Using Large Language Models for Qualitative Coding of Structured Data on Post COVID Research Based Instructional Strategies

Jacob Lineberry
Virginia Tech

Estrella Johnson
Virginia Tech

Naneh Apkarian
San Diego State University

In the spring of 2025, introductory Physics, Calculus, and Chemistry instructors were surveyed about their usage of active learning techniques and use of Research Based Instructional Strategies. Using open-source large language models running on consumer hardware, responses to open ended survey items data were analyzed to better understand the usage of specific strategies during and after the Pandemic Response Teaching period resulting from the COVID-19 pandemic, as well as for reasons how and why instructors adopted, maintained, or stopped RBIS. While the large language models were able to apply qualitative codes to the data, their effectiveness was significantly varied based on the dataset, initial prompt, large language model used, code specificity, and code prevalence.

Keywords: Large language models, Qualitative Analysis, Research Based Instructional Strategies, Pandemic response teaching

In 2025, we conducted a national survey of instructors of undergraduate chemistry, mathematics, and physics courses. The timing and selection of instructors were designed to understand if there were any lasting changes to how these courses are taught following the COVID-19 pandemic. In part, this survey asked instructors of first year Calculus, Chemistry, and Physics about their awareness and use of research based instructional strategies (RBIS), while providing an opportunity for respondents to explain how and why they use (or do not use) RBIS. The nature of this data offered us information into instructional trends and provided an opportunity to experiment with large language models in qualitative analysis.

AI, particularly in the form of large language models (LLMs), has become more common and powerful. Within education research, these tools have been applied to various forms of qualitative analysis, with varied success. However, their use has raised important questions about the nature of qualitative research, privacy, bias, and transparency. We attempted to address several of these questions by using an open source, locally run LLM to complete the first two steps of Braun and Clarke's (2012) thematic analysis with a graduate student researcher.

Here, we address two research questions. First, to what extent, and why, are instructors in first year STEM courses using, or shifting away from RBIS following the Pandemic Response Teaching (PRT) period? Second, to what extent can an open-source, locally hosted LLM effectively code a highly structured dataset using Braun and Clarke's thematic analysis?

Background Literature and Perspective

By comparing survey responses from a 2019 survey and this one from 2025, Apkarian & Johnson (2026) investigated the extent to which use of class-time and instructional practices changed following the COVID-19 pandemic. Analysis of that quantitative data revealed, overall, no significant changes in reports of how class time is spent with regards to listening to the instructor lecture or solve problems, working individually, working in small groups, and engaging in whole class discussion. Additionally, while there was increased usage of online office hours, supplemental videos, and recorded lectures, starting during PRT and continuing afterward, there were no major shifts to in-class instructional practices (e.g., think pair share,

clickers). Our analysis here of the open-ended survey response data seeks to provide additional nuance to the quantitative analysis completed by Apkarian & Johnson (2026).

Traditionally, analyzing open ended survey responses is carried out with any variety of qualitative analysis methods by an individual or a team. For example, thematic analysis, as developed by Braun and Clarke (2012), prioritizes six key steps, with each working towards identifying overarching themes across a dataset. By carrying out these steps, researchers can identify broader patterns key to understanding how, when, and why instructors prioritize and implement RBIS. That thematic analysis is a well-defined, commonly used approach to analyzing a well-structured dataset makes it both a practical choice for researchers and potentially a good fit for data analysis carried out by LLMs.

The robust and diverse capabilities of LLMs makes them increasingly more prevalent, including in qualitative research. Researchers have long used stochastic models for natural language processing (Chen at al., 2018; Gamieldien et al., 2023; Li et al., 2021; Rietz et al., 2021). While use of LLMs in qualitative research is still new, work on applying them within new methodologies and frameworks is growing rapidly (Naeem et al., 2018; Naeem et al., 2023; Xiao et al., 2023; Zhang et al., 2025). However, qualitative researchers have shown reasonable skepticism about their use. Some studies indicate researchers are generally concerned about hallucinations (when a model invents data and "facts"), data privacy, transparency, and explainability (Schroeder et al., 2025; Feuston & Brubaker, 2021; Springer & Whittaker, 2020; Zhang et al., 2025). Furthermore, results are generally mixed in using large language models for qualitative research. Many studies find significant variation in performance based on prompts used (Wang et al., 2024; Zamfirescu-Pereira et al. 2023), inductive or deductive methods (Bijker et al., 2024; Siiman et al., 2023), and researcher knowledge of LLMs (Chen et al., 2018; Zamfirescu-Pereira et al, 2023; Zhang et al., 2025).

