4-2025

Streamlining Field Note Analysis: Leveraging GPT for Further Insights

Saríah López-Fierro
Utah State University, sariah.lopezfierro@usu.edu

Umar Shehzad
Utah State University, umar.shehzad@usu.edu

Alireza (Sina) Zandi
Utah State University, alireza.zandi@usu.edu

Jody Clarke-Midura
Utah State University, jody.clarke@usu.edu

Mimi Recker
Utah State University, mimi.recker@usu.edu

Streamlining Field Note Analysis:
Leveraging GPT for Further Insights

Saríah López-Fierro
sariah.lopezfierro@usu.edu

Umar Shehzad
umar.shehzad@usu.edu

Alireza (Sina) Zandi
alireza.zandi@usu.edu

Jody E. Clarke-Midura
jody.clarke@usu.edu

Mimi M. Recker
mimi.recker@usu.edu

Utah State University
Department of Instructional Technology & Learning Sciences
2830 Old Main Hill
Logan, Utah 84322-2830 U.S.A.
+1 435-797-2694

Paper presented at the 2025 Annual Meeting of the American Education Research Conference
Denver, CO USA

Abstract

As an integral part of qualitative research inquiry, field notes provide important data from researchers embedded in research sites. However, field notes can vary significantly, influenced by the researchers' immersion in the field, prior knowledge, beliefs, interests, and perspectives. As consequence, their interpretation presents significant challenges. This study offers a preliminary investigation into the potential of using large language models to assist researchers with the analysis and interpretation of field notes data. Our methodology consisted of two phases. First, a researcher deductively coded field notes of six classroom implementations of a novel elementary-level mathematics curriculum. In the second phase, we prompted ChatGPT-4 to code the same field notes, using the codebook, definitions, examples, and deductive coding approach employed by the researcher. We also prompted Chatgpt to provide justifications of its coding decisions We then, calculated agreements and disagreements between ChatGPT and the researcher, organized the data in a contingency table, computed Cohen's Kappa, structured the data into a confusion matrix; and using the researcher's coding as the "gold standard", we calculated performance measures, specifically: Accuracy, Precision, Recall, and F1 Score. Our findings revealed that while the researcher and ChatGPT appeared to generally agree on the frequency in applying the different codes, overall agreement, as measured by Cohen's Kappa was low. In contrast, using measures from information science at the code level revealed more nuanced results. Moreover, coupled with ChatGPT justifications of coding decisions, these findings provided insights than can help support the iterative improvement of codebooks.

Introduction

As a part of qualitative research inquiry, field notes are written records that document observations (Montgomery & Bailey, 2007) and perceptions (Papen, 2019) of researchers embedded in their research site. Since the early 1900s, they have played an essential role in qualitative research (Johnson et al., 2024; Phillippi & Lauderdale, 2018). In these notes, researchers record the verbal and nonverbal actions of participants, the context in which these actions occur, as well as their personal reflections, emotions, and insights (Maharaj, 2016). These field notes can enhance qualitative data analysis by supporting triangulation, aid in assessing the transparency of findings, guide future data collection, and provide crucial support for informing ongoing research (Phillippi & Lauderdale, 2018).

However, field notes can vary significantly in their content, type, length, and style (Montgomery & Bailey, 2007). These variations are influenced not only by the extent of researchers' immersion in the field but also by their individual prior knowledge, beliefs, interests, and perspectives (Irwin et al., 2013; Tjora, 2006). In consequence, researchers' positionality affects **how** they write their notes (Papen, 2019) and **what** they subjectively decide to share for further qualitative analysis (Copland, 2018). Moreover, field notes are time consuming to collect and subsequently analyze. As such, generating effective field notes represents a challenging task. This is especially true for novice researchers (Maharaj, 2016).

Following the growing interest in the use of Large Language Models (LLM) to assist with qualitative analysis (e.g., Beltran et al., 2024; Combrinck, 2024; López-Fierro & Nguyen, 2024; Zambrano et al., 2023), this study aims to leverage the computational power and pattern recognition capabilities of large language models (Perkins & Roe, 2024) to help explore and evaluate the effectiveness of applying them (specifically the ChatGPT-4o model) to analyze field notes.

The context for our study consists of field notes taken by researchers as two elementary level teachers implemented a novel mathematics curriculum in their classrooms. In doing so, we address the following research question:

- To what extent can large language models (LLMs) engage in qualitatively coding researchers' field notes? How does LLMs' qualitative coding compare to human researchers?

Background

Qualitative Analysis of Field Notes

Field notes are a type of qualitative data source consisting of notes taken by researchers during fieldwork. These notes can contain observations, initial interpretations, and insights describing what researchers see and hear. In this way, field notes provide additional context and help bridge the gap between what researchers directly observe and how they interpret what they have observed (Johnson et al., 2024).

Field notes can be further analyzed for deeper understanding of the phenomena under study. This analysis can be performed inductively through open coding, where researchers review field notes line-by-line to identify emerging themes and patterns (Bussell, 2020; Chan et al., 2021). After open coding, researchers can then identify common categories, inductively associate themes, and cluster codes according to their similarity (Irwin et al., 2013). This method allows themes to emerge organically from the data itself, enabling researchers to develop a coding structure based on the content of the field notes (Chan et al., 2021).

Field notes can also be analyzed via deductive analysis, which involves applying pre-existing theoretical frameworks or categories (Flynn et al., 2024). These can be derived from specific research models, theoretical constructs (Eaton et al., 2019), or the objectives of the study (Bussell, 2020). This approach ensures that the analysis aligns closely with the study's goals and theoretical underpinnings.

Use of large language models (LLMs) in qualitative analysis

The advent of LLMs has opened new opportunities for integrating AI into qualitative analysis. Its rapid growth and widespread utility across various fields, particularly in streamlining content creation and understanding human instructions (Chavan et al., 2024), position LLMs as a powerful tool for enhancing qualitative research.

