Integrating Topic Modeling and LLM Prompt Engineering into a Human-driven Approach to Analyze Interview Transcripts

Teresa M. Ober
ETS
Princeton, NJ, USA
tober@ets.org

Karyssa A. Courey
Rice University
Houston, TX, USA
kac23@rice.edu

Michael Flor
ETS
Princeton, NJ, USA
mflor@ets.org

Topic modeling has become a widely used unsupervised machine learning method for extracting latent themes from large textual datasets. However, the interpretability of these themes often relies heavily on human judgment, which can limit transparency and reproducibility. Recent advances in large language models (LLMs) and prompt engineering offer new opportunities to enhance the interpretability and scalability of topic modeling outputs. This study presents a hybrid, human-in-the-loop methodological framework that integrates topic modeling, LLM prompting, and human-derived codes to support rigorous qualitative analysis. We apply this framework to focus group interviews with 13 U.S. teachers discussing the conceptualization and assessment of communication and digital literacy skills within competency-based education (CBE) contexts. The multi-stage process includes semantic clustering, LLM-assisted topic labeling, and iterative codebook refinement, enabling both scale and interpretive depth. Our findings demonstrate that this approach supports construct alignment, thematic stability, and methodological transparency, while preserving the contextual richness of qualitative data. We also highlight the importance of human oversight in guiding LLM outputs and ensuring theoretical coherence. This work contributes to emerging best practices for integrating AI tools into qualitative educational research by offering a replicable approach for analyzing complex, open-ended data that maintains both scalability and interpretability. The framework demonstrates how computational tools can augment human interpretive expertise while maintaining the epistemological integrity essential to qualitative inquiry. Supplemental materials are available at: https://doi.org/10.17605/osf.io/4q6w8.

Keywords: large language models, topic modeling, qualitative analysis, skills assessment, human-AI

1. INTRODUCTION

The integration of artificial intelligence (AI), particularly large language models (LLMs), into qualitative analysis has the apparent promise of enabling researchers to analyze large textual datasets and identify complex patterns reflecting variation in student learning with

unprecedented speed and scale (see Schroeder et al., 2025). One perceived advantage of using LLMs for qualitative data analysis is the promise of efficiency. Traditional approaches to qualitative analysis that rely on manual coding processes can be personnel and time-intensive and may produce codes that are subject to interpretive variability (Dai et al., 2023). These constraints become particularly problematic when analyzing large amounts of open-ended qualitative data.

Computational methods, particularly hybrid human-machine learning approaches, have shown promise in enhancing the efficiency and reliability of qualitative coding without displacing the interpretive depth of human analysis. For instance, one study found that integrating models like BERT into iterative coding processes can significantly reduce manual effort while maintaining or even improving coding reliability (Li et al., 2021). Digital tools and emerging technologies have also transformed qualitative research practices, enabling new forms of data collection, management, and analysis, though they introduce unique challenges that require careful methodological consideration (Magida, 2024). While computational methods can amplify the extraction of information from rich qualitative data, the emergence of meaning remains rooted in interpretive processes rather than analytical scale alone.

Qualitative research methods that involve hybrid human-machine learning approaches offer new avenues for theoretical exploration and data interpretation, allowing researchers to engage with qualitative data in more interactive and dynamic ways (Hayes, 2025). However, this computational augmentation introduces significant interpretability challenges that complicate the validation of findings within qualitative paradigms. LLMs operate as opaque systems, often producing outputs through processes that are not easily traceable or explainable, which can make it difficult to understand how themes are identified (Singh et al., 2024). This lack of transparency raises concerns not only about methodological reproducibility, but also about the epistemological integrity of qualitative inquiry, where understanding how interpretations are constructed is central to establishing validity and meaning.

The opacity of LLMs presents significant challenges for qualitative research, particularly in maintaining theme stability and interpretive coherence. Such opacity can obscure the traceability of themes back to original data sources, raising concerns about the consistency and trustworthiness of thematic structures across analytical iterations (Braun & Clarke, 2006). The methodological integrity in qualitative research depends on transparency and fidelity to the subject matter, both of which are undermined when the analytical process is not easily interpretable (Chatfield & Debois, 2022). While LLMs offer advantages in applying codes consistently, they often fall short of capturing the nuanced, context-dependent meanings that humans are adept at identifying without deliberate prompting (Gratsanis et al., 2025). This limitation becomes especially pronounced in interpretive tasks requiring deep contextual understanding, where human analysts demonstrate superior capacity for discerning subtle thematic distinctions (Morgan, 2023). Such challenges may be particularly notable in social and behavioral research, where understanding participant perspectives requires preserving contextual meaning while maintaining methodological rigor.

1.1. Collaborative Models of Human-AI Partnership in Qualitative Analysis

Collaborative frameworks that combine human interpretive expertise with computational capabilities, rather than replacing human analysis with automated systems, may offer a path forward toward AI-assisted qualitative research. However, these partnerships require careful attention to workflow design, quality assurance strategies, and the preservation of human interpretive authority throughout the analytical process to maintain interpretability and analytical validity

(Hitch, 2024). The goal is therefore to create synergistic relationships where computational tools enhance rather than replace human analytical capabilities.

Empirical studies have demonstrated that hybrid approaches can outperform both human-only and fully automated methods. For instance, the results of a study comparing four approaches to inductive codebook development, fully manual human-only, fully automated ChatGPT-only, and two hybrid approaches, indicate that hybrid methodologies developed codebooks more efficiently than human-only methods while demonstrating higher inter-coder reliability and quality ratings from human experts (Barany et al., 2024). Importantly, the fully automated approach produced lower quality results, confirming that human oversight remains essential for analytical rigor and interpretability, while hybrid approaches achieved the benefits of both computational efficiency and human interpretive sophistication.

Research using methodological approaches that include structured workflows for engaging with computational tools, such as LLMs in qualitative analysis, may provide guidance for future implementation of collaborative frameworks that maintain interpretability. For example, CollabCoder, a GPT-powered workflow, assists collaborative qualitative analysis at three critical stages: suggesting initial codes during open coding, identifying disagreements between human coders during refinement, and suggesting code merging strategies during finalization (Gao et al., 2023). Evaluation of this structured approach demonstrates how systematic integration of AI tools can support rather than replace human analytical work.

Integrating computational tools into established qualitative methodologies requires adherence to foundational methodological principles. For example, combining expert-developed codebooks with GPT-3 for deductive coding has been shown to produce substantial agreement with human-coded results, emphasizing the importance of clear and comprehensive codebook design (Prescott et al., 2024; Xiao et al., 2023). Similarly, structured frameworks for leveraging ChatGPT within grounded theory methodology have emphasized the essential role of human researchers in driving theory development while using AI to support systematic analysis (Sinha et al., 2024). These approaches demonstrate how computational tools can enhance established methodological frameworks without compromising their theoretical foundations.

At the same time, ensuring quality in collaborative frameworks also involves rethinking traditional approaches to trustworthiness and reliability. As qualitative methodologies evolve to accommodate new analytical possibilities, especially through the application of computational tools, it may be necessary to reconsider traditional concepts in qualitative analysis, such as sample adequacy. Rather than relying solely on the conventional notion of theme saturation to determine sample adequacy, the concept of "information power" offers a more dynamic framework. This approach posits that the more relevant information a sample holds for the study aim, the fewer participants are needed, which remains pertinent even as computational tools enable the analysis of larger datasets (Malterud et al., 2016).

Recent guidance advocates for treating LLMs as additional coders within a team-based framework, emphasizing reflexive discussion and comparison over purely statistical measures of inter-coder reliability (Nicmanis & Spurrier, 2025). In this light, the outputs of LLMs used for qualitative coding are subject to a similar level of scrutiny as that of human coders, maintaining the discursive processes that characterize rigorous qualitative research.

Innovative strategies further point to the potential for using human-AI hybrid approaches for qualitative data analysis. For example, using LLMs to generate counterarguments or challenge emerging themes can help researchers test the robustness of their interpretations, identify gaps in evidence, and refine their analyses (Hayes, 2025). Used in this way, hybrid approaches to qualitative data analysis could help researchers become aware of their own biases in the process. However, when LLMs are not used in complement with a more reflexive and human-driven

process, hybrid approaches may be subject to hallucinations and biases built into various AI systems (Williams, 2024).

1.2. Methodological Foundations: From Traditional to Hybrid Qualitative Approaches

Traditional qualitative inquiry is grounded in iterative interpretation, contextual sensitivity, and reflexive engagement with data, where the researcher is considered a central instrument in the meaning-making process (Trent & Cho, 2020). Conducting thorough qualitative research that is assisted by computational tools requires a nuanced understanding of both its theoretical underpinnings and the tools that can augment human interpretive expertise. Computational tools offer complementary capabilities that can enhance traditional qualitative methods when properly integrated.

