Use of Large Language Models to Aid Analysis of Textual Data

*Robert H. Tai, Ed.D.
University of Virginia

Lillian R. Bentley
University of Virginia

Xin Xia
University of Virginia

Jason M. Sitt
University of Virginia

Sarah C. Fankhauser
Oxford College of Emory University

Ana M. Chicas-Mosier
University of Kansas

Barnas M. Monteith
THInc AI Group

*Corresponding author

Abstract

The increasing use of machine learning and Large Language Models (LLMs) opens up opportunities to use these artificially intelligent algorithms in novel ways. In this article, we propose a methodology using LLMs to support traditional deductive coding in qualitative research. We began our analysis with three different sample texts taken from existing interviews. Next, we created a codebook and inputted the sample text and codebook into an LLM. We asked the LLM to determine if the codes were present in a sample text provided and requested evidence to support the coding. The sample texts were inputted 160 times to record changes between iterations of the LLM response. Each iteration was analogous to a new coder deductively analyzing the given text with the codebook information. In our results, we present the outputs for these recursive analyses, along with a comparison of the LLM coding to evaluations made by human coders using traditional coding methods. We argue that LLM analysis can aid qualitative researchers by deductively coding transcripts, providing a systematic and reliable platform for code identification, and offering a means of avoiding analysis misalignment. Implications of using LLM in research praxis are discussed, along with current limitations.

Keywords: *Qualitative methodology; Large Language Models (LLMs); Deductive qualitative coding; Reliability*

Introduction

Text is an important form of qualitative data and its analysis enables researchers to gain an understanding of complex social phenomena (Day, 2003). Traditionally, qualitative analysis of text involves meticulous, time-consuming, and labor-intensive manual coding. This subjective process depends on a number of different factors of a given researcher's background and experience, i.e. positionality, which may introduce bias in the analysis and interpretation of qualitative data. In order to mitigate bias, reliability measures often involve two or more researchers analyzing, i.e. "coding," the data. The themes, concepts, and phenomena that are identified are referred to as "codes." Similarities and differences in the coding can be identified and inter-rater reliability calculated (e.g. Cohen, 1960; Fleiss & Cohen, 1973). Recent advancements in natural language processing and machine learning have opened up new possibilities for assisting qualitative researchers in their analytical endeavors (Manning, 2022). In this paper, we outline a systematic methodology for using Large Language Models (LLM) as a qualitative research tool that can be used as an aid in the deductive coding of interview data.

Qualitative codes serve as interpretive labels or categories that describe various themes, concepts, or phenomena present in the textual data. These codes can be applied in an inductive or deductive manner, depending on the researchers' positionality and role in their work (Elo & Kyngäs, 2008). Inductive coding requires the researcher to read the textual data and identify codes within the text. This allows the researchers to build patterns from the data, eventually organizing the data into a comprehensive list of themes. For deductive analysis, researchers apply a series of codes to the text based on a theoretical framework. When performing deductive coding, researchers often follow a detailed "codebook" to ensure the coding matches the

predetermined characteristics or theoretical constructs that are being applied (Schwandt, 1997). This paper will focus on the use of an LLM in deductive coding.

Coding can be subjective, especially when only one researcher performs a textual analysis. Unintended bias may creep into analysis in several different ways. How a researcher approaches that data or how they define terms present in the text may have an influence on the findings of a researcher. This subjectivity can lead to variations in how different researchers interpret and assign codes to the data (Marques & McCall 2015).

To address potential researcher bias, some researchers create memos, writing and reflecting on their thoughts and ideas. If multiple researchers are coding the same data, they can apply an inter-rater statistic to quantify the degree of similarity between coders (Marques & McCall, 2015). When coders disagree, negotiations take place between them to refine the coding and reach an agreement. Codebooks are commonly re-defined and re-applied to the text. This iterative process helps to focus the coding and increases the reliability but is also highly labor intensive while remaining dependent on researcher subjectivity.

The purpose of this research article is to examine the systematic use of an LLM as a qualitative analysis tool that can supplement the deductive coding process. An LLM is a type of artificial intelligence algorithm. It is a trained deep-learning model that is able to understand and generate text, answer questions, and complete other language-related tasks with high accuracy (Kasneci et al. 2023). Some examples include ChatGPT®, Open AI®, Bing Chat®, and Microsoft®. By leveraging contextual understanding and LLMs, researchers may verify and enhance alignment in qualitative textual analysis, potentially improving the overall quality of the output.

Literature Review

To lay the groundwork for using an LLM as a deductive qualitative analysis tool, we begin by considering deductive coding. Next, we outline some of the current reliability measures traditional coding uses and make a case for the addition of LLM as a new research instrument. Our focus will then shift to current empirical work with an example of an LLM, ChatGPT® 3.5, which will anchor the discussion to the current academic landscape. Our review will conclude with opportunities for methodological expansion using LLMs as a reliability measure in qualitative research, specifically testing the reliability of deductive coding and application of the codebook in qualitative analysis.

Deductive Coding and Reliability

One popular approach to qualitative analysis in social science is the use of deductive codes. In deductive coding, researchers develop a qualitative codebook that contains a list of predetermined codes that researchers apply to their data. Oftentimes these codes are derived from theoretical frameworks and are used to test the presence or absence of certain characteristics or themes within the text. The codebook provides a list of codes, a code label for each code, a brief definition of the code, and an example of a quote that illustrates the use of the code (Creswell & Creswell, 2017).

To ensure the reliability of deductive coding, researchers can calculate different types of reliability statistics. For example, Cohen's (1960) kappa statistic can be used to measure the level of agreement between two raters. Cohen's kappa quantifies the extent to which raters agree on the relative ratings or coding and serves as a measure of accuracy (Linacre, 1989). Another example of an inter-rater statistic is Fleiss' kappa, which is used when there are more than two

coders (Fleiss, 1971). Given that Fleiss' kappa is a derivative of Cohen's kappa, it also uses the hypothetical probability of agreement by chance. Each of these measurements attempts to ensure the reliability of data analysis by comparing sets of codes to each other.

