Research Article

International Journal of Qualitative Methods
Volume 24: 1–14
© The Author(s) 2025
DOI: 10.1177/16094069251395943
journals.sagepub.com/home/ijq
Sage

Understanding Graduate School Through AI: A Scalable Approach to Thematic Coding

Miranda Shen[1], Jue Wu[2], Colette Patt[3], and Rodolfo Mendoza-Denton[3,4]

Abstract

This study explores the application of artificial intelligence (AI) in qualitative research, specifically examining how large language models (LLMs) can be utilized to code qualitative data and identify relationships among coder-defined themes. The approach is particularly useful for cases where researchers have previously-identified themes and hypotheses but lack the resources to code a large corpus of data manually. We outline a multi-step methodological framework grounded in qualitative research traditions, whereby researchers first conduct manual coding using a grounded theory approach (Charmaz, 2006; Glaser & Strauss, 1967) on a subset of the data. The resulting codes are then applied to the remaining data using a model-assisted process that integrates natural language processing, AI-based text classification (Noah et al., 2024), and topic identification. Lastly, this is followed by statistical analyses to test hypotheses and expected patterns, providing a robust approach to ensure reliability and accuracy. We illustrate this process through the systematic application of locally-run AI for coding interview transcripts related to graduate students' experiences in four Ph.D. programs at a large research university. We demonstrate how AI can improve the efficiency, consistency, and scalability of qualitative research without sacrificing confidentiality. This study highlights the potential for AI to enhance qualitative research processes while addressing challenges related to nuance and interpretation.

Keywords

artificial intelligence, large language models, qualitative analysis, thematic analysis, statistical validation, graduate student experiences

Introduction

In recent years, artificial intelligence (AI) has rapidly transformed a variety of fields, with advancements that were nearly unimaginable just a few years ago. Today, AI is embedded in almost every sector, including academia, where it is increasingly applied across disciplines ranging from engineering to the humanities (Whittaker et al., 2018). One of AI's most profound contributions lies in its capacity to process and analyze extensive amounts of data with high efficiency, enabling researchers to uncover patterns that might remain imperceptible to even the most experienced human analysts.

This capacity is particularly relevant to qualitative research, where one of the most promising applications of AI lies in analyzing large volumes of text-based data. Interviews, focus groups, and open-ended survey responses often produce rich but unwieldy datasets that require summarization and reduction–a process that is both methodologically demanding and labor-intensive. Traditionally, qualitative researchers have

[1]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA
[2]School of Human Development and Organizational Studies in Education, College of Education, University of Florida, Gainesville, FL, USA
[3]Division of Mathematical and Physical Sciences and College of Data Science, University of California, Berkeley, CA, USA
[4]Department of Psychology, University of California, Berkeley, CA, USA

Corresponding Author:
Rodolfo Mendoza-Denton, Department of Psychology, University of California, Berkeley, 2121 Berkeley Way, Room 3428, Berkeley, CA 94704, USA.
Email: rmd@berkeley.edu

had to manually code such data, developing categories or themes through iterative processes that require deep engagement and considerable time (Gamieldien et al., 2023). While foundational, manual coding is susceptible to challenges such as cognitive bias and interpretive inconsistency (Miles & Huberman, 1994). At the same time, it enables the recognition of nuance, contextual complexity, and meaning that automated systems alone may miss. In our approach, human coders first developed a codebook using a grounded approach (Charmaz, 2006; Glaser & Strauss, 1967), which was then applied to the full dataset using an AI-assisted classification process. The assumptions or oversights embedded in the human-coded framework thus influence AI outputs. Given that AI does not eliminate interpretive limitations, we incorporated manual validation and statistical testing to ensure the accuracy and reliability of AI-generated classifications.

The engine behind these advancements lies in large language models (LLMs), which are capable of generating and interpreting human-like text based on training on vast corpora (Brown et al., 2020; OpenAI, n.d.-a). These models hold the potential to revolutionize qualitative research by enhancing the speed, consistency, and scalability of thematic analysis. However, despite promising early results, their integration into established qualitative research methodologies remains in its infancy. Challenges persist around validation, contextual adaptation, and ensuring that the use of AI does not compromise analytical rigor or researcher intent.

Early evaluations of AI-assisted qualitative methods have demonstrated both potential and limitations. For example, Morgan (2023) explored ChatGPT's ability to replicate themes identified through manual coding and found that while the model was effective in extracting concrete, descriptive themes, it often struggled with capturing interpretive nuance. Similarly, Pattyn (2025) conducted a mixed-methods pilot study comparing generative AI tools (ChatGPT and Bard) with human coders. The study reported notable gains in efficiency, showing a fourfold reduction in coding effort and a fifteenfold reduction in throughput time, but also highlighted the model's challenges in handling ambiguity and subjectivity. These studies underscore the importance of human oversight when using AI in qualitative analysis.

Other scholars have assessed broader methodological implications. Gamieldien and colleagues (2023) conceptualized the potential of generative AI and natural language processing to expand the scope of qualitative inquiry. They emphasized gains in scalability and coding speed but warned of risks such as algorithmic opacity, model hallucination, and cultural bias. More recently, Dunivin (2025) proposed a field guide for integrating LLMs into interpretive work, suggesting workflows grounded in hermeneutic principles that maintain human insight during prompt design and code refinement.

Several new empirical studies further advance this area. Dai et al. (2023) introduced a "LLM-in-the-loop" method and found that hybrid human-AI teams could replicate human-coded themes with high reliability and substantial time savings. Noah et al. (2024) proposed a two-step LLM classification process for subjective experiential reports, which informed the design of our binary-to-topical approach. Similarly, Bennis and Mouwafaq (2025) evaluated multiple generative models for medical interviews, finding that performance varied significantly by domain, prompting calls for rigorous human validation across contexts. Katz et al. (2024) also contributed an open-source workflow (GATOS) that allows for inductive code generation using foundation models, pushing forward reproducibility in AI-augmented qualitative research.

Together, these studies point to key themes in the emerging literature: LLMs can dramatically improve efficiency in thematic analysis, but they require close alignment with researcher intent, subject matter expertise, and transparent workflows. While much of this research has emphasized AI's performance or technical characteristics, fewer studies have explored how these tools perform in socially complex or structurally ambiguous domains like graduate education, where context and interpretation are critical.

This study addresses that gap by demonstrating how LLMs can be thoughtfully integrated into the qualitative coding process while preserving the methodological integrity of human-driven research. Rather than relying on AI to generate themes independently, our approach applies AI solely to classify sentences according to a codebook developed through grounded theory. This maintains interpretive nuance while enabling scalability. To further support reliability, we incorporate statistical validation of AI-generated classifications, offering a replicable framework for researchers seeking both rigor and efficiency in high-volume qualitative analysis.

