

# Collaborative AI for Qualitative Analysis: Bridging AI and Human Expertise for Scalable Analysis\*

Grace C. Lin<sup>1\*</sup>, Emma Anderson<sup>1</sup>, Carúmeys Stevens<sup>1</sup>, Brandon Hanks<sup>1</sup>, Disha Chauhan<sup>1</sup>, Amelia Farid<sup>1†</sup>, Mic Fenech<sup>1†</sup>, and Eric Klopfer<sup>1</sup>

<sup>1</sup> Massachusetts Institute of Technology, Cambridge, MA 02139, USA

## Abstract

Qualitative research in learning settings often faces challenges of time consumption and iterative refinement. To address these issues, we developed CAILA (Collaborative AI for Learning and Analysis), a novel AI-assisted system designed to support researchers in thematic analysis and address challenges such as thematic saturation. Using a GPT-based model with adjustable parameters and stopping criteria, CAILA aids researchers in generating and refining themes efficiently while preserving the rigor of human oversight. Notably, CAILA's stopping criteria—three iterations with no new themes generated—ensures a balance between thoroughness and efficiency. We evaluated the CAILA tool by comparing the analysis of a set of student conversations (146 utterances) using CAILA with the thematic analysis conducted by two human researchers. While the human+CAILA approach found themes directly answering the question posed, the humans-only approach refined the research question, a staple in qualitative research. We discuss the implications of using AI-powered qualitative analytic tools.

## Keywords

Qualitative methods, Education/learning, Conversation analysis with AI, Collaborative and social computing ~ Collaborative and social computing design and evaluation methods

## 1. Introduction

Open up a book on qualitative research (e.g., [1]), and a plethora of methods will meet your eyes. It is said that there are as many methods of qualitative research as there are qualitative researchers; after all, the researcher is considered an instrument in the research [2], [3]. Yet, the field is surprisingly unified on one matter. Ask any qualitative researcher a pain point of their work. Undoubtedly the dominating winner will be the onerous burden and tremendous time it takes to iteratively code through pages after pages of texts, whether it be from interview transcripts, observation notes, online forum discussions, or even actual exchanges of text messages (see [4]).

This paper introduces a large language model (LLM)-based qualitative analysis tool named Collaborative Artificial Intelligence for Learning and Analysis (CAILA) meant to support and alleviate the time burden in qualitative analysis. Additionally, we aim to explore how analyzing with tools such as CAILA can differ with traditional (humans-only) methods of qualitative analysis.

---

\*Joint Proceedings of LAK 2025 Workshops, co-located with the 15th International Conference on Learning Analytics and Knowledge (LAK 2025), Dublin, Ireland, March 03–07, 2025.

\*Corresponding author.

†Please note that A Farid is now at the University of California, Merced, and M. Fenech is now at Gardner-Webb University; they were both at MIT when the work presented in this paper was conducted.

✉ gcl@mit.edu (G.C. Lin); eanderso@mit.edu (E. Anderson); carumeys@mit.edu (C. Stevens); bhanks@mit.edu (B. Hanks); disha31@mit.edu (D. Chauhan); mfarid@ucmerced.edu (A. Farid); mfenech@gardner-webb.edu (M. Fenech); klopfer@mit.edu (E. Klopfer)

ORCID 0000-0001-7552-2880 (G.C. Lin); 0000-0002-6561-9977 (E. Anderson); 0009-0003-4514-362X (B. Hanks); 0009-0007-3066-2100 (D. Chauhan); 0009-0005-4712-200X (A. Farid); 0000-0002-8719-7883 (M. Fenech); 0000-0002-6778-9478 (E. Klopfer)



Copyright © 2025 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

## 1.1. Positionality statements

As the researchers are the tools through which data is analyzed in qualitative research, it is imperative to know the researchers' stance. Here, we present the positionality statements of the first two authors GL and EA, who led the human+CAILA and the traditional humans-only approaches, respectively.

GL conducts mixed methods research and thinks very deeply about methods and methodology to the point that some colleagues consider her a methodologist. While she does not typically enjoy labels placed upon her, she does see their utility in demonstrating her approach to research. For example, while she conducts mostly qualitative research (or mixed methods with more emphasis on the qualitative nature of the work) at MIT, the position she holds at Harvard University is one of Lecturer in *Quantitative* Psychology. She is comfortable transforming qualitative data into quantitative ones for further analysis but at the same time realizes the limitations with such approaches. At times because of the divide she sees between the two worlds she walks, she tries to bridge the gap. In this work, she is the main researcher using the LLM tool in order to explore its potentials.

EA is also a mixed methods researcher. However, unlike GL, EA has a much deeper leaning toward the qualitative side with her undergraduate in Anthropology forming a foundation on how she approaches exploring human interactions. Over the last seven years she has primarily worked on qualitative research projects from deep analysis of classroom observational data, to interaction analysis of video data, to interview studies all in an attempt to understand how and in what ways learning is taking place. EA feels that in digging into what individuals are saying and the actions they are taking we can better understand how learning is taking place and how to better support learning.

## 2. Related Work

Using the newest technology and technique to analyze text is nothing new. Natural language processing (NLP) was born over half a century ago in an effort to automatically translate languages [5], [6]. Since its development in the late 1940s, researchers have continued to develop and use associated techniques for text analysis. For example, topic modeling uses text mining and unsupervised learning to extract key terms and topics represented in the document (see [7]). A number of algorithms were developed for topic modeling, such as latent Dirichlet allocation (LDA; [8]) or latent semantic analysis [9], [10]. These techniques have sped up the ability for researchers to process texts and apply statistical modeling to text-based data (e.g., [11], [12]).

### 2.1. AI for qualitative analysis

Though the use of emerging technology for analysis has been around for decades, the use of generative AI for qualitative analysis was still “in the state of being born” [6, p. 999]. In this very nascent stage, a number of research teams have started exploring the technique and the process in deductive coding (e.g., [14]) to thematic analysis [15], to even the development of computational grounded theory framework [12], [16].

#### 2.1.1. Methodological and ethical considerations

Computational Grounded Theory (CGT) is one of the integral human computer interaction (HCI) theories for Generative AI use in qualitative analysis. Specifically, CGT leverages the unique contributions of both the human analyst and computer through an iterative three-step process. It first conducts pattern detection through NLP and machine learning algorithms. This is followed by “pattern refinement,” in which human analysts interpret patterns that AI detected. The final step, “pattern confirmation,” aims to ensure the patterns detected and interpreted are applicable throughout the entire dataset [16]. Tschisgale and colleagues (2023) applied CGT in an (physics) educational research setting and found that CGT promotes efficiency by enabling researchers to

analyze large amounts of unstructured qualitative data at a faster pace. Moreover, it increases the rigor of qualitative research as the findings are more easily reproducible because it is encoded in the trained ML model [12].

These findings shed light on “mutual learning” frameworks which consider Generative AI as both a tool and partner to human researchers in the qualitative analysis process for data synthesis and codebook creation [17], [18]. Specifically, Barany and colleagues (2024) demonstrated this in an experimental study with multiple conditions including coding with human analysts only and ChatGPT only as well as collaboration between the two [17]. This study revealed that the hybrid conditions in which the computer and human coded collaboratively (in codebook development or refinement) resulted in the highest utility ratings, conceptual overlaps, and inter-rater reliability. This research highlights the need for human participation within the analyses process as the ChatGPT-only condition, or fully automated approach, was an outlier in relation to human analyst and AI partnered approaches resulting in errors, inconsistencies, and missed themes. The caution against overreliance on AI-powered analysis methods was echoed by researchers who developed LLoM, a concept induction algorithm that leverages LLMs to drive meaningful concepts from unstructured text [19]. Lam and colleagues warn that heavy reliance on LLoM outputs can result in gaps and misses in the concepts generated.

