Using AI to Understand Students' Self-Assessments of their Writing

Madeleine Sorapure, Seth Erickson, Sarah Hirsch & Kenny Smith

[1] UC Santa Barbara | US

Abstract: This study focuses on a generative AI approach to facilitate qualitative analysis in Writing Studies research. We gathered 13,336 one-sentence to one-paragraph responses written by 3,334 incoming students in a directed self-placement program administered at a large R1 U.S. university. In these responses, students describe their high school writing experience and college writing expectations. In stage one of the project, we pilot the use of Retrieval-Augmented Generation to expedite the selection of relevant responses for a topic—in this case, students' positive self-assessments as writers. The selected responses were then compared to a random sample and rated by three faculty with writing expertise. In stage two, these faculty generated codes and themes from a subset of the responses, incorporating ChatGPT-4 through the stages of thematic analysis. Results show that the use of AI expedites and enhances qualitative analysis, but human participation in the process is still essential. We suggest a machine-in-the-loop framework with which Writing Studies researchers can more readily integrate generative AI to study large corpora of student writing.

Keywords: artificial intelligence, qualitative analysis, writing assessment, writing placement

Sorapure, M., Erickson, S., Hirsch, S., & Smith, K.. (2026). Using AI to understand students' self-assessments of their writing. *Journal of Writing Research*, *17(3),* 555-579. DOI: https://doi.org/10.17239/jowr-2026.17.03.07

Contact: Madeleine Sorapure, 6824 Shadowbrook Drive, Goleta, CA 93117 | US - sorapure@ucsb.edu

1. Introduction

In the summer of 2020, UC Santa Barbara, a large R1 U.S. university, began offering a local writing placement program, Collaborative Writing Placement (CWP), to determine whether incoming first-year students had satisfied a university requirement and to place them in their initial first-year college writing course. Other U.S. colleges and universities have developed writing placement programs to introduce students to the local curriculum and, in some cases, to replace standardized tests like the SAT and ACT as placement mechanisms that may perpetuate biases against lower-resourced schools and students (e.g., Au, 2022). The CWP provides students with information about the first-year writing courses that we offer and asks seventeen sliding-scale questions and four open-ended questions designed to encourage students to reflect and write about their high school reading and writing experience as well as their expectations and sense of preparation for college-level writing (see Appendix A). Since 2020, more than 6,000 students have taken the CWP and nearly 5,000 have given us consent to use their data for research purposes. We thus have a very large repository of approximately 20,000 one-sentence to one-paragraph responses to open-ended questions from which we can learn about our students' high school writing experiences and college writing expectations.

This study focuses on 3,334 anonymized student responses gathered between summer 2020 and summer 2023, exploring this corpus using Retrieval-Augmented Generation (RAG) with two large language models (OpenAI and MXBAI) in the first stage of the study and using ChatGPT-4 in the second stage. To be sure, there are many topics we could explore in this corpus that would help us learn more about our students and inform pedagogical improvements. We could, for example, ask about the genres of writing that students are familiar with, their understanding of rhetorical analysis, their experience with writing from sources, their sense of the differences between high school and college writing, or their identification of weaknesses in their writing. For this study, we selected the topic of students' positive self-assessments as writers. This topic interests many writing instructors who use asset-based approaches to build on students' existing strengths. The topic also provides a challenge to AI's ability to select relevant responses since students identify different types of writing strengths and describe them using varied terminology. While we do not make specific pedagogical recommendations based on our AI-assisted thematic analysis, our findings provide a foundation for future studies to explore practical applications.

2. Literature review

2.1 Qualitative research on large corpora in writing studies

Several important projects involving the qualitative analysis of large corpora of student writing have helped to shape research and pedagogy in Writing Studies. The multi-year longitudinal studies at Harvard (Sommers, 2006; Sommers, 2008), Stanford (Lunsford, Fishman, and Liew,

2013), and Michigan (Gere, 2019) follow large numbers of students over their academic careers, collecting writing samples along with survey and interview data that is coded and interpreted to derive insights about student writing development. Studies of error in student writing (Connors and Lunsford, 1988; Lunsford and Lunsford, 2008) draw on many samples from multiple institutions. Baillargeon (2025) studied more than 1000 reflections from 278 dissertation writers to learn how graduate students manage the complex "multi-everything" genre of the PhD dissertation. Research on writing assignments (Britton,1975; Melzer, 2009) has employed qualitative methods on large datasets to identify patterns across institutions. Writing placement responses, like the ones we study here, have also been qualitatively studied in relatively large numbers. For example, Gere et al. (2013) and Tinkle et al. (2024) analyze collections from the University of Michigan's directed self-placement program to understand students' writing experience and skills.

Advances in digital technology have made it increasingly feasible for Writing Studies scholars to collect greater amounts of student writing and pursue longitudinal or cross-institutional qualitative research on larger scales (Licastro and Miller, 2021). In many writing courses, student work is submitted digitally and stored in a Content Management System (CMS); interviews can be recorded and transcribed automatically; surveys with open-ended questions can be sent and stored via Google, Qualtrics, and other programs; writing placement programs can collect responses and entire portfolios of student work digitally. In short, writing researchers can now more easily assemble large collections of diverse types of writing to help instructors understand their students and develop more effective pedagogies. In addition, advances in data storage and sharing create opportunities for cross-campus collaborations that support Writing Studies research aligned with Haswell's (2005) vision of scholarship that is RAD: replicable, aggregable, and data-supported.

However, traditional methods of manually coding and analyzing qualitative data may struggle with ever larger corpora. The sheer volume and complexity of textual data from writing courses—student drafts, essays, portfolios, reflections, discussion posts, peer feedback, instructor feedback, and interview and focus group transcripts—transforms the landscape of qualitative research by making possible many different studies about types and stages of student writing. This unprecedented access to diverse textual data presents both opportunities and methodological challenges. While qualitative analysis is powerful precisely because it can yield "deep understanding that reflects the richness, complexity, and, at times, contradictory nature of people and phenomena" (Feuston and Brubaker, 2021), it is already a time- and resource-intensive approach to research, and it is likely to become more complex and more costly when dealing with large corpora requiring multiple trained coders. The challenge is maintaining qualitative analysis's richness and nuance when scaled to handle larger volumes of data.

2.2 Using AI in qualitative analysis

A growing body of scholarship investigates challenges and opportunities of large datasets for qualitative scholarship by asking whether, when, and how to incorporate generative AI using

large language models (LLMs). Although the technology behind generative AI is relatively recent, qualitative researchers are already using it in commercial software. For instance, AI-assisted analysis is included in NVivo's Autocoding feature, MaxQDA's AI Assist, and ATLAS.ti's AI Coding (Nguyen-Trung, 2024; Paulus and Marone, 2024). Researchers are also developing tools specifically to facilitate AI-assisted qualitative research (e.g., Gao et al., 2023; Gebreegziabher et al., 2023; Hong et al., 2022; Rietz and Maedche, 2021). Researchers are also experimenting with generative AI LLMs such as ChatGPT to develop and test workflows incorporating AI (e.g., Dai et al., 2023; Hitch, 2024; Jalali and Akhaven, 2023; Mesec, 2023; Nguyen-Trung, 2024; Sinha et al., 2024; Turobov et al., 2024; Wachinger et al., 2024).