Central to a collaborative human-AI research paradigm for qualitative research is the recognition that different computational approaches offer distinct advantages for addressing specific analytical challenges. LLMs bring a unique set of strengths to qualitative research, namely, their capacity to summarize and label large volumes of text. Through prompt engineering, LLMs can generate topic summaries, suggest code labels and descriptors, and facilitate the interpretation of code categories, potentially augmenting human analysis and improving efficiency (Hayes, 2025; Player et al., 2025; Stammbach et al., 2023). When used collaboratively, such as in LLM-assisted codebook refinement or theme validation, prior research has shown that LLMs can improve inter-rater reliability and conceptual clarity, especially when paired with iterative human review (Barany et al., 2024; Chew et al., 2023). However, the effectiveness of using LLMs to support qualitative analysis depends on contextually framed prompts with clear construct definitions (H. Zhang et al., 2025), particularly for complex constructs (X. Liu et al., 2025). For example, research involving empirical comparisons of GPT-4 and human coders has shown that GPT-4 can achieve higher accuracy than human coders in identifying categories of complex constructs like student engagement on a task (McClure et al., 2024) or varying levels of knowledge in a particular domain (S. Zhang et al., 2025). However, more nuanced findings emerge from comparative analysis, revealing that GPT-4 exhibits superior performance on deductive tasks with clear codebooks and requires significant human refinement to achieve similar analytical depth on inductive tasks (X. Liu et al., 2025). These findings suggest that AI-assisted approaches can be effective but require a nuanced understanding of the coding task to ensure rigor. As such, the development of transparent, reflexive, and ethically grounded human-AI methodologies remains a critical frontier.

However, the pursuit of computational sophistication must be balanced with analytical coherence and interpretability. Topic modeling is an unsupervised machine learning technique that identifies latent themes within a collection of documents by grouping co-occurring words into topics that may represent the corpus content in a coherent and simplified manner (Churchill & Singh, 2022). This approach may provide a balance between computational scalability and interpretive clarity. Its mathematical foundations, particularly the use of probability distributions, enable researchers to systematically validate thematic structures and assess theme coherence through quantitative metrics (Blei et al., 2003; Maier et al., 2021). This structure contrasts with neural network-based approaches, which often function as opaque systems that limit interpretability. In contrast, topic modeling creates transparent pathways from data to analytical conclusions, supporting qualitative interpretation through reproducible and interpretable outputs.

A key advantage of topic modeling lies in the stability and consistency of theme identification, which is essential for iterative analytical refinement. This stability becomes especially important when integrating multiple computational tools, as it ensures that increasing computational complexity enhances rather than obscures analytical clarity. Traditional qualitative methods rely on human interpretation to iteratively develop themes and reach theoretical saturation, but this process can be undermined when AI-generated themes lack transparency. Topic modeling addresses this challenge by offering mathematically grounded, reproducible outputs that remain open to human interpretive refinement. This approach is particularly valuable in applied contexts such as educational research, where policy and practice decisions demand reliable analytical foundations and clear interpretive pathways that stakeholders can understand and validate.

Despite its strengths, topic modeling is not without limitations. Traditional approaches to topic modeling often struggle with short texts and sentence-level semantics, which can constrain interpretive depth (H. Wang et al., 2023). However, recent innovations, such as BERTopic and dynamic topic modeling, have improved topic coherence and contextual sensitivity, expanding the method's applicability (Jung et al., 2024). Even with these advances, topic modeling outputs require careful interpretation to ensure alignment with research questions and theoretical frameworks (Isoaho et al., 2021). These interpretive demands of topic modeling underscore the continued importance of human-guided decision-making in hybrid analytical frameworks. While topic modeling can identify patterns that may not be apparent through manual analysis, its outputs only become meaningful when contextualized by domain knowledge.

1.3. Toward a Model for Scalable and Interpretable Human-AI Collaborative Qualitative Analysis

Hybrid human-AI approaches may offer new pathways for addressing the analytical demands of large, complex datasets while preserving the interpretive depth that defines qualitative inquiry. Recent methodological innovations point to the potential of integrating topic modeling with LLMs to create scalable and interpretable frameworks for qualitative research. Topic modeling offers mathematical rigor, stability, and interpretability, while LLMs bring powerful natural language understanding and contextual reasoning capabilities. When integrated, these approaches can enhance both the stability and theoretical coherence of qualitative analysis.

Emerging hybrid frameworks illustrate how these complementary strengths can be operationalized. For instance, LLM-in-the-loop topic modeling (LLM-ITL) refines neural topic models by using LLMs to improve topic alignment and coherence through optimal transport-based alignment objectives (X. Yang et al., 2024). Similarly, *PromptTopic* uses LLMs to extract sentence-level themes, reducing the need for manual parameter tuning and enhancing topic relevance (H. Wang et al., 2023). These innovations demonstrate that successful AI integration in qualitative research requires more than computational scale, it demands theoretical grounding and methodological transparency (Forbus, 2019; Hayes, 2025).

Prior research has demonstrated that using multi-agent systems to deploy specialized LLM "agents" for tasks such as coding, synthesis, and refinement offers a potentially distributed approach to thematic analysis (Qiao et al., 2025; Sankaranarayanan et al., 2025; Yan et al., 2024). These architectures can manage large datasets while maintaining structured analytical processes, enabling the generation of more comprehensive thematic maps than single-system approaches. Topic modeling often serves as a foundational step in these pipelines, providing stable, mathematically grounded theme identification before LLM-assisted refinement (De Paoli, 2024). When combining human judgement and oversight, this collaborative model supports analytical

frameworks that are scalable, interpretable, and theoretically grounded. Such workflows may preserve the centrality of human judgment while leveraging computational efficiency (Yan et al., 2024; X. Yang et al., 2024). The present study builds on these emerging frameworks by demonstrating a practical integration of topic modeling, LLM prompting, and human coding that maintains interpretability while enabling analysis at scale. Rather than evaluating this approach against alternative methods, we focus on illustrating how these tools can be systematically integrated to support transparent, reproducible qualitative analysis in educational contexts where both scale and contextual sensitivity are essential.

1.4. Qualitative Analysis at Scale to Support an Understanding of Contextually Rich Educational Approaches

Educational research presents complex challenges for qualitative analysis, requiring methods that can capture deeply contextualized knowledge of classroom environments and subject areas. This is particularly challenging when research aims to operate at scales that may support decisions for policy and practice. Within a competency-based education (CBE) approach, assessing key skills exemplifies these challenges, as educators must develop reliable measures for complex skills that students are likely to demonstrate in unique ways.

Educational approaches adhering to CBE principles emphasize demonstrable mastery of competencies over traditional time-based progression, offering frameworks for preparing students for post-secondary education and careers (Levine & Patrick, 2019; O. Liu et al., 2023). However, widespread CBE adoption faces challenges in defining and assessing complex, non-cognitive skills in ways that are both reliable and sensitive to diverse educational contexts (Evans et al., 2020). State and district initiatives across Utah, North Carolina, and Colorado, among others, have emphasized skills like communication, digital literacy, and critical thinking (Atwell & Tucker, 2024). Developing usable assessments for these skills requires input from teachers, who provide critical insights into how competencies manifest in instruction and how they might be practically measured (Marion et al., 2020; Sturgis & Casey, 2018). This necessitates large-scale data collection, yet traditional analytical approaches may be insufficient for the scale and complexity of the analytical task. Such deeply contextualized circumstances surrounding CBE assessment pose a challenge that offers an ideal research opportunity for developing and testing a hybrid methodological approach that maintains interpretability while increasing analytical scale.

Central to this methodological challenge is the nature of teacher professional knowledge, which plays a pivotal role in shaping how competencies are interpreted, taught, and assessed in practice. Teachers' knowledge comprises explicit pedagogical theories, tacit experiential insights, and context-sensitive adaptations developed through practice (Elliott et al., 2011; Männikkö & Husu, 2019). This multifaceted nature of competencies presents analytical challenges that purely computational or decontextualized methods struggle to address. For instance, the persistent difficulty in evaluating competencies central to CBE highlights the limitations of such approaches (Evans et al., 2020). Thus, understanding contextually rich insights that are rooted in teachers' expertise requires methodologies capable of capturing both the explicit and tacit dimensions of professional knowledge, including the situated, intuitive, and often unarticulated aspects that guide instructional decision-making in real time.