Both Lam’s and Barany’s findings connect to research by Christou (2023) that warns against the overdependence on Generative AI in analyses [20]. While Generative AI has clear benefits for qualitative research analyses, it is important to consider and mitigate the biases and limitations LLMs and other algorithms and models are well-known to have [21], [22]. Christou has conducted research to begin to address the gap in critical perspectives and practical methodological guidelines regarding AI use in qualitative research analyses. In these guidelines, Christou emphasizes the importance of: (1) familiarity with one’s dataset to be able to identify biases, (2) transparency around AI usage and its limitations in analyses, and (3) cross-referencing in an effort to ensure accuracy and validate AI-generated insights through triangulation. These suggestions of best practices keep humans in the loop to mitigate some of the potential biases that LLMs may produce as partners or tools in qualitative analyses.

### **2.1.2. Supportive tools**

Along with the frameworks, guidelines, and recommendations, developments of tools and software are keeping pace. Commercially available qualitative analysis software that have partnered with OpenAI (e.g., Atlas.ti and MAXQDA) have touted the integration of AI to help with the analytic process and reduce time [23]. However, when our research team tried to use these tools, we found that they were lacking in flexibility and transparency. In terms of flexibility, we could not rerun the AI support or easily iterate on the codes it found. In terms of transparency, it was difficult to determine which parts of the data the AI was using to support the code it identified. Wanting greater insight and control over the processes, we concluded these off-the-shelf tools were not allowing us to do qualitative research in the way we felt honored our methodology.

Innovative tools have also come out of the research community. CoAI Coder and CollabCoder, for example, place the AI as a human collaborator in the process of qualitative coding [24], [25], [26]. CoAI Coder used classic NLP (e.g., SpacyNLP) models and the Dual Intent Entity Transformer (DIET) classifier [27]. The more recent CollabCoder integrated LLMs, specifically the GPT-3.5 model. The purpose of the tool is to enable researchers to develop codes with AI-generated code suggestions and more quickly resolve any disagreements [24], [25], [26]. LLoM, in contrast, is focused on concept induction. While very similar to thematic analysis [15], Lam and colleagues (2024) situated LLoM as a more advanced tool for extracting high-level concepts, comparing results from the tool to those from BERTopic [28], [29] as well as large language models alone [19].

The use of LLMs inevitably requires feeding preprompts to the large language model, and the integration of AI in the qualitative analysis process requires users to know how to write the preprompts and trust the AI in the process [18]. Therefore, the evaluation of the preprompt and the responses that result from the preprompts is essential. ChainForge enables the comparison of

multiple LLMs in how they make sense of the data [30], [31]. In fact, researchers have combined ChainForge’s capability with classic NLPs (e.g., term frequency-inverse document frequency [32] as well as a novel Positional Diction Clustering (PDC) algorithm in order to make sense of text data [33].

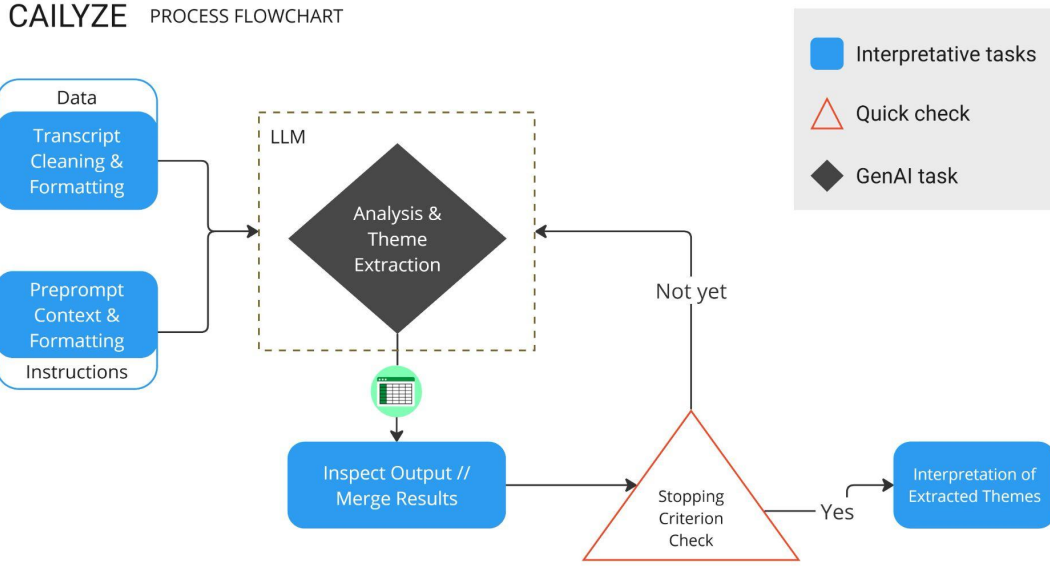
As using LLMs for qualitative analysis is still in its infancy, no technique has been established as the go-to. For the most part, the studies are concerned with speeding up the process of analysis, using LLMs, for example, to extract the codes before, for example, going into the next stage of deriving the themes (e.g., [13], [25], [26]). In general, there is still the tendency to stick with the previously established protocols in manual qualitative codebook development. This approach may be due to researchers’ understandable concerns in adhering to guidelines that ensures researchers’ analytic control and cognitive input [20] as well as following well-defined steps for analysis (e.g., [15]) in an otherwise subjective and fluid research methodology.

We present a slightly altered analytic order through CAILA, where human coders are still in the loop throughout the process but the themes are extracted without actively engaging in the first cycle of initial coding [1]. In essence, we skipped the development of the codebook with the LLMs, but aimed to see if we could still arrive at themes similar to those derived from the traditional manual coding process. Furthermore, instead of taking for granted that LLMs are acceptable tools for qualitative analysis and comparing various models (as ChainForge does), we take a step back to ask: **“How do the process and outcomes of thematic analysis using a generative AI-powered tool differ from those of traditional (human) analysis?”**

### 3. Collaborative AI for Learning and Analysis (CAILA) and its Process

Collaborative AI for Learning and Analysis (CAILA) is an LLM-powered system tool meant to support qualitative analysis. Its current capabilities are limited to inductive approaches such as thematic analysis. (See Appendix A for the current user interface and descriptions of how to use the tool.) We use the term “CAILyze” as a verb to indicate the process of using **CAILA** to **analyze** data. In other words, CAILyze is our approach to use LLMs to support the analysis process. Similar to Zhang et al.’s [18] framework, the process first involves cleaning the conversation into transcript files. The file is fed into the system that is connected to a LLM such as ChatGPT. On the backend, the system is primed with a number of preprompts. One preprompt sets the input and output expectations (e.g., “I’m going to give you a set of data from student group discussions” and “I want you to generate themes that would answer the questions I pose. Give the output in a spreadsheet format, containing the theme, the description and explanation of each theme”). The other allows the user to input their question of interest. Once the preprompts are set and the data is entered into the system, the user can run the program to “CAILyze” the data. The result of the CAILyze process should then be inspected by the user.

In contrast to the focus on editing and modifying the preprompt based on the output right away, we recognize that LLM outputs can be ephemeral. Even if one asks the same question, the response given by the LLM will differ each time. Therefore, we encourage the multiple iterations approach as demonstrated by Barany et al. [17]. However, instead of a fixed 13 iterations, the CAILyze process uses a stopping criterion common in many neuropsychological measurements. The user will start with one iteration, check the themes and their description, explanations, and examples. They will then run the program again. This time, they will check whether any generated themes are repeated or new. They will continue this process until they reach three consecutive iterations where no new theme emerges. The stopping criterion allows more flexibility as longer texts may result in more theme variations across the iterations than shorter texts. Additionally, this stopping criterion also serves to ensure that thematic saturation [34], [35], [36], [37] has been reached, such that the researcher can be more reasonably certain that no other codes or themes would emerge. Figure 1 illustrates the CAILyze flow.



