

Evaluation of LLMs and Other Machine Learning Methods in the Analysis of Qualitative Survey Responses for Accessible Engineering Education Research

Xiuhao Ding, University of Illinois at Urbana - Champaign

Meghana Gopannagari, University of Illinois at Urbana - Champaign

Kang Sun, University of Illinois at Urbana - Champaign

Alan Tao, University of Illinois at Urbana - Champaign

Delu Louis Zhao

Sujit Varadhan, University of Illinois at Urbana - Champaign

Sujit Varadhan is a Junior at the University of Illinois at Urbana-Champaign majoring in Computer Science. He is an undergraduate research assistant as well as a frontend developer on ClassTranscribe.

Bobbi Lee Battleson Hardy, University of Illinois at Urbana - Champaign

David Dalpiaz, University of Illinois at Urbana - Champaign

Dr. Chrysafis Vogiatzis, University of Illinois at Urbana - Champaign

Dr. Chrysafis Vogiatzis is a teaching associate professor for the Department of Industrial and Enterprise Systems Engineering at the University of Illinois Urbana-Champaign. Prior to that, Dr. Vogiatzis was an assistant professor at North Carolina Agricultural and Technical State University. His current research interests lie in network optimization and combinatorial optimization, along with their vast applications in modern socio-technical and biological systems. He is serving as the faculty advisor of the Institute of Industrial and Systems Engineers, and was awarded the 2019 and 2023 Faculty Advisor award for the North-Central region of IISE. Dr. Vogiatzis was awarded ASEE IL/IN Teacher of the Year in 2023.

Prof. Lawrence Angrave, University of Illinois at Urbana - Champaign

Dr. Lawrence Angrave is an award-winning computer science Teaching Professor at the University of Illinois Urbana-Champaign. He creates and researches new opportunities for accessible and inclusive equitable education.

Dr. Hongye Liu, University of Illinois at Urbana - Champaign

Hongye Liu is a Teaching Assistant Professor in the Dept. of Computer Science in UIUC. She is interested in education research to help students with disability and broaden participation in computer science.

Evaluation of LLMs and other machine learning methods in the analysis of qualitative survey responses for accessible engineering education research

Abstract

This research paper provides insights and guidance for selecting appropriate analytical tools in engineering educational research. Currently, educators and researchers face difficulties in gaining insights effectively from free-response survey data. We evaluate the effectiveness and accuracy of Large Language Models (LLMs), in addition to the existing methods that employ topic modeling, document clustering coupled with Support Vector Machine (SVM) and Random Forest (RF) approaches, and the unsupervised Latent Dirichlet Allocation (LDA) method. Free responses to open-ended questions from student surveys in multiple courses at University of Illinois Urbana-Champaign were previously collected by engineering education accessibility researchers. The data (N=129 with seven free response questions per student) were previously analyzed to assess the effectiveness, satisfaction, and quality of adding accessible digital notes to multiple engineering courses and the students' perceived belongingness, and self-efficacy. Manual codings for the seven open-ended questions were generated for qualitative tasks of sentiment analysis, topic modeling, and summarization and were used in this study as a gold standard to evaluate automated text analytic approaches. Raw text from open-ended questions was converted into numerical vectors using text vectorization and word embeddings and an unsupervised analysis using document clustering and topic modeling was performed using LDA and BERT methods. In addition to conventional machine learning models, multiple pre-trained open-sourced local LLMs were evaluated (BART and LLaMA) for summarization. The remote online ChatGPT closed-model services by OpenAI (ChatGPT-3.5 and ChatGPT-4) were excluded due to subject data privacy concerns. By comparing the accuracy, recall, and depth of thematic insights derived, we evaluated how effectively the method based on each model categorized and summarized students' responses across educational research interests of effectiveness, satisfaction, and quality of education materials. The paper will present these results and discuss the implications of our findings and conclusions.

Introduction

While open-ended questions in survey studies are very valuable in providing authentic and nuanced insights toward the respondents' opinions, it's often challenging to analyze such questions in large quantities. Not only the time and cost required in large open-ended questions' analysis make it difficult for researchers, but also the complexity of such free text poses significant difficulties [1]. For example, qualitative analyses need to consider factors such as

multiple data formats, contextual requirement for interpretation, subjectivity of the researchers as part of the instrument, and the holistic relationship between the design and the analysis.

Therefore, manual analysis of qualitative data requires substantial training for researchers and may need multiple passes of the analysis via several researchers for better accuracy [1, 2]. Even with the access to Qualitative Data Analysis Software, researchers can not fully automate their analysis of open-ended survey questions [3]. Because of these barriers, it's a common practice that researchers would opt not to design open-ended questions or simply quote a few sentences from their impression of the free responses.