Although Cohen's and Fleiss' kappa provide quantifiable statistics for qualitative research, there are limitations to applying these measures. For example, Block and Kraemer (1989) argued that because Cohen's kappa allows marginal distributions to differ, it actually measures the association between two sets of ratings instead of the agreement between two raters. Nicolas et al. (2011) argued that Fleiss' kappa does not apply weighting during analysis, which results in variation between coders for the same dataset, even if there is no variation. Using LLMs as another comparison tool could help qualitative researchers expand upon these different types of reliability measures that are currently being used.

LLMs and Empirical Research

ChatGPT® 3.5 is an autoregressive language model with more than 175 billion parameters, 10x more than any previous non-sparse language model (Brown et al. 2020). As ChatGPT® evolves online, different variations are released and the LLM subsequently becomes faster and more refined. At the time of writing, the free version of ChatGPT® was built on OpenAI's Generative Pre-Trained Transformer (GPT)- 3.5. GPT-4 is available via a paid subscription, all references to ChatGPT® will be to GPT-3.5 unless otherwise notated.

Researchers are starting to use LLMs in a variety of fields. Recent substantial improvements in efficient, accurate, and automatic determination of semantic equivalence between two narrative entities (concepts) in natural language processing have come to be incorporated into LLMs in both English and Chinese (Monteith & Sung, 2023). This ability to correlate semantics, largely enabled by parameter scaling, can then be used for generative AI

purposes. For example, ChatGPT® has been used in translation and text generation across linguistic research fields (Jiao et al., 2023 (GPT-4); Peng et al., 2023; Ubani & Nielsen, 2023). It has also been used in the medical field to write discharge summaries (Patel & Lam 2023), to help radiologists in their decision-making processes (Rao et al. 2023), and to help diagnose prostate cancer (Van Booven et al. 2021). This broad range of applications highlights the potential to use ChatGPT® in unprecedented ways.

Focusing on academic writing, Bhardwaz and Kumar (2023) evaluated the use of two LLM's, Google Bard® and ChatGPT® in writing literature reviews. The authors aimed to determine how well the LLMs could summarize the text without plagiarism. The authors asked Google Bard® and ChatGPT® to paraphrase the abstracts of ten articles. They demonstrated that text written by the study authors demonstrated low plagiarism rates, while the text written by Google Bard® and ChatGPT® had comparatively much higher plagiarism counts. Similarly, Rahman et al. (2023) used ChatGPT® 3.0 to write an academic paper and discovered that ChatGPT® 3.0 could generate problem statements and an outline, but artificial intelligence could not draft a validated literature review or quantify numerical data. The references for the literature review section of the academic paper were cited correctly using APA formats, but the citations were not connected to real empirical articles. The authors concluded that LLMs could be used to analyze qualitative data, but do not enumerate this potential data application.

Narrowing the focus on qualitative research, Xiao et al. (2023) used ChatGPT® 3.0 to investigate the potential use of the LLM for qualitative analysis. According to the authors, ChatGPT® 3.0 can be used to deductively code data, and this coding is comparable to traditional coding when the LLM is given a clear codebook. They also analyzed different prompt designs that were input into ChatGPT® 3.0. They concluded that codebook-type prompts, those that

provide structure and contextual information, are more reliable than example-type prompts, those that provide information without additional context. However, presenting both the codebook and at least one example prompt produced the most accurate results when compared to traditional coders. This work lays the foundation for exploring different applications of deductive coding with LLM.

There is a paucity of validated research examining the use of LLMs for qualitative analysis, specifically with multiple iterations of text being put into ChatGPT® 3.5 to test the consistency of results. A benefit of using LLMs in qualitative analysis is that it provides almost unlimited inter-rater measures. Each time a researcher logs into an LLM network, the prompt entered into the model becomes a new input text, and this action is analogous to a new rater coding the data. The LLM incorporates some randomness in the way it processes and weighs information. This randomness is measured and defaults to 1 (relatively random) in ChatGPT®. This random integration results in responses that may not agree with prior outputs (Gilardi et al. 2023). Researchers can then evaluate whether the imputed codebooks and examples sufficiently meet the assumptions of kappa statistics or if additional codebooks and examples are needed. This improves the efficiency of the researcher's time focusing on accurately depicting the coded data and reducing personnel needed for reliable evaluations.

Purpose

To date, there are no developed systematic practices in place for using LLMs for qualitative data analysis. As LLMs evolve, and presumably become more reliable, and as qualitative research applications based on LLMs arise, it is important to explore analytical practices and protocols. The purpose of this work is to outline a new method for using LLMs that

may contribute to the validity and reliability of qualitative analysis techniques of textual data. In this paper, we address the following research questions:

Research Questions

1. How does the output of a large language model respond to the recursive entry of a qualitative data analysis prompt across a range of iterations?
2. How do the responses of a large language model compare to the traditional qualitative analysis on the same three textual data sources?

Methodology

The focus of this study was to develop a procedure for using an LLM to analyze interview-based textual data and to test the reliability of the procedure. This methodology section will outline the participants that were part of the study, the sample that was used to run the iterations of codebook analysis in ChatGPT® 3.5, the data collection techniques, and data analysis.

Data Source

The data used in this study are archival interviews from *Project Crossover: A Study of the Transition from Student to Scientist* (NSF REC 0440002), a sequential, mixed-methods study exploring the transition of Ph.D. students to independent researchers. This project examined the experiences of individuals who had engaged in the process of becoming a scientific researcher in the fields of chemistry, physics, or chemical engineering or who were currently engaged in this process at the time (Dabney & Tai, 2014). As part of the project, 125 semi-structured interviews were collected from chemistry, physics, and chemical engineering graduate students, postdocs, scientists, and research engineers as well as individuals who had left the field of scientific

research. Each interview was recorded and transcribed. They ranged in length from 30 minutes to 2.5 hours.