Several studies have examined the ability of AI tools to automate qualitative coding tasks, with the goal of improving the efficiency and accuracy of workflows from open coding to codebook development (Gao et al., 2024; Sinha et al., 2024). Comparisons between AI tools and traditional human qualitative coding methods have demonstrated LLMs' ability in supporting nuanced interpretations of data (Amarasinghe et al., 2023; Zambrano et al., 2023). Additionally, the transparency provided by LLMs' explanations of coding decisions helps support greater consistency and validity when evaluating human coding of qualitative data (Zambrano et al., 2023). Furthermore, these automated processes help significantly reduce manual efforts and address bottlenecks in the qualitative research process (Barany et al., 2024).

Finally, the collaborative aspects of qualitative analysis have also benefited from AI integration. Systems like CollabCoder and CoAIcoder leverage AI to support independent open coding, iterative discussions of coding, and codebook refinement (Gao et al., 2023; Gao et al., 2024). Further, tools like PaTAT enable researchers to iteratively define and refine patterns in annotated data, thereby supporting the iterative and interpretive nature of thematic analysis (Gebreegziabher et al., 2023). LLMs have also played a role in supporting thematic analysis by enhancing coding efficiency, data exploration, and comprehension for researchers with varying levels of expertise (Yan et al., 2024). These advancements in tools can promote transparency, trustworthiness, and efficiency in the human-AI workflows, further supporting the collaborative research processes (López-Fierro & Nguyen, 2024).

To our knowledge, however, studies have not examined the role of LLMs in analyzing field notes. These written records are often a vital part of qualitative research inquiry, yet time consuming to analyze. Therefore, the purpose of this study is to compare large language models and human researchers in deductively coding field notes.

Methods

Study Context

The data analyzed in this study originated from a larger research project focused on developing instructional units as part of a curriculum for fifth grade students that integrate computer science (CS) concepts into elementary-level mathematics instruction (Shehzad et al., 2023).

Two teachers taught these lessons in a rural school district in the Western United States. They taught a total of six math lessons: five lessons on the topic of exponents and one lesson on the topic of polygons.

Large Language Model

For this work, we used ChatGPT-4o, OpenAI's late 2024 model, with twice the text generation speed of ChatGPT-4 Turbo (OpenAI, 2024a). All prompts were deployed within GPTs, which are custom versions of ChatGPT, allowing us to set up an environment with common data, instructions and parameters. By doing so, we limited the potential bias and out-of-context interpretations of the LLM by specifying the limited “sources” and describing a specific “context”. Moreover, with privacy in mind, none of our GPTs were published. Additionally, we used several different “chat windows” to analyze field notes separately, thereby mitigating potential cross-contamination between files.

Finally, during our analysis process, we randomly selected excerpts to confirm that they belonged to the appropriate file. We also prompted ChatGPT to provide “justifications”, “better answers”, or “wrong examples”, making it easier for us to assess its responses.

Data Sources

Field notes can be either structured—addressing discrete or predetermined categories for observations—or, unstructured—allowing for open-ended interpretation and observation (Mulhall, 2003). This study utilized a structured template where researchers recorded their field notes.

Three researchers observed the two teachers’ classrooms as they taught the math lessons. In our study, we were interested in observing the extent to which the teachers were able to implement the CS-integrated math lessons, made adaptations, used different instructional strategies, or needed support. During the researchers’ observations, they recorded field notes using a standardized template. The structured template included prompts for recording student engagement with the lessons, summarizing the lessons’ implementations, and describing strategies, supports, and needs of the teachers.

After the field notes were recorded, two researchers reviewed the field notes template and study goals to identify key themes. They then developed a codebook, composed of six main codes along with their definitions (see ***Appendix A***), to deductively code the field notes.

Deductive Coding Process

One researcher applied the deductive coding scheme to analyze all six field notes. In addition, based on how field researchers structured their notes, he also defined and used double coding to capture multiple themes within single excerpts, allowing for a more nuanced interpretation of the data. Further, the researcher provided examples to illustrate the application of each code.

Using the same codebook, definitions, and examples (see ***Appendix A***), ChatGPT-4o was prompted to perform a coding analysis. We began by setting up an instance of GPT Explore. In it, we (1) uploaded the six field notes as separate files, (2) provided basic context (such as data type, format, and organization of the data) and prompts (related to expected process outcomes and format; see ***Appendix B*** for this first prompt), and (3) unchecked the "Web browsing" and "Use conversation data in your GPT to improve our models" features.

Within this GPT, six different "new chats" were created to analyze each field note separately. Our process in each chat started by asking it to list all the file names to confirm that it was correctly reading the files. Next, we asked it to retrieve the content of the file for the field note we were about to analyze and confirmed that the information retrieved was as expected (we had cases where ChatGPT retrieved data from a different field note than the one requested or changed pieces of text). Finally, we asked it to code the file. Our prompt included the codebook consisting of a list of the codes, their definitions, and examples. Since we noticed that redundancy and repetition in the instructions were necessary to increase the accuracy of the results, the prompt also included the name of the file to be analyzed, details on how the field notes are organized, how to recognize the excerpt or "unit of analysis", and how we expected the results to be presented. Additionally, to understand the logic applied to the analysis, we also asked ChatGPT to justify its use of codes (see ***Appendix B*** for the second prompt).

To add an extra validation process, we also prompted each GPT to "redo the table" and add two new columns for: "Unrelated Codes" and "Justification for Unrelated Codes". See ***Appendix B*** for more details on this third prompt and ***Appendix C*** for examples of the results produced by ChatGPT.

Comparative Analysis

To contrast the performance of the researcher and the LLM, we began by identifying areas of agreement and disagreement in their coding of each excerpt from the field notes. To facilitate this comparison, we also organized the data into a contingency table, which displayed the frequency of code application for each excerpt and for each combination of codes.