To address some of these challenges, recently researchers have explored the use of machine learning (ML) tools to help automate the analysis of free responses, especially in the tasks of sentiment analysis and topic modeling. For example, Nawaz et al has used the Support Vector Machine (SVM) and Random Forest models to analyze the large free responses in the National Student Surveys in Northern England [4]. Nanda et al have utilized an unsupervised Latent Dirichlet Allocation (LDA) model to generate topics out of free responses to open-ended questions in large surveys about MOOC learning platform [5]. With the advent of large language models (LLMs), more and more efforts have been made in trying to analyze free text with such AI analytic tools [6, 7]. However, currently no studies have been devoted to applying LLM methods for the analysis of open-ended questions in large surveys in the field of engineering education research.

In this paper, we present the evaluation of the application of multiple LLM methods and other ML approaches in comparison with a manual analysis of free responses to open-ended survey questions based on the grounded theory. We used the free response data from a previous education research survey [8] where the questions were designed to collect college STEM students' feedback on an innovative digital note based pedagogy to support a Universal Design for Learning approach [9]. We did not include the popular LLM, ChatGPT due to subject data privacy concerns, and instead selected open-sourced local LLMs that could be run locally.

The rest of the paper is organized as follows. First, we summarize the background of this study including related work; in the method section we describe the research question, the manual coding and the ML methods for sentiment analysis, topic modeling and summarization respectively; for each of the three analysis types we report details of the comparison between our manual method and ML-based approach; next, we summarize our findings in the conclusion and discuss the advantages and limitations of these methods. Lastly, we present suggestions on potential future work and provide the resources from this study.

Background

Grounded theory and Manual coding approaches

Grounded theory is a structured, yet flexible methodology to construct theory from the ground up, by systematically obtaining and analyzing data inductively and comparatively [10]. When the subjective nature of human interpretation and the theory's inductive structured data analysis meet, grounded theory uses the constant comparative method to reach intersubjective consensus, approximating accurate interpretation, induce increasingly more abstract concepts, and create theories from the detailed and often messy qualitative data [11]. Such structured constant comparisons generate new interpretations and rounds of new comparisons between different

levels of data and codes as well as between different coders, making the grounded theory an iterative process of rigorous knowledge production [10].

Grounded theory requires three fundamental phases of coding, specifically, open, axial, and selective coding phases. Open coding usually includes verbatim quotes from the participants' words to simplify the qualitative data and reduce the conversational flow of the raw qualitative data. By comparing different open codes, axial coding focuses on finding connectable open codes to form categories or core meaningful chunks. Selective coding identifies major themes from axial coding as more general meaning sections, themes, or theories. Although different branches of grounded theory have different naming systems, what is shared in common is the inductive process for more concrete descriptive qualitative data to form increasingly abstract categories, themes, and theoretical threads [12, 13].

To make the analysis results accurate and valid, qualitative researchers employ different techniques including question triangulation, skills training of coders, multiple coders and rounds of discussions or meetings to reach intersubjective consensus or compromises by balancing out different perspectives. As the subjective instruments of the study, researchers also used intermediary steps such as memos, categorizing strategies, field notes, and narrative analysis to help them reach final results of the analysis, although such intermediary steps are often not reported. Since validity of findings is presented in narrative forms, the narrative often emphasizes alternative explanations to form tightly woven accounts against possible alternative explanations [14].

The data analysis method of grounded theory, through constant comparisons, memo keeping, reflection notes writing, fits the objectives of the present study of finding whether the machine learning-based data analysis resulting in similar and usable results as compared with the analysis results from the inductive process of the grounded theory. Raised as a theory-construction method that takes data as the basis for theories to emerge, grounded theory has a unique fit with the machine learning-based analysis approach in that both are inductive in nature.

Machine learning (ML)-based or mixed approaches

Previously researchers have conducted ML-based analysis on the sentiment of financial news reports or labeled information of survey questions [7]. Sentiment analysis is a classification task that can be handled by manual labeling of a small set of the responses or automatic labeling script using relevant libraries and then finished by the classification methods in ML [4, 7]. Nawaz et al has used SVM and RF to analyze the sentiment or labeled responses of large survey questions and found the results varied across different questions and SVM method performed best overall across the questions [4]. The unsupervised Latent Dirichlet Allocation (LDA) method is often used for topic analysis and the Cross-Validation (CV) coherence scores [26] is a standard approach to measure the relevance of a found topic or list of topics [5, 15].

LLM usage in open ended question analysis

Research in this area has been very recent. Tang et al have applied LLMs in the analysis of a number of medical review papers to help generate concise helpful medical reviews for diagnosis purposes [6]. They manually summarized fifty three medical reviews in six clinical domains and

applied ChatGPT and GPT3.5 to the summarization with either the abstracts or the main results section of the abstracts as input. The comparison between the LLM results and the expert coding demonstrated there is potential for harmful due to the misinformation in the LLMs' results. Further, the automatic scores of the LLM models do not correlate well with the quality of summary.