Sample

This study used three excerpts from two Project Crossover interviews. These interviews captured scientists' beliefs of what characteristics constitute a successful scientist. Interviews focused on graduate schools, additionally including early educational history, as well as relevant research and employment experiences. Project Crossover interviews were additionally screened for relevance based on the identification of five characteristics which will be referred to as codes: 1) autonomy or self-motivation to pursue research experiences, 2) persistence despite adversity or difficulty, 3) perception of researcher identity, 4) the desire to create novel knowledge, and 5) interest in engaging in research in a STEM-related discipline.

Data Collection

Codes and Prompts

For this study, we chose to examine five different codes: *Autonomy*, *Persistence*, "*Perception of Own Identity and Self*", *Novelty*, and *STEM Interests*. These codes were inductively created by the research team through the analysis of science fair participant interviews that were collected and analyzed for a separate research project. The definitions of the codes are shown below. We designed the LLM prompt for a "binary" query (i.e. a yes-or-no answer). For the first query to the LLM, we entered the prompt and interview sample, Text 1, below.

We defined the characteristics below:

1. Autonomy: is the ability to be self-driven in STEM research
2. Persistence: is the continuance in a course of action in spite of difficulty
3. "Perception of own identity or self": is the internal recognition as an individual who can talk about research with an expert in the discipline or confidence
4. Novelty: is the desire to create new knowledge or knowledge new to them
5. STEM interests: is having an interest in a STEM field.

Can you find the five characteristics in the transcript below, yes or no? If yes, give us the quote.

[Text 1]
It really was, I guess, working with the postdoc. I, of course, tried to be as independent as I could be and I did try a number of things, but then I would brainstorm with the postdoc and we met often with the advisor. I'm actually a very organized person and so if I can get a list of six things to try, I can work through them. So it was with the help of others.

For the second and third queries, we replaced Text 1, shown in italics above, with Text 2 and Text 3, respectively.

[Text 2]
In several ways. One, again, at the University, you're a scientist educator. And so that aspect of it, being successful is having my graduate students enjoy, have a passion for science, and find rewards in it, helping them to learn to think and analyze and become scientists of their own. That's a success, teaching the next generation. And being a successful scientist—wow. I mean I think it would be wonderful to be able to really make a major advance in terms of medicine or some really breakthrough area, but for me it's probably just discovering little bits about Mother Nature and helping it fit into the big picture. What I mentioned about the biochemistry earlier, in many ways I would love to be able to move a little bit that way where I think the ability to effect the well-being of people is stronger….Well, I certainly intelligence and creativity are necessary. More and more, a strong work ethic helps. I think an ability to communicate with others, both students and other scientists because, often, again, and what we were talking in the graduate school, that often we learn by interacting with others. So, yeah, that's certainly main,I mean, certainly a passion for the work. Science is, especially in academics now, one of the most highly paid professions, and so it really does take someone who is committed, who just really enjoys science.

[Text 3]
Whereas the veteran researcher is happy to be confused, because if you're not confused you're not going to learn anything that's worthwhile, because it must be already understood. And if you're confused there's a chance that when you finally figure out what's going on there's a chance you will have learned something that's of interest to other researchers. So that's the general thrust of what I have in mind.…It's a question of how thoughtful and original, and I say spunky or enterprising and all the students are. I cultivate their maturity, that sort of thing. If they, for example, show initiative in going to the literature, or going to talk to somebody else at another university, or somewhere, or finding something on the web, and then getting in touch with someone else all on their own initiative, it shows that they're really totally thinking of this as their thing, and carrying the ball and so forth….I think the kind of science I strive to do is that sort of spirit. It's a question of can you do something as an architect that opens people's eyes to new possibilities and gets them excited. I like to tell my students that the real value of an experiment is spiritual, in terms of how it arouses other people's interest and stimulates them, 'cause that's the way my experience is when I encounter some work of science, and I encounter lots of them, and they open my eyes, and I think 'Oh my gosh how beautiful. Wow!'

Experiments and Iterations

We will refer to the prompt with Text 1 as Experiment 1, the prompt with Text 2 as Experiment 2, and so forth. If the "experiment" prompt and the text to be analyzed were

submitted repeatedly to the LLM during the same session which means that the user did not logout and then login back, we can expect that a similar or even identical output would be returned. However, this narrowing of responses would not be a reflection of a general trend of the population, but would more likely be an artifact of the initial conditions set when the LLM algorithm was initiated (Urbani, 2023). One important characteristic of large language models in general is that the underlying algorithms are designed to include some variation to mimic variations within natural language processing. As a result, the same prompt entered by different users with different initial conditions into an LLM would be expected to return different responses that would trend toward a generally similar result that would be representative of the general population. This difference would depend on the random number captured by the LLM from a random number generator used by the model to set the initial conditions. In our case, the LLM used in this analysis used a pseudo-random number generator (PRNG). It is this variation in output that takes advantage of the power of an LLM. While differences of interpretations between people are expected, an LLM, using the vast internet archive data set harvested by the Common Crawl, offers some indication of a general interpretation for a population consisting of many thousands and possibly hundreds of thousands of individuals, when the same prompt is entered repeatedly. In our case, each time Experiment 1 text is entered by a different user, the responses may differ, but over a series of these entries, a trend in the responses was revealed. We call each repeated entry of an experiment text, an “iteration.” To take advantage of this effect, for each iteration, each time before we entered the text, we logged out of the LLM and then logged back in. For the purposes of this study, we performed 160 iterations for each of the three experiments.