**Figure 1:** The CAILYze Process Flowchart. The user inputs data and instructions into the system. The LLM processes the information and provides the output. The user inspects the output and merge results if any theme is repeated. They check against the stopping criterion. They iterate if the criterion is not yet reached. When the stopping criterion is reached (i.e., three consecutive iterations with no new themes), the user can perform a comprehensive interpretation of the extracted themes.

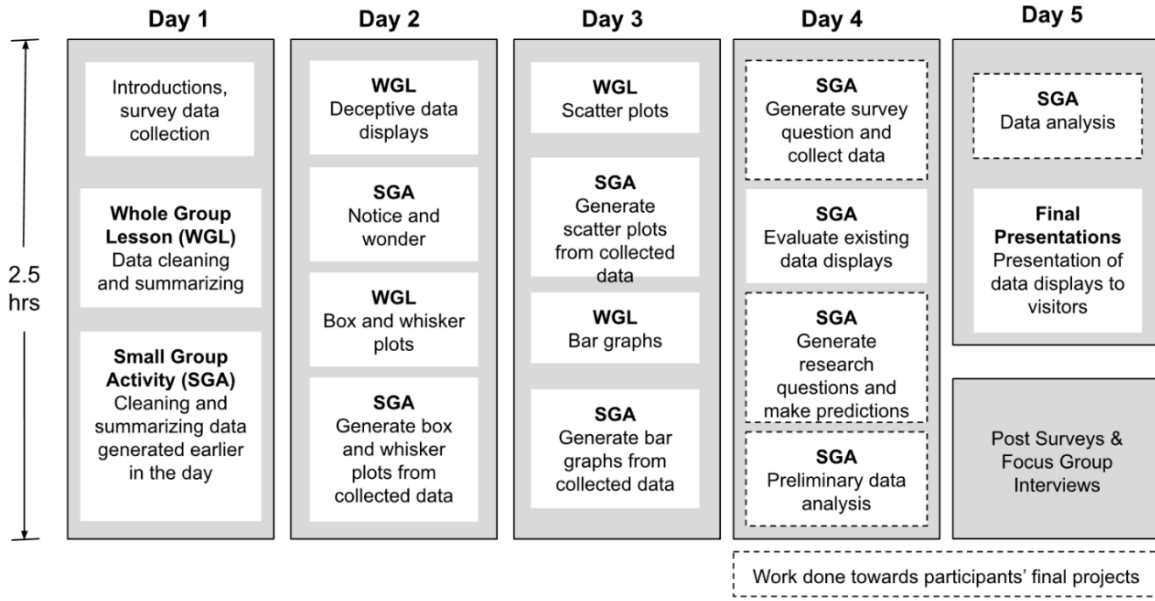
## 4. Method

We demonstrate the CAILYze process through a case study of texts from a data science workshop with high school students. The study has been approved by the authors’ institutional review board, and parental consent and student assent were obtained prior to the start of the study. All names used in this paper are pseudonyms.

### 4.1. Context

The data science workshop was held virtually over a week in February 2024 with nine high school students, five of whom attended a project-based learning charter school in the American South and the remaining four went to two urban public schools in a Northeastern city. Eight identified as female and one as male. The group included five 10th graders, two 11th graders, and two 12th graders. Four identified as African American or Black, four as White or Caucasian, and one as Asian. In a pre-survey, only one participant reported knowing how to engage in data science. The group included a mix of students from on-level and honors math classes, with three reporting some level of math anxiety. Five students typically received mostly A's in math, while the rest reported a mix of A's and B's.

In the five-day workshop, students worked in groups to examine data, organize and display information using Excel and Google sheets (e.g., creating graphs and pivot tables), generate their own research questions, and present their own findings. See Figure 2 for the workshop plan. The scenario with which we investigated the CAILYze process occurred on Day 4 of the workshop, when the students evaluated existing data displays from Our World in Data (<https://ourworldindata.org/>; [38]) by explaining what the display is showing and whether the information were accurately represented.



**Figure 2:** Workshop Schedule.

#### 4.2. Data source

The data source was the transcripts of the three student group conversations. With three groups combined, there were 146 lines of conversation utterances and 2,231 words. As we were interested in how and if the students have developed a sense of criticality when it comes to judging data displays, the preprompt we entered into the system was, “**generate themes about how students are critical of the graphs they were investigating.**”

#### 4.3. Analytic approach

While we understand that qualitative research results may not be perfectly comparable from one analysis to another, as researchers come in with their own lived experience that can shape their analytic lens (see [39]), we believe that illustrating CAILyze with a more traditional manual coding may lend support to this technique and offer additional insight beyond what other researchers have already shown (e.g., [17]). Therefore, as means of comparison, one researcher (GL) analyzed the data using the CAILyze process, and two other researchers (EA and AF) manually coded the data to answer the same question. All three researchers hold doctorate degrees in education and are experienced qualitative researchers. The positionality of the two leading researchers (GL and EA) can be found in the beginning of this article.

The researcher employing CAILA approached the analysis much like quantitative researchers handling large secondary datasets. She began by familiarizing herself with the data through an examination of overall speaker utterances—such as word counts per speaker during the activity—and network graphs. Notably, during dataset preparation, she also skimmed through the transcripts, a step that some quantitative researchers might skip. Following this initial phase, she applied the CAILA system and the CAILyze process to extract the themes, using GPT-4o—configured at its default temperature—as the LLM model.

In contrast, EA and AF were involved in data collection and thus possessed an inherent familiarity with the data from the outset. In this humans-only approach, they used process coding in their initial cycle and pattern coding [40] in the second cycle to identify the themes. They resolved any conflicts through social moderation [41].

## 5. Results

The CAILyze process resulted in 10 iterations. The human researcher labeled each theme after the first iteration as a new or repeated theme and continued the process. A total of 50 themes were generated. See Appendix B for the full table of each iteration's generated themes and descriptions/explanations. After reaching the stopping criterion, she organized the final 50 themes into 4 major themes and 6 subthemes (see Table 1). All generated explanations and example quotes were checked against the raw transcripts to ensure the LLM's accuracy; the researcher was able to verify that the LLM did not hallucinate any of the examples. The process took less than two working days.

On the other hand, the two other researchers went through two cycles of coding, and the human-only manual process resulted in themes beyond the way students were critical of the data. Instead, the final emerged themes captured the progression through which students may demonstrate criticality. Specifically, five themes emerged through this process with one theme, evaluation, consisting of four sub-themes (see Table 1). The process took the two researchers over two weeks to complete.

**Table 1**

Themes Generated through the Two Processes

Human + CAILA	Humans-Only
Demonstrated Criticality	Progression of Criticality Development
<ol style="list-style-type: none"> <li>Accurate/Misleading Representations of Data Elements               <ol style="list-style-type: none"> <li>Scale of variables</li> <li>Scale of the graph (e.g., starting point)</li> <li>Visual elements of the graph (e.g., size, shape)</li> </ol> </li> <li>Simplicity/Complexity for Comprehension</li> <li>Critical Evaluation and Interpretation with Contextual Information               <ol style="list-style-type: none"> <li>Data source</li> <li>Social historical context</li> <li>Time (e.g., outdated information, trend over time)</li> </ol> </li> <li>Graphical Comparison (which type of graph is better suited for which purpose)</li> </ol>	<ol style="list-style-type: none"> <li>Describe data displays</li> <li>Ask questions about the displays</li> <li>Interpret the data display</li> <li>Evaluate               <ol style="list-style-type: none"> <li>General accuracy</li> <li>Gaps in data</li> <li>Data representation</li> <li>Alternative displays / Graphical comparisons</li> </ol> </li> <li>Meta-Discussions</li> </ol>

## 6. Discussion

In this section, we summarize and discuss our findings by incorporating illustrative student quotes (with all names being pseudonyms). We then situate our results within the broader qualitative analysis literature to highlight and explain the key differences between the two approaches.