Another study used an LLM to analyze student feedback on the quality of teaching in a series of computer science classrooms [16]. The LLaMA model was used to analyze a sample of 2075 comments from students to conduct a deductive analysis. To keep student information confidential, the LLaMA model was run on a local machine. The model labeled each student's data with Accessibility, Collaboration, Communication, Clarity, Relevance, Feedback, Community, Engagement, or Quality based on which topic was the most relevant to the student comment. To validate the responses, the research team manually coded the responses independent of the LLaMA model, and the codings were compared. This approach revealed that although LLMs are not fully accurate, they can effectively tag specific comments with codes in deductive analysis. Another study also attempted to use large language models for deductive analysis, but they provided a codebook to help the LLM categorize data points [17]. The study's dataset was a collection of children's questions, and expert psychologists coded whether the questions had a high or low question complexity. Researchers separated the experiments into three groups: expert coding (manual coding), example-based coding (by LLM), and codebook-based coding. The results were assessed based on how similar the codings were with the expert results. The study found that the most accurate method of coding was using an expert codebook (61%) compared to example-based coding, but both results performed poorly compared to the expert analysis.

Data from a survey on UDL-based teaching pedagogy and intervention method

Universal Design for Learning (UDL) is an inclusive pedagogical framework and set of principles to improve learning for all students by emphasizing the importance of accommodating multiple modes of student learning, action, and engagement [18]. UDL recommends multiple modalities in the three key areas of learning: Representation, Engagement, and Action-expression.

Representation is the means through which content is presented to students [19]. An inclusive framework can benefit all students (e.g., by creating accurate captions that can be indexed and searched), but can be particularly valuable for students with disabilities. Engagement refers to the means through which students engage with the content. Action-expression refers to the manner in which students are assessed.

In this paper, we analyzed free responses of the open-ended survey questions collected from two Engineering courses where novel digital notes were implemented via the UDL based video teaching platform ClassTranscribe [8]. These digital notes consisting of both text and visual elements were automatically generated from lecture videos. The notes were then separated into distinct chapters and sub-chapters that include many images, image descriptions, text and hyperlinks that can be edited or shuffled by the instructor. Researchers have used it as a new accessible instructional tool and found it improved the course content accessibility and learning outcomes [8].

Methods

Research questions and aims

We conducted the analyses and experiments to address the following research questions:

1. How effectively can the current machine learning methods perform in the sentiment analysis of open-ended survey questions compared with manual analysis based on grounded theory?
2. How effectively can the current machine learning methods including the local LLM models perform in the topic modeling of open-ended survey questions compared with manual analysis based on grounded theory?
3. How effectively can the local LLM models perform in the summary generation of open-ended survey questions compared with manual analysis based on grounded theory?
4. What engineering methods are entailed in the above applications?

Open-ended Questions in a previous survey for a UDL pedagogy method

The qualitative data used in this paper were from a series of surveys in the aforementioned digital note project. In total, students were surveyed three times across different semester checkpoints. In this paper, the following subset of open-ended questions from the later two surveys were selected for analysis:

1. Evaluation 1: What benefits do you see from using digital notes?
2. Evaluation 2: What complaints do you have for digital notes in helping you learn new materials? (or frame as how satisfied are you?)
3. Errors: What errors in the digital notes have you noticed after using it for the past weeks?
4. Satisfaction: How do you describe your experience of using digital notes as a whole? And how do you describe your experience of using certain features of digital notes?
5. Adaptation: Have you developed new ways of learning with digital notes? If so, can you describe what that type of learning is and in what way is it different from your previous learning methods?
6. Social Learning: Do you feel digital notes promote, hinder, or do not do much to your interaction with instructors or fellow students? Can you explain how that works for you?
7. Effective: What parts/sections of the digital notes were the most effective in helping you learn? And how were they helpful?

Manual coding

Responses from the seven open-ended questions were compared with the course participation records. There were respondents who did not use the digital notes but chose to answer the questions, so we removed their responses which resulted in 129 responses in total. We chose to code the text with the inductive approach because the questions in the survey are rather focused. The inductive approach also fits well with the three levels of codings specified in grounded theory. The first coder started with 30 complete sets of data on all seven questions and completed the open, axial, and selective coding. Of the total study sample ($n = 129$), the selection of 30 exceeded the usually required percentage [20]. The second coder selected 2 questions and completed the open, axial, and selective coding on the whole set of data ($n = 129$). The selective coding results and summaries from two coders were compared through meetings. The discrepancies were analyzed. Instead of requiring strict uniformity in expression, the two coders reached the agreement to allow for differences in expression as long as the selective codings from two coders share the same themes. Such a strategy of allowing for semantic expression differences fits the overall comparison between manual coding and machine learning based method.

Other research team members were then trained in the grounded theory coding methods and completed the analysis of the remaining questions' all responses, each taking about two open-ended questions. In training meetings, three principles were emphasized to achieve rigorous coding results: First, during the coding process, reflective paragraphs on the assumptions, thoughts, and changes in open, axial, and selective coding processes were written by each coder to trace each moment and to show if there are overall changes in coding. Second, all coders were required to write out the summary from the responses to each question immediately after completing the coding process when their memories are fresh. Third, all coders were instructed to conduct the coding independently before running or seeing the results from the machine learning algorithm.