Data Analysis Approach

For data analysis, we coded the responses from the LLM as either positive when the code was reported as present (i.e. coded as 1), or negative when the code was reported as absent (i.e. coded as 0). Sometimes the response would be ambiguous as to the presence or absence of a characteristic, so we applied a conservative interpretation of the data. Only responses that were entirely affirmative were coded as positive. If the response implied a level of uncertainty (e.g., when the LLM responded, "Some level of the characteristic is present."), it was marked as negative.

Identifying an Outcome Measure

For each code in each iteration of each experiment, we recorded the results and after 5, 10, 20, 40, 80, and 160 iterations, we calculated the proportion of positive responses for the given number of iterations. The results are shown in Tables 1, 2, and 3 in the Results and Discussion section. We will refer to this value as the large language model quotient (LLMq), where the total number of positive responses in a set of iterations was divided by the number of iterations for that given set. The equation is presented below:

$$P(LLMq) = \frac{n}{N} = \frac{\textbf{Positive Responses}}{Total\ \textbf{Iterations}}$$

Comparison with Traditional Coding

The three sample interview excerpts were uploaded into a qualitative data analysis software package to facilitate data management and analysis.[1] The traditional coding was carried out by three research team members simultaneously and separately. None had access to the LLM coding results. All three coders are science educators and were not involved in collecting the

[1] The software package used in this analysis was Dedoose® (Los Angeles, CA).

interviews from which the quotes were extracted. These researchers only read and evaluated the selected quotes for the presence of the five conceptual codes used in this analysis. The coders used the same five codes included in the LLM prompts (i.e. *Autonomy*; *Persistence*; *Perception of Own Identity or Self*; *Novelty*; *STEM Interests*). They assess the transcripts line-by-line to identify text selection in the interview excerpts for the five codes. The coding results of the three traditional coders were compared, and any discrepancies or disagreements were discussed and resolved through group meetings.

Results and Discussion

Calculating the Large Language Model Quotient

The LLM data analysis began by performing 5 iterations and then calculating the Large Language Model quotient (LLMq) for this set of 5 iterations. Next, an additional 5 iterations were performed and the LLMq was calculated for a total of 10 iterations. This sequence was repeated to produce LLMq values for 20, 40, 60, 80, 100, 120, 140, and 160 total iterations. These results are shown in Tables 1, 2, and 3 for each of the three respective experiments. The results were also graphed to display how the LLMq-values changed with each successive set of 20 total iterations. The results displayed in the graphs indicated that LLMq values varied for each of the three different experiments across the five codes. However, it is important to note that the LLMq values for a specific code within a specific experiment appeared to be stable with the trace of the graph leveling out.

For the purposes of clarity, we offer the following examples of how LLMq values were calculated. For Text 1, with respect to the concept of *Autonomy*, the first 5 iterations produced one positive response. (See Table 1.) In this instance, the LLMq was equal to 0.2 indicating that 20% of the prompted responses from the large language model were positive. Next, for the set of

10 iterations with three positive responses, an LLMq value equal to 0.3 is calculated. For the third set which consisted of 20 iterations, an LLMq value of 0.35 was calculated. These results and the subsequent calculations of LLMq values for the other four codes, *Persistence*, *Perception of Own Identity and Self*, *Novelty*, and *STEM Interests* are also displayed in Table 1. The Table 1 results are graphed in Figure 1.

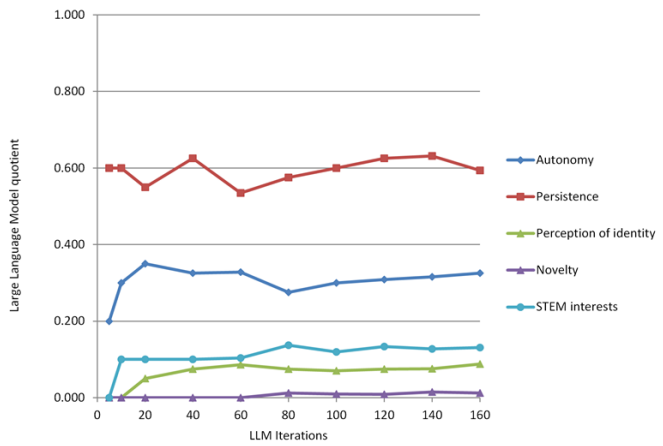
For Text 1, the LLMq results show low or very low levels of agreement across all five of the codes. Text 1 consists of 76 words and was selected for this analysis for its brevity and general lack of clear detail with respect to the five codes considered in the analysis. For three codes, *Perception of Identity and Self*, *Novelty*, and *STEM Interests*, the LLMq results showed that they were identified in less than 15% of the iterations. For example, in the results for *Perception of Own Identity and Self*, the LLM analysis results for 5 and 10 iterations, the LLMq was 0.000 or none of the iterations identified this code as present. As more iterations were carried out, the successive LLMq values were less than 0.100, which can be interpreted as the LLM identified this code in less than 10% of the iterations. More specifically, for 160 iterations, the result of LLMq = 0.088 is interpreted as the code *Perception of Own Identity or Self* is identified as present in only 8.8% of the iterations. The results for *Novelty* and *STEM Interests* may be similarly interpreted. The results for *Autonomy* show that this code is identified as present by the LLM in up to 35% of the iterations. After 160 iterations, the results show the LLMq = 0.325 indicating that *Autonomy* was identified as present by the LLM in 32.5% of the iterations. The results for *Persistence* show that the LLMq values from 5 iterations to 160 iterations, the LLM identified this code in more than half of the iterations. At 160 iterations, the LLMq = 0.594 which indicates that the LLM reported this characteristic present in 59.4% of the iterations. Overall, the LLM analysis of the Text 1 produced results indicating that none of the

five codes were clearly present. The strongest result was for the code *Persistence* found by the LLM in about 60% of each set of iterations.