### 6.1. Human with CAILA vs. Humans-only

While the CAILyze version only identified the themes that captured the ways students were critical of data, inspection and discussion of the humans-only version revealed that some of the "codes" contributing to the themes were conditions that are necessary but insufficient for criticality. That is, you need to satisfy the particular condition in order to reach critical thinking, but it in itself does not mean the student is being critical of the data. The perfect example is "describing data displays." Being able to describe the data displays is a necessary precursor to being critical of it. Lea's statement below illustrates a thorough description of a data display:

“So the data set that I chose, um, like I said, talks about pandemics like over the years, um, and it uses circles to represent the death toll because, like, they’re looking specifically at the number of people who died for each of these pandemics.”

In it, she identified the purpose of the display (pandemic over the years) and the visual elements used (e.g., circles). Without acquiring the ability to tell different elements of the data visualization apart, students will not be able to critically assess its soundness. The next phase of criticality is then to use the data they described and question their accuracy. For example, Mary demonstrated both the precursor “describing data displays” and critical thinking about data by “evaluating the data displays” when she commented,

“You could say one of the ways that it accurately portrays data is because it starts from zero and increases. While some other graphs might start from like a random number that isn’t zero.”

If we follow this line of reasoning, we can see that in the humans-only coding process, the researchers have shifted to answering a question about “*what are the conditions that are necessary for critical thinking around data*” and *how students are approaching the data displays* rather than the original question around how students were critical with data. This development demonstrated the flexibility in human thinking and coding, and the reflective process in qualitative research practice as research questions are refined and developed [42]. The refinement of research questions is not only acceptable, it is integral in qualitative research to ensure the question driving the study is in line with researchers’ increased understanding of the phenomenon they are investigating [43], [44]. In our case, this shift occurs because the human coders hold broader big picture objectives that are not privy to the AI system. That is, the researchers understand that eventually CAILA is meant to serve the ultimate purpose of automating analyses and displaying information that is helpful for teachers as part of a formative assessment of group discussions. Teachers may also be interested in the progression and development of the criticality that emerged in the data beyond merely the way that the students are critical of the data. The flexibility allowed the human researchers to expand the RQ to encompass the information and data (e.g., students are capable of describing the data displays in detail) that the CAILyze process treated as a given and neglected because it does not directly answer the question about how students were critical.

Despite the misalignment of research questions by the two approaches, the themes from the humans-only approach that are directly related to the criticality surrounding data inspection were similar to the themes derived with the CAILyze process. The “Evaluate” stage of criticality progression contained four sub-themes, which were similar to the four main themes focusing only on the criticality aspects of students’ data inspection. Both humans-only and CAILyzed results, for example, captured students’ evaluation of data accuracy, (mis)representation, and alternative graph types.

In sum, the CAILyze process was by far faster, which aligns with previous research results of leveraging LLMs for qualitative research [12], [17], [45], and the derived themes were directly reflective of the original question posed. In contrast, as human researchers coded through the raw data, the progressions of skills needed to eventually develop criticality emerged as the more important question. With only 10 iterations using the stopping criteria, the similarities in the “criticality” themes lends credence to the CAILyze approach.

## 6.2. Implications and recommendations

Our findings also pointed out the affordances and challenges of using LLMs to assist in qualitative coding. The model will only take in the questions asked and will not adapt the research question as it iterates. While incapable of the research question refining process recommended for qualitative research methodologies such as grounded theory [43], the inflexibility in changing the research question may align more with researchers with more positivist or post-positivist epistemology [46],



[47], [48]. For qualitative researchers who may be concerned with the inflexibility of the system and rigid RQ, we suggest that the researchers must have a deeper level of familiarity with the data (e.g., they are involved in the data collection or spent time living through the data with deep reading) such that the question posed to the LLM is already the refined one. Furthermore, they can also repeat the CAIlyze process multiple times with other modified, refined preprompts.

## 7. Conclusion

In this paper, we introduced the CAIlyze process and added a different approach—that with a stopping criterion to accommodate the ephemeral nature of LLMs, adapt to varying lengths of conversational transcripts, and ensure thematic saturation—to the nascent field of using AI for qualitative analysis. We demonstrated that the approach produced similar results as manual coding and illustrated that the key difference lies in the human flexibility in adjusting and refining the research question as they analyzed the textual data. This difference illustrates the circumstances under which the CAIlyze process may be suitable. Finally, we ended with suggestions for researchers to engage in qualitative analysis using generative AI.

## Acknowledgements

This work was funded by the Emerson Collective. We would like to thank Beatriz Familia Azevedo in reading earlier drafts of this paper and all of the high schoolers who participated in the data science workshop. Thank you also to Mary McCrossan for helping us organize the project page on the Scheller Teacher Education Program | The Education Arcade's website.

## Declaration on Generative AI

The work in this presentation uses OpenAI's GPT-4o as the model beneath the qualitative analysis. That is, generative AI was used in the human + AI qualitative analysis approach presented in this paper. In the preparation of this paper, we also used the model to ensure grammatical accuracy, spell check, and language clarity. Afterwards, we reviewed and edited the content as needed and take full responsibility for the publication's content.

## References

- [1] J. Saldaña, *The Coding Manual for Qualitative Researchers*, 4th ed. SAGE Publications, 2021.
- [2] S. Wa-Mbaleka, "The Researcher as an Instrument," in *Computer Supported Qualitative Research*, A. P. Costa, L. P. Reis, and A. Moreira, Eds., Cham: Springer International Publishing, 2020, pp. 33–41. doi: 10.1007/978-3-030-31787-4\_3.
- [3] M. A. Xu and G. B. Storr, "Learning the Concept of Researcher as Instrument in Qualitative Research," *TQR*, vol. 17, no. 42, pp. 1–18, 2012, doi: 10.46743/2160-3715/2012.1768.
- [4] M. S. Rahman, "The Advantages and Disadvantages of Using Qualitative and Quantitative Approaches and Methods in Language 'Testing and Assessment' Research: A Literature Review," *JEL*, vol. 6, no. 1, p. 102, 2017, doi: 10.5539/jel.v6n1p102.
- [5] P. Johri, S. K. Khatri, A. T. Al-Taani, M. Sabharwal, S. Suvanov, and A. Kumar, "Natural Language Processing: History, Evolution, Application, and Future Work," in *Proceedings of 3rd International Conference on Computing Informatics and Networks*, A. Abraham, O. Castillo, and D. Virmani, Eds., Singapore: Springer, 2021, pp. 365–375. doi: 10.1007/978-981-15-9712-1\_31.
- [6] K. S. Jones, "Natural Language Processing: A Historical Review," in *Current Issues in Computational Linguistics: In Honour of Don Walker*, A. Zampolli, N. Calzolari, and M. Palmer, Eds., Dordrecht: Springer Netherlands, 1994, pp. 3–16. doi: 10.1007/978-0-585-35958-8\_1.
- [7] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Information Systems*, vol. 94, p. 101582, Dec. 2020, doi: 10.1016/j.is.2020.101582.
- [8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. null, pp. 993–1022, Mar. 2003.