Current research identifies two areas where AI seems particularly helpful in qualitative analysis. First, AI can increase efficiency by automating labor-intensive coding tasks, allowing researchers to focus on complex analytical work in later stages of the process (Feuston and Brubaker, 2021; Turobov et al., 2024; Yan et al., 2024; Zhang et al., 2024). For initial understanding of large unstructured datasets, a first pass with generative AI can suggest key themes and initial coding frameworks, saving substantial time and effort. Generative AI tools like ChatGPT can efficiently synthesize information and rapidly generate initial coding schemes in inductive analysis, providing a solid foundation for refinement and further analysis (Hamilton et al., 2023; Perkins and Roe, 2024). Second, AI can improve insights and add value to qualitative analysis by identifying patterns, themes, and connections across data that human analysts might miss (Perkins and Roe, 2024; Sinha et al., 2024; Turobov et al., 2024). Ibrahim and Voyer (2023) suggest that AI can also serve as a robustness check with the human-generated codes "by corroborating and challenging the researcher's interpretations and conclusions."

Recent studies systematically comparing human and AI qualitative coding inform our project's second stage. Researchers have structured these comparisons in several ways. Hamilton et al. (2024) conducted direct comparisons with both human coders and AI analyzing the same dataset. Morgan (2023) applied AI coding to datasets that he had previously coded. Perkins and Roe (2024) explored a hybrid method where one researcher coded manually while another incorporated AI assistance. Dai et al. (2024) developed an innovative iterative approach, putting a human coder in "conversation" with the AI machine coder, creating a loop in which human and AI alternated tasks and refined results. Studies have also examined AI's capability in deductive coding using pre-existing codebooks (Kirsten et al., 2024; Xiao et al., 2024).

These comparative studies engage multiple qualitative methodologies, including thematic analysis (De Paoli, 2024; Nguyen-Trung, 2024; Perkins and Roe, 2024), reflexive thematic analysis (Hitch, 2024; Wachinger et al., 2024), phenomenological analysis (Hamilton et al., 2023), and grounded theory (Sinha et al., 2023). Across all methodologies and frameworks, a consistent finding emerges: generative AI can serve as an active participant in data interpretation with the potential to expedite and enhance qualitative analysis.

In our research's second stage, we follow De Paoli (2024), Hitch (2024), and Perkins and Roe (2024) in using ChatGPT-4 in the first five phases of Braun and Clarke's (2006) thematic

analysis: data familiarization, coding, developing initial themes, reviewing themes, and defining themes. The sixth phase of thematic analysis, writing the report, is not an area covered by this study or by other studies cited here. We compare human and AI coder results to understand strengths and weaknesses of the process.

2.3 Challenges with using AI in qualitative analysis

While AI can enhance efficiency and provide new insights, its use raises several issues requiring careful consideration. First, inherent biases in LLMs, resulting from biases in the data they are trained on, can affect research integrity. Unlike humans, ChatGPT and other GenAI tools can't reflect on their own biases or step outside of their own training data to recognize limiting preconceptions. In addition, AI tools that "learn" from interactions can bias subsequent analyses based on previous results (Wachinger et al., 2024). Inconsistent results from identical prompts force researchers to spend time comparing and verifying outputs (Sallam, 2023; Yan et al., 2024). Hallucinations are another challenge for LLMs, when they generate incorrect but plausible results beyond their training data that might not be apparent to researchers (DePaoli, 2023; Morgan, 2023; Zhao et al. 2024). More generally, qualitative researchers are skeptical about the black-box quality of AI because decision-making processes and sources remain unclear. As Ibrahim and Voyer (2023) point out, though, human coders are a bit of black box as well, with a sometimes incomplete understanding of the reasons for their own choices.

Finally and perhaps most importantly, AI can miss or misinterpret subtleties of human experience and subjective knowledge that are crucial in qualitative research (Yan et al., 2024). Wachinger et al. (2024) note "a sense that AI-assisted analysis might impede or undercut the human essence of qualitative research" by undermining researcher engagement with the data. While human researchers bring perspectives, experiences, and biases that can affect interpretation and coding, many qualitative researchers see this human element as positive and essential rather than limiting. The human researcher serves as an instrument in the process, with expertise needed to make the data meaningful. The challenge, then, is to develop frameworks that integrate AI assistance with human expertise while ensuring that qualitative analysis remains deeply interpretive and reflective.

2.4 Human-AI collaboration

Most researchers agree that frameworks for including AI in qualitative analysis must involve humans, although this involvement is conceptualized differently. AI is described alternately as a tool (Feuston and Brubaker, 2021), research assistant (Gebreegziabher et al., 2023; Ibrahim and Voyer, 2023), collaborator (Yan et al., 2024), co-pilot (Perkins and Roe, 2024), synergistic partner (Jiang et al., 2021; Sinha et al., 2024)--even an extension of the human cognitive process (Zhang et al., 2024).

For those who see AI more as a tool, the human analyst delegates tasks (Jiang et al., 2021; Lubars and Tan, 2019) and supervises (Wachinger et al., 2024). Those who see AI more as a collaborator describe conversational interactions facilitated by its natural language interfaces

(Dai et al., 2023; Zhang et al., 2024), with feedback loops where both human and AI adjust their responses during the process (Mesec, 2023; Yan et al., 2024). These approaches represent two ends of a spectrum: "machine-in-the-loop," where humans direct AI while maintaining control, versus "human-in-the-loop," where AI leads the analysis while humans oversee and refine. Overall it seems clear that incorporating generative AI fundamentally changes qualitative research. It is not the same practice with a different tool, but a different practice with new opportunities and challenges. As Perkins and Roe (2024) note, "The integration of GenAI tools into the academic realm signifies more than just technological advancement; it embodies a true paradigm shift in how research is conceptualized and executed."

2.5 RAG approach

Some weaknesses identified above in incorporating generative AI into qualitative analysis are addressed by Retrieval-Augmented Generation (RAG), a framework that enhances the capabilities of large language models by integrating external knowledge retrieval. RAG addresses the potential bias and knowledge gaps of LLMs by retrieving relevant information from specific external sources to improve the quality of the responses. Briefly, RAG directs an LLM to use only data from external authoritative sources when answering questions. RAG therefore avoids the bias, hallucinations, and inaccuracies from LLMs like ChatGPT that access a wide array of public data without specialized or up-to-date knowledge in every area.

The RAG model works by creating embeddings or numerical representations of the textual data in multidimensional vector space. The model can then perform relevancy searches to return the closest or most relevant responses to a specific prompt and can also generate answers that put relevant responses in context. Described initially by Lewis et al. (2020), rapidly growing research demonstrates RAG's potential to improve AI-generated content. Gao et al. (2024) survey the research on integrating RAG and LLMs, charting the evolution and anticipated future paths of RAG-assisted analysis. Zhao et al. (2024) describe methodologies, benchmarks, and limitations. Balaguer et al. (2024) compare RAG and fine-tuning, finding that both improve the quality of the LLM's responses, though fine-tuning requires costly computing resources and human expertise (Ovadia et al., 2024).

In the first stage of this study, RAG was developed in Python specifically for our student writing dataset, and we interacted with the program via Jupyter Notebook. Several LLMs — including OpenAI, Anthropic's Claude, Google's Gemini, and others —now feature native RAG functionality, making it easier for researchers to use this approach going forward.

2.6 Students' Self-Assessments

In our study of students' responses to open-ended questions in the CWP, we examine how students positively assess their writing abilities. This focus aligns with core principles of directed self-placement (DSP): students should have agency in course placement decisions, and institutions should value students' insights about their experience and their perceived needs for college writing instruction (Moos and Van Zanen, 2019). Tinkle et al.'s (2024) mixed-

methods study examined students' self-characterizations as writers in the UWrite University of Michigan placement program to learn how students "differentially construct their identities as writers" and understand SSP's broader "ecological" impacts. In the dataset of 5,422 responses, Tinkle et al. (2024) used both quantitative and qualitative measures, performing sentiment analysis computationally and thematic analysis manually. In the thematic analysis, they report that "all of the 5,422 responses were read by two project members to better understand the many themes present."