Table 1: Large language model quotients (LLMq) for five conceptual codes examining Text 1

Iterations	Autonomy	Persistence	Perception of identity	Novelty	STEM interests
5	0.200	0.600	0.000	0.000	0.000
10	0.300	0.600	0.000	0.000	0.100
20	0.350	0.550	0.050	0.000	0.100
40	0.325	0.625	0.075	0.000	0.100
60	0.328	0.534	0.086	0.000	0.103
80	0.275	0.575	0.075	0.013	0.138
100	0.300	0.600	0.070	0.010	0.120
120	0.308	0.625	0.075	0.008	0.133
140	0.316	0.632	0.075	0.015	0.128
160	0.325	0.594	0.088	0.013	0.131

Figure 1: Graph of LLMq values shown in Table 1 above



The traditional coding results were also mixed. The following extract of the excerpt was coded as demonstrating *Autonomy* by one of the three coders and *Perception of Own Identity and Self* by two of the three coders.

> *" tried to be as independent as I could be and I did try a number of things, but then I would brainstorm with the postdoc and we met often with the advisor"*

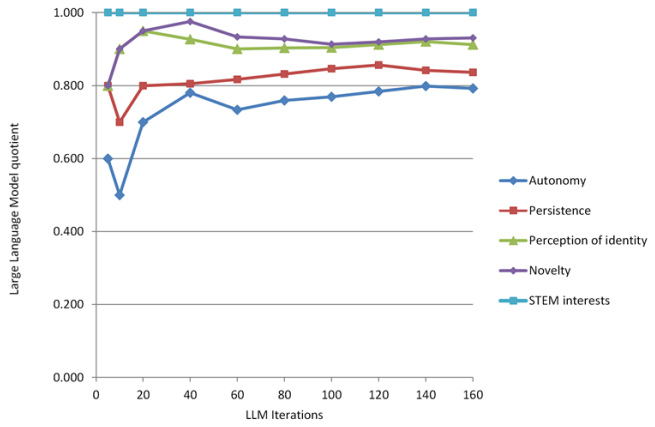
The three coders did not find evidence for the codes *Persistence*, *Novelty*, or *STEM Interest* in Text 1. The results from the three coders showed indications that for Text 1, the five codes were not likely or clearly present.

Next, we consider the results of the Text 2 LLM analysis. Here, the LLMq values for the codes *Perception of Own Identity or Self*, *Novelty*, and *STEM Interests* are 0.900 or greater across nearly all sets of iterations. The results also show that for the code *Persistence*, LLMq values are 0.800 or greater for all except one set of iterations. For the code Autonomy, LLMq values are 0.700 or greater for all except two iteration sets.

Table 2: Large language model quotients (LLMq) for five conceptual codes examining Text 2

Iterations	Autonomy	Persistence	Perception of identity	Novelty	STEM interests
5	0.600	0.800	0.800	0.800	1.000
10	0.500	0.700	0.900	0.900	1.000
20	0.700	0.800	0.950	0.950	1.000
40	0.780	0.805	0.927	0.976	1.000
60	0.733	0.817	0.900	0.933	1.000
80	0.759	0.831	0.904	0.928	1.000
100	0.769	0.846	0.904	0.913	1.000
120	0.784	0.856	0.912	0.920	1.000
140	0.799	0.842	0.921	0.928	1.000
160	0.792	0.836	0.912	0.931	1.000

Figure 2: Graph of LLMq values shown in Table 2 above



Overall, the LLM analysis identifies all five of the codes to be present in most of the iterations. In instances where the LLM identifies a code as present, the output included a

response that in some cases cited selections from the text. Examples are shown in Figures 2A and 2B. Note that the LLM outputs are consistent, but are not identical.

Figure 2A: Example of LLM output for Text 2

Yes, the characteristics mentioned in the transcript are:

1. Autonomy: The quote mentions "helping them to learn to think and analyze and become scientists of their own," indicating the desire for graduate students to develop their own independent research skills.
2. Persistence: The quote mentions "continuing to discover little bits about Mother Nature and helping it fit into the big picture," indicating a willingness to persevere in research despite challenges.
3. Perception of own identity or self: The quote mentions the satisfaction of being a successful scientist and making advancements in medicine or breakthrough areas, indicating a sense of confidence and recognition as an individual capable of contributing to the field.
4. Novelty: The quote mentions the desire to make a major advance in terms of medicine or some breakthrough area, indicating a motivation to create new knowledge.
5. STEM interests: The quote mentions the need for a passion for science and enjoying the work, indicating an interest in the STEM field.

Figure 2B: Example of LLM output for Text 2

Yes, the five characteristics can be found in the transcript:

1. Autonomy: The mention of "being a successful scientist" and "helping [graduate students] learn to think and analyze and become scientists of their own" implies a sense of autonomy in STEM research.
2. Persistence: The phrase "continuance in a course of action in spite of difficulty" is not explicitly mentioned in the transcript.
3. Perception of own identity or self: The statement "being successful is having my graduate students enjoy, have a passion for science, and find rewards in it" suggests a perception of the speaker's own identity as an educator and researcher who can inspire others in the field.
4. Novelty: The mention of "discovering little bits about Mother Nature and helping it fit into the big picture" indicates a desire for novelty in creating new knowledge and understanding.
5. STEM interests: The statement "it really does take someone who is committed, who just really enjoys science" highlights the importance of having an interest in STEM as a characteristic for success in the field.

The results from the traditional coding found that all five codes were present in Text 2. For example, the following excerpt was coded by two researchers as an example of *STEM Interests*, and one researcher coded for *Perception of Own Identity or Self*.

> *"being successful is having my graduate students enjoy, have a passion for science"*

The following excerpt was coded by all three coders as demonstrating autonomy.

> *"analyze and become scientists of their own"*

An example of novelty was coded by all three coders below.

> *" I think it would be wonderful to be able to really make a major advance in terms of medicine or some really breakthrough area, but for me it's probably just discovering little bits about Mother Nature and helping it fit into the big picture"*

Next, the following excerpt was coded by all three researchers as demonstrating persistence.