- [9] S. T. Dumais, "Latent Semantic Analysis," *Annual Review of Information Science and Technology (ARIST)*, vol. 38, pp. 189–230, 2004.
- [10] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Processes*, vol. 25, no. 2–3, pp. 259–284, Jan. 1998, doi: 10.1080/01638539809545028.
- [11] C. Hacking, H. Verbeek, J. P. H. Hamers, and S. Aarts, "Comparing text mining and manual coding methods: Analysing interview data on quality of care in long-term care for older adults," *PLoS One*, vol. 18, no. 11, p. e0292578, Nov. 2023, doi: 10.1371/journal.pone.0292578.
- [12] P. Tschisgale, P. Wulff, and M. Kubsch, "Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory," *Phys. Rev. Phys. Educ. Res.*, vol. 19, no. 2, p. 020123, Sep. 2023, doi: 10.1103/PhysRevPhysEducRes.19.020123.
- [13] S. De Paoli, "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*, vol. 42, no. 4, pp. 997–1019, Aug. 2024, doi: 10.1177/08944393231220483.
- [14] A. F. Zambrano, X. Liu, A. Barany, R. S. Baker, J. Kim, and N. Nasiar, "From nCoder to ChatGPT: From Automated Coding to Refining Human Coding," in *Advances in Quantitative Ethnography*, G. Arastoopour Irgens and S. Knight, Eds., Cham: Springer Nature Switzerland, 2023, pp. 470–485. doi: 10.1007/978-3-031-47014-1\_32.
- [15] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, Jan. 2006, doi: 10.1191/1478088706qp063oa.
- [16] L. K. Nelson, "Computational Grounded Theory: A Methodological Framework," *Sociological Methods & Research*, vol. 49, no. 1, pp. 3–42, Feb. 2020, doi: 10.1177/0049124117729703.
- [17] A. Barany *et al.*, "ChatGPT for Education Research: Exploring the Potential of Large Language Models for Qualitative Codebook Development," in *Artificial Intelligence in Education*, A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt, Eds., Cham: Springer Nature Switzerland, 2024, pp. 134–149. doi: 10.1007/978-3-031-64299-9\_10.
- [18] H. Zhang, C. Wu, J. Xie, Y. Lyu, J. Cai, and J. M. Carroll, "Redefining Qualitative Analysis in the AI Era: Utilizing ChatGPT for Efficient Thematic Analysis," Sep. 19, 2023. Accessed: Sep. 07, 2024. [Online]. Available: <https://arxiv.org/abs/2309.10771v3>
- [19] M. S. Lam, J. Teoh, J. A. Landay, J. Heer, and M. S. Bernstein, "Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLoom," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, in CHI '24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–28. doi: 10.1145/3613904.3642830.
- [20] P. Christou, "How to Use Artificial Intelligence (AI) as a Resource, Methodological and Analysis Tool in Qualitative Research?," *TQR*, vol. 28, no. 7, pp. 1968–1980, Jul. 2023, doi: 10.46743/2160-3715/2023.6406.
- [21] A. Acerbi and J. M. Stubbersfield, "Large language models show human-like content biases in transmission chain experiments," *Proceedings of the National Academy of Sciences*, vol. 120, no. 44, p. e2313790120, Oct. 2023, doi: 10.1073/pnas.2313790120.
- [22] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Reprint edition. Crown, 2017. Accessed: Sep. 12, 2024. [Online]. Available: <https://www.penguinrandomhouse.com/books/241363/weapons-of-math-destruction-by-cathy-oneil/>
- [23] "MAXQDA vs. ATLAS.ti | Best Qualitative Data Analysis Software," ATLAS.ti. Accessed: Sep. 09, 2024. [Online]. Available: <https://atlasti.com/maxqda-vs-atlasti-comparison>
- [24] J. Gao, K. T. W. Choo, J. Cao, R. K.-W. Lee, and S. Perrault, "CoAICoder: Examining the Effectiveness of AI-assisted Human-to-Human Collaboration in Qualitative Analysis," *ACM Trans. Comput.-Hum. Interact.*, vol. 31, no. 1, p. 6:1–6:38, Nov. 2023, doi: 10.1145/3617362.
- [25] J. Gao, Y. Guo, T. J.-J. Li, and S. T. Perrault, "CollabCoder: A GPT-Powered Workflow for Collaborative Qualitative Analysis," in *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, in CSCW '23 Companion. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 354–357. doi: 10.1145/3584931.3607500.
- [26] J. Gao *et al.*, "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*, Honolulu HI USA: ACM, May 2024, pp. 1–29. doi: 10.1145/3613904.3642002.
- [27] T. Bunk, D. Varshneya, V. Vlasov, and A. Nichol, “DIET: Lightweight Language Understanding for Dialogue Systems,” May 11, 2020, *arXiv*: arXiv:2004.09936. doi: 10.48550/arXiv.2004.09936.
  - [28] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” Mar. 11, 2022, *arXiv*: arXiv:2203.05794. doi: 10.48550/arXiv.2203.05794.
  - [29] M. Grootendorst *et al.*, *MaartenGr/BERTopic: v0.16.3*. (Jul. 22, 2024). Zenodo. doi: 10.5281/zenodo.12793147.
  - [30] I. Arawjo, C. Swoopes, P. Vaithilingam, M. Wattenberg, and E. L. Glassman, “ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing,” in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, in CHI ’24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–18. doi: 10.1145/3613904.3642016.
  - [31] I. Arawjo, P. Vaithilingam, M. Wattenberg, and E. Glassman, “ChainForge: An open-source visual programming environment for prompt engineering,” in *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, in UIST ’23 Adjunct. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 1–3. doi: 10.1145/3586182.3616660.
  - [32] D. Jurafsky and J. H. Martin, “Question Answering, Information Retrieval, and Retrieval-Augmented Generation,” in *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed., 2024. [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/14.pdf>
  - [33] K. I. Gero, C. Swoopes, Z. Gu, J. K. Kummerfeld, and E. L. Glassman, “Supporting Sensemaking of Large Language Model Outputs at Scale,” in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, in CHI ’24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–21. doi: 10.1145/3613904.3642139.
  - [34] M. Birks and J. Mills, *Grounded Theory: A Practical Guide*. SAGE, 2015.
  - [35] S. Rahimi and M. khatooni, “Saturation in qualitative research: An evolutionary concept analysis,” *International Journal of Nursing Studies Advances*, vol. 6, p. 100174, Jun. 2024, doi: 10.1016/j.ijnsa.2024.100174.
  - [36] B. Saunders *et al.*, “Saturation in qualitative research: exploring its conceptualization and operationalization,” *Qual Quant*, vol. 52, no. 4, pp. 1893–1907, Jul. 2018, doi: 10.1007/s11135-017-0574-8.
  - [37] C. Urquhart, *Grounded Theory for Qualitative Research: A Practical Guide*. SAGE Publications, Ltd, 2013. doi: 10.4135/9781526402196.
  - [38] M. Roser, “OWID Homepage,” Our World in Data. Accessed: Sep. 09, 2024. [Online]. Available: <https://ourworldindata.org>
  - [39] M. Pownall, “Is replication possible in qualitative research? A response to Makel et al. (2022),” *Educational Research and Evaluation*, vol. 29, no. 1–2, pp. 104–110, Feb. 2024, doi: 10.1080/13803611.2024.2314526.
  - [40] M. B. Miles and A. M. Huberman, *Qualitative Data Analysis*, 2nd ed. Thousand Oaks, CA: SAGE Publications, 1994.
  - [41] D. W. Shaffer, *Quantitative Ethnography*. Madison, WI: Cathcart Press, 2017.
  - [42] J. Agee, “Developing qualitative research questions: a reflective process,” *International Journal of Qualitative Studies in Education*, Jul. 2009, doi: 10.1080/09518390902736512.
  - [43] K. Charmaz, *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. SAGE, 2006.
  - [44] J. W. Creswell and C. N. Poth, *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. SAGE Publications, 2016.
  - [45] D. L. Morgan, “Exploring the Use of Artificial Intelligence for Qualitative Data Analysis: The Case of ChatGPT,” *International Journal of Qualitative Methods*, vol. 22, p. 16094069231211248, Jan. 2023, doi: 10.1177/16094069231211248.
  - [46] J. Armstrong, “Naturalistic Inquiry,” *Encyclopedia of research design*, vol. 2. SAGE Publications, Thousand Oaks, CA, pp. 880–885, 2010.
  - [47] J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 6th ed. SAGE Publications, 2022. Accessed: Sep. 09, 2024. [Online]. Available: <https://collegepublishing.sagepub.com/products/research-design-6-270550>

- [48] D. C. Phillips and N. C. Burbules, *Postpositivism and educational research*. in Postpositivism and educational research. Lanham, MD, US: Rowman & Littlefield, 2000, pp. ix, 101.

