQ2: What benefits and limitations emerge when generative AI is used to support qualitative coding with a theory-informed codebook?
RQ3: What reflexive and verification practices are necessary to ensure ethical, transparent, and rigorous AI-assisted qualitative analysis?

To address these questions, we designed a human-in-the-loop workflow anchored to a theory-informed codebook, used Copilot+ with structured prompts for first-pass coding, and verified AI outputs through dual human review, memo-ing, and reconciliation, with all prompt versions and decisions recorded [4], [6].

Methods

The data we used to model the AI processes in this study come from students in a two-semester introductory engineering course at a large, public, research-focused university in the mid-Atlantic. Students wrote short reflections in response to the prompt, "where do you views about work and life complement one another, where do they clash, and does one drive the other?" Our purpose in analyzing this data was to document how novice AI users can use generative AI tools to support qualitative coding while preserving researcher-led interpretation.

Students' written reflections were compiled into a data table (e.g., Table 1). Each student's written response was preserved as a distinct row in the table and associated with the student' a unique, anonymized ID. All AI interactions were conducted within Microsoft Copilot+, an institutionally sanctioned, protected platform.

We positioned generative AI as an assistant for early-stage organization and first-pass coding, while preserving interpretive authority with the research team [1], [2], [4]. All AI-supported steps were anchored to a theory-informed codebook that drew on Expectancy-Value Theory (Table 2).

Table 1: Sample of Compiled Reflection Data

ID	Reflection Text
567	Society won't always value what I value. I may find happiness in actions society doesn't. Sometimes work can take away time from the things that make you happy.
567	My life and work should produce the maximum amount of happiness for myself and society. They drive each other because both are motivated by the same principle.
569	My work view and lifeviews are similar in that I think they both need to serve the needs of people. I would like to help other people through my job and I want to care for people in my personal life as well.
569	For workview, I said I like the feeling of accomplishment, but I think that I enjoy spending time with friends and family more. It is hard to do both together because accomplishing things requires putting in the hours.

Table 2. Codebook for AI-Assisted First-Pass Coding

Category	Code	Abbreviated Definition
Motivational Orientation	Personal Passion	Motivation grounded in personal interest or enjoyment.
	Social Responsibility	Motivation tied to contributing to society or others.
	Economic Prestige	Emphasis on salary, financial security, or material outcomes.
	Social Prestige	Motivation related to status, recognition, or respect.
EVT Component	Expectancy	Anticipated success or confidence in ability.
	Attainment	Achievement framed as central to identity.
	Intrinsic	Enjoyment or fulfillment derived from the work.
	Utility	Usefulness of work for future goals.
	Cost	Tradeoffs in time, effort, or opportunity.
Work–Life Framing	Work Drives Life	Work is described as shaping life goals or values.
	Life Drives Work	Life goals or values shape career choices.
	Balanced	Work and life described as aligned or mutually supportive.

We constrained Copilot+ with a structured prompt to apply only predefined codes, avoid inference beyond the text, return "no code" when evidence was insufficient, and justify every suggested label with a verbatim quote and a one-sentence rationale aligned to the definition (Table 3) [3], [4]. Reflections were processed in small clusters rather than all at once.

Table 3: Final Prompt provided to Copilot+

You are assisting with qualitative coding of written student reflections. Apply only the predefined codes from the codebook provided below. Do not infer meaning that is not explicitly stated in the text. If no code clearly applies, return "no code." Context: Students responded to: 1) Where do your views on work and life complement one another? 2) Where do they clash? 3) Does one drive the other? If so, how? Input files and structure: - WorkviewLifeview.csv → Column A: ID; Column B: Reflection text (one response per row) - Codebook.csv → Column A: Code; Column B: Definition For each row in WorkviewLifeview.csv: 1) Assign applicable code(s) from the codebook (comma-separated). 2) Provide a short verbatim quote that warrants each code. 3) Provide a one-sentence rationale linking the quote to the definition. Output as CSV with columns: [ID] [Text (unchanged)] [Code(s)] [Quoted Evidence] [Rationale]

An abridged example of the AI output is shown in Table 4, illustrating how the participant ID, excerpt, codes and quoted evidence were given. The full audit trail includes the full original text, as well as additional columns for documenting reconciliation decisions.

Table 4. Example of AI Output

ID	Excerpt (abridged)	Final codes	Evidence cue (verbatim)
173	"Work drives life… a job will allow me to travel."	Work Drives Life; Utility	"work to live"; "allow me to travel"
176	"Desire for a slower pace but career-driven ladder climbing."	Cost; Social Prestige; Conflicting	"slower-paced lifestyle" vs "climbing the ladder"
181	"A task you do for compensation."	Utility	"compensation"
183	"Adhere to your moral compass in work and life… help the environment."	Social Responsibility	"moral compass" and "help the environment"

AI suggestions were treated as provisional and subjected to human-in-the-loop verification. Two researchers independently reviewed AI-assigned codes, checking that quoted evidence warranted the assigned codes and aligned with code definitions. Disagreements were resolved through discussion and documented in memos. Special attention was paid to constructs that researchers noticed were frequently conflated by the AI. For example, codes for Utility and Expectancy were frequently conflated [2], [4]. In the end, all student reflections were coded first by AI using the Final Prompt (Table 3) and the Final Codebook (Table 2), then reviewed and reconciled by two researchers.