We present a step-by-step approach to using AI for qualitative analysis that emphasizes flexibility and scalability. Unlike built-in AI tools in qualitative coding software, which often come with licensing or computational constraints and tie users to a specific vendor, our method is model-agnostic. This allows researchers to choose the most advanced or best-suited AI system (e.g., GPT-4 Turbo, Claude, open-source models) as technology evolves. By operating independently of any single software package, our approach supports cost-effective and adaptable analysis of large qualitative datasets without vendor lock-in.

The methodology we share here was borne out of our own need to systematically analyze a large corpus of interview data, the entirety of which we could not analyze manually due to personnel and financial constraints. We suspect many qualitative researchers find themselves in a similar position, and thus outline our steps in using a current LLM to help us tackle the task. We note at the outset that our goal was to use the AI to help us examine the relationships among the variables that we were interested in, rather than feeding the data into the AI without constraints and letting the model generate its own categorization. More specifically, our research team first extracted themes from a subset of interviews through a coder-driven systematic qualitative analysis. Only then did we ask the AI to extract these themes from the remaining data and analyze interrelationships. Our approach, therefore, allows researchers to maintain control over the research questions and to ask specific questions, by first going through a rigorous, human-driven qualitative analysis of a subset of the data. This process allows researchers to use the AI as a tool, rather than as a replacement for researchers.

Purpose Statement and Research Questions

The purpose of this study is to explore the integration of LLMs into qualitative research workflows, specifically examining how AI can be used to scale thematic coding while preserving methodological rigor. Grounded in a hybrid human-AI framework in which human researchers first define codes and validate outputs while AI models are used to apply these codes across large datasets, this research investigates whether AI-assisted classification can extend human-coded insights, especially in contexts where data volume exceeds manual processing capacity. To address these aims, we asked:

- How can AI-assisted thematic classification reliably replicate human-coded qualitative themes in large-scale textual data?
- What are the potential benefits and limitations of using LLMs (e.g., GPT-4 Turbo) to support scalable qualitative analysis?

And, specific to the content of our study:

- How do departmental structures, such as mentorship, norms, and expectations, relate to graduate students' perceptions of clarity and structure, as identified through AI-assisted coding?

The Current Study

The current study is part of a broader, ongoing research program investigating the factors that promote student success among graduate students pursuing doctoral degrees (Fisher et al., 2019; Mendoza-Denton et al., 2018; Wu et al., 2025). We briefly describe this broader program to contextualize the specific research question addressed here.

A consistent theme in the narratives of graduate students is that the process of getting through the Ph.D. is confusing and vague (Lorentz et al., 2022). It is not uncommon, for example, for faculty not to know the requirements for advancement, and for the program's expectations and rules to be implicit and ill-defined (Ardeljan, 2021; Mendoza-Denton et al., 2018). Indeed, graduate school has been referred to as "organized anarchy," due in part to the broad intellectual freedom granted to faculty in advising students as they see fit (Golde & Walker, 2006). This characterization remains relevant today, with recent studies and commentary continuing to highlight the lack of transparency, structural consistency, and clear expectations in doctoral education (Ardeljan, 2021; Lorentz et al., 2022; Wu et al., 2025). As a result of such distributed decision-making and rule-setting, graduate students may experience the norms, standards, and expectations of their department to be unclear and capricious.

Research suggests that students may in fact benefit from having structure within their graduate program. Studies suggest that programs with clear policies experience lower attrition rates, helping students understand expectations (Ehrenberg et al., 2007; Golde, 2005; Hirt & Muffo, 1998). Mendoza-Denton et al. (2018) hypothesized that three aspects of program structure may be particularly important for graduate student success: (1) clear norms, standards, and expectations; (2) their consistent and equitable application across students; and (3) the involvement of multiple faculty members in advising and supporting each graduate student.

To test these propositions, the research team conducted a series of semi-structured student interviews at a large public university on the West Coast of the United States. The interviews centered around students' expectations of graduate school, their perceptions of departmental clarity and structure, and their experiences throughout their graduate studies. As we detail below, for this case example, we will focus on clarity (and lack thereof) as experienced by the graduate students, with the expectation that perceived clarity would be related to positive student experiences. We delineate how we used a current LLM, GPT-4 Turbo, on the large dataset resulting from these interviews to classify and identify thematic patterns at a sentence-by-sentence level. GPT-4 Turbo was chosen for its enhanced processing speed, longer context window, improved accuracy in thematic classification, and cost-effectiveness for large-scale qualitative data analysis (OpenAI, n.d.-a). Beginning with text segmentation and classification, we used automated methods to differentiate sentences related to broad topics, followed by detailed topic identification based on predefined categories identified by a team of coders. By systematically testing and refining prompts (i.e., prompt engineering; Reynolds & McDonell, 2021), we ensured that the AI's classifications and insights were both relevant and comparable to those derived from manual analysis, supporting accuracy and reliability in automated coding. Further, to mitigate security risks in AI-based text analysis, we used OpenAI's Playground, which, at the time of analysis, did not retain or use user-entered data for ongoing training (OpenAI, 2024). This decision allowed us to benefit from advanced NLP capabilities without compromising the privacy of interview transcripts (Fung et al., 2010).

Methods

To provide a structured overview of the approach taken in this study, Figure 1 presents a flowchart summarizing the key methodological steps, from data preparation to statistical analysis. The subsequent sections detail each step in greater depth.

The interviews for this study were conducted across four academic departments during the year of 2020-2021; these were recorded and transcribed verbatim by Rev, a speech-to-

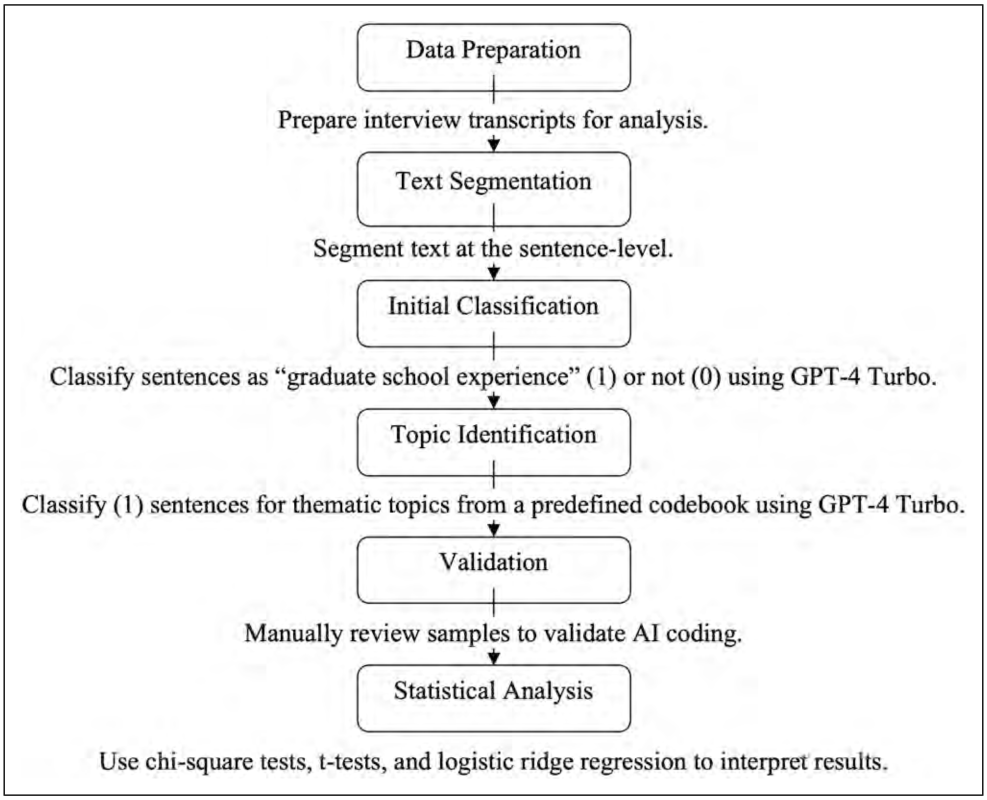
Data Preparation
Prepare interview transcripts for analysis.
Text Segmentation
Segment text at the sentence-level.
Initial Classification
Classify sentences as "graduate school experience" (1) or not (0) using GPT-4 Turbo.
Topic Identification
Classify (1) sentences for thematic topics from a predefined codebook using GPT-4 Turbo.
Validation
Manually review samples to validate AI coding.
Statistical Analysis
Use chi-square tests, t-tests, and logistic ridge regression to interpret results.