## A. Current CAILA User Interface

When the analytic work detailed in this paper was conducted, CAILA was accessible only through a Jupyter notebook interface. The team (mostly BH and DC) have subsequently made CAILA even more accessible. This appendix shows the current interface of the system, which is open source and accessible through github. The link to the github page can also be found on our project page: <https://education.mit.edu/project/collaborative-ai-for-learning-cail/>.

### Analyze New Data

OpenAI API Key:

Enter your OpenAI API key

Upload Data File

Preferred file format: CSV

Research Question:

How are the students being critical of the graphical displays?

#### Preset Information

Input Expectations:

I'm going to give you a set of data from student group discussions

Output Expectations:

I want you to generate themes that would answer the questions I pose. Provide the output in CSV format with the following columns: Theme, Description, Explanation. Do not include any additional text or explanations.

Select Additional Output Columns:

The user click on the Upload Data File section and a pop-up window will show up allowing the user to select their data file.

The user can also change the research question. The preset information is prefilled with suggestions such as output format, but the user can also adjust the input and output expectations.

One the file is uploaded, the upload data file section will turn green and change to “File Uploaded.”

### Analyze New Data

OpenAI API Key:

Enter your OpenAI API key

File Uploaded

Preferred file format: CSV

Research Question:

Enter your research question here

#### Preset Information

Input Expectations:

I'm going to give you a set of data from student group discussions

Output Expectations:

I want you to generate themes that would answer the questions I pose. Provide the output in CSV format with the following columns: Theme, Description, Explanation. Do not include any additional text or explanations.

Select Additional Output Columns:

Click on the CAILyze button on the bottom to have CAILA analyze the data.

File Uploaded

Preferred file format: CSV

**Research Question:**

What are the themes related to student criticality on the data visualizations?

**Preset Information**

**Input Expectations:**

I'm going to give you a set of data from student group discussions

**Output Expectations:**

I want you to generate themes that would answer the questions I pose. Provide the output in CSV format with the following columns: Theme, Description, Explanation. Do not include any additional text or explanations.

**Select Additional Output Columns:**

☒ Relevant Quotes

☒ Speaker Names

☐ Keywords

CAILyze!

If the user forgets to include the API key, they will get a reminder to fill out the field:

**OpenAI API Key:**

Enter your OpenAI API key

Please fill out this field.

Preferred file format: CSV

**Research Question:**

What are the themes related to student criticality on the data visualizations?

We are in the process of enabling selection of different models, and the new interface with model selection looks something like this:

## Analyze New Data

**OpenAI API Key:**

Enter your OpenAI API key

**Choose GPT Model / AI:**

ChatGPT-4o Latest: Latest model used in ChatGPT with real-time updates

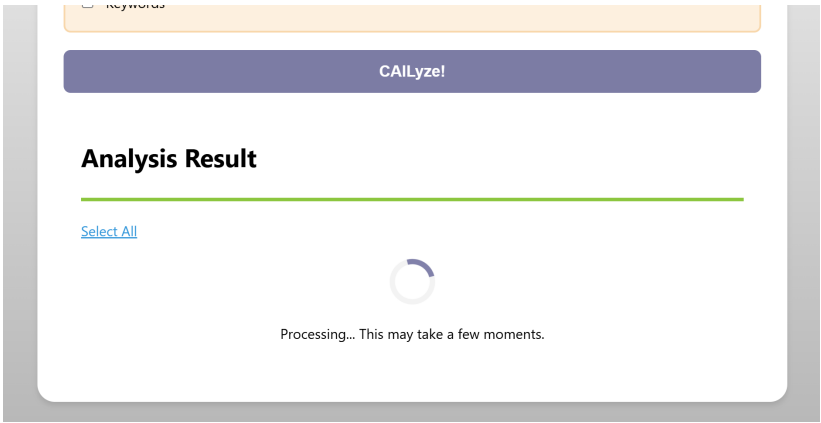
Upload Data File

Preferred file format: CSV

**Research Question:** ⓘ

E.g., What are common challenges in student group discussions?

After clicking on the "CAILyze!" button, a message will appear indicating that the system is processing:



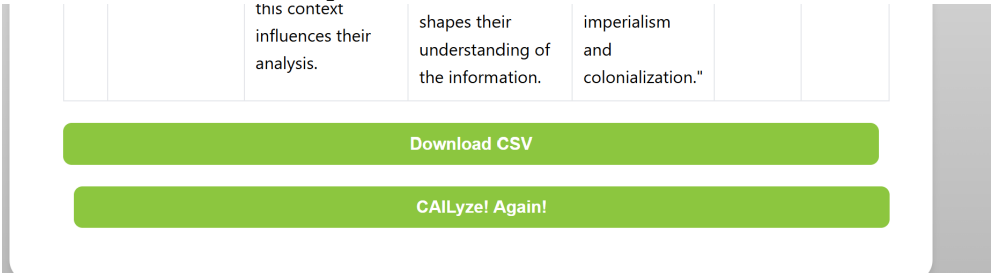
The results will appear as a table on the screen:

Analysis Result

[Select All](#)

	Theme	Description	Explanation	Relevant Quotes	Speaker Names	Actions
<input type="checkbox"/>	Evaluation of Data Accuracy	Students critically evaluate the accuracy of the data displays by considering factors such as representation, reliability, and potential biases.	This theme addresses how students assess the accuracy of the data displays they are investigating, focusing on aspects like representation and reliability.	"I want to say it's accurate. World population. Now that's that that makes. Yeah, I would say it's still pretty accurate because."	Teresa	<button>Delete</button>
		Students compare different data	This theme explores how students analyze			

The user can also download the output as a CSV file:



Based on the CAILA process, if the stopping criteria have not been met, click the “CAILyze! Again!” button to continue. If users notice repeated themes, they can select the checkboxes in the first column for the repeated themes, then click the “Merge Selected Rows” button above the table.

<input checked="" type="checkbox"/>	Critique of Data Representation	Students critique the representation of data in the displays, discussing aspects such as clarity, relevance, and potential misinterpretations.	This theme delves into how students evaluate the representation of data in the displays, considering factors like clarity, relevance, and the possibility of misinterpretations.	"And this is the literacy and this is the illiteracy. I don't. I think the one weird thing is the illiteracy is just, like, weirdly formatted how it's done, you know what I mean? But it works."	Greta	Delete
<input checked="" type="checkbox"/>	Critique of Graph Formatting	Students discuss the formatting of the graphs and how it impacts the interpretation of data.	This theme addresses how students are critical of the way graphs are formatted and how it affects the clarity and understanding of the data being presented.	"I think the one weird thing is the illiteracy is just, like, weirdly formatted how it's done, you know what I mean? But it works."	Greta	Delete

### Analysis Result

[Select All](#)

Merge Selected Rows

	Theme	Description	Explanation	Relevant Quotes	Speaker Names	Actions
--	-------	-------------	-------------	-----------------	---------------	---------

There is also an undo button available to correct mistakes:

### Analysis Result

[Select All](#)

Merge Selected Rows

Undo

	Theme	Description	Explanation	Relevant Quotes	Speaker Names	Actions
--	-------	-------------	-------------	-----------------	---------------	---------



Currently, merging simply combines the selected rows (see screenshot below). In future iterations, CAILA will synthesize the merged content and display the iteration number from which the content was drawn.