To compute Cohen's Kappa, we used the contingency table comparing the classifications of the researcher and ChatGPT. We calculated the observed agreement (p_0) as the proportion of times the raters agreed, and the expected agreement (p_e) as chance agreement. We then calculated Cohen's Kappa using its formula $\kappa = \frac{p_0 - p_e}{1 - p_e}$ (Conger, 2017).

We also organized the coding data into a confusion matrix (Heydarian et al., 2022). Using the researcher's coding as the "gold standard", we counted the frequency with which ChatGPT:

1) applied a code that agreed with the researcher's code (True Positive), 2) applied a code that disagreed with the researcher's code (False Positive), 3) did not apply a code that was applied by the researcher (False Negative), and 4) did not apply a code that was also not applied by the researcher (True Negative).

From the confusion matrix, we calculated several performance measures to compare the coding performed by ChatGPT and the researcher (Baldi et al., 2000; Galdi & Tagliaferri, 2018). We calculated the *Accuracy*, or the overall proportion of correct predictions, indicating how often ChatGPT's classifications matched the researcher's codes (Baldi et al., 2000). *Precision* measured the proportion of predicted positives that were correct, highlighting ChatGPT's ability to avoid false positives. *Recall* represents the proportion of actual positives that the LLM correctly identified, centering on its ability to minimize false negatives. Finally, we calculated the *F1 Score,* a combined measure of both precision and recall. This measure accounts for both false positives and false negatives, which is particularly useful when there is an uneven distribution of classifications. As these metrics are reported as proportions, they vary between .00 and 1.00, with values closer to 1.00 representing better alignment with the gold standard. Formula for each metric:

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$

Precision: $\frac{TP}{TP+FP}$

Recall: $\frac{TP}{TP+FN}$

F1: $\frac{2*Precision*Recall}{Precision+Recall}$

Positioning ChatGPT as a qualitative analysis partner

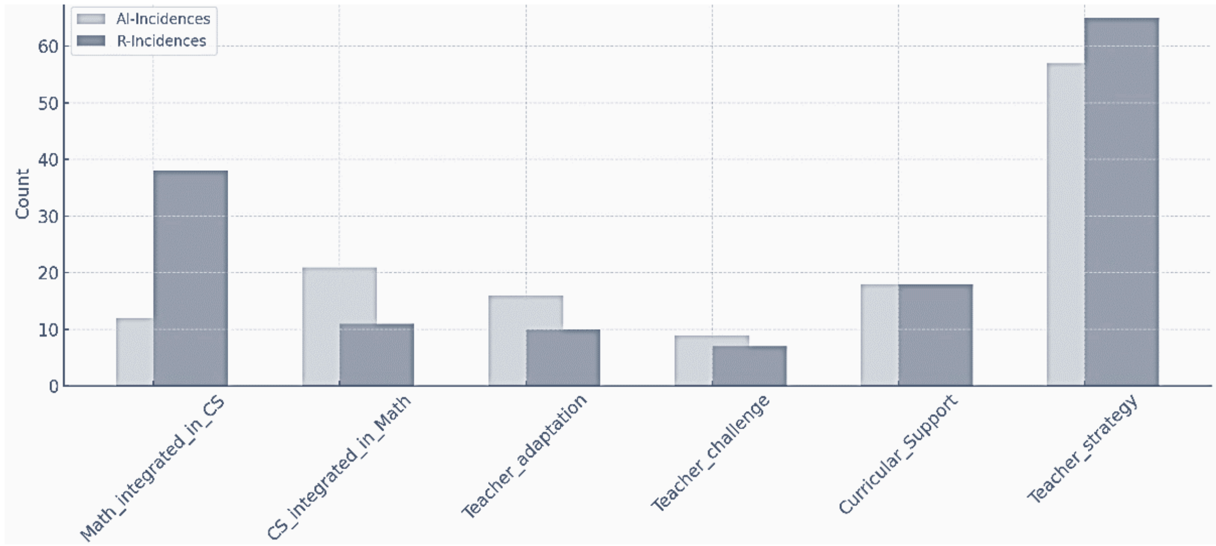
Leveraging the computational power of LLMs for recognizing patterns (Perkins & Roe, 2024), supporting nuanced interpretations of data (Amarasinghe et al., 2023; Zambrano et al., 2023), and providing explanations for coding decisions (Zambrano et al., 2023), we prompted ChatGPT to justify its code choices and analyze why other codes were not considered (which, on some occasions, included the codes chosen by the researcher). Additionally, we prompted it to infer the researcher's code choices, allowing us to gain a different perspective on disagreements between the researcher and ChatGPT.

Findings

We first conducted a general comparative analysis of code application in terms of agreements and disagreements between the researcher and the LLM. We noted that "Teacher strategy" was the most frequently used code by both, with a slightly higher frequency in the researcher's coding. "Teacher challenge" was the least frequently used code by both, with a slightly higher frequency in ChatGPT's coding (see ***Figure 1***). Thus, the researcher and ChatGPT appeared to agree on the frequency of applying these codes.

Figure 1

Frequency of code application by the researcher (R) and the LLM



We then organized our data in a contingency table (see ***Figure 2*** and ***Figure 3***). "Teacher strategy" showed the highest agreement in applying the code to the same excerpt (True negative n=43). "Math_integrated_in_CS", "Teacher_adaptation", and "Curricular_support" were the codes with the least overlap between ChatGPT and the researcher (N=3 each). Consequently, this general lack of agreement resulted in an overall very low Cohen's Kappa score of .06.