While we also ultimately develop codes and themes to discover the positive features that students identify in their writing, our study differs significantly from Tinkle et al. (2024). Most importantly, a key goal of our project is to test a generative AI method that would enable an accurate and effective overview of the responses without manually coding all of them, as Tinkle et al. (2024) do. Second, the UWrite placement directly asked students about their writing strengths and challenges, so researchers did not need to explore the corpus for relevant responses, as we do in the first stage of our project. Unless students misread or ignored the UWrite question, all of the responses to the question would be relevant. The four open-ended questions in the CWP, in contrast, elicited student commentary on a wide range of topics related to high school and college writing, requiring us to initially search for relevant responses about positive self-assessment. More broadly, researchers can overcome the limitations of targeted surveying—for instance, low response rates and response bias—by extracting relevant datasets from larger, more general corpora.

3. Research questions

1. How do relevancy results returned from the RAG model compare to results returned from random selection?
2. How do faculty with expertise in the CWP rate the relevancy results returned from the RAG model?
3. How do the positive self-assessment attributes indicated by expert human coders compare to those indicated by AI?

4. Methods

The CWP was administered 17 times from summer 2020 through summer 2023, with 4,001 students completing the placement during this period. We assembled a spreadsheet that collected these responses before deidentifying the data. 3,334 students taking the CWP during this period gave us consent to use their data for research purposes; 767 students did not give consent.

4.1 Stage 1

To answer our R1 and R2, in the first stage of this project we implemented a RAG approach to identify relevant responses via two LLMs, compared these results to randomly generated

responses, and assessed the results with expert human coders. A custom Python script, developed for this stage of the project by one of the authors, generated embeddings for each of the 13,336 responses to the four open-ended questions in the CWP (see https://osf.io/875up/). Two embeddings models were used: OpenAI's "text-embedding-3-small" (OpenAI, 2023) and "mxbai-embed-large-v1" (Lee et al., 2024). These models were selected because they represent the state-of-the-art and because they are suitable for short text fragments, like the student responses. We also wanted to compare results across LLMs; the models differ in size, training data, and architecture and those differences may affect their output.

Along with the responses, we generated embeddings for four different prompts:

- I am confident about my writing.
- I feel prepared for college writing.
- I am a strong writer.
- I can write well.

We used multiple prompts to explore the concept of positive self-assessment, measuring the results across each prompt and each LLM. Human readers would see each of these prompts as slightly different variations on the theme of positively assessing oneself as a writer. That is, the statements use different verbs (being, feeling, being able), different positive assessment categories (confidence, preparation, strength, quality), and different references to writing (my writing, college writing, writer, write), but they all indicate a positive assessment. We wanted to measure the extent to which RAG would return some of the same responses across these different categories given that they express similar content.

For each prompt in each LLM, the program was directed to return the 50 most relevant results. Following a RAG approach, "relevance" was calculated using the euclidean distance between the embedding generated from the prompt and those of the responses.[1] The most relevant responses are those closest to the embedded prompt in this vector space. The 400 responses generated from this process (50 from each of the four prompts in each LLM) were then checked for duplicates, leaving 260 responses. In addition, 50 responses were randomly selected from the spreadsheet containing all 13,336 responses and added to this set, creating a total of 310 non-duplicated responses. The random responses were included to assess whether a RAG-based relevancy search offers a more efficient method of dataset exploration than random sampling.

Three of the co-authors met to develop and apply a rating system for these 310 responses that would assess whether and to what extent they expressed a student's positive self-assessment of their writing. These three Writing Program faculty members initially developed the CWP and are very familiar with student responses to the open-ended questions by virtue of having read many such responses during placement processes. They also have a combined

[1] We used euclidean distance instead of cosine similarity because the embeddings from Ollama's implementation of mxbai-embed-large model were not normalized.

70+ years of experience teaching writing to undergraduate students. Table 1 shows the rating system they developed.

Table 1. Categories for rating responses, with examples

Rating	Explanation	Example
1	**no self-assessment:** student doesn't comment on their writing at all	Student2658: "The first reading is pretty similar to SAT readings, so I am familiar with it. Overall it's not too hard to understand. With the caption at the start of each new statement, the reading is easier to understand. The second statement is a bit harder to understand compared to the first one. Part of the reason is that the reading is longer, and you have to connect everything to the topic."
2	**negative self-assessment:** student states that they are a bad, poor, or weak writer, or that they are not confident at all about their writing	Student354: "I've had trouble in English classes time and time again. I started honors classes and dropped out 3 times. English isn't something that comes easy to me, and I constantly ramble on in my essays. I find it difficult to add evidence to my essays and often times grab something out of thin air that is not seen as substantial in comparison to other students. I would prefer starting out college in an environment where I can learn the absolute basics of college writing before continuing on further in my education."
3	**implicit self-assessment:** student describes writing they've done in the past without stating that they did it well or poorly	Student 1359: "The students' writing samples are very similar to those I have done. The writing style is formal-basic, the length of the texts is standard. I have a big experience writing those assignments because for the school those types are conventional."
4	**mixed self-assessment:** student states that they are a strong writer in some ways and weak in others, that they are confident or prepared about some aspects of their writing but not others	Student 3284: "I feel I am not yet able to write at the same level as the students used as examples yet. That being said I feel I have a good background to be able to learn how to write like that. I have always been good at creative writing in my English classes and have been able to get through academic writing even though it does not come as easily to me."
5	**positive self-assessment:** student states that they are generally a good,	Student 2428: "Overall, I believe I am a pretty solid writer. I've authored numerous stories, essays, and articles. I find writing to be enjoyable and therapeutic, writing for hours

	capable, confident, well-prepared writer; descriptions of past writing include some positive evaluation	while enjoying your favorite music or perhaps one of Beethoven's most notable sonatas. It can feel pretty exhilarating to finally complete a project you've been working on for hours on end. Like any other person, I have my uncertainties. Starting with Writing 2 may be difficult, but I'm willing to face the challenge."
6	**highly positive self-assessment:** student states that they are very confident about their writing, that they are a strong, capable, excellent writer and well prepared for college writing	Student 2784: "I am a skilled writer and am already competent in the skills outlined as goals of the writing 1 course. I have been an editor on two literary magazines, and am capable of writing well across a variety of genres. I am aware of what I'm doing as I write and edit, capable of engaging with complex texts, and have already developed a distinct voice as a writer. I will be bored and insufficiently challenged by the writing 1 course, and am ready to start writing 2."

Several of these rating categories required additional discussion. Specifically, the "implicit self-assessment" in rating 3 often took the form of a list of the students' previous writing projects and courses, likely meant to prove while not explicitly stating that the student is prepared and confident. The "mixed self-assessment" in rating 4 was considered a "correct" result because the prompts did not prohibit a negative assessment but only asked if a positive assessment was present. The distinction between "positive" and "highly positive" for ratings 5 and 6 was introduced to later assess the accuracy of the RAG distance measure.

The raters discussed and applied ratings together in this initial session, deliberating and making modifications to the rating descriptions and to their own assigned ratings. In several following weeks, they then each rated all 310 responses. They met again to resolve any differences in ratings and came to an agreement—an "expert rating"—on all responses. This expert rating was then used to assess the results returned by RAG for the eight prompt/LLM combinations and by the random sample.