<input type="checkbox"/>	Critique of Data Representation / Critique of Graph Formatting	Students critique the representation of data in the displays, discussing aspects such as clarity, relevance, and potential misinterpretations. / Students discuss the formatting of the graphs and how it impacts the interpretation of data.	This theme delves into how students evaluate the representation of data in the displays, considering factors like clarity, relevance, and the possibility of misinterpretations. / This theme addresses how students are critical of the way graphs are formatted and how it affects the clarity and understanding of the data being presented.	"And this is the literacy and this is the illiteracy. I don't. I think the one weird thing is the illiteracy is just, like, weirdly formatted how it's done, you know what I mean? But it works." / "I think the one weird thing is the illiteracy is just, like, weirdly formatted how it's done, you know what I mean? But it works."	Greta / Greta	<button>Delete</button>
--------------------------	--	---	---	---	---------------	-------------------------

The user also has the option to delete a row if it does not accurately capture the phenomenon under investigation. Finally, the user may download the CSV file again after the merging process is complete.

Please note that all names in this paper, including both appendices, are pseudonyms.

## B. Themes Generated Through the CAILyze Process

Iteration	Themes	Description	Example
1	Analyzing Data Representation	Students analyze how data is visualized for clarity and accuracy, considering whether graphs start from zero to accurately portray trends.	Okay. Oh yeah. You could say one of the ways that it accurately portrays data is because it starts from zero and increases. - Mary
1	Simplicity and Comprehension	Students express a desire for the data to be presented in a simple and clear manner, emphasizing the need for graphs to be understandable like explaining games to a fifth grader.	Let's keep it simple and clear. Like explaining games to a fifth grader. - Oscar
1	Misleading Visuals	Students critique graphs for potentially misleading representations, focusing on how the size of visual elements can distort the perception of data.	But if you know, at least 23 million people died, then it comes off like it kind of misleads. - Lea
1	Accuracy and Misrepresentation	Students question the accuracy of data representation, discussing how the use of visuals like box plots and scatter plots can either clarify or mislead the understanding of data trends.	Yeah, visuals can be super misleading. Size should match the stats, or it's kind of like lying with pictures. - Oscar
1	Historical and Regional Context	Discussion points include whether graphs consider historical and regional contexts adequately, particularly in showing trends over time and differences between regions.	Yeah. Focusing on why parts of Africa are still in the orange zone would tell a much bigger story. - Oscar

2	Concerns about Misrepresentation or Oversimplification	Students expressed concern that visual representations such as graphs and charts might oversimplify or misrepresent data, potentially misleading viewers about the true nature or scale of the data.	Lea's critique about how the pandemic data visualization with circles could be misleading, as it does not adequately represent the scale of death tolls, especially when exact numbers are not known for all pandemics (#74, #73).
2	Importance of Starting from Zero	The starting point of a graph can influence its interpretation. Graphs that start from zero are seen as providing a more accurate and less misleading portrayal of data, reflecting its natural progression.	Mary's approval of graphs starting from zero as a method for accurately portraying increases in data, suggesting that starting from a different number might distort the data's progression (#18).
2	Evaluation of Data Visualization Size Relevance	The size used in data visualization should accurately reflect the data's magnitude or significance. Inaccuracies or misalignment here could skew a viewer's understanding of the data's importance.	Lea's and Oscar's discussion about the misleading nature of representing death tolls with circles, pointing out that the visual size should match the statistics to avoid misinterpretation (#74, #76).
2	Accuracy and Reliability of Represented Data	Students focused on whether the data represented in graphs and charts was accurate and reliable, considering whether it gives a true and fair view of the information it purports to represent.	Teresa's assessment of the accuracy of a population graph, considering technology's impact on literacy and world population numbers (#121, #119).
2	Critical Analysis of Historical and Geographic Context	Students analyzed graphs through a critical lens of historical and geographical context, evaluating whether these visualizations account for significant external factors such as imperialism or economic changes.	Lea and James's discussion about poverty trends in Eastern Europe and how different regions' economic statuses are depicted in the data, stressing the necessity to account for historical and geopolitical influences (#94, #97, #99).
3	Accuracy and Representation	Students were critically assessing if the data was accurate and represented correctly. They examined whether graphs start from zero to accurately portray increases, which helps in showing true scales and trends without misleading.	Okay. Oh yeah. You could say one of the ways that it accurately portrays data is because it starts from zero and increases. While some other graphs might start from like a random number that isn't zero.
3	Simplicity and Clarity	The conversation highlighted the importance of keeping data presentations simple and clear for easier understanding, much like explaining complex concepts to a younger audience, ensuring the information is accessible to all.	Let's keep it simple and clear. Like explaining games to a fifth grader.
3	Visual Misrepresentation	There was concern over how visuals might mislead the audience, specifically how the size of elements (like circles) used in the graphs could fail to match the statistics they're meant to represent, potentially distorting the perceived impact of the data.	Yeah, visuals can be super misleading. Size should match the stats, or it's kind of like lying with pictures.
3	Outdated Information	The students critiqued graphs for using outdated data, pointing out the importance of presenting the most current data available to make accurate and relevant conclusions.	2018 data trying to talk 2023 stuff. That's like using floppy disks for homework.
3	Overlook of Socio-economic Factors	There was a consideration of how graphs might overlook significant socio-economic contexts, such as the effects of imperialism and colonialism on poverty rates, suggesting a need for more nuanced presentations that account for underlying causes.	I would say I just think that is a little misleading because to me, like, I think it's very representative of what we classify as like a first world and a third world country.
4	Critical Analysis of Data Representation	Students critically analyze the accuracy and representation of data in graphs, questioning whether the data is presented in a misleading manner or if it accurately reflects the information.	Yeah, visuals can be super misleading. Size should match the stats, or it's kind of like lying with pictures.

4	Consideration of Data Starting Points	Some students focus on the importance of graphs starting from zero to accurately portray increases or changes in the data, noting that starting from a non-zero number could misrepresent the data.	Okay. Oh yeah. You could say one of the ways that it accurately portrays data is because it starts from zero and increases. While some other graphs might start from like a random number that isn't zero.
4	Evaluation of Graphical Trends	Participants evaluate how trends are shown in graphs, debating their effectiveness in displaying relationships between variables or time-based changes.	Scatterplot could show how cases change over time. Box plot good for comparing months maybe.
4	Understanding of Graph Purpose and Clarity	The students discuss whether the purpose of the graph is clear and if the graph successfully conveys its intended message in a simple and understandable manner.	Let's keep it simple and clear. Like explaining games to a fifth grader.
4	Discussion on Completeness and Misleading Elements	There's an awareness among the students about the presence of certain elements in the graphs that could mislead the audience or omit important information.	But if you know, at least 23 million people died, then it comes off like it kind of misleads. It comes off misleading only because, like, that's half of the amount of people that died in the Black Death.
4	Comparison and Contextual Analysis	Students compare graphs not only internally for consistency or accuracy but also contextually, considering whether they represent broader truths effectively and are current.	2018 data trying to talk 2023 stuff. That's like using floppy disks for homework.
5	Critical Analysis of Graph Accuracy	Students critically analyzed the accuracy of the graphs to ensure the data represented was accurate and not deceiving. For example, Mary and Alice found their graph to start from zero and cover data from at least 14 years ago, demonstrating increases throughout the year, which they considered accurately represented the data.	
5	Evaluation of Graph Clarity and Simplicity	Students emphasized the importance of clarity and simplicity in representing data. Oscar suggested keeping explanations simple and clear, akin to explaining games to a fifth grader, highlighting the need for easily understandable data representation.	
5	Graphs Representing Change Over Time	The students assessed how well graphs demonstrate changes over time. Lea discussed how her chosen dataset on pandemics visually represented the death tolls over the years but found it misleading when exact numbers were not available, indicating a need for accurate temporal representation.	
5	Critical Consideration of Graph Scales and Proportions	The students were critical about how graph scales and proportions could mislead or accurately depict the data. James and his group discussed how the colors represented different income levels and how the distribution changed over time, which required careful scrutiny to avoid misinterpretation.	
5	Assessing Relevance and Misleading Visuals	Students were attuned to the potential for graphs to mislead through their visuals. Lea critiqued her graph for potentially misleading viewers due to disproportionate circle sizes representing deaths from pandemics, showing a critical approach to the relevance and accuracy of visual aids.	