After the machine learning based analysis is completed, the results are compared with the qualitative data, open, axial, and selective codes to evaluate the machine learning based analysis results. We adopted Tang et al's (2023) evaluation scheme by comparing the manual coding and machine learning based coding results in three dimensions, namely, coherence, factual consistency, and comprehensiveness. Coherence means the capability to summarize qualitative data input into a coherent piece of information with cohesion. Factual consistency evaluates whether each meaning unit in the summary is backed up by the qualitative data. Importantly, we also added whether information found in the source qualitative data is represented in the summary. Comprehensiveness evaluates the extent to which the summary reached the comprehensiveness of the source qualitative data [6]. We dropped "harmfulness" from Tang et al.'s evaluation scheme since the data in this project does not have the clear physiological harms in the biomedical studies. We adopted a 5-point Likert scale with 1 being "the least satisfied" and 5 being "the most satisfied." In addition, we created a column for coders to jot down their short reflections in evaluating each prompt and response. Reflections typically included how we assigned our evaluation points, with what backup information, what implications and connections we see during the coding period. Using just-in-time reflections offers information to explain and justify the Likert scale points. Since all researchers conducted their evaluations separately, such reflections provide insights in our interpretations of the quantitative results and in informing the

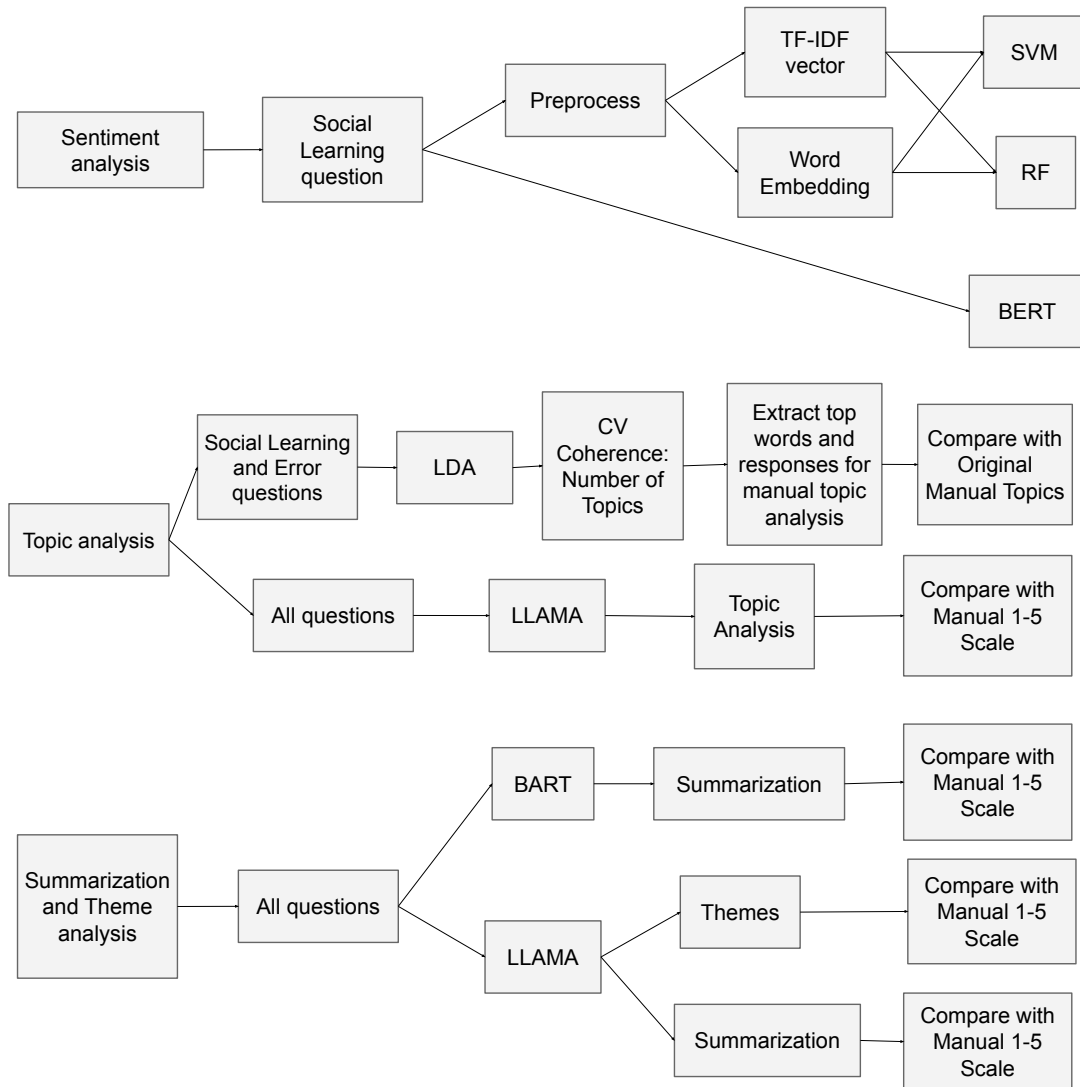


Figure 1: Flowchart of methods for 3 tasks: sentiment analysis, topic analysis, and summarization.

discussions section. The machine learning based analysis is detailed below.