Large-scale data collection in educational research introduces significant methodological challenges, particularly in balancing analytical scale with interpretive depth. As noted previously, traditional manual coding methods offer rich, context-sensitive insights but are time-consuming and subject to interpretive variability, which limits their scalability in studies of

teacher perspectives and instructional practice (Dai et al., 2023). Conversely, purely computational approaches, such as topic modeling, often sacrifice the contextual nuance essential for educational applications (Hayes, 2025), especially given approaches such as CBE, where meaning is deeply embedded in practice. This methodological tension underscores the need for hybrid approaches that can preserve interpretability while enabling large-scale analysis.

Recent systematic reviews and empirical studies have documented the growing use of LLMs in qualitative educational research. These models have been applied to tasks such as transcription, thematic coding, and pattern recognition, offering efficiency gains and new interpretive possibilities (Xing et al., 2025). However, prior research also highlights critical limitations, including inconsistent reporting standards, limited validation of AI-generated findings, and a lack of research on LLM use in resource-constrained educational settings. These gaps point to the need for methodological frameworks that are both theoretically grounded and practically feasible, particularly in highly contextualized domains like CBE. The goal then of creating such a methodological framework is to ensure it is simultaneously more capable and more interpretable than either purely human or purely computational approaches (Singh et al., 2024). This framework may enable researchers to analyze teacher perspectives at a scale that captures regional and disciplinary variation, while preserving the contextual nuance essential for understanding how competencies are demonstrated by students differently across classroom settings.

2. Research Aims

The purpose of the investigation described in subsequent sections is to develop a multi-stage analytical framework that integrates grounded human coding, machine learning-based topic modeling, and LLM prompt engineering to support qualitative analysis at scale. The methodological aim is to demonstrate how these tools, when used iteratively, can enable scalable and conceptually grounded analysis of open-ended qualitative data while maintaining interpretability and thematic stability. The proposed framework positions human decision-making as the theoretical anchor to ensure alignment with research constructs and interpretive goals. As described in the previous section, topic modeling provides mathematically grounded, reproducible theme identification, while LLMs contribute to pattern recognition and interpretive refinement. This integration addresses key challenges in AI-assisted qualitative research by supporting analytical transparency, stability, and systematic validation of thematic structures.

The framework is applied to analyze teachers' perspectives on the conceptualization and assessment of key CBE skills, including communication and digital literacy, which are the focus of the present study. This emphasis serves two objectives:

1. To examine how educators define, prioritize, and assess 21st-century competencies.
2. To demonstrate how integrating topic modeling and LLM prompting within a human-in-the-loop framework can support scalable, interpretable thematic analysis while maintaining construct alignment and methodological transparency.

In this way, the findings contribute both methodological guidance for integrating AI tools into qualitative workflows and substantive insights into the implementation of highly contextual pedagogical approaches such as CBE. The application of this framework illustrates how hybrid approaches can support rigorous, interpretable analysis of complex qualitative data at scale, while preserving the contextual richness essential to educational inquiry.

3. METHODS

3.1. DATA COLLECTION

Data were collected during the 2023–2024 school year through six semi-structured focus group interviews involving 13 middle and high school teachers from across the United States. Prior to data collection, Institutional Review Board (IRB) approval was secured. The interview protocol explored teachers' experiences and insights related to the conceptualization and assessment of core 21st-century skills. All focus groups were held via Zoom, and each included 2-4 middle school or high school instructors currently teaching a variety of subjects with diverse experience ranging from one year to 32 years. A total of 13 teachers participated in the focus groups (see Table 1).

Table 1: Background characteristics of teachers interviewed in the focus groups (*N*=13).

Category	Subcategory	Counts and Percentages (%)
Subject area	English and Language Arts	2 (15%)
	Foreign Languages	1 (8%)
	Mathematics	3 (23%)
	Natural Science	5 (38%)
	Social Sciences	2 (15%)
Teaching experience (in years)	0-10	2 (15%)
	11-20	4 (31%)
	21-30	3 (23%)
	31+	4 (31%)
Gender	Female	9 (69%)
	Male	4 (31%)
Region	West	0
	Midwest	4 (31%)
	South	2 (15%)
	Northeast	7 (54%)

During each focus group, the facilitator presented the framework followed by example assessment tasks for each skill (i.e., communication, digital literacy), inviting teachers to share their perspectives, opinions, and experiences (see Appendix A in Supplemental Materials for the interview protocol). In the context of the present discussion, we focus primarily on the sections of the interview that pertain to the skills of communication and digital literacy as illustrative examples of our methodological approach. During these interviews, teachers provided valuable insights on the feasibility and relevance of assessing the competencies as well as contextual considerations for implementing them in various educational environments.

The transcripts, which were automatically recorded using the Zoom software, were then manually cleaned to correct speaker attribution and remove facilitator turns. Content was organized into two main sections: one focusing on skill conceptualization and the other on skill assessment. Each speaker's turn was further segmented into individual sentences, which served as the unit of analysis for subsequent computational processing. Given the semi-structured nature of the interview, each conversation turn tended to focus exclusively on a particular skill. Prior to analysis, we tagged sections of the transcript with the respective skill (i.e., communication [tagged as "*COM*"], digital literacy [tagged as "*DL*"]) and the issue being discussed (i.e., framework,

assessment design). Analyses were conducted separately for each skill area with the goal of building an overarching coding framework across all skills.

3.2. Analytic Approach

To explore teachers' perspectives on the conceptualization and assessment of key competencies, this study employed a hybrid qualitative analysis framework that integrates both human-led and AI-assisted methods. This iterative approach allows for systematic identification, refinement, and synthesis of themes from complex qualitative data.

3.2.1. Preliminary Human Coding and Codebook Development

While topic modeling can identify latent themes, interpretation is inherently dependent on human expertise to resolve conceptual ambiguity. To ensure the interpretability of the topic modeling solution, we conducted an initial review of the interview transcripts, which were organized by focus group and segmented into sections concerning skill conceptualization and skill assessment. We used grounded theory approaches (see Glaser & Strauss, 2017) to identify high-level thematic codes based on the goals of the study and the interview protocol. This manual coding process allowed the research team to synthesize common patterns and generate a structured set of codes that reflected both anticipated and emerging themes in the data (see Table 2).

3.2.2. Topic Modeling

We employed a multi-step topic modeling process beginning with semantic embedding. Each sentence was converted into a dense vector using SentenceBERT (Reimers & Gurevych, 2019). These sentence embeddings were then clustered using Affinity Propagation (Frey & Dueck, 2007). We selected Affinity Propagation over k-means because it does not require pre-specifying the number of clusters, allowing themes to emerge organically from the data.

The topic modeling process resulted in two levels of clustering: specific clusters and higher-level superclusters (H-clusters). The algorithm identified 78 first-level clusters and 32 H-clusters for communication, and 79 and 32, respectively, for digital literacy. Cosine similarity was used as the distance metric, and sentences with a cosine similarity below 0.5 to their nearest neighbor were excluded as thematic outliers. This threshold was chosen to filter out sentences with weak semantic alignment, ensuring that only thematically coherent content was retained for analysis. For each cluster, we calculated a centroid vector that represented the average semantic content of all sentences within that cluster. We then identified representative keywords, or "bestwords," for each cluster by embedding individual words and computing their cosine similarity to the centroid vector using the same SentenceBERT model. The weights assigned to each word reflect the cosine similarity score, bounded between 0 and 1, with higher scores indicating stronger representativeness. The higher-level clusters of sentences were produced by treating specific clusters as units, representing them with centroid vectors and then using the Affinity Propagation algorithm to cluster those units. Figure 1 provides a schematic of this process, resulting in representative sentences and bestwords for each cluster. The goal of this process was to gather related clusters into more general thematic groups.

For each H-cluster, we then reviewed the sentence text and bestwords for each topic to determine whether they carried a distinct meaning. As a result of this process, we identified 18 H-clusters for communication and 16 for digital literacy that did not reflect meaningful or distinct themes. For example, such topics often reflected the use of common phrases used in spoken communication, such as "What?", "Okay," "Right, exactly," and "Yeah, I would agree," that

are primarily used to seek clarification or convey acknowledgement rather than a distinct concept. These H-clusters were dropped from the subsequent analysis, leaving 14 H-clusters reflecting meaningful topics for communication and 16 for digital literacy.

Table 2: Human-derived preliminary themes, descriptors, and example quotes or messages.