Figure 1. Flowchart Summarizing the Methodology. *Note*. The process begins with Data Preparation and Segmentation, followed by Initial Sentence Classification, Topic Identification, Validation, and Statistical Analysis

text company that provides closed captioning, subtitles, and transcription services. Each interview lasted approximately 1 hour. A total number of 89 transcripts were used in this study. Table 1 shows the average number of words and sentences per transcript, per department. To protect the anonymity of participants and their departments, we have assigned numerical IDs (Department 1, Department 2, etc.) instead of using department names. These numerical labels were randomly assigned and do not correspond to any ranking or inherent characteristics of the departments.

Text Segmentation

The interview transcripts from Rev included speaker titles, distinguishing between "Interviewer" and "Interviewee." To prepare the transcripts for automated analysis, we extracted only the "Interviewee" portions and segmented the text at the sentence level using a Python-based approach. We chose to work at the sentence level because it provides a balance between granularity and context, allowing for clearer analysis of participants' responses. This approach ensures that each meaningful statement remains intact while making it easier to classify and interpret different aspects of the graduate school experience. The text was divided at periods followed by a space, a common indicator of sentence boundaries. Specific measures were implemented to preserve ellipses as single units and to recognize patterns like email addresses, preventing incorrect segmentation (Bird et al., 2009). Department 1 had 20,516 total sentences, Department 2 had 7,816 total sentences, Department 3 had 3,495 total sentences, and Department 4 had 14,303 total sentences.

Prior Research

The classification process used in this study builds on methodologies developed in prior research by Noah et al. (2024), which introduced a two-step classification approach for analyzing subjective experience reports. In that study, LLMs were employed to first distinguish relevant from

Table 1. Summary Statistics per Average Departmental Interview

	Number of words per interview		Number of sentences per interview	
Department	M	SD	M	SD
Department 1	8,316	533	472	55
Department 2	7,207	775	455	46
Department 3	7,841	841	469	53
Department 4	8,697	888	483	21

irrelevant sentences (binary classification) and then categorize relevant sentences into thematic topics. Their method was designed to handle highly subjective, unstructured qualitative data, ensuring that the AI could accurately extract meaningful patterns while minimizing noise. This study adapts that approach; our primary classification task involved determining whether each sentence in the transcripts described aspects of the graduate school experience.

Model Selection, Parameters, and Rationale

For the classification task, we utilized GPT-4 Turbo, a variant of GPT-4 optimized for performance, consistency, and cost-effectiveness. This model features a 128k-token context window, allowing for the processing of approximately 300 pages of text in a single prompt. GPT-4 Turbo is also significantly more affordable than its predecessor, approximately three times cheaper for input tokens and twice as inexpensive for output tokens, making it a scalable choice for large-scale qualitative analysis (OpenAI, n.d.-a).

To ensure reproducibility, we set the temperature parameter to 0, which minimizes randomness in generation. This makes the model's classifications consistent for the same input in nearly all cases. We also limited the output to a single token, allowing the model to classify each sentence as either relevant (1) or not relevant (0) to the graduate school experience. Below are example classifications.

- 1: “I would say that the style of the relationship, it's informal, but comfortably professional I would say.”
- 0: “Yeah, I think it was pretty clear.”

This model was deployed through OpenAI's Playground, which, at the time of analysis, adhered to a no-data-retention policy. This was a critical factor in protecting participant confidentiality and meeting institutional data privacy standards. Additionally, our approach is model-agnostic and does not rely on vendor-specific software, allowing flexibility for future research using other platforms such as Claude, Gemini, or open-source LLMs.

From a methodological standpoint, this AI-assisted process allowed us to apply a human-developed codebook to more than 18,000 sentences with consistency and speed. While a team of coders might take several weeks to manually code this volume of data, the AI-based approach completed the task in under 24 hours. In terms of cost, processing 1 million tokens on GPT-4 Turbo (2024 pricing) costs roughly $0.003 per 1,000 tokens for input and $0.006 per 1,000 tokens for output, making the method several-fold more cost-efficient than human labor at typical research assistant rates (OpenAI).

Importantly, this process retains human judgment in theme development: the AI did not generate categories but instead extended an existing human-coded framework across a larger dataset. We validated the model's performance through both manual review and statistical methods including chi-square tests, t-tests, and ridge logistic regression to assess classification reliability and thematic associations.

Overall, this method offers a reproducible, scalable framework for qualitative researchers facing large textual datasets and limited time or personnel. It demonstrates how LLMs can increase the speed and consistency of coding while preserving the interpretive control central to qualitative inquiry.

Topic Creation

The process of identifying topics for classification followed a structured, iterative approach to ensure accuracy and comprehensiveness. As part of a larger project examining departmental structures in STEM graduate education, we leveraged an existing codebook from that project as the foundation for topic selection and refinement in this research.

The larger project team employed a rigorous grounded theory approach to coding transcript data, adhering to methodological principles established by Glaser and Strauss (1967) and later refined by Charmaz (2006). In the initial phase, all four team members independently reviewed the complete set of transcripts, taking detailed memos–an essential practice for capturing analytic insights, as emphasized by Corbin and Strauss (2014). The team then convened to discuss these memos, identifying recurring themes and potential coding categories through a process of constant comparison. Next, the transcripts were divided among the researchers, each conducting detailed open coding on their assigned sections, following the initial coding stage outlined by Saldaña (2021). After this deeper analysis, the team collaboratively developed an initial codebook, defining proposed codes and providing representative examples. Through an iterative cycle of application, discussion, and refinement–reflecting the theoretical sampling approach in grounded theory–the team gradually shaped the codebook to ensure it accurately captured the complexity of the data while remaining consistent and applicable across all transcripts.