Figure 2

Contingency table and heat map of frequency of code applications

		Human					
		Math_integrated_in_CS	CS_integrated_in_Math	Teacher_adaptation	Teacher_challenge	Curricular_Support	Teacher_strategy	Total AI
AI	Math_integrated_in_CS	3	4	1	0	4	7	19
	CS_integrated_in_Math	11	8	2	0	5	15	41
	Teacher_ adaptation	7	1	3	3	2	12	28
	Teacher_ challenge	0	0	2	5	3	4	14
	Curricular_ Support	9	2	1	1	3	15	31
	Teacher_ strategy	30	7	3	2	8	43	93
	Total Human	60	22	12	11	25	96	65

Note. The researcher and ChatGPT applied codes to each field note excerpt. Darker cells indicate higher agreement, while lighter cells indicate lower or no agreement.

Figure 3

Contingency table and heat map of percentages of code applications

AI \ Human	Math_integrated_in_CS	CS_integrated_in_Math	Teacher_adaptation	Teacher_challenge	Curricular_Support	Teacher_strategy	Total AI
Math_integrated_in_CS	3.33%	4.44%	1.11%	0.00%	4.44%	7.78%	21.11%
CS_integrated_in_Math	12.22%	8.89%	2.22%	0.00%	5.56%	16.67%	45.56%
Teacher_adaptation	7.78%	1.11%	3.33%	3.33%	2.22%	13.33%	31.11%
Teacher_challenge	0.00%	0.00%	2.22%	5.56%	3.33%	4.44%	15.56%
Curricular_Support	10.00%	2.22%	1.11%	1.11%	3.33%	16.67%	34.44%
Teacher_strategy	33.33%	7.78%	3.33%	2.22%	8.89%	47.78%	103.33%
Total Human	66.67%	24.44%	13.33%	12.22%	27.78%	106.67%	72.22%

Note. The researcher and ChatGPT applied codes to each field note excerpt. Darker cells indicate higher agreement, while lighter cells indicate lower or no agreement.

To examine the performance of ChatGPT's coding in comparison to the researcher, we generated a confusion matrix and computed several performance measures for each code (accuracy, precision, recall, and F1 score; see ***Table 4***). Results showed that although "Teacher_challenge" was the code applied with the lowest frequency, (resulting in a high number of True negative instances), it had the highest accuracy (.93), a high recall (.71) and precision (.56) value, resulting in a medium F1 score (.63). The "Teacher strategy" code showed the highest F1 score (.70), also with a high precision value (.75).

Table 4

Confusion Matrix Results and Performance Metrics (Accuracy, Recall, Precision, F1) for 6 Original Codes

	TP	FP	TN	FN	Accuracy	Precision	Recall	F1
Math_integrated_in_CS	3	9	43	35	.51	.25	.08	.12
CS_integrated_in_Math	8	13	66	3	.82	.38	.73	.50
Teacher_adaptation	3	13	67	7	.78	.19	.30	.23
Teacher_challenge	5	4	79	2	.93	.56	.71	.63
Curricular_Support	3	15	57	15	.67	.17	.17	.17
Teacher_strategy	43	14	11	22	.60	.75	.66	.70
Total	62	59	280	49	.76	.41	.51	.45

For the code “CS integrated in Math”, ChatGPT applied the code 21 times (see ***Figure 1***), while the researcher only applied it 11 times. However, as there were many instances where the code was not found by both (*True negative*), that resulted in a high accuracy value of .82, a recall value of .73 and an overall F1 score of .5. The “Math Integrated in CS” code was applied by ChatGPT 17 times while the researcher applied it 38 times (see ***Figure 1***), yet they only agreed 3 times. Not surprisingly, this code had the lowest F1 score (.12).

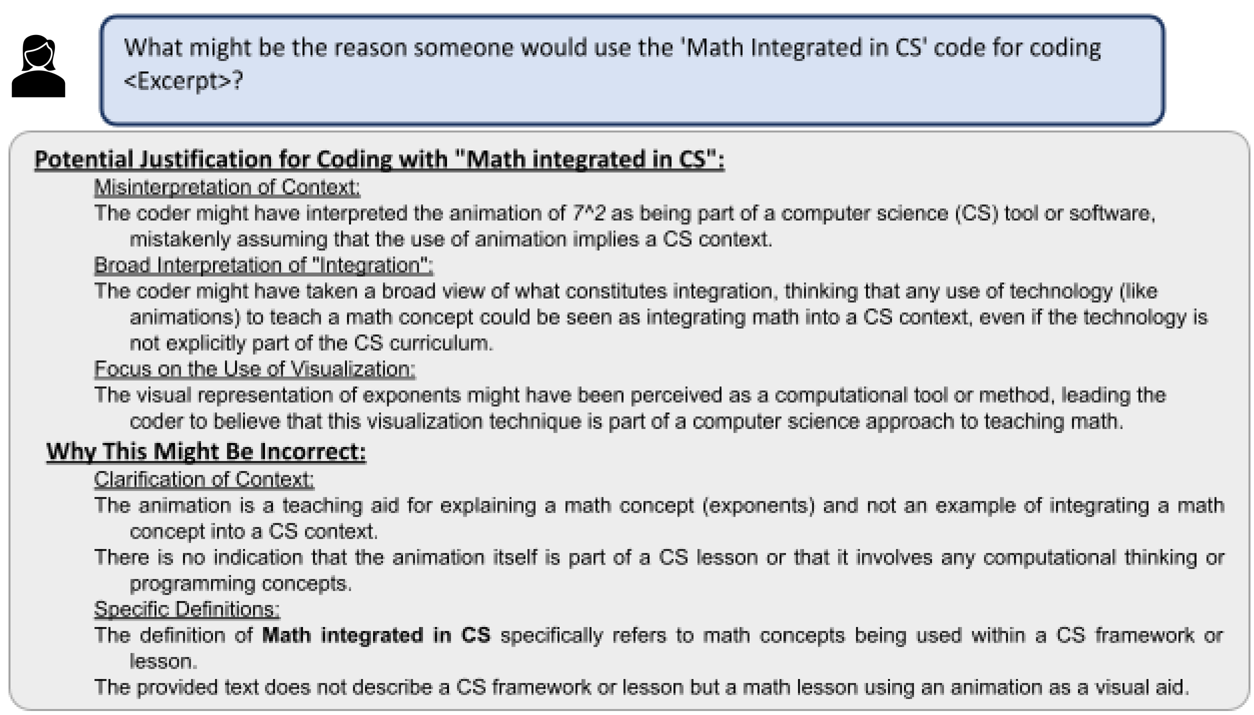
To further examine disagreements between code applications by the researcher and ChatGPT, we prompted the LLM to justify its coding decisions for both codes selected and not selected. For example, ChatGPT applied the code “Teacher_strategy” and the researcher selected “Teacher_challenge” (see from ***Appendix C, excerpt 5***) for the following field notes excerpt:

> “*some hesitation when problems like 4x10^2 were introduced. The Scratch program does not allow for an expression like this but teacher could have done 10^2 and then discussed…The teacher was able to talk them through it but there was some hesitation on the students’ part.”*

ChatGPT justified its choice of the code “Teacher_strategy”, stating that there is “*need for more support and real-world examples*”. It justified its choice of not applying “Teacher_challenge” because there is “*no specific issues faced by the teacher. The hesitation is on the students’ part, not a problem faced by the teacher*.”

Figure 4

Example of how the LLM justified discrepancies between researcher and its coding results



What might be the reason someone would use the 'Math Integrated in CS' code for coding <Excerpt>?

Potential Justification for Coding with "Math integrated in CS":

Misinterpretation of Context:
The coder might have interpreted the animation of 7^2 as being part of a computer science (CS) tool or software, mistakenly assuming that the use of animation implies a CS context.

Broad Interpretation of "Integration":
The coder might have taken a broad view of what constitutes integration, thinking that any use of technology (like animations) to teach a math concept could be seen as integrating math into a CS context, even if the technology is not explicitly part of the CS curriculum.

Focus on the Use of Visualization:
The visual representation of exponents might have been perceived as a computational tool or method, leading the coder to believe that this visualization technique is part of a computer science approach to teaching math.

Why This Might Be Incorrect:

Clarification of Context:
The animation is a teaching aid for explaining a math concept (exponents) and not an example of integrating a math concept into a CS context.
There is no indication that the animation itself is part of a CS lesson or that it involves any computational thinking or programming concepts.

Specific Definitions:
The definition of **Math integrated in CS** specifically refers to math concepts being used within a CS framework or lesson.
The provided text does not describe a CS framework or lesson but a math lesson using an animation as a visual aid.

As an additional approach to validating ChatGPT's potential as a contributor to qualitative research, we prompted it to infer on the researcher’s code choice.

Another example of disagreement (see from ***Appendix C*, *excerpt 1***) is where the code "Math_integrated_in_CS" was applied by the researcher*,* but not by ChatGPT. We prompted it to comment on the possible reasons for "mistakenly" coding that excerpt differently that the researcher (see ***Figure 4***). Part of its response was:

> "*coder might have interpreted the animation of 7^2 as being part of a computer science (CS) tool or software, mistakenly assuming that the use of animation implies a CS context [...] the scenario is purely mathematical […] The animation is a teaching aid for explaining a math concept (exponents) and not an example of integrating a math concept into a CS context. […] The definition of Math integrated in CS specifically refers to math concepts being used within a CS framework or lesson.*"

This response highlighted that these two similarly worded codes ("Math_integrated_in_CS" and "CS_integrated_in_Math") could lead to misinterpretations when coding by both ChatGPT and the researcher. As a result, we decided to collapse the codes into a new 'Math_CS_integration'. After the combination, the number of times that ChatGPT and the researcher agreed (*True Positive*) increased, also resulting in a higher precision value of .72 and a higher F1 score of .58 (see ***Table 5***).

Table 5

Confusion Matrix Results and Performance Metrics (Accuracy, Recall, Precision, F1) for Combined Code

	TP	FP	TN	FN	Accuracy	Precision	Recall	F1
Math_CS_Integration	23	9	34	24	.63	.72	.49	.58

Note. "Math_integrated_in_CS" and "CS_integrated_in_Math" are combined into the code "Math_CS_Integration".

These findings demonstrate that while ChatGPT can engage in deductive coding, it, as well, can provide insights to better understand the data and coding decisions. Moreover, with additional validation prompts, researchers can leverage deductive coding in conjunction with LLMs to enhance their analysis of field notes, ensuring that both, human insights and LLMs capabilities contribute to a more comprehensive understanding of the data.

Discussion and Limitations

Field notes taken during research field work form an important qualitative data source, yet their analysis remains time consuming and complex. This study examined the extent to which ChatGPT can engage in qualitative coding of field notes and to what extent the automated analysis of deductive coding is synergistic with a human coder.

Our findings suggest that in some cases, the LLM's coding showed congruence with the researcher, but in many cases it did not. In particular, the "Teacher_strategy" code showed the highest level of agreement, with the other codes showing much lower levels. Further analysis revealed that the researcher and the LLM appeared to generally agree on the frequency in applying the different codes. However, as measured by the more traditional Cohen's Kappa metric, overall agreement was very low. Nevertheless, when examining the individual code level agreements using a human coder as the “gold standard” and establishing measures in information science (accuracy, precision, recall, and F1 scores), a more nuanced picture emerged.

Justifications provided by the LLM for different code applications also revealed potential confusions in codebook definitions, perhaps contributing to the low level of agreements. In particular, collapsing two codes with overlapping definitions helped raise the agreement values. This additional level of insight can perhaps help support the iterative development of codebooks.

While this study offers insights into the role of LLM-assisted coding in enhancing qualitative analysis, it has several limitations. First, the sample size of field notes analyzed was relatively small, which may limit the generalizability of the findings to larger datasets or different research contexts. Furthermore, the study focused on a specific set of codes which were deductively applied. This can constrain the analysis of the field notes by not representing all their content and limiting nuanced interpretation. Additionally, the deductive coding was performed by a researcher without a follow-up on inter-rater reliability, which may have compromised the consistency and accuracy of the coding process. Finally, while the LLM was employed to assist in the coding process, it still relies on human input for training and validation, which introduces the potential for bias in the model's performance.