Theme	Short Descriptor	Example Quotes or Messages
Reaction to Skills Frameworks	Teachers' initial impressions of skills frameworks, including clarity, relevance, and applicability in classrooms.	*"[T]his definition makes total sense. But [...] this definition seems a little squishy and a little bit difficult to assess. Seeing some of these sub skills does make it easier, but I still feel pieces of it are a little more difficult to assess especially."*
Prioritizing Certain Skills	Teachers' perspectives on which skills are most essential for students and why.	*"I feel like [digital literacy] might be the most important skill for students in the current world that they live in. They really need to be able to evaluate that information critically …."*
Assessment of Skills	Covers multiple subthemes about how skills are assessed, including teacher experience, assessment features, equity, and challenges.	- *Experience assessing skills*: Teachers reported limited experience and shared practical challenges and successes. - *General assessment features:* Effective assessments were seen as valid, reliable, fair, and often dynamic. - *Skill-specific assessment features*: Different skills require tailored approaches (e.g., using constructed responses for certain skills). - *Equity and accessibility*: Assessments should be inclusive and avoid disadvantaging students with diverse needs. - *Assessment challenges*: Barriers included time, resources, and difficulty measuring complex skills.
Unique Disciplinary Considerations for Skills Assessments	Highlights the differences in assessing skills across subject areas and the need for discipline-specific approaches.	*"[P]articularly in a social studies classroom, these skills are really critical. I think in the past 5 years we've shifted to emphasize information management and evaluation a lot more than early in my career, just because the kids are so inundated with different digital resources just in their day-to-day lives."*
Progression of Certain Skills along a Continuum	Focuses on tracking students' skill development over time and how it varies among learners.	*"So many students […] have no clue or very little clue, and then others who […] talk about the bots and AI, and some of them, even, you know, do the coding."*
Shifts in Social and Cultural Contexts Affecting Skills Assessment	Examines how societal, technological, and cultural changes impact the value and assessment of skills.	- "*I feel like AI, in a sense, is almost like forcing us to go more that way as opposed to […] only submitting things in writing*." - Emphasis now on "presentation and public speaking skills."

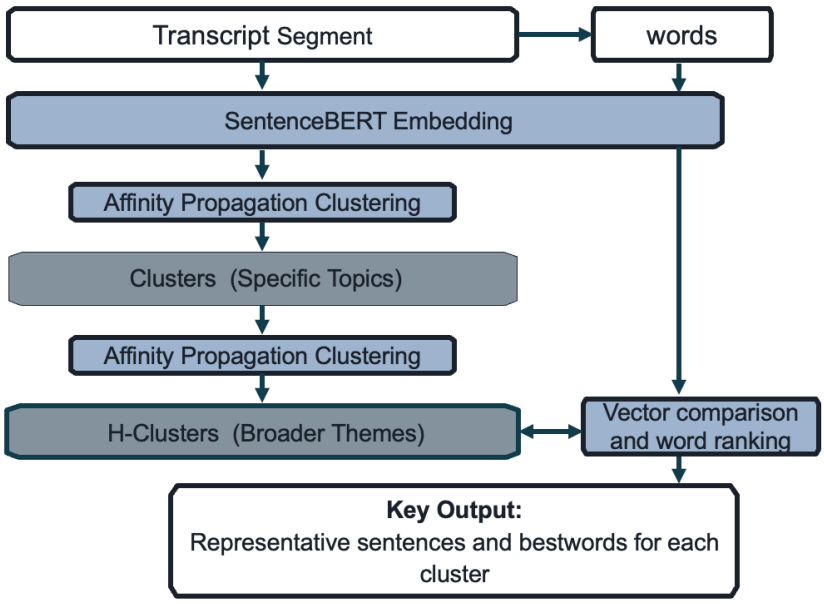

Figure 1: Schematic of the topic modeling process, including inputs and outputs.

3.2.3. LLM prompting for topic labeling

Once clustering was complete, we used a prompting strategy adapted from Barany et al. (2024) to generate topic labels and summaries (see Appendix B in Supplemental Materials). All prompt iterations were made by the research team. Representative sentences and bestwords (referred to as "keywords" in the prompt given the more common usage of the term) from each cluster were used to construct structured prompts for two GPT-4-based models: ChatGPT-4o (OpenAI) and Copilot (Microsoft). We selected two GPT-4-based models to assess consistency within the same model family while using commercially accessible tools. Both models were selected for their accessibility to educational researchers and their demonstrated capabilities in natural language understanding tasks. Furthermore, the selection of these two LLMs reflects a deliberate methodological choice grounded in accessibility, consistency, and replicability in practical contexts. Both ChatGPT-4o and Copilot are powered by the GPT-4 architecture, which offers a state-of-the-art balance between language understanding capabilities and widespread availability for educational researchers. While both models share the same underlying architecture, they differ in their implementation and interface, allowing us to assess the consistency of outputs across two commercially accessible GPT-4-based systems. This choice prioritizes practical replicability: researchers in educational settings can access these tools without specialized infrastructure or computational resources. Recent findings suggest that when paired with strategic prompt engineering and human-in-the-loop refinement, GPT-4-based models can produce interpretable and analytically robust outputs suitable for qualitative research applications (Zhao et al., 2023). By using two implementations of the same model family, we aimed to evaluate the stability of LLM-generated labels while maintaining a methodological approach that other educational researchers can readily adopt.

The objective of this step was to apply LLMs to generate concise, interpretable topic labels that could be compared with human-generated codes. As emphasized by X. Liu et al. (2025), the accuracy of LLM-generated labels improves when constructs are clearly defined and contextualized within the prompts. Outputs from both models were evaluated for semantic

consistency and minimal variation was observed, suggesting some stability in the LLM-based summaries.

3.2.4. LLM configuration and quality assurance

To ensure replicability and transparency, we documented information about the specific configurations and procedures used in the LLM-assisted topic labeling process. Two GPT-4-based models were employed: ChatGPT-4o (OpenAI, accessed October 2023–January 2024) and Copilot (Microsoft, powered by GPT-4, accessed during the same period). The default commercially available settings were used for both models and are shown in Table 3.

Table 3: Default model parameters.

Parameter	ChatGPT-4o (OpenAI API)	Copilot (Microsoft Web Interface)	Notes
Temperature	1	~1.0 (platform-managed)	Balances deterministic and creative outputs; moderate randomness.
Top-p	1	1	Inclusive sampling; considers full token probability distribution.
Frequency Penalty	0	0	No discouragement for repeated token usage.
Presence Penalty	0	0	No bias against new vs. existing content.
Max Tokens (output)	Up to 16,384	~4,096 (typical Azure default)	ChatGPT-4o supports longer responses; Copilot limits are managed internally and may vary by deployment.

To evaluate the consistency of LLM-generated labels, we compared outputs from ChatGPT-4o and Copilot for all clusters. Of the labels for the H-clusters analyzed, 33.3% (7 out of 21) received identical labels from both models for communication and 43.8% (7 out of 16) for digital literacy. Cosine similarity scores were computed using the same SentenceBERT model mentioned previously to assess semantic alignment between labels and descriptors assigned by each LLM relative to the other. Findings showed strong semantic alignment between models. For communication clusters, the average cosine similarity between the labels assigned by each LLM was 0.835 and for each description the average was 0.827. For digital literacy clusters, label similarity averaged 0.848, and the average similarity for descriptions was 0.880. These findings suggest strong similarity overall between the labels and descriptions assigned by each model. Of those that differed, there generally appeared to be minor variations in phrasing, but the core meaning was otherwise preserved (see Tables S2 and S3 in Supplemental Materials). All LLM-generated labels were further reviewed by the research team for accuracy, relevance, and alignment with the underlying sentence examples and bestwords. No clear instances of hallucination (i.e., fabricated content not grounded in the data) were observed.

3.2.5. Mapping of Human-Derived Themes and Data-Derived Codes for Further Refinement

The final step involved integrating the themes derived from human coding (Version 1) with codes generated through topic modeling, resulting in a refined and more comprehensive codebook (Version 2). Using additional LLM prompts (see Appendix B), we mapped the human-derived codebook onto the data-derived clusters, which reflected the results of a latent Dirichlet allocation (LDA) topic modeling analysis that produced words and sentence examples most representative of the key topics identified. Through this mapping process, we attempted to evaluate the alignment, overlap, and divergence between the two approaches.

The transition from Version 1 to Version 2 of the codebook represents a shift from an exploratory to a more structured and analytic approach to qualitative coding (see Table S1 in Supplemental Materials). Version 1 was developed early in the research process, based on initial impressions and preliminary review of teacher focus group transcripts. Following the topic modeling and LLM-prompting approach described above, topics that emerged were then integrated into the codebook to provide greater clarity, consistency, and codability. Codes were refined to reflect clearer distinctions between themes, and language was updated to be more specific and actionable. For example, general terms like "definitions" were expanded to include more precise questions, such as what components should be included in a skills framework. Subcategories were added to capture finer-grained differences, such as separating prioritization of whole skills from subskill prioritization. In addition, new categories such as "teaching" were introduced to reflect themes that emerged more strongly in the data.