Even with these challenges, LLMs can be a powerful resource. Many highlight the potential for LLMs to serve as an additional coder, where a human researcher and LLM each code the data and compare (Bijker et al., 2024). Other uses include academic writing (Tu et al., 2024), developing a codebook (Bijker et al., 2024), coding data (Marathe & Toyama, 2018; Tai et al., 2024), identifying themes (Castellanos et al., 2025), and supporting new researchers in developing qualitative analysis skills (Gao et al., 2024). In addition, several processes have been developed for using LLMs, usually ChatGPT, for qualitative analysis, like thematic analysis (Dai et al., 2023; De Paoli, 2024; Naeem et al., 2018; Naeem et al., 2023; Zhang et al., 2025).

Methods

The 2025 survey was sent to roughly 17,000 names, the same roster that was used for the 2019 survey. The goal was to capture those who taught before and after the pandemic-induced disruption from 2020-2023. To be eligible, instructors must have taught one of the target introductory courses in the 2023-24 and/or 2024-25 academic year. In total, 933 participants provided usable responses to our survey: 339 chemistry (99 new; 240 repeaters), 279 mathematics (112 new; 167 repeaters), and 315 physics (86 new; 229 repeaters) instructors. Included in the 2025 survey were the following free response questions:

1. What, if any, instructional practices did you start using during the pandemic response period (roughly March 2020-June 2022), that you had not used previously? Include anything that you tried, even if you did not maintain that practice throughout or afterward.
2. Of the instructional practices you mentioned in the previous question, which (if any) do you continue to use?

3. You may use the space below to add any additional comments about your experiences during the pandemic response period, if you wish.
4. You may use the space below for any additional comments about the context and climate of your department, if you wish.

Anonymized data was stored in a Google Sheet. Of the 933 participants, a random sample of 450 were selected to have their responses coded manually by a human coder (the first author). Instructor responses were considered holistically, even if answers to any of the four questions above were missing. Then, the data was coded using three different large language models: Gemma3:12b, Qwen3:14b, and Deepseek-R1:14b. The large language models were run locally using the Python programming language and the Ollama software package. All models were run on the researcher's consumer desktop with an AMD RX 6900 XT GPU. Each model was run using default Ollama settings, and data was input as 30 sample chunks to ensure the context window of each model was not exceeded. The prompt used was carefully designed using common prompt engineering techniques outlined in the literature and was revised 3 times before the final implementation. The prompt carefully defines the role of the LLM, explains the task in multiple ways, and provides an output template. However, the authors specifically did not extensively tune the initial prompt or implement specific techniques to improve performance.

Each run of the models output a text file consisting of codes for each row of data. This text file was then converted to a CSV using Python. This conversion expected consistent output from the LLMs and is likely undercounting codes from the models. If the model provided a synonymous code, instead of a code from the codebook, we would not count that towards our results. Afterwards, Cohen's Kappa and Krippendorff's Alpha were calculated using Python for each run, and an average for each model was calculated for each code. After this programming pipeline was created, we revised our initial prompt to better understand why instructors implemented, maintained, or moved away from RBIS in their instruction. This prompt used a "oneshot" technique, providing examples and non-examples for how the data should be coded. Only the LLMs coded the dataset for this research question and analysis is still ongoing.

Results and Discussion

Each LLM coded the entire dataset, while the first author only coded a random sample of responses from 450 instructors. We present the percentage of rows coded with each code in table 1. Each LLM coded the data 3 times, and an average for each code is reported. Since each row of data could have more than one code, the total percent frequency adds to more than 100%.