Machine learning based analysis

1. ML methods used in this study

We analyzed our survey questions related to the user experience of digital notes through three distinct dimensions: sentiment analysis, topic modeling, and summary generation. Figure 1 presents the details of our workflow. In the realm of sentiment analysis, we explored the task of label prediction suitable for supervised learning models such as SVM, RF, and BERT. This exploration aims to assess the feasibility of training on a limited subset of survey responses to predict the sentiment of the majority of the remaining responses at a larger scale. Secondly, for tasks utilizing unsupervised learning models, such as topic modeling and summary generation,

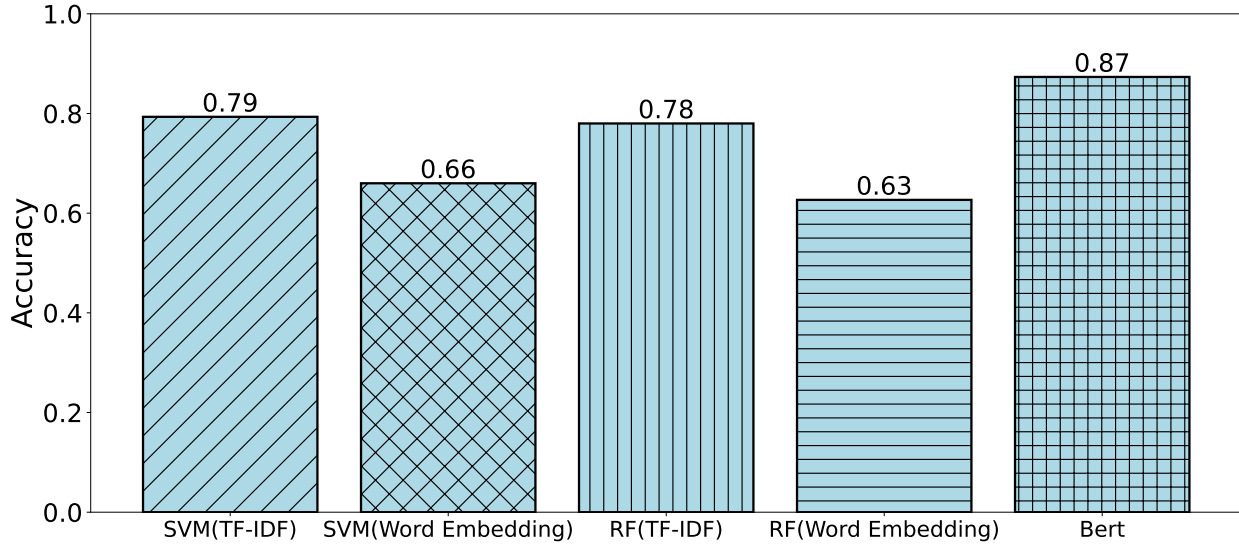


Figure 2: Average accuracy of different models for the 'Social Learning' question.

we have evaluated the LDA, BART, and LLaMA models for the potential to generate thematic insights and comprehensive summary paragraphs within a substantial corpus of survey responses.

2. Sentiment analysis

In this subsection, we focus on responses to the “Social Learning” question. Among the open-ended questions from our survey, this is the question that has responses closest to the ordinal data type with categories. This question investigates whether using digital notes by students promotes, hinders, or does not change their interactions with professors or other students. We chose this question for label prediction within our sentiment analysis because the responses can be distinctly categorized into three types: promoting, hindering, and having no influence. This categorization enables us to label each response effectively manually for training and testing in our label prediction analysis. Following the exclusion of irrelevant or invalid responses, we randomly selected 100 student responses from the “Social Learning” question. These were then labeled as promoting (positive), hindering (negative), or having no influence (neutral). The final dataset included 53 neutral, 29 positive, and 17 negative responses, providing a well-balanced foundation for our label prediction in sentiment analysis.