Redundant or overlapping codes were combined or restructured for clarity. Specific examples included in Version 1 were removed from the main list in Version 2 to support generalizability. Contextual dimensions of the data, such as social, political, or historical considerations, were made more explicit in Version 2. Figure 2 illustrates this process. This iterative, human-in-the-loop strategy is supported by recent research showing that hybrid methods lead to more valid and interpretable codebooks (Barany et al., 2024; Dai et al., 2023).

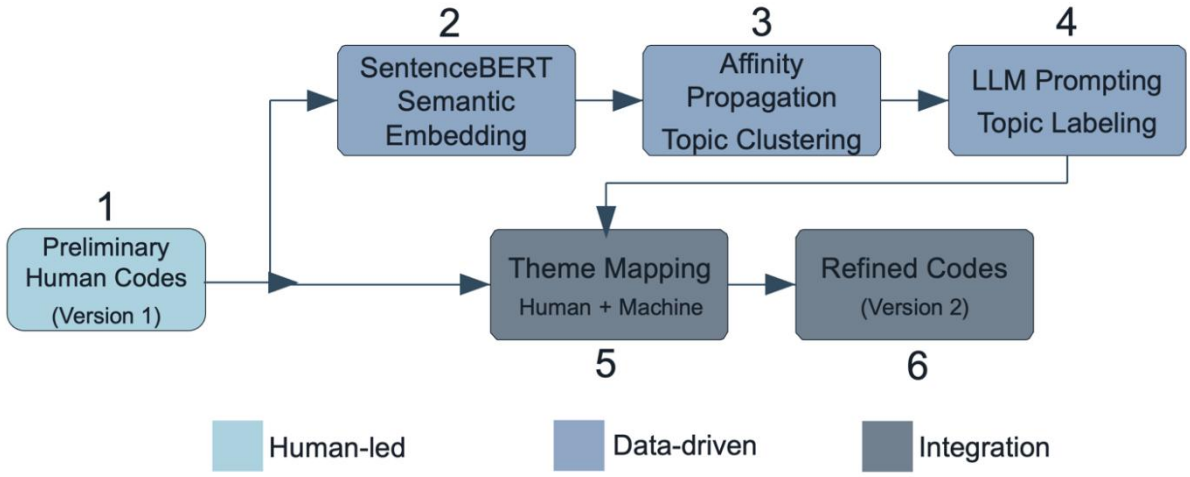

Figure 2: Schematic of the process for integrating insights from topic modeling and LLM topic labeling into the revised thematic codes.

4. RESULTS

As previously noted, we focus our results on the findings in relation to the skills of communication and digital literacy. Tables S2 and S3 in Supplemental Materials provide more information about each of the topics that were grouped in H-clusters, reflecting the respective parts of the interviews pertaining to each skill. We chose only topics that were grouped into H-clusters, given that any specific first-level clusters that were not gathered into H-clusters can be considered as thematic outliers. In this section, we present an analysis of the major themes that emerged with respect to the two competencies, aligned with the refined coding framework.

4.1. SKILLS FRAMEWORKS

Teachers emphasized that both communication and digital literacy require clear definitions and must be framed in ways that are pedagogically and contextually relevant. For communication, the concept was described as inherently multimodal, encompassing speaking, listening, writing, and presenting (Topic *COM3*; see Table S2 in Supplemental Materials). Teachers expressed concerns about vague or "squishy" frameworks, calling for clearer delineations and concrete examples (*COM13*). Audience awareness and social-emotional dimensions were also noted as integral to the definition (*COM26, COM11*).

For digital literacy, teachers stressed the importance of framing it as a fluid skill, shaped by the rapid evolution of digital tools. They highlighted the value of information evaluation, digital citizenship, as well as other skills such as critical thinking (Topics *DL1, DL11, DL17*; see Table S3 in Supplemental Materials). Definitions often emphasized the ability to manage, interpret, and ethically engage with digital information in everyday contexts.

Across both domains, educators consistently prioritized these skills as essential for student success. Communication was framed as foundational across grade levels and subject areas, with specific attention given to active listening and the ability to adapt communication strategies for different contexts (*COM14*, *COM7*, *COM4*). Digital literacy was similarly viewed as an essential skill, especially in an age of misinformation and AI-generated content (*DL5, DL18*). Teachers referenced the urgency of equipping students to navigate digital environments critically and responsibly.

4.2. ASSESSMENT

Teachers shared various assessment strategies, often describing both innovative and conventional formats. In terms of the skill of communication, assessment formats ranged from script writing and oral presentations to comparative evaluations of sample tasks (*COM21, COM29*). Teachers noted the value of assessing communication through authentic, contextualized activities. For digital literacy, teachers discussed video-based assessments, task comparisons, and performance-based evaluation tools (*DL11, DL13, DL16*). Teachers emphasized the importance of clear task instructions, with attention to digital tools that align with classroom realities.

Despite enthusiasm for assessment, teachers voiced concerns about validity, subjectivity, and feasibility. Communication posed challenges due to the nuanced, contextual, and interpersonal nature of the skill. Teachers noted the difficulty of assessing subtleties such as tone, audience alignment, or emotional intelligence in standardized formats (*COM16*). Similarly, digital literacy assessments were described as difficult to design in ways that capture critical reasoning and ethical use of digital tools (*DL15, DL3*). Teachers questioned how to ensure that assessments reflect real-world digital practices, especially when students vary widely in digital fluency.

Educators provided both positive and critical reflections on their experience assessing each skill. In communication, some teachers shared success stories with student engagement in presentations and collaborative work (*COM12, COM6*). Others reflected on efforts to refine rubrics or observation strategies to better capture communication growth. For digital literacy, teachers described designing authentic performance tasks that aligned with classroom instruction (*DL9*, *DL7*). They noted that assessing digital literacy worked best when embedded into content-based learning, such as research projects or digital source analysis.

4.3. Variation in Instruction, Skill Development, and Contexts

Instructional strategies were varied, but teachers described both competencies as deeply integrated into daily teaching. For communication, teachers referenced discipline-specific expectations, especially in STEM vs. humanities classrooms (*COM5*). Activities like presentations and collaborative discussions were common (*COM8*). Digital literacy was often taught through project-based and inquiry-driven instruction (*DL2, DL24*), with some content-specific adaptations (e.g., in social studies classes, *DL6*). Teachers emphasized that digital skills were most effectively taught in tandem with content-based tasks.

Teachers also noted significant variability in student skill development, often highlighting gaps and growth areas. In communication, students struggled particularly with oral communication, confidence, and audience engagement, while thriving with structured activities like script writing and video creation (*COM9*, *COM18*). In digital literacy, student skill levels ranged from advanced to underdeveloped. Teachers observed strong engagement with tools but weaker performance in evaluating sources or understanding digital ethics (*DL9*). They emphasized the need for repeated practice and feedback to foster progress.

In addition, teachers underscored how both competencies are shaped by broader contextual factors such as school culture, social norms, and technological shifts. Communication was framed within modern digital environments, with teachers noting students' heavy use of texting and social media as both a barrier and a resource for instruction (*COM17*, *COM22*). Teachers also emphasized the importance of teaching social and ethical norms (*COM4*). In digital literacy, the rapid rise of AI tools, increasing school diversity, and evolving media landscapes were cited as key challenges (*DL18*, *DL19*). Teachers expressed concern that curricula often lag behind technological change, underscoring the importance of adaptability.

5. Discussion

This study introduces and demonstrates a hybrid, human-in-the-loop framework that integrates topic modeling, LLM prompting, and human coding to support scalable and interpretable qualitative analysis. By applying this framework to focus-group interviews with teachers, we illustrate its capacity to surface nuanced insights into how educators conceptualize and assess communication and digital literacy within CBE contexts. The findings underscore the value of systematically combining computational tools with human interpretive depth, offering a methodological contribution that may help to address the challenge of maintaining both analytical scale and contextual rigor.

Our findings suggest that teachers consistently framed communication and digital literacy as foundational competencies for student success, echoing broader calls for 21st-century skills integration in education (Levine & Patrick, 2019; O. Liu et al., 2023). Communication was described as inherently multimodal, encompassing speaking, listening, writing, and presenting, with a strong emphasis on audience awareness and social-emotional dimensions (see *COM3, COM14, COM26*). These findings align with Morreale et al. (2024), who emphasize the importance of instructional communication competence and social presence in digital learning environments. Teachers' emphasis on ethical interaction and audience adaptation reinforces the need for assessments that capture these complex, relational dimensions of communication.