4.2 Stage 2

To answer our R3, in the second stage of this project we created a new dataset with the 212 responses that were returned by prompt/LLM combinations and that received "correct" ratings of 4, 5, and 6. We randomly selected 100 of these responses to use in developing codes and themes; because responses rated 4 included both positive and negative self-assessments, we selected only from the 5 and 6 categories so as to focus on coding only positive statements. Three of the authors then worked through the first five phases of thematic analysis as defined by Braun and Clarke (2006) and enacted in De Paoli (2024), Hitch (2024), and Perkins and Roe (2024). We followed Hitch (2024) in particular by interacting with ChatGPT-4 at each phase in

the analysis, rather than performing code and theme development separately by humans and by AI, as other researchers have done. Integrating and interacting with the AI results during each phase of thematic analysis maintains human control over the process while integrating AI as a tool for identifying new elements and potential revisions. This method aligns with the advice of Sinha et al. (2024) who recommend that an AI workflow incorporate a "constant comparison—a comparison of researchers' codes with that of the GenAI model's codes."

The three authors began the initial data familiarization phase with existing knowledge of the dataset, having previously rated responses from the RAG and random selection methods. To deepen this familiarity and gain new perspectives, we engaged with ChatGPT-4 to generate summaries at varying lengths (50, 150, and 250 words) of the 100 selected responses. Next, following a process similar to that used by Sinha et al. (2024), the authors individually conducted open coding of the 100 responses using Dedoose (https://dedoose.com). We then met to synthesize and reconcile our codes, arriving at a consolidated set, after which we engaged ChatGPT-4 to independently generate its own codes from the 100 responses. We compared its output with our own and made modifications before moving on to the next phase. We followed this same iterative process—human collaboration followed by AI comparison and refinement—through the subsequent phases of developing initial themes, reviewing themes, and naming/defining themes. Using a machine-in-the-loop approach, the authors first worked together in each phase to generate a shared output before obtaining and comparing AI-generated results, using these insights to refine our work before proceeding to the next phase.

5. Results

5.1 Stage 1

Regarding our R1 in stage one of our project, the results returned from the RAG model were significantly more likely to express a positive self-assessment of writing than the results returned from a random sample. Table 2 shows the percentage of "correct" results in the 4, 5, and 6 rating categories that were returned for each prompt/LLM combination and for the random sample.

Table 2. Percentage of "correct" results returned

LLM/prompt	percentage in 4, 5, and 6 rating categories
MXBAI/confident	96
MXBAI/prepared	82
MXBAI/strong	74
MXBAI/well	80
OpenAI/confident	96
OpenAI/prepared	92
OpenAI/strong	86
OpenAI/well	82
random	30

RAG clearly outperforms the random sample in returning relevant results for our query. Figure 1 provides more detail; responses in the 4, 5, and 6 categories in our expert rating constitute the majority of the results from all four prompts in both LLMs, whereas the random sample returned results that were mostly not relevant for the topic. In other words, RAG effectively generated mostly relevant results across all prompts and LLMs. Whereas a random sample of responses yielded 30 % in which students positively assessed their writing, the prompt/LLM combinations yielded an average or 86 % relevant responses, with OpenAI averaging 89 % correct and MXBAI averaging 83 % correct. In the best case, using the prompt "I feel confident about my writing" yielded 96 % relevant responses from both OpenAI and MXBAI. This suggests that qualitative researchers using RAG would be able to explore a large corpus with a specific question or set of questions much more efficiently than if they used a random sample.

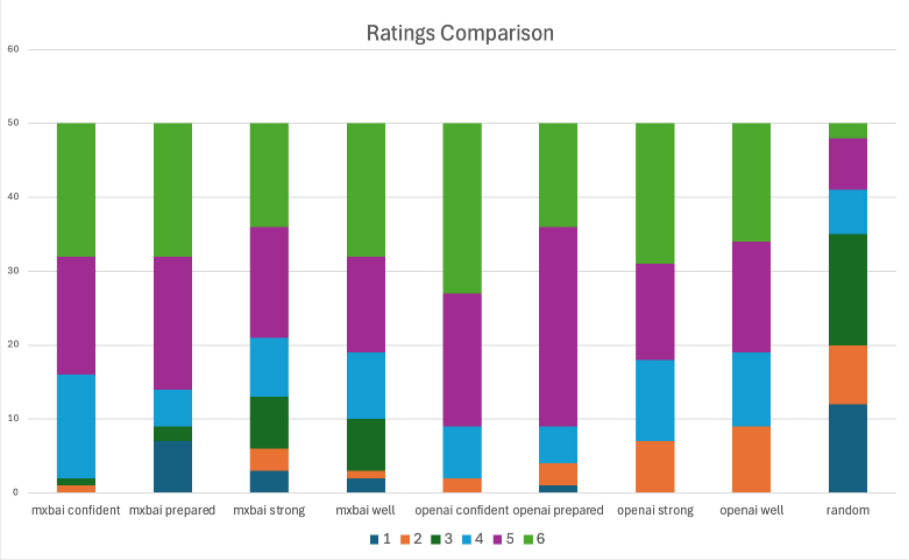

Figure 1: Ratings comparison.

Regarding our R2 in stage one of our project, Table 3 shows the percentage of results for each expert rating that was delivered via the two LLMs and via random sampling. For instance, 6 % of the responses returned by MXBAI across all four prompts were given a rating of 1 (no self-assessment present), while .5 % of the responses returned by OpenAI across all prompts and 24 % of the responses returned by a random sample were given a rating of 1. There are some notable differences between the LLMs. OpenAI returned more negative assessments than did MXBAI, with 10.5 % of responses receiving a rating of 2, or negative self-assessment. For instance, for the prompt "I am a strong writer," OpenAI returned this result from Student426: "I feel like my writing abilities are not very strong. I have never really enjoyed English nor have I ever really been a strong writer. I feel like taking writing 1 will provide me the best opportunity to learn valuable skills and prepare nicely in my transition to writing 2. Yes it will be hard but I am here to learn and help sharpen my abilities in all courses. Writing included." While the passage ends with an overall positive tone, the negation in the first two sentences clearly indicates that this is not a positive self-assessment. MXBAI, in contrast, returned few negative assessments but more non-assessments (6 %) and more implicit assessments (8.5 %). These differences between LLMs, particularly at the lower rating levels, suggest the importance of having human coders check for edge cases and errors in RAG-generated results. Comparing the efficacy of each of the four prompts, we see in Figure 2 that they returned roughly the same numbers of responses for each expert rating, with *confidence* and

preparation being the most effective prompts for the positive and highly positive ratings of 5 and 6.

Table 3. Ratings comparison, LMs vs. random, in percentages

	1	2	3	4	5	6
mxbai	6	2.5	8.5	18	31	34
openai	.5	10.5	0	16.5	36.5	36
random	24	16	30	12	14	4

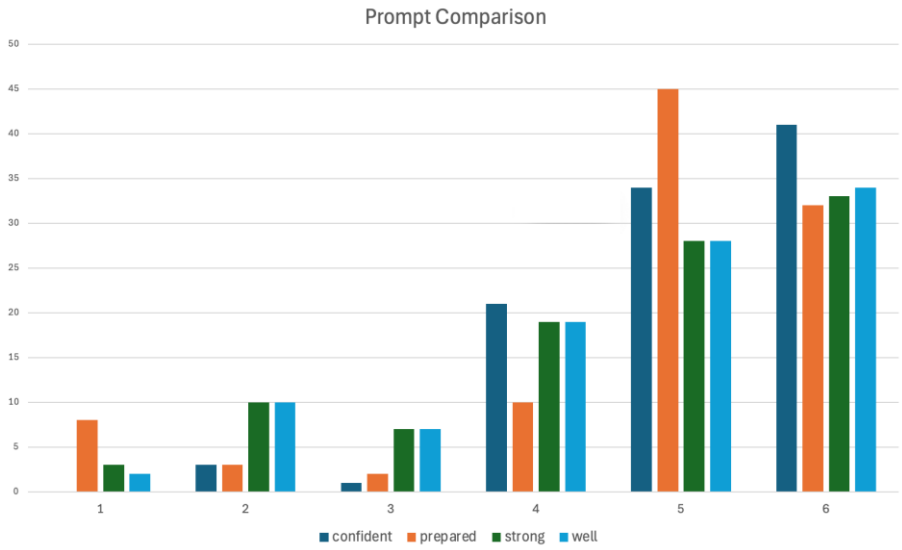

Figure 2: Prompt comparison.