> *"it really does take someone who is committed, who just really enjoys science"*

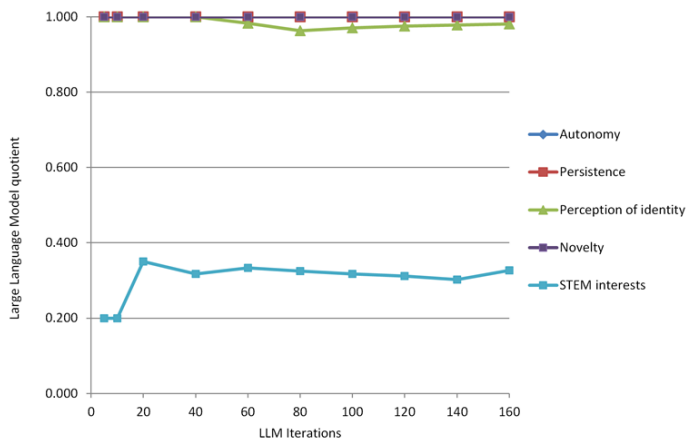
Comparing the LLM output and the traditional coding results, we find that results agree and in some instances are identical.

Finally, the LLM analysis found that Text 3 had LLMq values definitive or nearly definitive for four codes, *Autonomy*, *Persistence*, *Perception of Own Identity or Self*, and *Novelty*.

Table 3: Large language model quotients (LLMq) for five conceptual codes examining Text 3

Iterations	Autonomy	Persistence	Perception of identity	Novelty	STEM interests
5	1.000	1.000	1.000	1.000	0.200
10	1.000	1.000	1.000	1.000	0.200
20	1.000	1.000	1.000	1.000	0.350
40	1.000	1.000	1.000	1.000	0.317
60	1.000	1.000	0.983	1.000	0.333
80	1.000	1.000	0.964	1.000	0.325
100	1.000	1.000	0.971	1.000	0.317
120	1.000	1.000	0.976	1.000	0.312
140	1.000	1.000	0.978	1.000	0.302
160	1.000	1.000	0.981	1.000	0.327

Figure 3: Graph of LLMq values shown in Table 3 above



For the code *STEM Interest*, the results were mixed. The LLMq values were stable as shown in Figure 3 with the graphed curve flattening out as iterations increased, but the LLM analysis results are inconsistent, showing that the LLM analysis reported code *STEM Interest* was present in only one-third of the cases for 60 iterations and only 32.7% of cases for 160 iterations. This result indicates that the code *STEM Interest* was not clearly present in Text 3.

Turning to the traditional coding analysis, two examples of LLM output responses for Text 3 are shown in Figures 3A and 3B. Comparing these two examples, the responses for the three codes of *Autonomy*, *Perception of Own Identity or Self*, and *Novelty* are the same in the results of these two iterations of Experiment 3. However, the output identifies the presence of the code *Persistence* but cites two different selections from Text 3. Also, while both iterations do not

identify the code *STEM Interest* in Text 3, the output in Figure 3B offers the following additional insight, “However, the overall context of the discussion revolves around STEM research, experiments, and encountering scientific work, indicating a focus on STEM interests.”

Figure 3A: Example of LLM output for Text 3

Yes, I can find some of the characteristics in the provided transcript. Here are the quotes that correspond to each characteristic:

1. Autonomy: "If they, for example, show initiative in going to the literature, or going to talk to somebody else at another university, or somewhere, or finding something on the web, and then getting in touch with someone else all on their own initiative, it shows that they're really totally thinking of this as their thing, and carrying the ball and so forth."
2. Persistence: "And if you're confused there's a chance that when you finally figure out what's going on there's a chance you will have learned something that's of interest to other researchers."
3. Perception of own identity or self: "I like to tell my students that the real value of an experiment is spiritual, in terms of how it arouses other people's interest and stimulates them, 'cause that's the way my experience is when I encounter some work of science, and I encounter lots of them, and they open my eyes, and I think 'Oh my gosh how beautiful. Wow!'"
4. Novelty: "It's a question of can you do something as an architect that opens people's eyes to new possibilities and gets them excited."
5. STEM interests: The transcript does not provide a direct quote related to STEM interests.

Figure 3B: Example of LLM output for Text 3

Yes, I can find some of the characteristics in the given transcript. Here are the quotes that correspond to each characteristic:

1. Autonomy: "If they, for example, show initiative in going to the literature, or going to talk to somebody else at another university, or somewhere, or finding something on the web, and then getting in touch with someone else all on their own initiative, it shows that they're really totally thinking of this as their thing, and carrying the ball and so forth."
2. Persistence: "It's a question of how thoughtful and original, and I say spunky or enterprising and all the students are. I cultivate their maturity, that sort of thing."
3. Perception of own identity or self: "I like to tell my students that the real value of an experiment is spiritual, in terms of how it arouses other people's interest and stimulates them, 'cause that's the way my experience is when I encounter some work of science, and I encounter lots of them, and they open my eyes, and I think 'Oh my gosh how beautiful. Wow!'"
4. Novelty: "It's a question of can you do something as an architect that opens people's eyes to new possibilities and gets them excited."
5. STEM interests: The specific term "STEM interests" is not explicitly mentioned in the transcript. However, the overall context of the discussion revolves around STEM research, experiments, and encountering scientific work, indicating a focus on STEM interests.

Please note that these interpretations are based on the provided characteristics and the given transcript.

The traditional coding results found that Text 3 contained examples of all five codes. For example, all three coders coded the following excerpt for *Autonomy*.

> *"If they, for example, show initiative in going to the literature, or going to talk to somebody else at another university, or somewhere, or finding something on the web, and then getting in touch with someone else all on their own initiative, it shows that they're really totally thinking of this as their thing, and carrying the ball and so forth.*"

This result is consistent with the two LLM output examples shown above. A second excerpt from Text 3 shown below was coded by two coders as *Novelty* and as both *Persistence* and *STEM Interests* by a third coder.