To further refine the topic list, an iterative process among the core research team was employed. This involved testing the codebook's classifications on sampled interview data and making adjustments to ensure the categories were neither too broad nor too restrictive. The primary goal was to balance specificity with flexibility, allowing the AI model to accurately classify sentences while minimizing the risk of hallucination–where the model generates misleading or incorrect classifications.

The finalized set of topics was informed by emerging themes from the interviews and agreement among the research team. These topics encompass key aspects of the graduate school experience, including.

- Graduate student communities: Positive and negative experiences related to peer networks and support systems.
- Work-life balance: Statements reflecting either a healthy balance or struggles in managing academic responsibilities alongside personal life.
- Mental health: Mentions of well-being, stress, or psychological challenges, categorized as positive or negative.
- Clarity and lack of clarity: Statements indicating whether expectations, departmental structures, or academic requirements were well-communicated or ambiguous.
- Mentorship: Experiences with mentors, categorized as approachable or unapproachable.
- Departmental culture and norms: Statements reflecting the overall environment within a department, including positive and negative aspects.
- Expectations: Whether academic or professional expectations were met, exceeded, or fell short.
- Identity threat and belonging: Experiences related to feeling included or excluded within the academic community.
- Trust and mistrust: Indicators of confidence or skepticism in institutional processes, faculty, or peers.
- Career goals: Aspirations and professional trajectories discussed within the academic setting.
- COVID-19 effects: The impact of the pandemic on graduate students' experiences.
- Student outcomes: Reflections on academic progress, categorized as positive or negative.

Each of these topics was structured to reflect both positive and negative dimensions where applicable. Mentorship, for example, can be a source of support and encouragement, but it can also feel distant or unhelpful. Similarly, departmental culture can create a sense of belonging and collaboration, or it can lead to feelings of isolation and competition. By distinguishing between these variations, we ensure that our analysis reflects the full spectrum of experiences rather than oversimplifying them into a single category. Making this distinction also allows us to explore how different experiences interact. For example, we can see whether clear expectations are linked to greater confidence and well-being, while unclear expectations might contribute to stress or frustration. This approach helps us uncover meaningful patterns in the data, ensuring that AI-driven analysis doesn't just classify statements but also provides insights that reflect the complexity of real student experiences. We also included a "none of the above" category for statements that didn't clearly fit into any predefined topics. This helps keep the analysis flexible and adaptable, ensuring that important nuances aren't forced into categories where they don't belong.

Topic Identification

Following the initial classification, only sentences identified as describing aspects of the graduate school experience (i.e., classified as 1) were selected for further topic analysis. GPT-4 Turbo was then used to categorize each sentence based on the key themes above related to the graduate student experience. Again, by setting the temperature to 0, we ensured that the AI produced the same output for the same input every time, eliminating unnecessary variability. Table 2 shows a sample of 1-classified sentences, alongside their topic classification. This process aligns with the "Topic Identification" phase shown in Figure 1, serving as the foundational step that enables AI-based classification of relevant sentences.

Validation Process

To assess the accuracy of the GPT-4 Turbo topic classifications, we conducted a manual review of a random sample of 100 sentences from each department that had been classified as describing an aspect of the graduate school experience followed by a subsequent classification of one or more topics (including "none of the above"). Discrepancies between the automated classifications and the manual review were minimal, with a disagreement rate of 5% for Department 1 and 4% for Department 2, Department 3, and Department 4, suggesting that the model's classifications were generally accurate. Nonetheless, these discrepancies highlight areas where the automated process did not fully align with human judgment. All identified discrepancies in topic agreement were corrected across the assessed sample in each department.

Key Analytical Topics

We focused on "clarity" and "lack of clarity" as key analytical categories based on our initial research questions, which

Table 2. Sentence Topic Classification Sample

Sentence	Topic classification
I think I've had the good fortune of having good mentors in the department so far	Approachable mentors
There's a lot more professional development kind of seminars and things that I don't know about	Lack of clarity
We don't really have a visit due to COVID.	COVID-19 effects
It is segmented, disjointed, the overall feel of the department	Negative culture

Note. Classification of each sentence is defined by representing aspects of the graduate school experience (1) followed by a topic classification.

emphasized the role of clear communication in shaping graduate student experiences. Specifically, we examined how clarity–regarding academic expectations, mentorship, and departmental norms–related to students' sense of confidence, belonging, and academic success, as supported by prior research (Austin, 2009; Lovitts, 2001). Our manual review of interview transcripts confirmed that these dimensions of clarity were salient for students and consistently present in their accounts. Conversely, a lack of clarity was frequently associated with confusion, frustration, and increased stress, aligning with concerns about student well-being and retention (Golde, 2005; Weidman et al., 2001). By analyzing clarity as a central variable, we aimed to identify patterns in how transparent or ambiguous communication within graduate programs influenced students' experiences and progression.

Statistical Analyses

Chi-Square Tests. To explore the associations between topic variables and the constructs of "clarity" and "lack of clarity," we applied the chi-square test of independence (McHugh, 2013). The chi-square test is used to determine whether there is a statistically significant association between categorical variables by assessing the difference between observed frequencies and expected frequencies in contingency tables. If the p-value is less than the chosen significance level (0.05), we reject the null hypothesis, indicating a statistically significant association between the variables.

Contingency tables were created to explore the relationships between specific topics and the primary constructs. Each table organizes data into rows and columns, with each cell representing the frequency of a particular combination of the variables. For instance, if we were examining the topic of "positive mental health" in relation to "clarity," the contingency table might display how often sentences classified under "positive mental health" were also tagged as "clarity" versus "lack of clarity." The rows of the contingency table could represent the presence or absence of a specific topic (e.g., "positive mental health" present or not present), while the columns could represent whether the sentence was categorized as "clarity" or "lack of clarity". The values within the table represent the number of occurrences for each combination.

By comparing the observed frequencies within the contingency table to the expected frequencies under the assumption of no association, the chi-square test evaluates whether there is a statistically significant association between the topic and the construct of interest.

Independent t-tests. To assess differences in the prevalence of topics between responses classified as "clarity" versus "lack of clarity," we conducted independent two-sample *t*-tests (Cressie & Whitford, 1986). These tests evaluate whether the mean prevalence of a topic differs significantly between two independent groups. For example, we examined whether sentences tagged as "clarity" contained positive mental health topics at a higher average rate than those tagged as "lack of clarity." The *t*-tests provided evidence of whether such differences in topic prevalence were statistically significant, offering insights into how clarity-related responses diverged in thematic content.

Cross-Department Analysis Using Ridge Logistic Regression. To compare the prevalence of "clarity" across academic departments, we conducted a cross-department analysis using ridge logistic regression (Hoerl & Kennard, 1970). This approach was chosen over traditional logistic regression to mitigate potential multicollinearity among predictor variables and improve model stability. For instance, topics such as "positive department norms" and "positive mentorship" may co-occur, leading to correlated predictors when included in the model. Ridge regression applies L2 regularization, shrinking coefficients toward zero to reduce variance inflation and prevent unstable estimates. This regularization improves the robustness and generalizability of the model by reducing overfitting, allowing for more reliable inferences about departmental differences in "clarity". This analysis involved several steps.