To support reproducibility, all prompt versions, cluster IDs, and codebook versions were archived alongside outputs. We logged reconciliation decisions in a table that documented the AI-suggested label, the final label, and a one-sentence rationale, as shown in Table 5 [4], [6]. This documentation supported transparency and communication amongst the research team and re-running a small cluster with slight prompt variants.

Table 5: Illustrative examples of AI-assisted coding and human reconciliation decisions

ID	Excerpt (abridged)	AI-suggested code(s)	Final code(s)	Reconciliation rationale
174	"Work is a tool to achieve the life I want."	Expectancy	Utility	The statement frames work instrumentally rather than expressing confidence in success; Utility better aligned with the code definition.
567	"Society won't always value what I value."	Social Prestige	no code	The excerpt critiques societal values but does not reference status, recognition, or prestige; no predefined code was sufficiently warranted.
569	"I want to help other people through my job."	Social Responsibility; Attainment	Social Responsibility	Helping others is explicit, but the excerpt does not frame this motivation as identity-defining; Attainment was removed due to insufficient evidence.
567	"Work can take away time from the things that make you happy."	Cost; Conflicting	Cost	The excerpt describes a tradeoff in time and effort but does not indicate an unresolved tension between work

We processed data in small clusters so that two researchers could review all AI code suggestions and finalize codes for each cluster in a single sitting, which improved recall of decision rules and increased consistency across excerpts. The "no code" option was normalized to discourage over-assignment. When the model returned "no code," we recorded a brief memo indicating whether the absence reflected insufficient evidence or an ambiguity to revisit during later theme development. Treating "no code" as a valid outcome reinforced the principle that codes should be warranted by the student reflection data, not by model confidence [3], [4].

Limitations

In this work in progress paper, we report lessons about workflow rather than claims about student outcomes. Our use of Copilot+ was shaped by institutional data-protection requirements, which limited experimentation with alternative tools. As novice AI users, our prompting strategy may be rudimentary. Large language models can produce confident but misaligned suggestions for adjacent constructs; we mitigated this risk through evidence requirements, small-cluster processing, and dual review, though residual misclassification remains possible [3], [4].

Positionality

The authors are novices in the use of AI-assisted research tools but acknowledge their growing ubiquity and potential to transform analytic processes. The authors also recognize that AI-generated outputs can be incomplete and flawed and therefore require human review and interpretation. The authors further acknowledge the ethical, sustainability, and broader socio-political tensions surrounding the use of AI technologies, including the substantial environmental and material resources required to support their operation. Within this context, the goals of this study included developing a more informed understanding of AI-assisted methods and critically examining how such tools may shape research practices within engineering education. We used collaborative coding and reconciliation meetings, required verbatim textual evidence for every code assignment, and memo-ed analytic decisions, including when AI suggestions were accepted, revised, or rejected [6]. These practices preserved researcher interpretation and supported transparency while using AI for organizational support.

Work-in-Progress Observations

Copilot+ aided with first-pass organization and initial code application when constrained by a theory-informed codebook and structured prompt; however, AI-assigned code suggestions frequently required human review, especially when coding passages that involved interpreting relationships between students' beliefs, values, and actions or distinguishing among closely related EVT constructs [3], [4]. Code clarity and prompt specificity were central to rigor; requiring verbatim evidence for each code application reduced over-assignment and made adjudication more efficient. Common conflations such as Utility vs. Expectancy and Intrinsic vs. Attainment reinforced the need for simple decision rules and human oversight. Prompt refinements and small-cluster processing was an effective method for our team to approach the use of AI in qualitative coding because most errors reflected systematic boundary ambiguities rather than random variation. Iterative prompt refinements clarified construct distinctions and evidence requirements, reducing recurring misclassifications, while small, consistent clusters supported reviewer recall. In contrast, re-running unchanged prompts often reproduced the same decision patterns and imposed additional verification burden without improving alignment. Independent review, reconciliation, and memo-ing supported researcher reflexivity and ensured decisions made by researchers were grounded in evidence from the dataset. We do not claim accuracy or efficiency at scale; our observations suggest that accessible AI tools may support organizational aspects of qualitative coding when embedded in a transparent, theory-guided, researcher-led workflow.

Typical places where AI failed included labeling instrumental statements as expectancy, interpreting generalized expressions of satisfaction as attainment rather than intrinsic value (which, per EVT, refers to enjoyment or interest in the activity itself, whereas attainment value concerns the importance of doing well because it affirms one's identity), and treating rhetorical contrasts between work and life as evidence of conflict, even when the student later resolved the tension. Each case required returning to the code definitions and the participant's written words. In practice, most corrections were straightforward once the AI was required to provide quoted evidence, and its outputs were checked against the original data. This pattern suggests that improved accuracy of AI-assigned codes depend less on model switching and more on maintaining theory-aligned definitions, evidence-first prompts, and disciplined verification.