The distances returned by RAG for the responses from the LLMs did not show any substantial differences. One would expect that responses with a rating of 5 or 6 would be closer to the prompts in RAG's multidimensional vector space and thus more relevant. However, as Figure 3 shows, the distances are more or less the same in both LLMs (for OpenAI, there were no results in category 3, implicit self-assessment). For MXBAI, responses in category 1 (no assessment) are slightly closer than for category 6 (highly positive assessment); for OpenAI,

the average distances from the prompts are the same. This indicates that the distance measurement is not only insignificant but also perhaps slightly misleading.

Finally, 35 % of the responses were returned by two or more prompt/LLM combinations, as Table 4 shows. For 221 out of the 400 responses returned by the two LLMs across four prompts, there were no duplicates; 56 responses were returned by two prompt/LLM combinations. At the other end of the spectrum, two responses were returned by six prompt/LLM combinations. This shows that there is some overlap across the LLMs and across the prompts that express positive self-assessment in different ways.

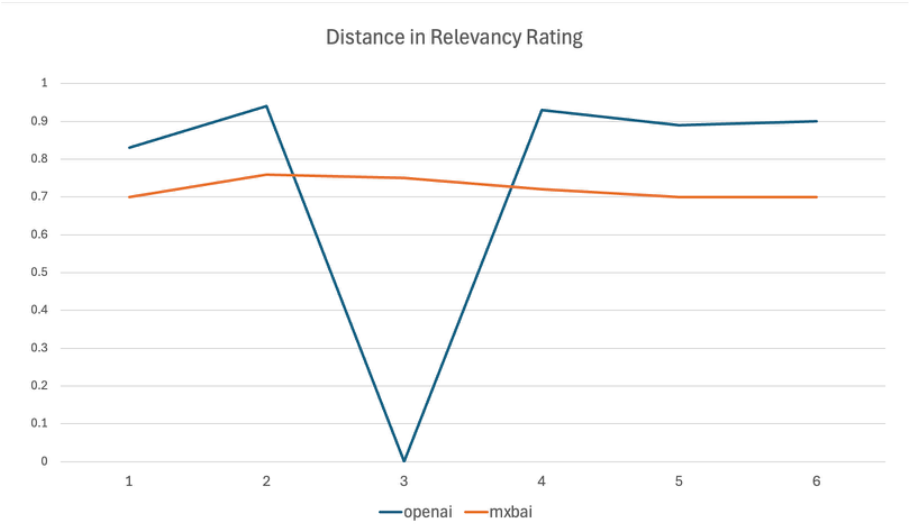

Figure 3: Distance in relevancy rating.

Table 4. Number of duplicates per expert rating

Number of duplicates	none	two	three	four	five	six
	221	56	22	6	3	2

In Table 5, we see that responses returned more frequently had higher expert ratings. For instance, the 221 responses that had no duplicates had an average rating of 4, whereas the 2 responses that were returned by six prompt/LLM combinations had an average rating of 6.

Table 5. Average expert rating of duplicates

Number of duplicates	none	two	three	four	five	six
Average rating of duplicates	4	4.9	5.1	5.3	5.7	6

Given that the RAG distance rating is questionable in terms of establishing relevance among returned results, we conclude that using multiple prompts and noting duplicate results is a more effective method of identifying the most relevant responses in a corpus.

5.2 Stage 2

Regarding our R3 in stage two of our project, asking how the positive self-assessment attributes identified by expert humans compare to the attributes identified by ChatGPT-4, we found similarities and differences at each phase in the thematic analysis process. Table 6 summarizes our observations of the responses ChatGPT-4 provided.

In the data familiarization phase, the three authors were already familiar with students' responses by virtue of having read them while developing the expert ratings in the first stage of this project. We prompted ChatGPT to generate summaries of varying lengths: 50, 150, and 250 words. These summaries revealed a consistent thematic organization. Key themes—personal experiences, challenges, and advanced coursework—appeared in all three summaries, with longer summaries simply providing additional detail rather than introducing new themes. The summaries seemed generally accurate and mostly descriptive rather than interpretive. However, in the 250-word summary, the model offered one interesting interpretive insight, noting a "mix of optimism and realism" in the responses. This observation suggests that prompting for longer summaries or explicitly prompting for interpretation could enhance AI's usefulness in the data familiarization phase. Still, as Hitch (2024) comments, over-reliance on AI at this phase "would prevent researchers from gaining the deep understanding of the richness in their data which underpins the rest of this analytical approach."

In phase 2 of the thematic analysis process, the coding phase, the three authors generated 183 initial codes altogether in their individual readings of the 100 responses. As Braun and Clarke (2006) explain, codes "identify a feature of the data (semantic content or latent) that appears interesting to the analyst." In a shared Google spreadsheet, we sorted and combined our codes, noting substantial overlaps along with differences in naming and assigning codes. For instance, one author assigned the codes "confident" and "prepared" to any students who used those words in their response, whereas the other two authors used that code for students who stated that they were confident or prepared without providing any supporting evidence. We ultimately decided on the latter approach. This collaborative engagement with coding reflects Kantor's (2024) advice: "given that the coding process is itself a way for the researcher to engage with the data and refine hypotheses, it is critical to avoid over-automating this approach in the name of efficiency, thereby undermining the researcher's ability to engage with their data adequately and deeply."

Table 6. Phases of thematic analysis and observations of ChatGPT-4 responses

Phases	Observations of ChatGPT-4 responses
1. Data familiarization	• summaries of different length all mentioned the same key points • accurate but mostly descriptive rather than interpretive
2. Coding	• missed all of the contextual codes related to the CWP • otherwise very similar to human-generated codes • no outliers or hallucinations • picked up elements we considered less important (e.g., knowing citation formats)
3. Developing initial themes	• substantial overlap with human-developed themes • helpful differences in interpreting and classifying key concepts (e.g., challenges, confidence, experience) • introduced new themes (e.g., transferable/real-world skills) • some questionable code assignment in theme categories
4. Reviewing themes	• substantial overlap with human-developed summaries of each theme • some overlap in selection of representative responses for each theme, especially in the less frequent themes • different representative responses helped us see the edges and contours of the theme
5. Defining themes	• substantial overlap with theme definitions • weakness with contextualizing theme descriptions and suggesting effective theme titles

We then prompted ChatGPT to generate codes for the 100 responses and we discussed the similarities and differences in the 75 codes provided by AI. While we found substantial overlap between human and AI-generated codes, two significant differences emerged. First, ChatGPT consistently missed contextual elements in student responses: their comparisons to CWP writing samples, reflections on writing experience in relation to CWP assignments, and self-assessments framed within the context of the placement materials. Even when provided with a more detailed prompt that included contextual information, ChatGPT's codes did not capture these references. Second, ChatGPT identified granular details in the responses that the human coders had deemed less significant and therefore hadn't recorded or even remembered—for instance, mentions of MLA and other citation formats, vocabulary skills, and experience with peer feedback. We wondered if these were examples of hallucinations on the part of ChatGPT, but a subsequent check confirmed that these details were present in

the responses. ChatGPT was highlighting elements we had considered too minor to code. It is also worth noting that each author spent approximately 3-4 hours coding, while ChatGPT generated its code set in approximately 20 seconds.