Table 1. Percent frequency of each code by "coder"

Code	Deepseek-R1	Qwen3	Gemma3	Human
Adapted existing strategies to Pandemic	14.29%	21.33%	40.09%	1.33%
Alternative Exams	1.61%	3.00%	3.47%	2.67%
Did not use RBIS	0.75%	1.46%	0.43%	0.22%
Experimented	1.57%	2.57%	15.18%	0.22%
Flipped	11.08%	10.86%	9.54%	8.22%
Group Work	17.97%	20.40%	19.08%	9.11%
Increase in tech use	14.22%	20.04%	9.50%	5.11%
Inquiry/Modeling	6.72%	4.86%	3.14%	1.11%
Lab Kits	1.04%	0.54%	1.32%	0.89%

Learning Assistants/Peer to Peer Teaching	0.82%	2.57%	1.68%	1.33%
None	15.93%	31.80%	30.01%	17.11%
Online Assignments	5.72%	8.54%	18.65%	7.56%
Online discussion tools	1.04%	2.64%	7.18%	1.78%
Online Exams/Proctoring	3.39%	6.47%	8.40%	7.78%
Online Lecture	8.82%	9.97%	24.01%	12.22%
Online Office Hours	4.22%	8.29%	9.72%	8.89%
Online simulations	2.00%	6.15%	6.36%	7.56%
Polling/Clicker	6.29%	5.93%	7.43%	4.44%
Posting resources online	1.04%	6.04%	10.97%	5.78%
Recording/Posting Videos	23.94%	25.76%	27.05%	26.22%
Reflection/Exam Wrappers	1.04%	0.39%	2.29%	0.44%
Used new RBIS	5.79%	1.54%	0.32%	0.00%

We calculated Cohen's Kappa and Krippendorff's Alpha for each of the codes, comparing the interrater reliability between each LLM and the human coder (Table 2). While the averages for Kappa and Alpha were between 0.27 and 0.47, we see significant differences in interrater reliability between various codes and across models. Qwen3 had the highest interrater reliability, while Deepseek-R1 had the lowest. Furthermore, we see a significant range in potential Kappa and Alpha values. For example, when using Qwen3, the codes Online Office Hours and Recording/Posting Videos had a Kappa and Alpha value above 0.815, while the codes Experimented and Did Not Use RBIS had a Kappa and Alpha value of approximately 0.

Table 2. Average Cohen's Kappa and Krippendorff's Alpha for each LLM across three runs

Model	Kappa	Alpha
Deepseek-R1	0.285	0.278
Qwen3	0.463	0.454
Gemma3	0.363	0.344

The prevalence of "online" codes aligns with instructors of these introductory courses transitioning from in-person lecture to Zoom. Additionally, we see fewer codes related to the instructional features of the class, such as "group work," "flipped," "polling/clicker," and "learning assistants/peer to peer teaching" than for "recording/posting videos." These relative prevalence indicators broadly match the quantitative analysis from Apkarian & Johnson (2026).

The AI framework developed was meant to be a proof of concept, involving as little programming or AI specific knowledge as possible. This simplicity, coupled with less than perfect human coding (i.e., data was first coded by a graduate student new to qualitative analysis with no second coder, the initial codebook was quite large) could have resulted in specific struggles for the LLM to code the data. However, overall, the models did relatively well with those codes used frequently by the human coder. When removing codes that were used fewer than 10 times by the first author, we find Kappa increased to 0.362, 0.576, and 0.508 for Deepseek, Qwen, and Gemma respectively. In addition to challenges around infrequently used codes, the LLMs struggled to apply the code "None." The "None" code was used relatively frequently by the human coder, 77 times, but was consistently the worst for Kappa and Alpha.

While the interrater agreement between a human coder and the LLMs used varied greatly in the first two phases of Braun and Clarke's Thematic Analysis, our data and learnings suggest, with software and hardware adjustments and technological improvements, open source, locally hosted LLMs can provide valuable support to a qualitative researcher and should be further explored.

Conclusions

We are not arguing for a total reliance on AI or LLMs for qualitative research (Christou, 2023), especially given our interrater reliability results. However, to completely avoid LLMs is likely misguided. Qualitative research has used stochastic tools (Tausczik & Pennebaker, 2010), natural language processing (Bryda & Sadowski, 2024; Gamieldien et al., 2023), and digital tools (MaxQDA, nVivo, etc.) for a long time now. However, LLMs raise questions about bias, accuracy, transparency, and privacy to address within the context of qualitative research.