> *"And if you're confused there's a chance that when you finally figure out what's going on there's a chance you will have learned something that's of interest to other researchers."*

Finally, one coder identified the code *Perception of Own Identity or Self* in the following text which matches with the two LLM outputs cited above.

> *"I like to tell my students that the real value of an experiment is spiritual, in terms of how it arouses other people's interest and stimulates them, 'cause that's the way my experience is when I encounter some work of science, and I encounter lots of them, and they open my eyes, and I think 'Oh my gosh, how beautiful. Wow!'"*

This study explored the use of an LLM to perform qualitative coding analysis leading to the calculation of the large language model quotient that represents the percentage of positive outcomes returned from a set of analytical iterations. The results generated from the analysis of the iterations were compared to traditional coding and were found to be consistent. Recent work by Xiao, et al. (2023) demonstrates the feasibility of using an LLM in the capacity of "another rater" for qualitative analysis. However, in that study, only single LLM outputs were examined.

Given that LLMs are stochastic and are designed to generate variable responses from one output to the next, drawing conclusions from a single LLM response is not the best approach. Multiple iterations of a prompt produce a better representation of LLM outputs. For this reason, we have chosen to explore the results across a range of output iterations for three different text examples.

In addition, these results show that the application of an LLM in performing qualitative data analysis produces consistent results with a sufficient number of repetitions. Similar consistency was found by Gilardi, Alizadeh, & Kubli (2023) in which ChatGPT® 3.5 produced accurate and reliable repeated classifications of Twitter posts regardless of randomness (temperature range 0.2-1). In their study, ChatGPT outputs were more reliable than traditional coders (>84%). An LLM can be used as a research instrument because it is designed to pull in vast amounts of data and then to converge to a solution. This characteristic of large language models works well with the considerations of qualitative research which aims to identify meaning and trends within non-ordinal data. The LLMq values calculated for each of the three Texts examined in this analysis and graphed in Figure 1, 2, and 3 indicate that LLMq values appear fairly stable beyond 40 iterations for each of 5 codes with the trajectories for each of the graphed curves flattening out. In addition, the results from the traditional analysis also appear to be reflected in the LLMq values, even in the case of Text 1 for the code Persistence. Here, the traditional analysis showed a fairly low interrater agreement among the coders and this lack of agreement is reflected in the LLMq values reported in Table 1 and Figure 1. The LLMq value after 160 iterations was 0.594 or 59.4% agreement from the LLM analysis that the code Persistence was found in Text 1, an inconsistent result from the LLM analysis reflect in the inconsistence results from the traditional coding. In this case, the LLMq values show that the LLM reports finding the code Persistence with a probability only 9.4% greater than a coin flip.

These results show that the LLM analysis was stable even in cases when the results are not consistent, and when the results are not consistent, this lack of consistency is reflected by the findings from traditional coding. These results offer some strong support for the application of large language models as an analytical tool for qualitative data.

Limitations

While LLM applications have shown great potential in aiding qualitative data research, they also come with inherent limitations. The biggest limitation of the use of LLM is the algorithm itself. LLMs rely on patterns and structures present in the training data, and if certain linguistic nuances or subtleties are absent, the model's understanding may be limited. The use of the data generated by the Common Crawl discussed earlier in this paper mitigates this limitation, drawing from the vast and growing body of content generated through the internet.

The quality of the input data also plays a vital role in the effectiveness of LLMs in qualitative research. If the input data is of poor quality or contains biases or inaccuracies, the LLM's outputs will likely reflect those shortcomings. Noisy data can also pose a challenge for LLMs in qualitative research. Noisy data refers to data that is unstructured, inconsistent or contains errors, such as transcription errors. LLMs may struggle to accurately interpret and assign codes to such data, leading to potential inaccuracies or misinterpretations.

The length of the input data is limited by the LLM. For example, at the time of data collection for this study, ChatGPT® 3.5 had a default limit of 2048 characters. Each character is referred to as a token and is a measurement for the length of an inputted text. For some LLMs, it is possible to pay for extra tokens, increasing the length of the text that may be entered into each iteration. For example, with ChatGPT® 3.5, the maximum number of tokens may be increased to 4096 by paying a subscription fee.

Conclusions

These results offer some support that LLMs may be used as a tool to streamline qualitative research data analysis. Far from suggesting the replacement of researchers, it is clear that decisions about how to employ the LLM and how to interpret its output are essential. LLMs offer researchers the autonomy to design prompts aligned to their own interests, and then test their ideas with an artificial intelligence engine designed to bring to bear massive data harvested from the internet. Creating a prompt that is codebook-driven may mimic traditional coding, whereas prompts that simply ask the LLM to generate a list of themes for a given excerpt may lead to the generation of themes not initially considered by the researcher. Currently, LLMs are far from autonomous and should be prompted carefully to generate useful data.

In this paper, we examined how the output of an LLM behaved when queried recursively across three different sets of texts using the same prompt. The results led to the development of a simple representation of these results in the form of the quotient of the number of positive responses divided by the total number of queries which we have referred to as the large language model quotient (LLMq). This simple calculation offers results that may be used for some helpful applications.

One potential application is using an iterative LLM analysis as a screening tool. Since the LLMq offers quantification of the degree of positive identification of a code within a text, entering a lengthy interview as a series of short excerpts would likely produce results highlighting sections of text where particular codes are present. Conversely, excerpts with low LLMq suggest a paucity of codes. While entering the type of prompt described in this paper would produce a binary outcome—that codes are either entirely present or absent in an excerpt—applying the prompt in an iterative manner produces results with more nuance, asserting that

codes may exist across a spectrum. Effectively, the LLMq optimizes the efficiency of researchers, who can hone in on areas of particular interest.