(1) For each department, we excluded irrelevant columns and included a categorical variable representing the department.
(2) Rows where both "clarity" and "lack of clarity" were zero were excluded to focus the analysis on relevant data points.
(3) We utilized ridge logistic regression to model the probability of "clarity" as a function of department and the other topics. The logistic regression model is represented by:

$$logit(P(Y=1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

where $logit(P(Y=1))$ is the log odds of the binary outcome Y being 1 (clarity), β_0 is the intercept, and $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients for the predictor variables $X_1, X_2, \ldots, X_p$. In ridge logistic regression, the model is estimated by minimizing the penalized log-likelihood function:

$$L(\beta) = -\sum_{i=1}^{n} [Y_i logP(Y_i) + (1-Y_i) log(1-P(Y_i))] + \lambda \sum_{j=1}^{p} \beta_j^2$$

where λ is the regularization parameter that controls the strength of the penalty.

(4) The model coefficients were analyzed to understand how different departments and other features influence the likelihood of a response being classified as "clarity."

Results

Classification and Topic Identification

Out of 46,127 total sentences, 18,482 sentences were classified as relevant to the graduate school experience. The most frequently identified topics were “none of the above” (N = 9,322), “negative expectations” (N = 2,861), “negative department norms” (N = 2,020), “positive department norms” (N = 1,891), and “approachable mentors” (N = 1,296). The count of topics across the classified sentences is presented in Table 3. Sample sentences which were classified as “none of the above” include.

- So I haven’t had any funding issues.
- If I see them at some seminar or something, I’ll make conversation with them.
- I rotated in both and really liked both.
- And in the fall, I wasn’t teaching.
- So you need to choose two major and one minor topics.

Statistical Analyses

Chi-Square Tests. Chi-square tests of independence were conducted to assess the relationships between categorical factors and the outcomes of clarity and lack of clarity across departments. A p-value of less than .05 was considered statistically significant.

Clarity. Significant factors associated with clarity varied across departments.

- Department 1: Negative department norms, χ^2 (1, N = 4,029) = 25.32, $p < .001$; negative expectations, χ^2 (1, N = 4,029) = 15.22, $p < .001$; and trust, χ^2 (1, N = 4,029) = 7.54, $p < .01$.
- Department 2: Negative expectations, χ^2 (1, N = 3,531) = 16.55, $p < .001$; negative department norms, χ^2 (1, N = 3,531) = 9.91, $p < .01$; and positive student outcomes, χ^2 (1, N = 3,531) = 4.52, $p < .05$.
- Department 3: No significant variables were identified for clarity.
- Department 4: Approachable mentors, χ^2 (1, N = 5,753) = 18.70, $p < .001$; negative department norms, χ^2 (1, N = 5,753) = 17.38, $p < .001$; and negative expectations, χ^2 (1, N = 5,753) = 17.07, $p < .001$.

Table 3. Topic Classifications per Department

Topic	Department 1	Department 2	Department 3	Department 4	Total
Positive graduate student communities	206	50	338	184	778
Negative graduate student communities	96	16	80	37	229
Positive work-life balance	32	7	56	51	146
Negative work-life balance	71	18	174	115	378
Positive mental health	10	2	16	18	46
Negative mental health	90	22	127	63	302
Lack of clarity	332	101	419	232	1,084
Approachable mentors	311	78	556	351	1,296
Unapproachable mentors	255	66	381	236	938
Positive culture	97	17	132	98	344
Negative culture	265	25	196	85	571
Department norms	322	53	265	222	862
Expectations	345	70	396	377	1,188
Identity threat	76	15	157	56	304
Belonging	102	28	242	101	473
Exclusion	196	28	296	95	615
Trust	33	3	20	10	66
Mistrust	111	4	62	14	191
Career goals	259	42	347	192	840
COVID-19 effects	61	18	76	58	213
Student outcomes	104	26	110	55	295
Positive department norms	526	97	784	484	1,891
Negative department norms	759	109	792	360	2,020
Positive expectations	403	58	446	345	1,252
Negative expectations	779	194	1,144	744	2,861
Positive student outcomes	242	41	384	211	878
Negative student outcomes	297	65	337	180	879
None of the above	2,253	960	3,748	2,361	9,322
Clarity	112	26	156	114	408

Note. The table presents the distribution of topics identified in 18,482 sentences classified as relevant to the graduate school experience. The topics are categorized based on the frequency of their occurrence across four academic departments.

Lack of Clarity. For lack of clarity, negative expectations and negative departmental norms were significant across all departments.

- Department 1: Negative expectations, χ^2 (1, N = 4,029) = 166.01, p < .001; negative department norms, χ^2 (1, N = 4,029) = 48.20, p < .001; and approachable mentors, χ^2 (1, N = 4,029) = 29.10, p < .001.
- Department 2: Negative expectations, χ^2 (1, N = 3,531) = 136.00, p < .001; positive department norms, χ^2 (1, N = 3,531) = 14.58, p < .001; and approachable mentors, χ^2 (1, N = 3,531) = 25.80, p < .001.
- Department 3: Negative expectations, χ^2 (1, N = 1,086) = 35.58, p < .001; positive department norms, χ^2 (1, N = 1,086) = 3.30, p < .01; and approachable mentors, χ^2 (1, N = 1,086) = 7.35, p < .01.
- Department 4: Negative expectations, χ^2 (1, N = 5,753) = 355.83, p < .001; positive department norms, χ^2 (1, N = 5,753) = 26.54, p < .001; and approachable mentors, χ^2 (1, N = 5,753) = 46.39, p < .001.

Independent t-tests. Independent-samples t-tests were conducted to compare the frequency of topics between sentences classified as "clarity" versus "lack of clarity." In this analysis, a positive t-value indicates that a topic occurred more frequently in clarity-related sentences, whereas a negative t-value indicates greater frequency in lack-of-clarity sentences. A p-value of less than .05 was considered statistically significant.

Clarity. Sentences classified under clarity were significantly more likely to mention positive expectations, positive department norms, career goals, and approachable mentors.

- Department 1: Positive expectations, *t* (442) = 6.86, p < .001; career goals, *t* (442) = 4.89, p < .001.
- Department 2: Positive department norms, *t* (344) = 4.17, p < .001; career goals, *t* (344) = 4.25, p < .001.
- Department 3: Career goals, *t* (125) = 1.99, p = .0483; approachable mentors, *t* (125) = 4.25, p < .001.
- Department 4: Approachable mentors, *t* (573) = 10.18, p < .001, showed the strongest association with clarity.

Lack of Clarity. Sentences classified under lack of clarity were significantly more likely to mention negative expectations, negative department norms, exclusion, and negative student outcomes.