In our approach for the label prediction, we applied two widely used traditional ML methods, Support Vector Machine (SVM) and Random Forest (RF), alongside a more innovative method, the BERT (Bidirectional Encoder Representations from Transformers) language model. SVM is renowned for its efficiency in high-dimensional spaces, making it ideal for text classification tasks [21]. At the same time, Random Forest is a powerful ensemble method known for its high accuracy [22]. BERT represents a breakthrough in natural language processing, utilizing deep learning to grasp the context and subtleties of language [23]. For the SVM and RF models, we utilized two primary methods for preprocessing the data: TF-IDF vectorization and GloVe word embedding. TF-IDF (Term Frequency-Inverse Document Frequency) vectorization is a technique that reflects the importance of words in a document relative to a corpus [24]. In our TF-IDF implementation, we customize the list of stopwords, notably excluding the word “not” to preserve

SVM/BERT	Precision		Recall		F1-score		Support
	SVM	BERT	SVM	BERT	SVM	BERT	
Negative responses	1.00	0.67	0.29	0.86	0.44	0.75	7
Neutral responses	0.82	1.00	0.96	0.93	0.89	0.96	28
Positive responses	0.87	0.93	0.87	0.93	0.87	0.93	15
Accuracy					0.84	0.92	50
Macro avg	0.89	0.87	0.71	0.91	0.73	0.88	50
Weighted avg	0.86	0.93	0.84	0.92	0.82	0.92	50

Table 1: Experience/Usefulness of aggregated UDL practice types – All students

its critical negation context. Word embedding, on the other hand, provides a dense representation of words in a multi-dimensional space, capturing semantic relationships between them [25]. Finally, we divided the dataset into a 50% randomized training set and a 50% testing set.

We compared the accuracy of the different ML methods including methods that utilized different preprocessing approaches in the sentiment analysis and found the BERT Model had the highest overall accuracy score. The two traditional classification methods SVM and RF performed similarly regarding accuracy. Figure 2 presents the average accuracy scores of different models on the question of social learning, utilizing two preprocessing methods—TF-IDF vectorization and GloVe word embedding (notably, the BERT model does not require preprocessing), obtained by averaging the accuracy scores from multiple random splits of the train-test sets. The employment of TF-IDF with SVM and RF models yielded higher accuracy rates than using word embedding.

We also calculated the F1-scores for these methods while investigating the number of negative responses for the question for imbalanced datasets are a common occurrence in survey research. The number of negative responses was significantly lower than that of positive and neutral responses across the overall responses, as well as within both the training and testing sets. Table 1 demonstrates that compared with the traditional statistical models, the large language model BERT is better equipped to handle imbalanced datasets. For example, for the negative responses, BERT had a higher recall than SVM. Also, with a minimal amount of training data, there was a notable improvement in the F1 score when compared to traditional models such as SVMs.

3. Topic modeling

We conducted a topic modeling analysis on at least two questions' responses among the seven open-ended questions. These included responses to a question regarding social learning: 'Do you feel digital notes promote, hinder, or do not do much to your interaction with instructors or fellow students? Can you explain how that works for you?' Another question focused on errors encountered, 'What errors in the digital notes have you noticed after using them for the past weeks?'

We first employed Latent Dirichlet Allocation (LDA) models for topic modeling, adapting our methodology based on the nature of the survey questions. Specifically, for questions with evident

Error question	Top-10 Topic Words	Manual Topic Theme
Topic 1	error, noticed, grammatical, significant, subtitle, digital, note, sometimes, grammar, formatting	Noticed grammatical, subtitle, formatting errors
Topic 2	error, nothing, really, usually, see, word, note, would, like, sentence	Nothing really stood out for errors
Topic 3	mathematical, sometimes, read, wrong, equation, format, pause, occasionally, confusing, everything	Mathematical formatting sometimes wrong
Topic 4	typo, transcript, slide, sentence, page, poll, word, lecture, theta, estimation	Typos in transcript
Topic 5	dialogue, would, paragraph, also, mistake, lecture, typo, note, noticed, problem	Problem with understanding lecture dialogue, typos
Topic 6	lecture, note, error, likelihood, better, yet, happens, function, example, slide	Content specific
Topic 7	sometimes, transcript, equation, incorrectly, interpreted, would, confusing, math, completely, accurate	Content specific
Topic 8	none, lecture, written, math, spoken, sentence, place, mathematical, convenient, notation	Confusing sentences and mathematical notation from spoken lectures
Topic 9	obvious, two, note, viewed, observed, digital, error, none, typo, verbal	Content specific
Topic 10	many, sometimes, material, error, verbal, helpful, speaking, alright, naturally, okay	Verbal expressions from speaking are not helpful in transcriptions
Topic 11	error, note, sometimes, grammar, digital, really, would, lecture, helpful, think	Sometimes digital notes from lecture have grammar errors
Topic 12	much, error, sometimes, take, mainly, largely, inconsequential, figure, time, wrong	Mainly inconsequential errors
Topic 13	sometimes, note, work, major, link, besides, issue, really, none, grammar	Content specific

Table 2: Topics identified by LDA and phrased by an author for the “Errors” question.

subgroups, such as the “Social learning” question, we used our manual methods to categorize responses according to sentiment (promote, hinder, neutral) before applying LDA to each subgroup individually. This nuanced approach prevented the conflation of responses from different subgroups into an indistinct topic [4]. For questions that do not exhibit such subgroup distinctions, we applied LDA across the entire set without prior categorization.

To determine the appropriate number of topics for each run of LDA, we used Cross-Validation (CV) coherence scores, setting our target range from 5 to 15 topics. This decision was based on our a priori belief that fewer than 5 topics might oversimplify the analysis, while more than 15 topics could lead to unnecessary excessive detail.