In phase 3 of the thematic analysis process, we began developing themes, which Braun and Clarke (2006) describe as broad interpretive patterns of shared meaning, united by a central concept or idea. Our iterative process began with sorting codes into preliminary categories, carefully distinguishing between nuanced concepts like writing experience (variety and volume) and writing competencies (specific skills and general traits). One author introduced the concept of "folk learning theories"—informal beliefs about learning acquired from teachers, parents, and peers—as a framework for understanding how students conceptualize and represent the process of becoming good writers. For instance, building a foundation, being driven by passion and enjoyment, taking on challenges, growing and developing skills, and employing natural talent provide frameworks of learning that students use in their responses. Working in a shared Google doc, we identified potential themes from the patterns and relationships we observed in the codes. In our subsequent discussion, we consolidated overlapping themes and refined theme names to incorporate each other's contributions.

We then turned to ChatGPT to generate themes from the 100 responses and 75 codes it had assigned. We crafted a detailed prompt explaining thematic analysis principles and we provided example themes to guide the AI's work. Table 7 below shows a comparison of human-generated and AI-generated themes, with overlapping themes aligned in rows. The first two themes mirror those that we had put in the prompt to serve as examples. For almost all other themes, the three authors and ChatGPT expressed similar ideas but organized or "sliced" them somewhat differently. For instance, where human coders associated overcoming writing challenges with developing confidence, ChatGPT connected challenge-related codes with academic rigor and growth opportunities. These different interpretive frameworks highlight how both humans and AI can derive meaningful but different insights from the same qualitative data. ChatGPT provided no equivalent for two themes that human coders identified: students described themselves as naturally talented or innately good writers, and students simply asserted that they are good writers without any evidence or explanation. Conversely, there was no equivalent among human coders for the theme of "Transferable skills and real-world applications" identified by ChatGPT. In this sense, then, the AI results confirmed our initial identification of themes, encouraged us to clarify and reframe important concepts in these themes, and augmented our findings by suggesting a theme we had not identified.

Following our theme revision incorporating ChatGPT's input, we conducted a thorough review to ensure the themes were both grounded in the data and coherently expressed—phase 4 of Braun and Clarke's (2006) thematic analysis process. During this validation phase, one author reviewed the coded responses to assess theme fit and identify potential exemplars. This review led to the identification of an additional theme: "support from others." Although this theme appeared in only six of the hundred responses, all three authors agreed that it captured

a distinct and significant element of students' writing assessment. For example, responses ranging from acknowledging "incredible English teachers" to describing help from a tutor showed that some students attributed their writing ability to external support in ways not captured by other themes. After validating the themes' presence in the data, we developed summaries and selected representative examples for each theme.

We then sought to explore our themes, summaries, and examples further with ChatGPT, prompting it to provide summaries of each theme. The prompt was: "Please summarize the data in the input below relevant to the following theme: [theme]. Your summary should be one or two sentences. Also extract two or three responses that best illustrate this theme." The summaries generated by humans and by ChatGPT were quite similar, but there were differences in the examples selected. In those instances where ChatGPT selected examples that we had not coded with a particular theme, the three authors reviewed ChatGPT's explanation for why it had chosen those examples along with the examples themselves, and we discussed whether to modify the theme's definition. Prompting ChatGPT for illustrative examples proved particularly valuable, as it highlighted nuances in the themes that we hadn't previously considered. For instance, the two examples that ChatGPT identified for the "support" theme explicitly detailed forms of support, such as personalized help and unique course experiences, whereas we chose examples that praised individuals without identifying a particular contribution. This process of comparing human and AI-selected examples helped us refine our understanding of each theme's distinctive characteristics. Moreover, we noted that ChatGPT tended to select longer responses as examples, where there might be additional information presented less concisely and with other factors, whereas human coders tended to choose examples that were more brief and that isolated the theme in question.

In the fifth phase of thematic analysis, we developed theme definitions and names following Braun and Clarke's (2006) guidance to articulate "the story that each theme tells" within the broader context of our research question. We developed definitions that we thought met these criteria and decided to call on ChatGPT to help formulate names for each of the themes based on the definitions and following Braun and Clarke's (2006) advice that theme names should be "concise, punchy, and immediately give the reader a sense of what the theme is about." Here ChatGPT was notably weak, offering theme names that were either dry summaries or oddly failed punchy phrases (see Appendix B).

Table 7. Comparison of Themes Developed in Stage 3

Human-generated themes	ChatGPT-generated themes
1. Assessing writing in context—in relation to other students, benchmarks (e.g., AP, IB tests), or other external references (e.g., teachers)	A. Benchmark against others and external validation: Students assess their writing in relation to their peers, external recognition, and competitive experiences
2. Being driven by passion, enjoyment, and intrinsic motivation	B. Passion, enjoyment, and intrinsic motivation: Students see writing as an activity they enjoy, a personal passion, or a way to express themselves, rather than merely an academic requirement.
3. Gaining confidence in writing through taking on challenges	C. Mastery through academic rigor: Students view their writing skills as a result of intensive academic coursework, high expectations, and exposure to challenging assignments
4. Acquiring experience and cataloging skills that demonstrate writing ability	D. Writing as a developed and refined skill: Students attribute their writing confidence to continuous improvement, mastery of technical aspects, and a structured approach to writing.
5. Employing natural talents and innate abilities	
6. Developing a strong foundation and building on it through sequential learning	E. Overcoming challenges and embracing growth: Students recognize challenges in writing but view them as opportunities for growth, improvement, and adaptation. See also 3 above.
7. Assertion without evidence or explanation	
8. Miscellaneous: multiple languages, unique perspectives, understands the value of writing	F. Miscellaneous (for now): Self-perceived unique writing style; excitement for interdisciplinary tasks; support from excellent teachers; experience integrating sources into writing. G. Transferable skills and real-world applications: Students believe their writing abilities extend beyond the classroom and are useful in professional, research, or real-world contexts.

For instance, in naming the "challenge" theme, ChatGPT suggested several dry summary titles, such as "Writing Resilience: Past Trials, Future Triumphs" and "Conquering Adversity, Showcasing Writing Excellence." When prompted for more "punchy" titles, ChatGPT

suggested titles such as "Bold Writing. Tough Trials" and "Writing Forged in Challenge." In contrast, the human coders settled fairly quickly on the concept of "grit," a common way of describing how learners in general improve by sustaining effort through challenges and adversity, and we settled on the title "The Grit Mindset," paralleling the "The Growth Mindset" title of another theme. Unlike ChatGPT, the authors were able to draw on a broader cultural knowledge and an understanding of concepts (including "mindset") that resonate in terms of our research question and the context of education.

6. Discussion

As we demonstrate in the first stage of this study, RAG provides clear advantages over random sampling as a method of extracting relevant responses from a large corpus of open-ended answers. For qualitative researchers analyzing surveys where respondents cover multiple topics, using RAG can significantly reduce the time-intensive and tedious process of identifying relevant responses for specific queries. Since it is not always possible to ask targeted questions and generate robust response rates on surveys, selecting relevant responses from broader corpora may provide a viable alternative for researchers.