- Department 1: Negative expectations, *t* (442) = −8.57, p < .001; negative department norms, *t* (442) = −7.43, p < .001.
- Department 2: Negative expectations, *t* (344) = −9.36, p < .001; exclusion, *t* (344) = −2.77, p = .0502.
- Department 3: Negative expectations, *t* (125) = −3.63, p < .001; negative department norms, *t* (125) = −2.03, p = .0446.
- Department 4: Negative expectations, *t* (573) = −11.70, p < .001; negative student outcomes, *t* (573) = −2.10, p = .0365.

Across departments, negative expectations and negative department norms consistently occurred more often in sentences classified as lack of clarity, with exclusion also emerging in Department 2.

Cross-Department Analysis Using Ridge Logistic Regression. Ridge logistic regression was used to model the probability of a sentence being classified as clarity (1) versus lack of clarity (0) across the four academic departments. Department affiliation was represented with dummy variables, with Department 1 as the reference category. Rows in which neither clarity nor lack of clarity was coded were excluded to ensure the analysis focused on sentences with meaningful classifications.

Clarity. Key predictors of clarity across departments include.

- Approachable mentors (Department 4: $\beta = 2.05$, Department 2: $\beta = 1.44$)
- Positive department norms (Department 4: $\beta = 1.03$, Department 2: $\beta = 0.93$)
- Career goals (Department 1: $\beta = 1.34$, Department 3: $\beta = 0.35$)

Negative department norms had the strongest negative association across all models.

Lack of Clarity. Key predictors of lack of clarity across departments include.

- Negative expectations (Department 4: $\beta = 6.04$, Department 1: $\beta = 4.65$)
- Negative department norms (Department 4: $\beta = 2.32$, Department 2: $\beta = 2.54$)

Exclusion and mistrust were particularly significant in Department 2.

Discussion

The goal of this study was twofold: (1) to assess the alignment between LLM-based thematic analysis and human-coded qualitative classifications, and (2) to uncover key factors related to clarity and lack of clarity in graduate school experiences. To achieve this, we developed a structured, AI-assisted approach that first used binary classification to filter relevant qualitative data, followed by thematic categorization based on predefined topics. We ensured reproducibility by employing a model-agnostic framework that allows for flexibility in AI selection and by validating AI-generated classifications against human-coded analyses. Statistical validation techniques, including chi-square tests, *t*-tests, and ridge logistic regression, provided further confirmation of the reliability of AI-driven classifications.

Our findings show that AI-assisted qualitative coding can be both efficient and reliable, especially when combined with human oversight and statistical validation. By blending AI with established qualitative research methods, we have shown that LLMs can be powerful tools to help researchers analyze data more quickly and thoroughly without sacrificing accuracy or depth.

Key Findings on Clarity in Graduate School

The statistical analyses consistently identified mentorship, departmental norms, and expectations as the strongest factors associated with clarity. Chi-square tests revealed significant associations between clarity and positive elements such as approachable mentors, positive department norms, and, in some departments, positive expectations. Conversely, negative expectations and negative departmental norms were consistently associated with a lack of clarity across all departments.

These findings are consistent with Anderson and Louis (1994), who argued that departmental climate, structure, and mentorship shape graduate students' adherence to academic norms and their professional integration into academia. Their research emphasizes that graduate student experiences vary across disciplines, which aligns with our finding that mentorship and departmental norms do not have a uniform impact across departments. While Anderson and Louis (1994) focus on socialization into academic norms rather than clarity itself, their work supports the broader idea that departmental structure and faculty interactions influence graduate students' academic development.

Similarly, Lechuga (2011) highlights the critical role of faculty mentorship in fostering graduate students' professional identity and success. Our findings align with this perspective, as students with approachable mentors were more likely to express clarity in their academic experiences. Lechuga's framework of faculty suggests that mentorship extends beyond academic guidance to include professional and personal development, which is consistent with our finding that mentorship plays a central role in shaping students' graduate school experiences. Additionally, our results show that negative departmental norms and negative expectations were strongly associated with a greater likelihood of students expressing a lack of clarity, reinforcing the idea that departmental culture and faculty interactions impact graduate student success.

Expanding on this, Pollard and Kumar (2021) discuss the evolution of graduate student mentorship, particularly in online and hybrid settings. Their research suggests that mentorship models emphasizing transparency, shared expectations, and holistic support contribute to positive student outcomes. While our study focuses on in-person graduate student experiences, our findings are consistent with the broader theme that structured, supportive mentorship enhances students' academic clarity and sense of direction.

Notably, the impact of graduate student communities and belonging varied across departments. While these factors were associated with clarity in some departments, their effects were inconsistent or weaker in others, suggesting that their influence may be context-dependent. This reflects broader patterns in how disciplinary structures, departmental climate, and faculty-student relationships shape graduate student experiences. Future research should explore how differences in departmental structures and social dynamics mediate the role of community support in shaping students' academic clarity.

Validation of Findings Through Statistical Methods

Multiple statistical approaches supported these patterns. Independent-samples *t*-tests showed that sentences classified as clarity were significantly more likely to mention positive expectations, career goals, and positive departmental norms. In contrast, sentences classified as lack of clarity more often referenced negative expectations, negative student outcomes, and negative departmental culture.

The ridge logistic regression analysis further indicated that mentorship and departmental culture were central to clarity. Positive departmental culture and career goals were moderately associated with increased clarity, while the absence of approachable mentorship and the presence of negative departmental norms showed the strongest negative associations. These findings highlight that although multiple factors shape clarity, the combination of supportive mentorship and a positive departmental environment plays a particularly critical role.

Human-Coded Validation and LLM Classification

Lastly, our results align with human-coded analyses, particularly the manual sample review process used to assess the accuracy of LLM classifications. This alignment suggests that LLMs can effectively identify thematic patterns in qualitative data, producing classifications that parallel those made by human coders. While computational approaches cannot replace human interpretation, these findings highlight their potential as complementary tools in large-scale qualitative research.

Implications

This study has implications for both qualitative research methodology and applied practice. Our findings demonstrate that AI can be meaningfully integrated into qualitative workflows without supplanting human interpretation. Specifically, we show that LLMs can reliably apply a human-developed codebook rooted in grounded theory to a large corpus of text data. This hybrid approach preserves methodological rigor by ensuring that coding categories originate from human expertise, while also enabling the consistent application of those codes across thousands of data points. By removing coder subjectivity in code application (i.e., reducing coder drift and inter-rater variability), AI supports uniform interpretation across datasets. Moreover, our model-agnostic methodology, applicable to various LLMs, offers a reproducible and transparent framework for other researchers seeking to integrate AI with traditional qualitative strategies.

Practically, this study provides a scalable solution for research teams with limited time, personnel, or funding. Automating the application of qualitative codes significantly reduces the manual burden of coding large textual datasets, enabling teams to scale their analyses without compromising analytic precision. Our use of secure tools, such as OpenAI's Playground, demonstrates that LLMs can be deployed in ways that align with data confidentiality needs. In addition to OpenAI's models, other platforms such as Claude (Anthropic) or Gemini (Google DeepMind) can also be adapted for similar purposes depending on institutional or budgetary constraints. Importantly, researchers should ensure that their use of AI adheres to institutional review board (IRB) requirements and complies with applicable privacy regulations when handling sensitive qualitative data.