With regards to questions about bias (Barany et al., 2024; Chen et al., 2018; Pangakis et al., 2023; Schroeder et al., 2025), by relying on open-source models that reveal their weights, training data, and more, we can better understand what potential biases may exist within the model and its training data. With regards to accuracy concerns, many studies have found a benefit and necessity for human checks and interaction with any AI tool in qualitative analysis (Feuston & Brubaker, 2021; Pangakis et al., 2023). For our work, instances of inaccuracy were mostly the model incorrectly assuming the data contained more information than was actually present and thus applying a code where one did not belong. We believe we can minimize this by improving prompts with one-shot techniques, averaging several runs, and other formatting techniques. With regards to concerns about data privacy (Byun, 2023; Schroeder et al., 2025), the use of local models avoids sharing IRB protected data with private companies who are explicit about their use and storage of user chats (e.g., chatGPT). Finally, with regards to transparency concerns, using an open source model allows researchers to both understand and choose the model best suited to their work, as well as better communicate to others what is happening "behind the scenes" with the models used.

Preliminary Nature of this Report

This work is still preliminary, as we believe we can further tune our use of LLMs in thematic analysis to better understand our dataset. While the models struggled with certain types of codes, our analysis suggested a second phase of coding focused on why instructors implemented, maintained, or moved away from RBIS in their instruction. We have completed an initial round of coding using the LLM pipeline, but need to compare and validate with a human coder to determine how well LLM's handled a narrower codebook consisting of more nuanced codes.

Acknowledgments

This research was supported by National Science Foundation award numbers #2416741, #2416742. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Apkarian, N., Henderson, C., Stains, M., Raker, J., Johnson, E., & Dancy, M. (2021). What really impacts the use of active learning in undergraduate STEM education? Results from a national survey of chemistry, mathematics, and physics instructors. PLOS ONE, 16, 2. https://doi.org/10.1371/journal.pone.0247544

Apkarian, N. & Johnson, E. (accepted). STEM teaching practices: Before, during, and after Pandemic Response Teaching. Proceedings of the 28th Annual Conference on Research in Undergraduate Mathematics Education. Alexandria, VA.

Barany, A., Nasiar, N., Porter, C., Zambrano, A., Andres, A., Bright, D., Shah, M., Liu, X., Gao, S., Zhang, J., Mehta, S., Choi, J., Giordano, C., & Baker, R. (2024, July). ChatGPT for education research: Exploring the potential of large language models for qualitative codebook development. 25th International Conference on Artificial Intelligence in Education.

Bijker, R., Merkouris, S. S., Dowling, N. A., & Rodda, S. N. (2024). ChatGPT for automated qualitative research: Content analysis. J Med Internet Res, 26, e59050. https://doi.org/10.2196/59050

Braun, V., & Clarke, V. (2012). Thematic Analysis. APA Handbook of Research Methods in Psychology, Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological., 2(2), 57–71. APA PsycNet. https://doi.org/10.1037/13620-004

Bryda, G., & Sadowski, D. (2024). From words to themes: AI-Powered qualitative data coding and analysis. In A. Pedro Costa, L. Paulo Reis, F. Neri de Sousa, A. Moreira, & D. Lamas (Eds.), Computer Supported Qualitative Research (pp. 309–345). https://doi.org/10.1007/978-3-031-65735-1_19

Byun, C. L. (2023). Dispensing With Humans in Human-Computer Interaction Research. https://scholarsarchive.byu.edu/etd/10158/

Castellanos, A., Jiang, H., Gomes, P., Meer, V., & Castillo, A. (2025). Large language models for thematic summarization in qualitative health care research: Comparative analysis of model and human performance. JMIR AI, 4, e64447. https://doi.org/10.2196/64447

Chen, N.-C., Drouhard, M., Kocielnik, R., Suh, J., & Aragon, C. (2018). Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. ACM Transactions on Interactive Intelligent Systems, 8, 1–20. https://doi.org/10.1145/3185515

Christou, P. (2023). How to Use Artificial Intelligence (AI) as a Resource, Methodological and Analysis Tool in Qualitative Research? The Qualitative Report, 28(7), 1968–1980. https://doi.org/10.46743/2160-3715/2023.6406