Upon analyzing the top words associated with each topic generated by LDA, we often found these words challenging to interpret due to their abstract nature [27]. To enhance interpretability, we pinpointed the responses that were the most indicative of each topic by selecting those with the highest probability scores. These scores reflected the likelihood that a given response is strongly associated with a specific topic, thereby allowing us to identify the most representative responses for each topic accurately. Based on these responses and the top words, we then manually crafted concise topic phrases that more clearly conveyed the thematic content of each topic [5].

Table 2 presents our analysis of the “Error” question responses using the LDA method. The table enumerates both the words with the highest probability for each topic and the manually defined topic phrases based on the highest probability words and responses (we do not show the original responses with the highest probability in this table). Notably, certain topic phrases were designated as “content-specific” because, for some topics, only one or two responses exhibited a significantly high probability of association with the topic, while the probabilities for other top possibility responses sharply declined (e.g., falling below 0.2). Consequently, these were not treated as separate topics. We further conducted a comparative analysis of these LDA-based manually defined topic phrases against our original manual coding (axial coding), assessing them across coherence, factual consistency, and comprehensiveness dimensions. The coherence of these topics was ensured by their manual curation. Their consistency was guaranteed as they were synthesized from the responses with the highest topic probabilities, ensuring factual alignment with the original submissions. Upon comparison for comprehensiveness, it was observed that the LDA-derived topics substantially cover all topics from the original manual coding. This comparison attests to the high quality of topics manually generated from LDA, albeit necessitating considerable manual work.

For the question “Social Learning” responses, our decision to conduct Latent Dirichlet Allocation (LDA) on sub-groups (positive, negative, neutral) rather than on the entire set of responses was inspired by methodologies outlined in previous research [4]; upon implementation, we indeed found this approach to be superior. Specifically, analyzing the dataset as a whole with LDA often led to the amalgamation of responses from different sentiment sub-groups into identical topics, thereby diluting the clarity of thematic distinctions. Conversely, applying LDA to each sentiment-based sub-group separately allowed for a clearer identification of theme-specific phrases for each topic, as predicted.

Next, we employed LLaMA, a large language model, to generate topics for each question’s responses in the survey. To ensure that our analysis could be replicated on other common computers, we used the LLaMA 2, 7B model for our experiments. To keep the user information

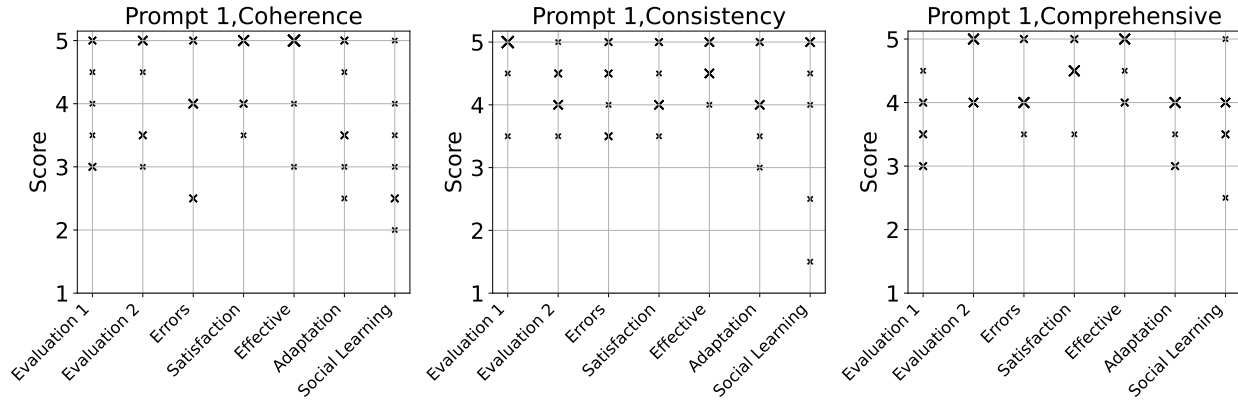


Figure 3: Evaluation of results using LLaMA Prompt 1 (Topic modeling).

private, we used a local version of the model using Ollama open source software [28]. Additionally, each prompt in the experiment was run twice to gain a better understanding of any variation between responses and to reduce the impact of randomness on the results. After the result of each trial was recorded, the model was restarted to prevent older trials from impacting later trials.

To conduct topic modeling using the LLaMA model, we asked the model to categorize the students' responses into groups. The prompt for this task (prompt1) was "Group the responses into categories" We compared the topics from the manual coding with LLaMA's reports for the categories of topics. We used three numeric scores to rate each LLaMA response on a scale of 1 to 5 with respect to three dimensions: Coherence, Factual Consistency, and Comprehensiveness as mentioned previously [6]. Seven members of the research team independently analyzed each LLaMA response, and the ratings were averaged to generate an overall set of ratings for each response. In contrast with the LDA method, we implemented topic modeling with LLaMA on all seven question responses.