However, the effectiveness of RAG-based selection varies across different LLMs. In our comparison, OpenAI and MXBAI exhibited distinct patterns of errors. OpenAI tended to return more self-assessments that were entirely negative, indicating that it sometimes failed to see positive terms (e.g., confident, prepared, strong) in context (e.g., not confident, poorly prepared, not strong). MXBAI results skewed toward responses that either lacked any self-assessment or that only implicitly suggested positive self-assessment by listing many different types of writing experience. The variation in results targeting positive self-assessments across different prompt/LLM combinations suggests that researchers should either test prompts to optimize selection or deliberately use multiple prompts to capture broader semantic patterns.

Overall, the methodological implications of the first stage of our study suggest that the involvement of human experts is still essential when using RAG to facilitate the selection of relevant text. Integrating humans at the stage of selection also aligns well with the first phase of thematic analysis (Braun and Clarke, 2006), as it offers a chance for researchers to familiarize themselves with the data through a more focused lens. Unlike sorting through a random sample to find relevant responses, reviewing RAG-selected responses provides greater exposure to relevant content, potentially improving both efficiency and depth at the phase of data familiarization. For researchers with limited coding backgrounds, the built-in RAG capabilities of many AI programs makes it feasible to implement this approach. Researchers can explore their dataset of student writing using AI to identify relevant passages for inquiries on specific topics and select text for further examination and coding.

In the second stage of this study, we integrated ChatGPT-4 into the five phases of thematic analysis for data engagement, coding, and theme development. At each phase and in different ways, the model augmented and extended the insights we were able to develop from the data. While there were similarities and differences in the codes and themes developed by the authors and by ChatGPT, our machine-in-the-loop approach allowed us to carefully select

from the results it provided and to integrate AI-generated insights. In Hitch's study (2024), she observed that "ChatGPT supported iterative analysis by encouraging a deliberative approach and the development of shared meaning." However, we found that the process differed significantly between human-human and human-AI collaboration. Among human collaborators, disagreements led to productive discussions and revisions until consensus was reached. In contrast, our interaction with ChatGPT was necessarily unilateral: we selected valuable insights without engaging in back-and-forth dialogue about its "reasoning" or our integration choices. ChatGPT's lack of human characteristics—e.g., opinions, feelings, the need for consensus—streamlined the integration of its contributions to our analysis.

Ultimately we were able to learn about why students assess their writing positively, and we developed a thematic framework for identifying underlying beliefs about learning that often tacitly inform students' evaluations of themselves as writers. However, these findings are peripheral to our primary methodological goal of exploring how Writing Studies researchers might incorporate generative AI to expedite and even improve qualitative analysis of large corpora. The two-stage approach we developed—combining RAG-facilitated selection and generative AI-assisted thematic analysis—offers a practical framework for analyzing large collections of open-ended responses while maintaining the crucial elements of human interpretation and expertise. This methodology preserves the nuanced human understanding central to qualitative analysis while employing AI to expand the scope and depth of analysis possible with large datasets.

This study has several limitations that suggest directions for future research. First, while we drew from a substantial initial corpus of over 13,000 student responses to the CWP prompts, our analysis ultimately focused on a relatively small subset: 310 responses for expert rating and 100 responses for thematic analysis. Though these responses were strategically selected using both AI and human evaluation to ensure high relevance to our research question, a larger sample might have revealed additional patterns, codes, and themes to analyze. Second, our decision to limit RAG-generated results to the top 50 responses for each prompt/LLM combination may have artificially constrained our findings. A larger initial selection could have yielded different patterns in the relevancy ratings and potentially captured more nuanced expressions of positive self-assessment. Third, while our two-stage methodology brought promising results, we did not systematically test different prompting strategies that might have improved RAG's performance or ChatGPT's contribution to the thematic analysis. Other researchers (e.g., Meng et al., 2024; Turobov et al., 2024; Zhang et al., 2024) have emphasized the importance of crafting and fine-tuning prompts to achieve better results in qualitative analysis.

Our focus on students' positive self-assessments and our use of thematic analysis means that our findings about the effectiveness of our AI-assisted methodology might not generalize to other topics or types of qualitative analysis in Writing Studies research. In addition, while we identified meaningful themes in students' self-assessments, we did not explore the pedagogical implications of these findings. Future research could build on this foundation in several valuable directions. Incorporating quantitative performance measures such as grades,

direct evaluation of student writing samples, or in-depth qualitative interviews would provide an understanding of how different types of positive self-assessment impact educational outcomes. Including student demographic data would also help identify contextual factors that influence self-assessment patterns. Importantly, such follow-up studies need not rely on AI tools but could instead leverage the thematic insights we developed here, demonstrating how AI can serve as an effective analytical tool within a machine-in-the-loop framework for generating research directions that extend beyond computational methods.

7. Conclusion

This study demonstrates the potential of integrating AI into qualitative analysis through two distinct approaches: using RAG to identify relevant responses from large, varied corpora, and using ChatGPT to facilitate thematic analysis. Our findings support a machine-in-the-loop framework where subject-matter experts maintain control of the research process while integrating AI-generated insights at each phase. This approach creates what Perkins and Roe (2024) describe as a workflow that "can expedite the analytical process without diminishing the essential role of the researcher's expertise and critical engagement."

Beyond improving efficiency, our results suggest that AI can meaningfully enrich qualitative analysis. In our thematic analysis of student writing self-assessments, ChatGPT identified patterns and perspectives that complemented human observations, leading to richer interpretations through the collaborative coding process. This synergy between human expertise and AI capabilities offers promising directions for scaling up qualitative research while maintaining analytical depth.

While some researchers express concerns that AI-assisted analysis might "impede or undercut the human essence of qualitative research" (Wachinger et al., 2024), the integration of AI tools into qualitative methods appears inevitable. Rather than resisting this change, the academic community should focus on developing and testing frameworks for responsible AI integration. As Perkins and Roe (2024) argue, "By understanding how the benefits of these tools can enhance the research process, the academic community can harness their strengths more effectively while simultaneously limiting the potential negative impacts of their challenges and weaknesses." Our study provides one such framework, specifically tailored for Writing Studies research, that demonstrates how AI can augment and improve rather than replace human qualitative analysis.

References

Au, W. (2022). Unequal by design: High-stakes testing and the standardization of inequality. Routledge. https://doi.org/10.4324/9781003005179

Baillargeon, K. (2025). Reflective orientations and genre navigation in dissertation writing: Insights from seven years of doctoral writing retreats. *Journal of Writing Research*. https://www.jowr.org/jowr/article/view/1670

Balaguer, A., Benara, V., de Freitas Cunha, R. L., Estevão Filho, R. D. M., Hendry, T., Holstein, D., ... & Chandra, R. (2024). RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. arXiv e-prints, arXiv-2401. https://doi.org/10.48550/arXiv.2401.08406

Barany, A., Nasiar, N., Porter, C., Zambrano, A. F., Andres, A. L., Bright, D., ... & 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* (pp. 134-149). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64299-9_10

Bartholomae, D. (2005). Inventing the university. *Writing on the margins: Essays on composition and teaching*, 60-85. https://doi.org/10.37514/JBW-J.1986.5.1.02

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

Britton, J. (1975). The Development of Writing Abilities (11-18).