Limitations

This study has several limitations that merit discussion, spanning technical, conceptual, cultural, and practical dimensions. These limitations also provide guidance for researchers seeking to adapt or extend this approach in their own work.

Model and Technical Limitations. Although GPT-4 Turbo produced accurate classifications, the model remains a black box in many respects. It is often difficult to determine how or why a decision was made, especially when analyzing ambiguous or context-dependent language. This opaqueness presents a challenge to interpretability and underscores the importance of "checking the work" of the model through manual spot-checking, validation samples, and statistical consistency tests to ensure that AI-generated classifications align with human standards. Additionally, because LLMs are trained on large-scale internet data, they may reflect biases embedded in that data. This poses a risk when analyzing experiences of academic, cultural, or identity groups that may be underrepresented or misrepresented in the training corpus. Such bias may influence which themes are surfaced, overlooked, or misclassified.

To mitigate these risks, we recommend careful prompt design, configuring the model to produce more deterministic outputs (using low-temperature settings), and ongoing human review to ensure alignment with research goals. As LLM capabilities evolve, so too should protocols for transparency, error analysis, and interpretability.

Conceptual Trade-offs

Our choice to analyze data at the sentence level prioritized scalability and computational tractability, but came at the cost of reduced contextual nuance. In qualitative research, meaning often unfolds across paragraphs or through the interplay of multiple ideas. Sentence-level segmentation may oversimplify complex, emotionally layered, or evolving narratives, particularly around sensitive topics like mental health or identity. Similarly, working from a predefined codebook while ensuring consistency limited our ability to detect novel or emergent themes that did not align with our initial research focus.

These trade-offs reflect a broader tension between efficiency and interpretive richness. We recommend that researchers using similar methods consider supplementing AI-driven analysis with more traditional qualitative methods such as open coding or narrative synthesis to capture the complexity of participant voices.

Cultural and Personal Limitations. As discussed in the following section on researcher positionality, our team's academic backgrounds and institutional affiliations shaped both the framing of research questions and the construction of our coding framework. Constructs such as "structure," "clarity," and "support" were informed by our disciplinary lenses and institutional norms, which may not reflect how students from other cultural, socioeconomic, or educational contexts experience graduate education. Additionally, because the interview data were drawn from a single large U.S. research university, findings may not generalize to students in other institutional or cultural settings.

We also acknowledge that classification outputs were co-constructed between model behavior and researcher input. Prompts, code definitions, and the selection of themes were all guided by our assumptions and goals, shaping how the LLM interpreted student responses. We encourage future researchers to explicitly consider how such cultural, linguistic, and institutional factors might influence the design and application of AI-assisted qualitative methods.

Limitations on Generalizability Stemming from Researcher Positionality. Although the use of LLMs reduces some variability associated with manual, multi-coder qualitative analysis, it remains essential to acknowledge that the thematic framework guiding this analysis was developed by humans and is therefore shaped by the backgrounds, perspectives, and biases of our research team. In our case, prior work on the relationship between program structure and graduate student success directly influenced not only the design of our interview protocol but also the development of the coding schema applied in this study.

Our team brings interdisciplinary expertise from psychology, education, and computational fields, and includes individuals with experience as graduate students, postdocs, administrators, and faculty members. Several members have direct experience navigating and supporting graduate students within departmental and divisional structures. This range of experiences contributed to our group-based development of the initial codebook and enriched our interpretations of the data. At the same time, the team shares a long-standing interest in how structural features of graduate education, such as clarity, mentorship, and departmental norms, shape student outcomes (e.g., Fisher et al., 2019; Mendoza-Denton et al., 2018; Wu et al., 2025). Our comfort with programming, statistical modeling, and data science also made us more inclined to explore AI as a tool for qualitative research.

We recognize, however, that disciplinary expertise can narrow interpretive lenses and limit openness to novel patterns or unanticipated meanings (Darley & Gross, 1983; Sackett, 1983). While we attempted to mitigate this risk through interdisciplinary collaboration, complete neutrality is neither possible nor claimed. Our institutional context (research-intensive universities) may also have shaped our assumptions about what constitutes "support" or "structure" in graduate education. These assumptions likely influenced not only how we interpreted the data, but also how we designed our questions and framed our hypotheses–biases that inevitably shaped the input data the LLM was trained on in this study.

In that sense, our approach is not immune to the same subjectivities present in traditional team coding. We emphasize that using an LLM, at least as we have done here, does not remove the influence of researcher positionality. Rather, it shifts the point of influence upstream to the construction of the coding framework and the design of prompts guiding AI classification.

Ethical and Practical Considerations

While our use of OpenAI's Playground ensured a no-retention data policy, privacy protections may vary across platforms. Researchers must be cautious when using commercial APIs or third-party platforms, and should ensure compliance with institutional guidelines, IRB approvals, and applicable data protection laws (e.g., FERPA, GDPR). In addition, although our method does not require advanced programming, it does assume a basic level of technical fluency. This requirement may present a barrier for some qualitative researchers or teams without access to technical support.

To support broader adoption, we encourage the development of open-source tools, user-friendly workflows, and interdisciplinary collaborations that bridge qualitative expertise with technical capability.

Future Directions for AI in Qualitative Research

Despite these challenges, the future of AI in qualitative research presents exciting possibilities. Multi-modal AI systems, which integrate text, images, and other data types, could enhance qualitative research by analyzing non-verbal cues in video interviews or behavioral observations alongside textual data. Additionally, specialized AI models tailored to specific disciplines could improve precision by incorporating domain-specific language and contextual understanding. However, it is essential to ensure that AI's implementation is ethical, sensitive, and remains aligned with human oversight.

Conclusion

This study highlights the potential of AI-driven methods, particularly LLMs, in enhancing qualitative research. By automating text classification and thematic analysis, AI offers a scalable and consistent approach to identifying patterns in qualitative data. Importantly, our findings reaffirm the critical role of mentorship, departmental norms, and expectations in shaping graduate students' academic clarity.

The statistical analyses, particularly ridge logistic regression, indicated that mentorship and departmental norms were significant predictors of students' sense of clarity, underscoring the importance of structured support systems in graduate education. These results suggest that targeted interventions in mentorship and departmental communication could meaningfully improve student experiences.

While AI offers advantages in efficiency and scalability, challenges remain in capturing contextual nuances and complex themes. Future research should focus on enhancing AI's interpretative capabilities and developing hybrid approaches that integrate human oversight. As AI continues to evolve, its role in qualitative research will likely expand, providing new opportunities to increase analytical depth and rigor. However, ensuring ethical, transparent, and thoughtful implementation will be essential to maintain the integrity of qualitative inquiry while emphasizing AI's strengths.

Acknowledgements

We are grateful to the anonymous peer reviewers for their thoughtful and constructive feedback, which significantly improved the clarity and rigor of this manuscript. We also acknowledge the support of our colleagues and institutions throughout the development of this project.