Conclusions

Field notes can play a crucial role in qualitative research by providing a rich description of in-situ activities. However, their analysis can be a time consuming and challenging endeavor. This study investigated the extent that LLMs can provide additional support in analyzing and validating these data sources. It suggests ways that LLMs can support research processes, critique researchers’ work, and offer insights that complement traditional methods. Moreover, the study also highlighted differences between how a LLM and the researcher approached a task.

In addition, acknowledging the complexities of qualitative research for novice field note takers, this work contributed methods and prompts to LLMs that could help guide and organize analysis, thereby aiding in qualitative inquiry. In future work, we plan to test LLM’s ability to engage in the generation of field notes by directly analyzing transcriptions of classroom implementations. This will help assess the capabilities of a LLM to act as a "field note taker,” also a time consuming, challenging, yet vital part of qualitative inquiry.

Acknowledgments

This work was supported by the National Science Foundation under grants Nos. 2031382 and 2031404. The opinions expressed herein are those of the authors and do not necessarily reflect those of the funding organization.

References

Amarasinghe, I., Marques, F., Ortiz-Beltrán, A., & Hernández-Leo, D. (2023, August). Generative pre-trained transformers for coding text data? An analysis with classroom orchestration data. In *European Conference on Technology Enhanced Learning* (pp. 32-43). Cham: Springer Nature Switzerland.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, *16*(5), 412-424.

Barany, A., Nasiar, N., Porter, C., Zambrano, A. F., Andres, A. L., Bright, D., Shah, M., Liu, X., Gao, S., Zhang, J., Mehta, S., Choi, J., Giordano, C., & 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.*

Beltran, M. A., Ruiz Mondragon, M. I., & Han, S. H. (2024, June). Comparative Analysis of Generative AI Risks in the Public Sector. In *Proceedings of the 25th Annual International Conference on Digital Government Research* (pp. 610-617).

Bussell, J. (2020). Shadowing as a Tool for Studying Political Elites. *Political Analysis, 28*(4), 469–486. doi:1.1017/pan.202.14

Chan, E., Small, S. S., Wickham, M. E., Cheng, V., Balka, E., & Hohl, C. M. (2021). The utility of different data standards to document adverse drug event symptoms and diagnoses: mixed methods study. *Journal of Medical Internet Research, 23*(12), e27188.

Chavan, J. D., Mankar, C. R., & Patil, V. M. (2024). Opportunities in Research for Generative Artificial Intelligence (GenAI), Challenges and Future Direction: A Study. *International Research Journal of Engineering and Technology*, *11*(02), 446-451.

Combrinck, C. (2024). A Tutorial for Integrating Generative AI in Mixed Methods Data Analysis. *Discover Education*, *3*(1), 116.

Conger, A. J. (2017). Kappa and rater accuracy: Paradigms and parameters. *Educational and psychological measurement*, *77*(6), 1019-1047.

Copland, F. (2018). Observation and Fieldnotes. In: Phakiti, A., De Costa, P., Plonsky, L., Starfield, S. (eds) *The Palgrave Handbook of Applied Linguistics Research Methodology*. Palgrave Macmillan, London. https://doi.org/1.1057/978-1-137-59900-1_12

Eaton, K., Stritzke, W. G., & Ohan, J. L. (2019). Using scribes in qualitative research as an alternative to transcription. *The Qualitative Report, 24*(3), 586-605.

Flynn, N., Teemant, A., Viesca, K. M., & Perumal, R. (2024). Effective Teachers of Multilingual Learners: A Mixed-Method Study of UK and US Critical Sociocultural Teaching Practices. *TESOL Quarterly, 58*(1), 195-221.

Galdi, P., & Tagliaferri, R. (2018). Data mining: accuracy and error measures for classification and prediction. *Encyclopedia of bioinformatics and computational biology*, *1*, 431-436.

Gao, J., Choo, K. T. W., Cao, J., Lee, R. K. W., & Perrault, S. (2023). CoAIcoder: Examining the effectiveness of AI-assisted human-to-human collaboration in qualitative analysis. *ACM Transactions on Computer-Human Interaction, 31(1), 1-38.*

Gao, J., Guo, Y., Lim, G., Zhang, T., Zhang, Z., Li, T. J. J., & Perrault, S. T. (2024, May). CollabCoder: a lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1-29).*

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).*

Johnson, A. H., Taylor, J. L., Caudillo, L., Hwang, H., Gill, E., & Harrison, T. C. (2024). Addressing Race in Fieldnotes in Qualitative Health Research: A Methodological Critique*. International Journal of Qualitative Methods, 23*, 16094069231225372.

Heydarian, M., Doyle, T. E., & Samavi, R. (2022). MLCM: Multi-label confusion matrix. *IEEE Access*, *10*, 19083-19095.

Irwin, R. E., Houck, N. M., Kramer, C. N. P., Zoucha, R., Martin, M. B., BC, C., & Turk, M. T. (2013). Fieldwork as a Way of Knowing: An Italian Immersion Experience. *Online Journal of Cultural Competence in Nursing and Healthcare 3(3*):1-15.

López-Fierro, S., & Nguyen, H. (2024). Making Human-AI Contributions Transparent in Qualitative Coding. In *Proceedings of the 17th International Conference on Computer-Supported Collaborative Learning-CSCL 2024, pp. 3-1.* International Society of the Learning Sciences.

Maharaj, N. (2016). Using field notes to facilitate critical reflection. *Reflective Practice, 17*(2), 114–124.

Montgomery, P., & Bailey, P. H. (2007). Field notes and theoretical memos in grounded theory. *Western journal of nursing research, 29*(1), 65-79.