6	Accuracy and Representation	Students critically assessed how accurately and fairly the data was represented in the graphs, focusing on whether the graphical elements like starting points, size of elements, and overall design gave a true picture of the underlying data.	Mary highlighted the importance of graphs starting from zero for accuracy in representation, expressing that it more accurately portrays data compared to graphs that start from arbitrary non-zero values.
6	Simplicity vs. Complexity	There was a discussion about the balance between simplicity and complexity in data representation, with students weighing the need for graphs to be easily understandable while still capturing the full scope and nuances of the data.	Oscar suggested keeping the explanation simple, akin to explaining games to a fifth-grader, highlighting the need for clarity in data presentation to make it accessible to all viewers.
6	Misleading Visuals	Students were critical of visual elements that could potentially mislead the viewer about the data's true story, such as the size of elements not matching the scale of the data they represent or the selection of visual types like circles or colors.	Lea talked about how the representation of death tolls using circles might mislead viewers, especially if the size of the circles doesn't correspond accurately to the numbers they're supposed to represent.
6	Data Timeliness and Relevance	The students assessed the relevance and timeliness of the data being presented, understanding that outdated or non-current data could distort or diminish the utility of the information being conveyed.	James brought up concerns about using data from 2018 to talk about current situations, comparing it to using outdated technology like floppy disks - underlying the need for up-to-date information in making relevant analyses.
6	Underlying Socio-Economic Factors	A theme emerged around the critical analysis of socio-economic factors not being represented in the data, where students expressed concern over the graphs not showing the 'why' behind patterns or distributions, particularly in representing global issues.	Lea and Oscar discussed the importance of considering underlying factors such as imperialism and colonization in the analysis of poverty rates across different regions, implying that data without context could provide a misleading or incomplete narrative.
7	Theme: Analyzing Data Representation	Description: Students critically evaluate how well graphs display data, focusing on accuracy and ease of understanding. They discuss whether graphs start at zero to accurately represent growth and whether the representation correctly portrays increases over time.	Example: Mary comments on how a graph accurately portrays data because it starts from zero, highlighting a critical analysis of graph initiation and its impact on data interpretation (#00:01:55#).
7	Theme: Comparing Graph Types	Description: Discussing different types of graphs, students assess their effectiveness in showing trends, relationships between variables, or changes over time. They weigh the pros and cons of box plots versus scatter plots for visualizing data.	Example: Oscar points out the utility of scatter plots in showing changes over time compared to box plots, which might not show the relationship between variables (#00:00:38#, #00:01:34#).
7	Theme: Clarifying Data's Story	Description: Students express concern over how certain visual representations might mislead or fail to capture the full story behind the numbers. They stress the importance of visual aids that match statistical data accurately to avoid confusion or misinterpretation.	Example: Lea discusses how the visualization of pandemic data might be misleading due to its representation of death tolls through circles and triangles, suggesting that visuals can distort the perceived impact of pandemics (#00:07:30#).
7	Theme: Historical and Geographical Context	Description: Critical analysis extends to how graphs incorporate or neglect historical and geographical context, affecting the viewer's understanding of data trends over time or across different regions.	Example: James and Lea discuss a graph's depiction of poverty rates in Eastern Europe, debating whether it misleadingly portrays economic progress without considering outer context or regions (#00:10:57#, #00:13:04#).
7	Theme: Assessing Reliability and Sources	Description: Evaluating the credibility of the data sources behind graphs, students consider whether the information provided can be trusted, emphasizing the role of	Example: Teresa regards a graph as reliable because the data source is from UNESCO, showing an awareness of source validity in

		authoritative sources in ensuring data accuracy.	assessing graph credibility (#00:04:14#).
8	Awareness of Misleading Visual Representations	Students expressed concern over how the visual representation of data might mislead viewers. For instance, Lea noted that the representation of the death toll in pandemics using circles could be misleading if not scaled accurately to reflect the magnitude of the data, as it could minimize the perceived impact of significant events.	
8	Understanding the Importance of Starting Points in Graphs	The discussion on the importance of graphs starting at zero to accurately portray data increases was highlighted. Mary mentioned that one of the ways data is accurately portrayed is by graphs starting from zero, as opposed to starting from a random number which could misinterpret data trends.	
8	Critical Evaluation of Data Representativeness	Students critically evaluated whether the data presented was representative and accurately depicted. For example, James discussed how the distribution of population across different poverty thresholds might not be misleading but highlighted the lack of updated data might pose issues for current applicability.	
8	Insights on Data Presentation and Clarity	The need for simplicity and clarity in presenting data was emphasized, with students suggesting that data should be explained in a manner that is understandable to individuals without expertise in the field. Oscar mentioned keeping explanations simple and clear, analogous to explaining games to a fifth grader.	
8	Concerns Over Data Currency and Relevance	Students showed concern for the relevance of the data based on its currency, noting that using outdated data for current analysis can be misleading. James's critique of using data from 2018 to discuss poverty in 2023 exemplifies this concern, likening it to using floppy disks for modern homework.	
9	Evaluating data representation accuracy	Students critiqued the effectiveness and accuracy of the data representations in conveying information.	Mary remarked on how one of the ways data is accurately portrayed is by starting from zero, suggesting awareness of how graph starting points can affect interpretation.
9	Complexity and clarity of visualization	Students reflected on the importance of keeping data presentations simple and understandable, highlighting that complexity may hinder comprehension.	Oscar suggested keeping explanations simple and clear, like explaining games to a fifth grader, emphasizing the need for clarity in data visualization.
9	Critical analysis of visual elements	Students critically analyzed the use of visual elements in graphs, such as size and color, and how they can mislead or accurately represent data.	Lea discussed how the use of circles to represent the death toll in pandemics could be misleading, especially when the size of the circles does not correspond with the numbers they represent.
9	Contextual relevance and updating of data	The relevance and timeliness of the data were considered, with students questioning how current the data was and whether it reflects recent changes or conditions.	James critiqued a distribution of population between poverty thresholds graph for using data up to 2018, pointing out its lack of updation to reflect 2023 circumstances.

10	Critical Analysis of Data Representation	Students critically analyze how effectively data and trends are represented in graphs, scrutinizing the clarity and accuracy of the portrayal.	Yeah, visuals can be super misleading. Size should match the stats, or it's kind of like lying with pictures.
10	Identification of Misleading Visuals	Participants identify and discuss how certain visual elements can be misleading, emphasizing the importance of an accurate match between visuals and statistical data.	But if you know, at least 23 million people died, then it comes off like it kind of misleads.
10	Concerns Over Historical Accuracy and Relevance	The conversation includes concerns regarding the historical accuracy and the relevance of the data depicted, discussing how out-of-date or lacking information affects the understanding of the subject matter.	2018 data trying to talk 2023 stuff. That's like using floppy disks for homework.
10	Challenges in Comparing Variables	Discussing the effectiveness of different types of graphs in comparing variables or showcasing trends clearly to enhance understanding.	It's easier to see the trend box plots won't show the relationship between variables.
10	Authenticity and Reliability of Data Sources	Evaluating the authenticity and reliability of the data sources represented in the graphs, ensuring the information is based on credible research.	Yeah. Okay. That's nice. I also saw like the credentials or like the data source below by Unesco. Yeah, I'd say like. Yeah, I say this is reliable.