Dai, S.-C., Xiong, A., & Ku, L.-W. (2023). LLM-in-the-loop: Leveraging large language model for thematic analysis. In H. Bouamor, J. Pino, & K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 9993–10001). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.669

Feuston, J., & Brubaker, J. (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, 1–25. https://doi.org/10.1145/3479856

Gamieldien, Y. G., Case, J. M., & Katz, A. (2023). Advancing Qualitative Analysis: An Exploration of the Potential of Generative AI and NLP in Thematic Coding. Social Science Research Network. https://doi.org/10.2139/ssrn.4487768

Gao, J., Guo, Y., Lim, G., Zhang, T., Zhang, Z., Li, T. J.-J., & Perrault, S. T. (2024, May 11). CollabCoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative

analysis with large language models. CHI ’24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3613904.3642002
Li, Z., Dohan, D., & Abramson, C. M. (2021). Qualitative Coding in the Computational Era: A Hybrid Approach to Improve Reliability and Reduce Effort for Coding Ethnographic Interviews. Socius: Sociological Research for a Dynamic World, 7. https://doi.org/10.1177/23780231211062345
Marathe, M., & Toyama, K. (2018). Semi-automated coding for qualitative research: A user-centered inquiry and initial prototypes. CHI ’18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–12. https://doi.org/10.1145/3173574.3173922
Naeem, M., Ozuem, W., Howell, K., & Ranfagni, S. (2023). A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. International Journal of Qualitative Methods, 22. https://doi.org/10.1177/16094069231205789
Naeem, M., Smith, T., & Thomas, L. (2025). Thematic Analysis and Artificial Intelligence: A Step-by-Step Process for Using ChatGPT in Thematic Analysis. International Journal of Qualitative Methods, 24. https://doi.org/10.1177/16094069251333886
Pangakis, N., Wolken, S., & Fasching, N. (2023). Automated Annotation with Generative AI Requires Validation. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2306.00176
Paoli, S. D. (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. https://doi.org/10.1177/08944393231220483
Rietz, T., & Maedche, A. (2021, May). Cody: An AI-Based System to semi-automate Coding for Qualitative Research. CHI ’21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3411764.3445591
Schroeder, H., Quéré, L., Randazzo, C., Mimno, D., & Schoenebeck, S. (2025). Large language models in qualitative research: Uses, tensions, and intentions. https://doi.org/10.1145/3706598.3713120
Siiman, L., Rannastu, M., Pöysä-Tarhonen, J., Häkkinen, P., & Pedaste, M. (2023). Opportunities and challenges for AI-Assisted qualitative data analysis: An example from collaborative problem-solving discourse data. In Innovative Technologies and Learning (pp. 87–96). https://doi.org/10.1007/978-3-031-40113-8_9
Springer, A., & Whittaker, S. (2020). Progressive disclosure: When, why, and how do users want algorithmic transparency information? ACM Trans. Interact. Intell. Syst., 10, 4. https://doi.org/10.1145/3374218
Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. International Journal of Qualitative Methods, 23, 16094069241231168. https://doi.org/10.1177/16094069241231168
Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, 29, 1. https://doi.org/10.1177/0261927X09351676
Tu, J., Hadan, H., Wang, D. M., Sgandurra, S. A., Reza Hadi Mogavi, & Nacke, L. E. (2024). Augmenting the author: Exploring the potential of AI collaboration in academic writing. https://arxiv.org/abs/2404.16071
Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., Li, Q., & Li, J. (2024). Prompt engineering in consistency and reliability with the evidencebased guideline for LLMs. Npj Digital Medicine, 7(1), 41. https://doi.org/10.1038/s41746024010294

Xiao, Z., Yuan, X., Vera, L. Q., Abdelghani, R., & Oudeyer, P.-Y. (2023). Supporting Qualitative Analysis with Large Language models: Combining Codebook with GPT-3 for Deductive Coding. 75–78. https://doi.org/10.1145/3581754.3584136

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. https://doi.org/10.1145/3544548.3581388

Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., & Carroll, J. M. (2025). Harnessing the power of AI in qualitative research: Exploring, using and redesigning ChatGPT. Computers in Human Behavior: Artificial Humans, 4, 100144. https://doi.org/10.1016/j.chbah.2025.100144