Digital literacy, in turn, was characterized as a dynamic skill set involving critical evaluation of information, ethical digital citizenship, and adaptability to emerging technologies (*DL1, DL5, DL17*). This characterization of the skill aligns with a growing recognition of the need for AI literacy curricula that prepare students for a rapidly evolving digital landscape without

undermining the opportunity to develop essential competencies (Wang & Lester, 2023). Teachers in this study echoed this urgency, highlighting the need for assessments that reflect real-world digital practices and support students in navigating AI-infused environments.

Despite recognizing the importance of these skills, teachers expressed concern about the feasibility and validity of assessing them. Challenges included capturing interpersonal nuance, evaluating digital ethics, and ensuring equity across diverse student populations (*COM16, DL15, DL3*). These concerns mirror longstanding difficulties in assessing non-cognitive skills, as documented by Merchant et al. (2018), who found substantial variability in how such skills are defined and reported across Canadian provinces. Similarly, Cheng and Zamarro (2018) demonstrated that teacher non-cognitive traits, such as conscientiousness, are predictive of student outcomes yet are often overlooked in traditional measures of teacher quality. The current study builds on this work by showing how teacher perspectives can be systematically analyzed to inform the design of more valid and context-sensitive assessments.

Teachers also described embedding these competencies into instruction through project-based learning and inquiry-driven tasks, while noting wide variability in student proficiency, particularly in oral communication and digital evaluation (*COM6, COM18, DL2, DL24*). These findings align with Hall and Trespalacios (2019), who found that personalized professional learning significantly improved teachers' self-efficacy in integrating technology into instruction. This study also revealed patterns of disciplinary variation in communication assessment (*COM5*) that might not emerge from smaller samples. For instance, STEM teachers emphasized technical precision while humanities teachers prioritized audience adaptation, insights that inform the need for discipline-specific assessment rubrics within CBE frameworks. The present study extends this insight by illustrating how hybrid analytic tools could support professional learning communities in reflecting on instructional strategies and assessment practices.

5.1. Limitations and Future Directions

Several limitations of the study warrant consideration. First, while topic modeling offers a stable foundation for theme identification, its outputs are sensitive to parameter tuning and may struggle with short or ambiguous text segments. Similarly, LLM-generated labels, though efficient, are subject to variability and may reflect anchoring biases or model-specific artifacts (Blackwell et al., 2024). The use of two closely related LLMs (ChatGPT-4o and Copilot) may also limit the diversity of interpretive perspectives. Future work should explore the use of alternative architectures, confidence scoring, and prompt variation to enhance robustness and reduce redundancy.

Second, while the human-in-the-loop design mitigates some risks of hallucination and misalignment, it does not eliminate them. The interpretive process still relies on the researcher's judgment to validate and refine outputs, which introduces subjectivity. Additionally, the sample was limited to 13 teachers, who were predominantly from the Northeast U.S. (54%) and female (69%). Expanding the participant pool to include a broader range of educators would strengthen the framework's applicability. Future research should further test this framework across diverse educational contexts and content areas, including other complex competencies such as perseverance and critical thinking.

Finally, our study design does not include comparative conditions (human-only or AI-only coding) necessary to empirically evaluate whether the hybrid approach outperforms alternatives. While we demonstrate the framework's application, claims about improvement relative to other methods remain conceptual rather than empirically validated in this study. In addition, researchers' conceptual understanding, with and without LLM assistance, was noted through discussion

only, and therefore, it is speculative. Future studies should build on this demonstration by conducting systematic comparisons across different analytical approaches (human-only, AI-only, and various hybrid configurations), measuring efficiency and validity outcomes, and testing the framework's applicability across diverse educational contexts and research questions. Such comparative work will be essential for establishing evidence-based guidelines for when and how to integrate AI tools into qualitative educational research.

5.2. IMPLICATIONS

Despite these noted limitations, this study makes several specific contributions to the emerging literature on AI-augmented qualitative research, while also acknowledging the boundaries of what can be claimed without comparative experimental design. This study may also contribute to the growing body of research on AI-augmented methods for qualitative data analysis by demonstrating how topic modeling and LLMs can be used in tandem to support rigorous, scalable, and interpretable thematic analysis. We empirically demonstrated an approach to integrate three distinct analytical approaches, grounded human coding, probabilistic topic modeling, and leveraging LLM-assisted labeling, while maintaining thematic stability and construct alignment throughout the analytical process.

Topic modeling provided a foundation for identifying latent themes (Maier et al., 2021), while LLM prompting facilitated the generation of concise, interpretable topic labels (Barany et al., 2024; Chew et al., 2023). Human oversight remained essential throughout, ensuring contextual alignment and mitigating risks of hallucination or conceptual drift (Nicmanis & Spurrier, 2025; Than et al., 2025). The transition from Codebook Version 1 (preliminary human-derived themes) to Version 2 (refined through integration with topic modeling outputs) illustrates how computational tools can support systematic codebook refinement. Specifically, the topic modeling process identified thematic clusters that were not explicitly represented in the initial human coding, leading to the addition of codes. Conversely, some preliminary codes were consolidated or refined based on the semantic clustering patterns, resulting in clearer construct boundaries and more actionable code definitions (see Table S1 in Supplemental Materials).

Beyond these methodological contributions, this study offers substantive insights into how educators conceptualize and assess complex competencies within CBE frameworks, insights that may have direct implications for assessment design, professional development, and policy implementation. The framework, and the flexibility it affords, is itself a potential contribution to future research. Teachers consistently identified tensions between the complexity of these competencies and the feasibility of assessing them in standardized formats (*COM16*, *DL15*). The framework's capacity to systematically analyze these concerns across multiple focus groups revealed that educators value assessments that are authentic, embedded in content-based tasks, and sensitive to developmental progression. These findings suggest that CBE assessment systems should prioritize performance-based measures that capture competencies as they manifest in real instructional contexts, rather than relying solely on decontextualized standardized tasks. There is also potential to integrate the framework with adaptive learning systems or personalized instruction platforms, enabling real-time feedback loops between qualitative insights and instructional design. Institutional policies and training programs are also needed to support responsible and ethical use of AI in qualitative research, particularly in education, where interpretive integrity is paramount.

Several themes that emerged regarding instructional strategies and student skill development (*COM8*, *DL2*, *DL24*) point to opportunities for targeted professional learning. Teachers expressed both confidence in embedding these competencies into daily instruction and uncertainty

about how to assess them reliably. Professional development initiatives could leverage the insights generated through this framework to create a shared understanding of competency definitions, develop discipline-specific assessment rubrics, and build communities of practice around evidence-based assessment strategies. The hybrid analytical approach demonstrated here could be applied iteratively to analyze teacher reflections and refine professional learning resources based on emerging needs. While applied here to understand perspectives for implementing CBE, it may be extended to other domains where rich and deeply contextualized phenomena may be explored through qualitative data analysis.

The concerns expressed by teachers about equity, accessibility, and the rapid evolution of digital tools (*DL18*, *DL19*) underscore the need for policy frameworks that support adaptive, context-sensitive implementation of CBE. The scale of analysis enabled by this framework makes it feasible to gather and synthesize teacher perspectives across diverse educational settings, informing state and district policies. As CBE initiatives continue to expand, systematic analysis of educator input at a scale that captures regional, demographic, and disciplinary variation will be essential for developing assessment systems that are both rigorous and equitable.

5.3. Conclusion

In the present study, we demonstrate a hybrid human-AI framework that supports rigorous, scalable, and interpretable qualitative research in education. By systematically integrating topic modeling, LLM prompting, and human coding, the framework offers a replicable method for analyzing complex qualitative data while preserving the contextual richness and theoretical grounding essential to educational inquiry. The application to teacher perspectives on CBE skills illustrates how this approach maintains thematic stability, construct alignment, and methodological transparency across the analytical process. Within the context of CBE, where skills like communication and digital literacy are both essential and difficult to assess, this approach provides a pathway for systematically incorporating teacher voice into assessment design. More broadly, it contributes to the evolving landscape of qualitative methodology by showing how computational tools can augment, rather than replace, human interpretation, judgement, and decision-making. As educational systems continue to grapple with the demands of scale, equity, and complexity, hybrid frameworks such as this one offer a promising direction for future research and practice.

6. Supplemental Material

Supplemental materials, including tables, referenced in this manuscript are available in the Open Science Framework repository here: https://doi.org/10.17605/osf.io/4q6w8. Data are available upon reasonable request.

Use of Generative AI

This study employed ChatGPT-4o (OpenAI) and Copilot (Microsoft) for topic labeling as described in Section 3.1.3. ChatGPT-4o was also used to assist with refining prompt wording, under direct human supervision and review. All prompts are provided in Appendix B (see Supplemental Materials). Generative AI was not used in manuscript writing, prompt development, or initial human coding.