Another use of LLMq is as a post-hoc analysis tool. LLMq calculations can provide confirmatory or additional data to the traditional coding. The quantification of traits can justify the inter-rater reliability found, and perhaps, the LLM may discover other traits not identified originally. A significant advantage of the LLMq is the fact that a rater can independently and efficiently perform a “check” on their own coding without relying upon other raters of equal competence. Essentially, the LLM can provide supplementary or even new data to complement a researcher’s findings.

The future of LLM development holds immense promise for significant improvements in LLM performance that would increase the validity and reliability of this tool for qualitative research. As data sets used to train LLMs continue to expand, there is an increasing focus on domain-specific and refined data. This type of data will expose LLMs to a wider range of linguistic patterns and contextual information. This form of enhanced artificial intelligence training will enable LLMs to better reflect the nuances and complexities of qualitative data, leading to improved accuracy in code identification and interpretation. With access to larger and more diverse training data, future LLMs will be increasingly accurate in linguistic interpretations, thereby reducing the need for extensive iterations. This advancement will not only expedite the qualitative analysis process but also facilitate more reliable and insightful findings. Researchers may use LLMs in an inductive process to search for themes and codes arising from the textual data. The application of an LLM as an inductive tool is beyond the scope of this current paper, but is a topic for future investigation. By leveraging the power of future

LLMs, researchers can unlock new possibilities for deeper understanding and exploration of qualitative data in an efficient and accurate manner.

We envision the primary application of LLMq would be for the coding of qualitative data. In this study, we purposefully did not propose an LLMq “cutoff” value, choosing to leave this open for researchers to consider moving forward. It does, however, bear mentioning that in statistical analysis, the lower limit of significance is considered as having a p-value of 0.05, which is equivalent to an LLMq of 0.950 or 95%.

By leveraging LLM applications, researchers can augment their expertise while benefiting from multiple data sources to analyze transcripts and identify patterns and connections that might otherwise go unnoticed.

References

Banerjee, M., Capozzoli, M., McSweeny, L., Sinha, D.: Beyond kappa: a review of interrater agreement measures. Can. J. Stat. 27, 3–23 (1999)

Bhardwaz, S., & Kumar, J. (2023). An Extensive Comparative Analysis of Chatbot Technologies-ChatGPT, Google BARD and Microsoft Bing. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 673-679). IEEE.

Bloch, D. A., & Kraemer, H. C. (1989). 2 x 2 kappa coefficients: measures of agreement or association. *Biometrics*, 269-287.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46 (1960)

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Dabney, K. P., & Tai, R. H. (2014). Factors associated with female chemist doctoral career choice within the physical sciences. *Journal of Chemical Education*, *91*(11): 1777-1786.

Dey, I. (2003). *Qualitative data analysis: A user friendly guide for social scientists*. Routledge.

Elo, S., & Kyngäs, H. (2008). *The qualitative content analysis process. Journal of advanced nursing, 62(1), 107-115.*

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5): 378-382.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd-workers for text-annotation tasks. arXiv preprint arXiv:2303.15056v1.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274.

Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. *arXiv preprint arXiv:2301.08745*.

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics *33*, 159-174

Liljequist, D., Elfving, B., Skavberg Roaldson, K. (2019). Intraclass correlation- a discussion and demonstration of basic features. *PLoS One*, *14*(7): e0219854.

Linacre, J. M. (1989). *Many-faceted Rasch measurement* (Doctoral dissertation, The University of Chicago).

Marques, J. F., & McCall, C. (2005). The Application of Interrater Reliability as a Solidification Instrument in a Phenomenological Study. The Qualitative Report, 10(3), 439-462.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3): 276-282.

Monteith, B; Sung, M. (2023). Unleashing the Economic Potential of Large Language Models: The Case of Chinese Language Efficiency. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.23291831.v1

Nichols, T. R., Wisner, P. M., Cripe, G., Gulabchand, L. (2011). Putting the kappa statistic to use. *The Quality Assurance Journal*, *13*(3-4): 57-61.

Pan, W., Chen, Q., Xu, X., Che, W., & Qin, L. (2023). A preliminary evaluation of chatgpt for zero-shot dialogue understanding. *arXiv preprint arXiv:2304.04256*.

Patel, S. B., & Lam, K. (2023). ChatGPT: the future of discharge summaries?. *The Lancet Digital Health*, *5*(3), e107-e108.

Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., … & Tao, D., (2023). Towards making the most of the ChatGPT for machine translation. arXiv preprint arXiv:2303.13780v1.

Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, *13*(9), 5783.

Rao, A., Kim, J., Kamineni, M., Pang, M., Lie, W., & Succi, M. D. (2023). Evaluating ChatGPT as an adjunct for radiologic decision-making. *medRxiv*, 2023-02.

Schwandt, T. (1997). Qualitative Inquiry: A dictionary of terms. Newbury Park, CA: Sage. (ISBN: 02546).

Stojanov, A. (2023). Learning with ChatGPT 3.5 as a more knowledgeable other: an autoethnographic study. *International Journal of Educational Technology in Higher Education*, *20*(1), 35.

Sun, S. (2011). Meta-analysis of Cohen's kappa. *Health Services and Outcomes Research Methodology*, *11*, 145-163.

Ubani, S., Polat, S. O., & Nielsen, R. (2023). ZeroShotDataAug: Generating and Augmenting Training Data with ChatGPT. *arXiv preprint arXiv:2304.14334*.

Van Booven, D. J., Kuchakulla, M., Pai, R., Frech, F. S., Ramasahayam, R., Reddy, P., ... & Arora, H. (2021). A systematic review of artificial intelligence in prostate cancer. Research and reports in urology, 31-39.

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., ... & Han, W. (2023). Zero-shot information extraction via chatting with ChatGPT. *arXiv preprint arXiv:2302.10205*.