Connors, R. & Lunsford, A.A. (1988). Frequency of Formal Error in Current College Writing, or Ma and Pa Kettle Do Research." College Composition and Communication, 39.4: 395–409. https://doi.org/10.2307/357695

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*. https://doi.org/10.18653/v1/2023.findings-emnlp.669

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

Feuston, J. L., & Brubaker, J. R. (2021). Putting tools in their place: The role of time and perspective in human-AI collaboration for qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1-25. https://doi.org/10.1145/3479856

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*. https://doi.org/10.48550/arXiv.2312.10997

Gebreegziabher, S. A., Zhang, Z., Tang, X., Meng, Y., Glassman, E. L., & Li, T. J. J. (2023, April). Patat: Human-ai collaborative qualitative coding with explainable interactive rule synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-19). https://doi.org/10.1145/3544548.3581352

Gere, A. (2019). *Developing writers in higher education: A longitudinal study* (p. 385). University of Michigan Press. https://doi.org/10.3998/mpub.10079994

Gere, A. R., Aull, L., Escudero, M. D. P., Lancaster, Z., & Vander Lei, E. (2013). Local assessment: Using genre analysis to validate directed self-placement. *College Composition & Communication*, *64*(4), 605-633. https://doi.org/10.58680/ccc201323661

Hamilton, L., Elliott, D., Quick, A., Smith, S., & Choplin, V. (2023). Exploring the use of AI in qualitative analysis: A comparative study of guaranteed income data. *International Journal of Qualitative Methods, 22*, https://doi.org/10.1177/16094069231201504

Haswell, R. H. (2005). NCTE/CCCC's recent war on scholarship. *Written Communication*, *22*(2), 198-223. https://doi.org/10.1177/0741088305275367.

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

Hong, M. H., Marsh, L. A., Feuston, J. L., Ruppert, J., Brubaker, J. R., & Szafir, D. A. (2022, October). Scholastic: Graphical human-AI collaboration for inductive and interpretive text analysis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (pp. 1-12). https://doi.org/10.1145/3526113.3545681

Ibrahim, E. I., & Voyer, A. (2024). The Augmented Qualitative Researcher: Using Generative AI in Qualitative Text Analysis. *SocArXiv Preprints*. https://doi.org/10.31235/osf.io/gkc8w.

Jalali, M. S., & Akhavan, A. (2024). Integrating AI Language Models in Qualitative Research: Replicating Interview Data Analysis with ChatGPT. *System Dynamics Review*, 40(3), e1772. https://doi.org/10.1002/sdr.1772

Jiang, J. A., Wade, K., Fiesler, C., & Brubaker, J. R. (2021). Supporting serendipity: Opportunities and challenges for Human-AI Collaboration in qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1-23. https://doi.org/10.1145/3449168

Kantor, J. (2024). Best practices for implementing ChatGPT, large language models, and artificial intelligence in qualitative and survey-based research. *JAAD international*, *14*, 22-23. https://doi.org/10.1016/j.jdin.2023.10.001.

Lee, S., Shakir, A., Koenig, D., & Lipp, J. (2024). Open Source Strikes Bread—New Fluffy Embeddings Model. https://www.mixedbread.ai/blog/mxbai-embed-large-v1

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474. https://doi.org/10.5555/3495724.3496517.

Licastro, A., & Miller, B. M. (Eds.). (2021). *Composition and big data*. University of Pittsburgh Press. https://doi.org/10.2307/j.ctv22tnmg2

Lubars, B., & Tan, C. (2019). Ask not what AI can do, but what AI should do: Towards a framework of task delegability. *Advances in neural information processing systems*, *32*. https://doi.org/10.5555/3454287.3454293

Lunsford, A. A., Fishman, J., & Liew, W. M. (2013). College writing, identification, and the production of intellectual property: Voices from the Stanford study of writing. *College English*, *75*(5), 470-492.

Lunsford, A. A., & Lunsford, K. J. (2008). "Mistakes are a fact of life": A national comparative study. *College Composition & Communication*, *59*(4), 781-806. https://doi.org/10.58680/ccc20086677.

Melzer, D. (2009). Writing assignments across the curriculum: A national study of college writing. *College Composition & Communication*, *61*(2), W240-W261. https://doi.org/10.58680/ccc20099487.

Meng, H., Yang, Y., Li, Y., Lee, J., & Lee, Y. C. (2024). Exploring the Potential of Human-LLM Synergy in Advancing Qualitative Analysis: A Case Study on Mental-Illness Stigma. *arXiv preprint arXiv:2405.05758*. https://doi.org/10.48550/arXiv.2405.05758

Mesec, B. (2023). The language model of artificial inteligence chatGPT-a tool of qualitative analysis of texts. Authorea Preprints. https://doi.org/10.22541/au.168182047.70243364/v1.

Moos, A., & Van Zanen, K. (2019). Directed self-placement as a tool to foreground student agency. *Assessing Writing*, *41*, 68-71. https://doi.org/10.1016/j.asw.2019.06.001.

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.

Nguyen-Trung, K. (2024). ChatGPT in Thematic Analysis: Can AI become a research assistant in qualitative research?. *OSF Preprint*.

OpenAI. (2023). *text-embedding-3-small*. Retrieved from https://platform.openai.com/docs/guides/embeddings

Ovadia, O., Brief, M., Mishaeli, M., & Elisha, O. (2024). Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. *arXiv preprint arXiv:2312.05934*.

Paulus, T. M., & Marone, V. (2024). In Minutes Instead of Weeks: Discursive Constructions of Generative AI and Qualitative Data Analysis. *Qualitative Inquiry*, 30(1), https://doi.org/10.1177/10778004241250065

Perkins, M., & Roe, J. (2024). The use of Generative AI in qualitative analysis: Inductive thematic analysis with ChatGPT. *Journal of Applied Learning and Teaching*, *7*(1), Article 1. https://doi.org/10.37074/jalt.2024.7.1.22

Rietz, T., & Maedche, A. (2021, May). Cody: An AI-based system to semi-automate coding for qualitative research. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-14). https://doi.org/10.1145/3411764.3445591

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. Healthcare, 11(6), Article 887. https://doi.org/10.3390/healthcare11060887

Sinha, R., Solola, I., Nguyen, H., Swanson, H., & 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 Symposium on Learning, Design and Technology* (pp. 17–25). https://doi.org/10.1145/3663433.3663456

Sommers, N. (2006). Across the drafts. *College Composition & Communication*, *58*(2), 246-266. https://doi.org/10.58680/ccc20065899

Sommers, N. (2008). The call of research: A longitudinal view of writing development. *College Composition and Communication*, *60*(1), 152-164.

Tinkle, T., Godfrey, J., Hammond, J. W., & Moos, A. (2024). Self-Characterization in the Self-Placement Assessment Ecology: Complicating the Stories We Tell about DSP's Effects and Effectiveness. *Journal of Writing Assessment*, *17*(1). https://doi.org/10.5070/W4jwa.1625.

Turobov, A., Coyle, D., & Harding, V. (2024). Using ChatGPT for thematic analysis. *arXiv preprint arXiv:2405.08828*.

Wachinger, J., Bärnighausen, K., Schäfer, L. N., Scott, K., & McMahon, S. A. (2024). Prompts, Pearls, Imperfections: Comparing ChatGPT and a Human Researcher in Qualitative Data Analysis. *Qualitative Health Research*, 34(7), 10497323241244669. https://doi.org/10.1177/10497323241244669

Yan, L., Echeverria, V., Fernandez-Nieto, G. M., Jin, Y., Swiecki, Z., Zhao, L., Gašević, D., & Martinez-Maldonado, R. (2024). Human-AI Collaboration in Thematic Analysis using ChatGPT: A User Study and Design Recommendations. *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3613905.3650732

Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., & Carroll, J. M. (2023). Redefining Qualitative Analysis in the AI Era: Utilizing ChatGPT for Efficient Thematic Analysis. *arXiv preprint arXiv:2309.10771*. https://doi.org/10.48550/arXiv.2309.10771

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., ... & Cui, B. (2024). Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.