Mulhall, A. (2003). In the field: notes on observation in qualitative research. *Journal of advanced nursing, 41*(3), 306-313.

OpenAI. (n.d.). Safety best practices - openai API. *https://platform.openai.com/docs/guides/safety-best-practices*

OpenAI. (2024a). Models - openai API. https://platform.openai.com/docs/models/gpt-4o

OpenAI. (2024b). Prover-Verifier Games improve legibility of language model outputs. https://openai.com/index/prover-verifier-games-improve-legibility/

Papen, U. (2019). Participant observation and field notes. In *The Routledge handbook of linguistic ethnography (pp. 141-153). Routledge.*

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).*

Phillippi, J., & Lauderdale, J. (2018). A guide to field notes for qualitative research: Context and conversation. *Qualitative health research, 28(3), 381-388.*

Shehzad, U., Clarke-Midura, J., Beck, K., Shumway, J. & Recker, M. (2023). Co-Designing Elementary-Level Computer Science and Mathematics Lessons: An Expansive Framing Approach. In *Proceedings of the 2023 Annual Meeting of the International Conference of the Learning Sciences.* Montreal, Canada: International Society of the Learning Sciences*.*

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 ACM the Symposium on Learning, Design and Technology (pp. 17-25).*

Tjora, A. H. (2006). Writing small discoveries: an exploration of fresh observers' observations. *Qualitative research, 6(4), 429-451*.

Tomitza, C., Schaschek, M., Straub, L., Winkelmann, A. (2023) "What is the Minimum to Trust AI?—A Requirement Analysis for (Generative) AI-based texts. *Wirtschaftsinformatik 2023 Proceedings.* 35.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu C., Xiong Z., Dutta R., Schaeffer R., Truong S., Arora S., Mazeika M, Hendrycks D., Lin Z., Cheng Y., Koyejo S., Song D, Li, B. (2023, June). DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In

NeurIPS 2023. Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track
Yan, L., Echeverria, V., Fernandez-Nieto, G. M., Jin, Y., Swiecki, Z., Zhao, L., Gašević, D., & Martinez-Maldonado, R. (2024, May). Human-AI Collaboration in Thematic Analysis using ChatGPT: A User Study and Design Recommendations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1-7).
Zambrano, A. F., Liu, X., Barany, A., Baker, R. S., Kim, J., & Nasiar, N. (2023, October). From nCoder to ChatGPT: From automated coding to refining human coding. In *International conference on quantitative ethnography (pp. 470-485). Cham: Springer Nature Switzerland.*

Appendix A
Code Book for Analyzing Field Notes

Code	Definition	Example: Excerpts from field notes
Math integrated in CS	Math and CS integration highlighting CS concepts within math: Code is used when a math concept is integrated in a computer science (CS) context	Activity 3 is relatively new – only a brief exposure in computer lab. Put in base of 7 and exponent of 2 Teacher effectively used Scratch to show 7^2
CS integrated in Math	Math and CS integration highlighting math concepts within CS: Code is used when a CS concept is integrated in a math context	<Teacher_name> showed the visualizations for repeated addition and repeated multiplication side by side and explained one is repeated addition and one is repeated multiplication even though they are using the same numbers but the output is different. <Teacher_name> zoomed into the outputs of the two codes,which showed that the first line of cats was the same number but it started changing and exponents started becoming much bigger very quickly. Many students exclaimed in wonder while <Teacher_name> emphasized that exponents are much different from multiplication.
Teacher_adaptation	Teacher adaptations or extensions of the curriculum: Code is used when the teacher applies an adaptation to the lesson or extends it	This is obviously an experienced and skilled teacher. Students were led through classroom procedures effortlessly. They felt free to make comments and participate.
Teacher_challenge	Challenges faced by teachers during implementation: Code is used when the teacher runs into a problem while implementing	On slide with the cat, <Teacher_name> used the clicker to go forward in the slides. When the 72 and the explanatory text showed up in the slide, <Teacher_name> said "we should show the animation first". She also had problems pulling up the correct animation. She pulled one and then said "oh it was the other animation". This suggests that the order of the slides did not match <Teacher_name>'s expectation. She wanted to show the animation before discussing the answer of 72.
Curricular Support	Use of curricular supports by teachers: Code is used when the teacher uses a support embedded in the curricular materials	<Teacher_name> begins the lesson. She asks what the 6 in 6^4 is. A student responds exponent. <Teacher_name> explains that the 4 is the exponent. <Teacher_name> prompts her to phone a friend. The student asks a friend to help her with the answer.
Teacher strategy	Strategies or supports influencing student interest or engagement: Code is used when the teacher uses a strategy or support aimed at improving students' interest in or engagement with the curriculum	<Teacher_name> finishes the lesson by reading the closing statement. Students take the SEET. As she is collecting SEETs one student mentions how cool it would be to create "Where's Waldo" in Scratch. <Teacher_name> encouraged him to figure it out and then come show her how he did it.