A theory-informed codebook, constrained prompting, and systematic human verification enabled novice AI-users to use an accessible AI tool in first-pass coding while preserving interpretive rigor. The workflow was most helpful for organization and candidate code surfacing; finer distinctions and theme development remained researcher led.

Discussion

The findings highlight how deliberate design choices in an AI-assisted, researcher-led workflow can support theory-informed, evidence-based qualitative coding, particularly for student teams and early-career researchers. In this study, a shared prompt template and a "no code" policy structured AI use to support theory-informed, evidence-based coding without delegating interpretive decisions to the system [6], [13]. Together, these practices keep interpretation with the researcher, align with common expectations for qualitative rigor, and make it feasible to handle larger text sets without expanding team size. This approach offers a practical pathway for novice AI users to integrate generative AI into qualitative analysis. Treating AI as infrastructure rather than interpreter enabled faster early-stage organization while preserving researcher authority through evidence requirements and transparent verification. Simple prompt guardrails that apply only predefined codes, allow "no code," and justify each label with a quote were effective. EVT distinctions were prone to conflation, underscoring the importance of a strong codebook and human oversight. Accessibility and data protection led us to use Copilot+, which limited configuration but encouraged disciplined prompting, clustering, and audit trails. The process invited reflexivity; memo-ing and collaborative reconciliation helped examine both AI behavior and researcher assumptions.

Conclusions

This WIP presents an accessible, reproducible workflow for using generative AI to support qualitative coding while preserving rigor. With a theory-informed codebook, constrained prompts that require quoted evidence, and systematic human verification, novice AI users were able to integrate AI into their qualitative analysis processes. The approach was most effective for organization and candidate code surfacing; higher-order interpretation remained researcher led. These provisional results suggest that careful workflow design and transparent documentation are essential.

References

[1] A. Katz, M. Norris, A. M. Alsharif, M. D. Klopfer, D. B. Knight, and J. R. Grohs, "Using natural language processing to facilitate student feedback analysis," in *Proc. ASEE Annu. Conf. Exposition*, Virtual, 2021.

[2] T. Rietz and A. Maedche, "Cody: An AI-based system to semi-automate coding for qualitative research," in *Proc. ACM CHI Conf. Human Factors Comput. Syst. (CHI)*, Yokohama, Japan, 2021, Art. no. 394, doi: 10.1145/3411764.3445591.

[3] Y. Gamieldien, J. M. Case, and A. Katz, "Advancing qualitative analysis: An exploration of the potentials of generative AI and NLP in thematic coding," *SSRN preprint*, 2023, doi: 10.2139/ssrn.4487768.

[4] D. L. Morgan, "Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT," *International Journal of Qualitative Methods*, vol. 22, pp. 1–10, 2023, doi: 10.1177/16094069231211248.

[5] Y. Katz, "Manufacturing an artificial intelligence revolution," *SSRN preprint*, 2017, doi: 10.2139/ssrn.3078224.

[6] M. Nicmanis and H. Spurrier, "Getting started with artificial intelligence assisted qualitative analysis: An introductory guide to qualitative research approaches with exploratory examples from reflexive content analysis," *International Journal of Qualitative Methods*, vol. 24, pp. 1–14, 2025, doi: 10.1177/16094069251354863.

[7] J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 5th ed. Thousand Oaks, CA, USA: SAGE Publications, 2018.

[8] S. Knaflic and C. M. Richards, "Using qualitative data analysis software to support analytic rigor," *Qualitative Research Journal*, vol. 12, no. 2, pp. 101–113, 2012, doi: 10.1108/14439881211248385.

[9] J. Saldaña, *The Coding Manual for Qualitative Researchers*, 3rd ed. Thousand Oaks, CA, USA: SAGE Publications, 2016.

[10] A. Wigfield, "Expectancy–value theory of achievement motivation: A developmental perspective," *Educational Psychology Review*, vol. 6, no. 1, pp. 49–78, Mar. 1994, doi: 10.1007/BF02209024.

[11] A. Wigfield and J. S. Eccles, "Expectancy–value theory of achievement motivation," *Contemporary Educational Psychology*, vol. 25, no. 1, pp. 68–81, 2000, doi: 10.1006/ceps.1999.1015.

[12] A. Wigfield, K. Muenks, and J. S. Eccles, “Achievement motivation: What we know and where we are going,” *Annual Review of Developmental Psychology*, vol. 3, pp. 87–111, Dec. 2021, doi: 10.1146/annurev-devpsych-050720-103500.

[13] K. Stolpe and J. Hallström, “Artificial intelligence literacy for technology education,” *Computers and Education Open*, vol. 6, p. 100159, 2024, doi: 10.1016/j.caeo.2024.100159.

Data Governance Note

All AI interactions occurred within Microsoft Copilot+ under institutional data protection. Raw reflections were anonymized before use, and no data was entered into public tools.