REFERENCES

ATWELL, M. N.,AND TUCKER, A. 2024. Portraits of a Graduate: Strengthening career and college readiness through social and emotional skill development. *Collaborative for Academic, Social, and Emotional Learning*. https://files.eric.ed.gov/fulltext/ED641286.pdf

BARANY, A., NASIAR, N., PORTER, C., ZAMBRANO, A. F., ANDRES, A. L., BRIGHT, D., SHAH, M., LIU, A., GAO, S., ZHANG, J., MEHTA, S., CHOI, J., GIORDANO, C., AND BAKER, R. S. 2024, July. ChatGPT for education research: exploring the potential of large language models for qualitative codebook development. In *International conference on artificial intelligence in education*, 134–149. Cham: Springer Nature Switzerland. https://link.springer.com/chapter/10.1007/978-3-031-64299-9_10

BLACKWELL, R. E., BARRY, J., AND COHN, A. G. 2024. Towards reproducible LLM evaluation: Quantifying uncertainty in LLM benchmark scores. *arXiv preprint arXiv:2410.03492*. https://doi.org/10.48550/arXiv.2410.03492

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

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

CHATFIELD, S. L., AND DEBOIS, K. A. 2022. *Engaging students in socially constructed qualitative research pedagogies*, Strategies for collaborative classroom practice in qualitative data analysis. 234–252. https://doi.org/10.1163/9789004518438_015

CHENG, A., AND ZAMARRO, G. 2018. Measuring teacher non-cognitive skills and its impact on students: Insight from the Measures of Effective Teaching Longitudinal Database. *Economics of Education Review*, 64, 251–260. https://doi.org/10.1016/j.econedurev.2018.03.001

CHEW, R., BOLLENBACHER, J., WENGER, M., SPEER, J., AND ANAND, A. 2023. LLM-assisted content analysis: Using large language models to support deductive coding. *Behavior Research Methods*, 55(4), 1485–1497. https://arxiv.org/abs/2306.14924

CHURCHILL, R., AND SINGH, L. 2022. The evolution of topic modeling. *ACM Computing Surveys*, *54*(10s), 1–35. https://doi.org/10.1145/3507900

DAI, W., LIN, J., JIN, F., LI, T., TSAI, Y. S., GAŠEVIĆ, D., AND CHEN, G. 2023. Can large language models provide feedback to students? A case study on ChatGPT. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, N.-S. Chen, G. Rudolph, D. G. Sampson, M. Chang, R. Kuo, & A. Tlili, Eds., 323–327. https://ieeexplore.ieee.org/abstract/document/10260740

DE PAOLI, S. 2024. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, *42*(4), 997–1019. https://doi.org/10.1177/08944393231220483

ELLIOTT, J. G., STEMLER, S. E., STERNBERG, R. J., GRIGORENKO, E. L., AND HOFFMAN, N. 2011. The socially skilled teacher and the development of tacit knowledge. *British Educational Research Journal*, *37*(1), 83–103. https://doi.org/10.1080/01411920903420016

EVANS, C. M., LANDL, E., AND THOMPSON, J. 2020. Making sense of K-12 competency-based education: A systematic literature review of implementation and outcomes research from 2000 to 2019. *The Journal of Competency-Based Education*, *5*(4), e01228. https://doi.org/10.1002/cbe2.1228

FORBUS, K. D. 2019. *Qualitative representations: How people reason and learn about the continuous world*. MIT Press. https://direct.mit.edu/books/monograph/4167/Qualitative-RepresentationsHow-People-Reason-and

FREY, B. J., AND DUECK, D. 2007. Clustering by passing messages between data points. *Science*, *315*(5814), 972–976. https://doi.org/10.1126/science.1136800

GAO, T., DANG, A., AND REINECKE, K. 2023. CollabCoder: A GPT-powered workflow for collaborative qualitative analysis. In *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), M. D. Choudhury, X. Ding, S. Guha, A. F. P. de Carvalho, H. Kuzuoka, K. Reinecke, H.-C. Wang, & N. Yamashita, Eds., 1–27. https://arxiv.org/pdf/2304.07366

GLASER, B., AND STRAUSS, A. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge. https://doi.org/10.4324/9780203793206

GRATSANIS, P., KARYDIS, I., SIOUTAS, S., AND VONITSANOS, G. 2025. Human-AI Co-creation: LLMs, contextual hints, performance. In *Artificial Intelligence Applications and Innovations. AIAI 2025 IFIP WG 12.5 International Workshops,* A. Papaleonidas, E. Pimenidis, H. Papadopoulos, & I. Chochliouros, Eds. *AIAI 2025. IFIP Advances in Information and Communication Technology*, vol 754, 81–94. Springer, Cham, 7. https://doi.org/10.1007/978-3-031-97313-0_7

HALL, A. B., AND TRESPALACIOS, J. 2019. Personalized professional learning and teacher self-efficacy for integrating technology in K–12 classrooms. *Journal of Digital Learning in Teacher Education*, *35*(4), 221–235. https://doi.org/10.1080/21532974.2019.1647579

HAYES, A. S. 2025. "Conversing" with qualitative data: Enhancing qualitative research through Large Language Models (LLMs). *International Journal of Qualitative Methods*, *24*, 16094069251322346. https://doi.org/10.1177/16094069251322346

HITCH, D. 2024. Artificial intelligence augmented qualitative analysis: the way of the future?. *Qualitative Health Research*, *34*(7), 595-606. https://doi.org/10.1177/10497323231217392

ISOAHO, K., GRITSENKO, D., AND MÄKELÄ, E. 2021. Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal*, *49*(1), 300–324. https://doi.org/10.1111/psj.12343

JUNG, H. S., LEE, H., WOO, Y. S., BAEK, S. Y., AND KIM, J. H. 2024. Expansive data, extensive model: Investigating discussion topics around LLM through unsupervised machine learning in academic papers and news. *Plos One*, *19*(5), e0304680. https://doi.org/10.1371/journal.pone.0304680

LEVINE, E., AND PATRICK, S. 2019. *What is competency-based education? An updated definition.* Aurora Institute. https://aurora-institute.org/wp-content/uploads/what-is-competency-based-education-an-updated-definition-web.pdf

LI, Z., DOHAN, D., AND ABRAMSON, C. M. 2021. Qualitative coding in the computational era: A hybrid approach to improve reliability and reduce effort for coding ethnographic interviews. *Socius*, *7*, 23780231211062345. https://doi.org/10.1177/23780231211062345

LIU, O. L., KELL, H. J., LIU, L., LING, G., WANG, Y., WYLIE, C., SEVAK, A., SHERER, D., LEMAHIEU, P., AND KNOWLES, T. 2023. *A new vision for skills-based assessment*. Educational Testing Service. https://www.ets.org/pdfs/rd/new-vision-skills-based-assessment.pdf

LIU, X., ZAMBRANO, A. F., BAKER, R. S., BARANY, A., OCUMPAUGH, J., ZHANG, J., PANKIEWICZ, M., NASIAR, N., AND WEI, Z. 2025. Qualitative coding with GPT-4: Where it works better. *Journal of Learning Analytics*, *12*(1), 169–185. https://doi.org/10.18608/jla.2025.8575

MAGIDA, A. 2024. The Use of Digital Tools and Emerging Technologies in Qualitative Research—A Systematic Review of Literature. In *World Conference on Qualitative Research*, 257-269. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-65735-1_16

MAIER, D., WALDHERR, A., MILTNER, P., WIEDEMANN, G., NIEKLER, A., KEINERT, A., PFETSCH, B., HEYER, G. , REBER, U. , HÄUSSLER, T. , SCHMID-PETRI, H., AND ADAM, S. 2021. Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. In *Computational methods for communication science* (1st Ed), W. van Atteveldt & T.-Q. Peng, Eds., 13–38. New York: Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781003082606-2/applying-lda-topic-modeling-communication-research-toward-valid-reliable-methodology-daniel-maier-waldherr-miltner-wiedemann-niekler-keinert-pfetsch-heyer-reber-h%C3%A4ussler-schmid-petri-adam

MALTERUD, K., SIERSMA, V. D., AND GUASSORA, A. D. 2016. Sample size in qualitative interview studies: Guided by information power. *Qualitative Health Research*, *26*(13), 1753–1760. https://doi.org/10.1177/1049732315617444

MÄNNIKKÖ, I., AND HUSU, J. 2019. Examining teachers' adaptive expertise through personal practical theories. *Teaching and Teacher Education*, *77*, 126–137. https://doi.org/10.1016/j.tate.2018.09.016

MARION, S., WORTHEN, M., AND EVANS, C. 2020. *How systems of assessments aligned with competency-based education can support equity*. Aurora Institute and Center for Assessment. https://files.eric.ed.gov/fulltext/ED603989.pdf