ORCID iDs

Miranda Shen https://orcid.org/0009-0002-0864-1083
Jue Wu https://orcid.org/0000-0002-2931-7146
Rodolfo Mendoza-Denton https://orcid.org/0000-0002-7965-7309

Ethical Considerations

This research was reviewed and approved by the University of California at Berkeley's Committee for the Protection of Human Subjects, protocol #2019-04-1219.

Consent to Participate

All participants consented to participate in this research.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by Grant 1954923 from the National Science Foundation, USA

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

The data analyzed during the current study are not publicly available due to confidentiality agreements with research participants.

References

Anderson, M. S., & Louis, K. S. (1994). The graduate student experience and subscription to the norms of science. *Research in Higher Education*, *35*(3), 273–299. https://doi.org/10.1007/bf02496825

Ardeljan, J. M. (2021). *Navigating graduate school: It's all about the process*. Inside Higher Ed. https://www.insidehighered.com/advice/2021/10/18/how-navigate-unwritten-rules-graduate-school-opinion

Austin, A. E. (2009). Cognitive apprenticeship theory and its implications for doctoral education: A case example from a doctoral program in higher and adult education. *International Journal for Academic Development*, *14*(3), 173–183. https://doi.org/10.1080/13601440903106494

Bennis, I., & Mouwafaq, S. (2025). Advancing AI-driven thematic analysis in qualitative research: A comparative study of nine generative models on cutaneous leishmaniasis data. *BMC Medical Informatics and Decision Making*, *25*(1), 1–14. https://doi.org/10.1186/s12911-025-02961-5

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage Publications.

Corbin, J., & Strauss, A. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage Publications.

Cressie, N. A. C., & Whitford, H. J. (1986). How to use the two sample t-test. *Biometrical Journal Journal*, *28*(2), 131–148. https://doi.org/10.1002/bimj.4710280202

Dai, S. C., Xiong, A., & Ku, L. W. (2023). LLM-in-the-loop: Leveraging large language model for thematic analysis. arXiv preprint arXiv:2310.15100.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20–33. https://doi.org/10.1037//0022-3514.44.1.20

Dunivin, Z. O. (2025). Scaling hermeneutics: A guide to qualitative coding with LLMs for reflexive content analysis. *EPJ Data Science*, *14*(1), 28. https://doi.org/10.1140/epjds/s13688-025-00548-8

Ehrenberg, R. G., Jakubson, G. H., Groen, J. A., So, E., & Price, J. (2007). Inside the Black box of doctoral education: What program characteristics influence doctoral students' attrition and graduation probabilities? *Educational Evaluation and Policy Analysis*, *29*(2), 134–150. https://doi.org/10.3102/0162373707301707

Fisher, A. J., Mendoza-Denton, R., Patt, C., Young, I., Eppig, A., Garrell, R. L., Richards, M. A., & Nelson, T. W. (2019). Structure and belonging: Pathways to success for underrepresented minority and women PhD students in STEM fields. *PLoS One*, *14*(1), Article e0209279. https://doi.org/10.1371/journal.pone.0209279

Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, *42*(4), 1–53.

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

Glaser, B., & Strauss, A. (1967). *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Golde, C. M. (2005). The role of the department and discipline in doctoral student attrition: Lessons from four departments. *The Journal of Higher Education*, *76*(6), 669–700. https://doi.org/10.1080/00221546.2005.11772304

Golde, C. M., & Walker, G. E. (Eds.), (2006). *Envisioning the future of doctoral education: Preparing stewards of the discipline-Carnegie essays on the doctorate*. John Wiley & Sons.

Hirt, J. B., & Muffo, J. A. (1998). Graduate students: Institutional climates and disciplinary cultures. *New Directions for Institutional Research*, *25*(2), 17–33.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Katz, A., Fleming, G. C., & Main, J. (2024). Thematic analysis with open-source generative AI and machine learning: A new method for inductive qualitative codebook development. arXiv preprint arXiv:2410.03721.

Lechuga, V. M. (2011). Faculty-graduate student mentoring relationships: Mentors' perceived roles and responsibilities. *Higher Education*, *62*(6), 757–771. https://doi.org/10.1007/s10734-011-9416-0

Lorentz, K. G., Mallinson, D. J., Hellwege, J. M., Phoenix, D. L., & Strachan, J. C. (2022). *Strategies for navigating graduate school and beyond*. American Political Science Association.

Lovitts, B. E. (2001). *Leaving the ivory tower: The causes and consequences of departure from doctoral study*. Bloomsbury Publishing PLC.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia Medica*, 143–149.

Mendoza-Denton, R., Patt, C., & Richards, M. (2018). Go beyond bias training. *Nature*, *557*(7705), 299–301. https://doi.org/10.1038/d41586-018-05144-7

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage.

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

Noah, S., Shen, M., Erowid, E., Erowid, F., & Silver, M. (2024). A novel method for quantitative analysis of subjective experience reports: Application to psychedelic visual experiences. *Frontiers in Psychology*, *15*, Article 1397064. https://doi.org/10.3389/fpsyg.2024.1397064

OpenAI. (2024). *Privacy policy*. OpenAI. https://openai.com/policies/row-privacy-policy/

OpenAI. (n.d.-b). GPT-4 turbo in the OpenAI API. OpenAI. https://help.openai.com/en/articles/8555510-gpt-4-turbo-in-the-openai-api

OpenAI. (n.d.-a). *Fine-tuning*. OpenAI. https://platform.openai.com/docs/guides/fine-tuning/hyperparameters

Pattyn, F. (2025). The value of generative AI for qualitative research: A pilot study. *Journal of Data Science and Intelligent Systems*, *3*(3), 184–191. https://doi.org/10.47852/bonviewjdsis42022964

Pollard, R., & Kumar, S. (2021). Mentoring graduate students online: Strategies and challenges. *International Review of Research in Open and Distance Learning*, *22*(2), 267–284. https://doi.org/10.19173/irrodl.v22i2.5093

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems* (pp. 1–7). Association for Computing Machinery.

Sackett, D. L. (1983). Proposals for the health sciences-I compulsory retirement for experts. *Journal of Chronic Diseases*, *36*(7), 545–547. https://doi.org/10.1016/0021-9681(83)90132-7

Saldaña, J. (2021). *The coding manual for qualitative researchers*. Sage.

Weidman, J. C., Twale, D. J., & Stein, E. L. (2001). *Socialization of graduate and professional students in higher education: A perilous passage?* (ASHE-ERIC Higher Education ReportJossey-Bass, Vol. *28*, No. 3).

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., & Schwartz, O. (2018). *AI now report 2018* (pp. 1–62): AI Now Institute at New York University.

Wu, J., Guzman, L., Patt, C., Eppig, A., & Mendoza-Denton, R. (2025). Can program structure advance equity in graduate education? *Innovative Higher Education*, 1–23. https://doi.org/10.1007/s10755-025-09808-x