Appendix B

Prompts used in the study

Type	Prompt
For GPT Explore Instance To provide context and general instructions to the GPT	These documents are part of the academic records of a school math class, using CS concepts. The dialogues belong to a teacher and her students. The teacher's name is anonymized with initials (ABC or XYZ). -Work only with the document(s) listed in the instructions -Review the entire document, including the table (Time, Narrative: What you observed Notable moments: Interest and engagement with CS concepts; Ways math highlights CS or CS highlights math), the "Observer Summary," and the "Strategies and Supports that worked" -All responses should be excerpts from the document -For each response you give, put the name of the document in parentheses and, as an index of where you take the information, put in brackets what is in the "Time" column, in addition to the text excerpt. -If a table is required, all responses should be given in table format. ALWAYS put the codes at the top and the extracts in the following rows -For all the answers given, explain at the end your rationale or the reason behind each answer you provide.
For AI-deductive coding To request to code the field notes	I have 6 codes and their definitions: (1) code: Math integrated into CS, definition: This code is used when a math concept is integrated into a CS context (2) code: CS integrated into math, definition: This code is used when a CS concept is integrated into a math context (3) code: Teacher_adaptation, definition: This code is used when the teacher applies an adaptation to the lesson or extends it (4) code: Teacher_challenge, definition: This code is used when the teacher encounters a problem during implementation (5) code: Curriculum support, definition: This code is used when the teacher uses a support integrated into the curriculum materials (6) code: Teacher strategy, definition: This code is used when the teacher uses a strategy or support intended to enhance student interest or engagement in the curriculum From file: "<file_name>", based on the codes and definitions provided, code each block (each block starts with the value that is in the "Time" column, and includes the columns "Time," "Narrative: What You Observed," and "Notable Moments: Interest and Engagement with CS Concepts"; ways math highlights CS or CS highlights math"; each block ends before a new "Time" column value.) and sections ("Observer Summary," "Strategies and Supports That Worked," "Strategies and Supports Still Needed," and "Lesson Plan Notes"). Each block or section can have more than one code. FINALLY, generate two tables. (1) a single table for all blocks and sections (each value from the block should be combined into a single cell, bolding what is in the "Time" column). Put all the data (block and sections) from this table in the first column. In the second column put all the codes you find. The third column, justifications for all those codes. Fourth column verbs or nouns from the text (block or sections) that would justify your codes. (2) In the second table, the first row put all the codes in horizontal order, at the top; And below each ALL the complete blocks or sections that were coded (same as the table above, the ones you generated at the beginning). Put each block or section in a different row. Put the file name in the first column. If there is no data in any of the files for the codes, put "N/A"
For validating AI codes To request to code again the field notes	Review the blocks and sections again, and in this case put a new column with the codes that are not related to them, and in a following column justify your answer

Appendix C

Examples of ChatGPT output

E	Block/Section	R	AI	Justifications for AI Codes	Unrelated Codes	Justifications for Unrelated Codes
1	<Teacher_name> showed the ladybugs animation of 7^2. A student called them "mustache bugs." <Teacher_name> asked if it is a multiplication problem or exponents problem. Notable Moments: Starting from a visual representation of exponents and asking what operation it represents is another way the concept can be transferred between representations.	Math_integrated_in_CS	Teacher_strategy	Exponents shown via animation, a strategy for engagement	Math_integrated_in_CS Teacher_adaptation Teacher_challenge Curricular_Support	Math concepts were not taught through computer science activities. The lesson did not involve changes or additions from the standard plan. No issues or obstacles were encountered by <Teacher_name>. <Teacher_name> did not rely on predefined aids or materials from the curriculum during this part of the lesson.
2	After taking some answers for the 7^2, <Teacher_name> changed the numbers in the code to 2^5. She then changed it to 6^3. When the ladybug said 216 before printing, the students were audibly surprised. When asked about what they noticed, a student said there was a 6 in every bubble. Notable Moments: The student noticed the 6 in every bubble, which is a correct observation, but not the main takeaway of the activity. The ladybug stopped before reaching the ceiling. The code can be changed so they keep stamping over each other starting from the bottom again.	Teacher_strategy, CS_integra	Math_integrated_in_CS, Curri	Using math concept (exponents) in a coding activity; feedback on coding issue provided	CS_integrated_in_Math Teacher_adaptation Teacher_challenge Teacher_strategy	Computer science ideas were not brought into the math instruction No deviations or extensions from the original lesson were applied. <Teacher_name> did not experience any problems or setbacks. No specific methods aimed at increasing engagement were noted.

		ted_in_Math	cular_Support			
3	"Reads computer lab statement from lesson"	Curricular_Support	Curricular_Support	Teacher reads directly from curricular materials	Math_integrated_in_CS CS_integrated_in_Math Teacher_adaptation Teacher_challenge Teacher_strategy	No math concept was integrated into a CS context. No CS concept was integrated into a math context. No adaptation or extension of the lesson was mentioned. No challenges were encountered during this activity. No specific strategy aimed at improving engagement was mentioned.
4	<Teacher_name> told students that <Researcher_name> is here to observe "how smart you are so show him your best". It's hard to say if <Teacher_name> did this as a way to manage the classroom or if some students had previously asked her questions about the USU people being in the class.	Teacher_strategy	Teacher_strategy	<Teacher_name> motivated students by mentioning an observer, aiming to improve engagement.	Math_integrated_in_CS CS_integrated_in_Math Teacher_adaptation Teacher_challenge Curricular_Support	Math_integrated_in_CS: No math concept is being applied in a CS context. CS_integrated_in_Math: No CS concept is being applied in a math context. Teacher_adaptation: No evidence of adaptation or extension of the lesson. Teacher_challenge: No problem or difficulty faced by the teacher. Curricular_Support: No specific curricular materials were referenced or used.

5	**Strategies and Supports that are still needed** Some hesitation when problems like 4x10^2 were introduced. The Scratch program does not allow for an expression like this but teacher could have done 10^2 and then discussed how the 4 affects the answer. However,there seemed to be a small disconnect once the students transitioned from expressions with one base/exponent to expressions like 4x10^2. The teacher was able to talk them through it but there was some hesitation on the students' part.	Teacher_challenge	Teacher_strategy	Need for more support and real-world examples	Math_integrated_in_CS, CS_integrated_in_Math, Teacher_adaptation Teacher_challenge	Math_integrated_in_CS: No integration of math concepts in a CS context. The focus is on understanding math expressions CS_integrated_in_Math: No CS concepts are used in this explanation. The emphasis is on math Teacher_adaptation: No modifications or extensions mentioned. The teacher follows the planned lesson. Teacher_challenge: No specific issues faced by the teacher. The hesitation is on the students' part, not a problem faced by the teacher.