MCCLURE, C., SMYSLOVA, O., HALL, A., AND JIANG, Y. 2024. Deductive coding's role in AI vs. human performance. In *Proceedings of the 17th International Conference on Educational Data Mining (EDM 2024)*, C. Demmans Epp, B. Paaßen, & D. Joyner, Eds.. https://educationaldatamining.org/edm2024/proceedings/2024.EDM-posters.91/

MERCHANT, S., KLINGER, D., AND LOVE, A. 2018. Assessing and reporting non-cognitive skills: A cross-Canada survey. *Canadian Journal of Educational Administration and Policy*, (187), 2–17. https://journalhosting.ucalgary.ca/index.php/cjeap/article/view/43135

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

MORREALE, S., LOWENTHAL, P., THORPE, J., AND OLESOVA, L. 2024. Instructional communication competence and instructor social presence: enhancing teaching and learning in the online environment. *Frontiers in Communication*, *9*, 1397570. https://doi.org/10.3389/fcomm.2024.1397570

NICMANIS, M., AND SPURRIER, H. 2025. 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*, *24*, 16094069251354863. https://doi.org/10.1177/16094069251354863

PLAYER, L., HUGHES, R., MITEV, K., WHITMARSH, L., DEMSKI, C., NASH, N., PAPAKONSTANTINOU, T., AND WILSON, M. 2025. The use of large language models for qualitative research: The Deep Computational Text Analyser (DECOTA). *Psychological Methods*. https://doi.org/10.1037/met0000753

PRESCOTT, M. R., YEAGER, S., HAM, L., RIVERA SALDANA, C. D., SERRANO, V., NAREZ, PALTIN, D., DELGADO, J., MOORE, D., AND MONTOYA, J. 2024. Comparing the efficacy and efficiency of human and generative AI: Qualitative thematic analyses. *JMIR AI*, *3*, e54482. https://ai.jmir.org/2024/1/e54482

QIAO, T., WALKER, C., CUNNINGHAM, C., AND KOH, Y. S. 2025. Thematic-LM: a LLM-based multi-agent system for large-scale thematic analysis. In *Proceedings of the ACM on Web Conference 2025*, G. Long, M. Blumestein, Y. Chang, L. Lewin-Eytan, H. Huang, & E. Yom-Tov, Eds., 649–658. https://doi.org/10.1145/3696410.3714595

REIMERS, N., AND GUREVYCH, I. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, R. Huang & S. Padó, Eds., 3982–3992. Association for Computational Linguistics, Hong Kong, China. https://arxiv.org/abs/1908.10084

SANKARANARAYANAN, S., BORCHERS, C., SIMON, S., TAJIK, E., ATAŞ, A. H., CELIK, B., AND BALZAN, F. 2025. *Automating thematic analysis with multi-agent LLM systems*. *OSF preprint*. https://osf.io/preprints/edarxiv/kq8zh_v1

SCHROEDER, H., AUBIN LE QUÉRÉ, M., RANDAZZO, C., MIMNO, D., AND SCHOENEBECK, S. 2025, April. Large Language Models in qualitative research: Uses, tensions, and intentions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, N. Yamashita, V. Evers, K. Yatani, X. Ding, B. Lee, M. Chetty, & P. Toups-Dugas, Eds., 1–17. https://doi.org/10.1145/3706598.3713120

SINGH, C., INALA, J. P., GALLEY, M., CARUANA, R., AND GAO, J. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*. https://arxiv.org/abs/2402.01761

SINHA, R., SOLOLA, I., NGUYEN, H., SWANSON, H., AND LAWRENCE, L. 2024, June. The role of generative AI in qualitative research: GPT-4's contributions to a grounded theory analysis. In *Proceedings of the 2024 Symposium on Learning, Design and Technology*, G. Arastoopour Irgens & H. Swanson, Eds., 17–25. https://doi.org/10.1145/3663433.3663456

STAMMBACH, D., ZOUHAR, V., HOYLE, A., SACHAN, M., AND ASH, E. 2023. Revisiting automated topic model evaluation with large language models. *arXiv preprint arXiv:2305.12152*. https://doi.org/10.48550/arXiv.2305.12152

STURGIS, C., AND CASEY, K. 2018. *Designing for equity: Leveraging competency-based education to ensure all students succeed.* CompetencyWorks Final Paper. iNACOL. https://files.eric.ed.gov/fulltext/ED589907.pdf

THAN, N., FAN, L., LAW, T., NELSON, L. K., AND MCCALL, L. 2025. Updating "The Future of Coding": Qualitative coding with generative Large Language Models. *Sociological Methods & Research*, *5*(3), 849–888. https://doi.org/10.1177/00491241251339188

TRENT, A., AND CHO, J. 2020. Interpretation in qualitative research: What, why, how. In *The Oxford Handbook of Qualitative Research* (2nd Ed.), P. Leavy, Ed., 956–982, Oxford Handbooks. https://doi.org/10.1093/oxfordhb/9780190847388.013.35

WANG, H., PRAKASH, N., HOANG, N. K., HEE, M. S., NASEEM, U., AND LEE, R. K. W. 2023. Prompting large language models for topic modeling. In *2023 IEEE International Conference on Big Data (BigData)*, A. Cuzzocrea & R. Agrawal, Eds., 1236–1241. https://doi.org/10.1109/BigData59044.2023.10386113

WANG, N., AND LESTER, J. 2023. K-12 Education in the age of AI: A call to action for K-12 AI literacy. *International Journal of Artificial Intelligence in Education*, *33*(2), 228-232. https://doi.org/10.1007/s40593-023-00358-x

WILLIAMS, R. T. 2024. Paradigm shifts: Exploring AI's influence on qualitative inquiry and analysis. *Frontiers in Research Metrics and Analytics*, *9*, 1331589. https://doi.org/10.3389/frma.2024.1331589

XIAO, Z., YUAN, X., LIAO, Q. V., ABDELGHANI, R., AND OUDEYER, P. Y. 2023, March. Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Proceedings of the 28th international conference on intelligent user interfaces*, F. Chen, M. Billinghurst, & M. Zhou, Eds., 75–78. https://doi.org/10.1145/3581754.3584136

XING, W., NIXON, N., CROSSLEY, S., DENNY, P., LAN, A., STAMPER, J., AND YU, Z. 2025. The use of Large Language Models in education. *International Journal of Artificial Intelligence in Education*, 35, 1–5. https://doi.org/10.1007/s40593-025-00457-x

YAN, L., ECHEVERRIA, V., FERNANDEZ-NIETO, G. M., JIN, Y., SWIECKI, Z., ZHAO, L., GAŠEVIĆ, D., AND MARTINEZ-MALDONADO, R. 2024. Human-AI collaboration in thematic analysis using ChatGPT: A user study and design recommendations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, N. Yamashita, V. Evers, K. Yatani, & X. Ding, Eds., 1–7. https://doi.org/10.1145/3613905.3650732

YANG, J., JIN, H., TANG, R., HAN, X., FENG, Q., JIANG, H., ZHONG, S., YIN, B., AND HU, X. 2024. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *ACM Transactions on Knowledge Discovery from Data*, *18*(6), 1–32. https://doi.org/10.1145/3649506

YANG, X., ZHAO, H., XU, W., QI, Y., LU, J., PHUNG, D., AND DU, L. 2024. Neural topic modeling with Large Language Models in the loop. *arXiv preprint arXiv:2411.08534*. https://arxiv.org/abs/2411.08534

ZHANG, H. E., WU, C., XIE, J., LYU, Y., CAI, J., AND CARROLL, J. M. 2025. Harnessing the power of AI in qualitative research: Exploring, using and redesigning ChatGPT. *Computers in Human Behavior: Artificial Humans*, *4*, 100144. https://doi.org/10.1016/j.chbah.2025.100144

ZHANG, S., MESHRAM, P. S., GANAPATHY PRASAD, P., ISRAEL, M., AND BHAT, S. 2025, February. An LLM-based framework for simulating, classifying, and correcting students' programming knowledge with the SOLO taxonomy. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education v.2*, J. A. Stone, T. Yuen, L. Shoop,

S. A. Rebelsky, & J. Prather, Eds., 1681–1682. https://dl.acm.org/doi/abs/10.1145/3641555.3705125

ZHAO, W. X., ZHOU, K., LI, J., TANG, T., WANG, X., HOU, Y., MIN, Y., ZHANG, B., ZHANG, J., DONG, Z., DU, Y., YANG, C., CHEN, Y., CHEN, Z., JIANG, J., REN, R., LI, Y., TANG, X., LIU, Z., … WEN, J. R. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, *1*(2). https://doi.org/10.48550/arXiv.2303.18223