For the comparison of LLaMA topics with the manually identified topics, the seven team members scored the answers of LLaMA using Prompt-1 as presented in Figure 3. The average scores of the referees for coherence ranged from 3.21 ("Social Learning") to 4.57 ("Effective") depending on the question. The median of the scores for coherence was 4.0. The average scores of the referees for consistency ranged from 3.93 ("Social Learning") to 4.71 ("Evaluation1") across the questions. The median of the scores for consistency was 4.29. The average scores of the referees for comprehensiveness ranged from 3.64 ("Adaptation") to 4.64 ("Adaptation") depending on the question. The median of the score for comprehensiveness was 4.21. The question "Social Learning" had the worst-scored response from the LLaMA query. When we investigated into the reflections of the referees, we found that one referee identified that the result "M18" (a result by machine code) was not a negative response, but LLaMA categorized it as a negative; and most of the responses that LLaMA identified as negative were not negative, they were either positive or neutral, other two referees noticed LLaMA missed the information about negative or neutral responses.

4. Summary generation

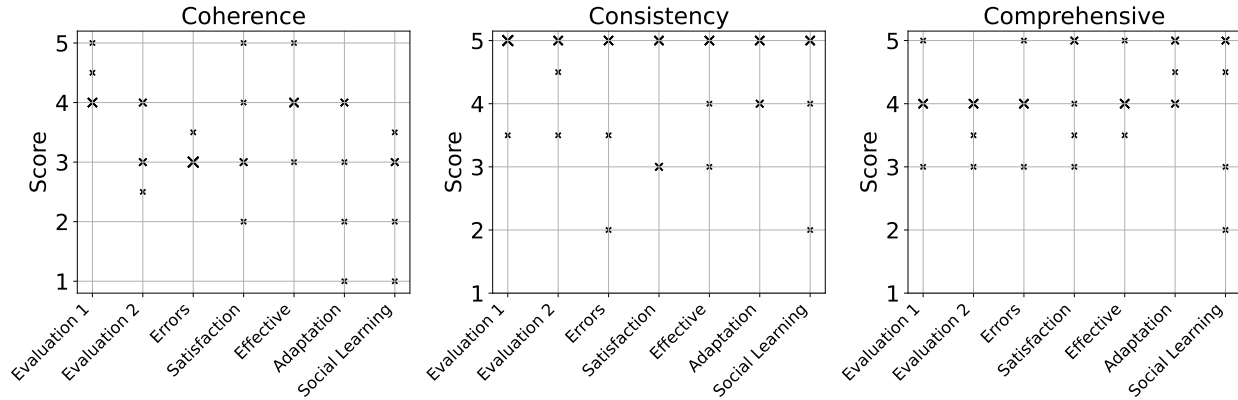


Figure 4: The scores for the evaluation of BART summary results.

For the task of summary generation, we tested whether local LLM models could accurately provide summaries analogous to the summaries manually coded based on the grounded theory method, applying this test across all the seven questions in our surveys. Two LLM models were selected for this task: BART and LLaMA. We used BART because of its strengths in generating contextually relevant summaries [29]. LLaMA is a large language model, and it's well known for its ability to comprehend text to generate summaries [30]. Additionally, it can be quickly run on many machines using the application Ollama [28]. To protect the confidentiality of student responses, we chose these models because they could be run locally without sending data to APIs or to the cloud.

First, we used the BART model for the task of summary generation. The BART model is incapable of recognizing prompts; it interprets prompts as part of the text requiring summarization. Consequently, for the BART model, we refrained from using prompts and instead utilized the original summary model. To enhance the accuracy of the article summarization, we opted not to apply sampling in BART's text generation process. Additionally, due to the BART model's limitation on the number of tokens that can be inputted at a single query, and given the large number of responses for each question in our survey, we divided the responses into smaller clusters to comply with this limitation. Subsequently, we processed each cluster through the BART model and ultimately amalgamated the summaries of each cluster.

Five referees gave scores of the BART query results for all seven questions regarding summary generation using the manual code as the reference. The details are illustrated in Figure 4. The average scores of the referees for coherence ranged from 2.5 (“Social Learning”) to 4.3 (“Evaluation1”) depending on the question. The median of the scores for coherence is 3.3. The average scores of the referees for consistency ranged from 4.1 (“Errors”) to 4.7 (“Evaluation 1”) across the questions. The median of the scores for consistency was 4.4. The average scores of the referees for comprehensiveness range from 3.7 (“Evaluation2”) to 4.5 (“Adaptation”) depending on the question. The median of the score for comprehensiveness was 4.0.

The LLaMA experiments for the summarization task were conducted in the same manner as the LLaMA experiments for the topic analysis task described in the previous section. For each survey question, each prompt was asked twice, and the results were recorded. We used two prompts focused on summarization (Prompt2) and gathering themes (Prompt3). First, for each survey

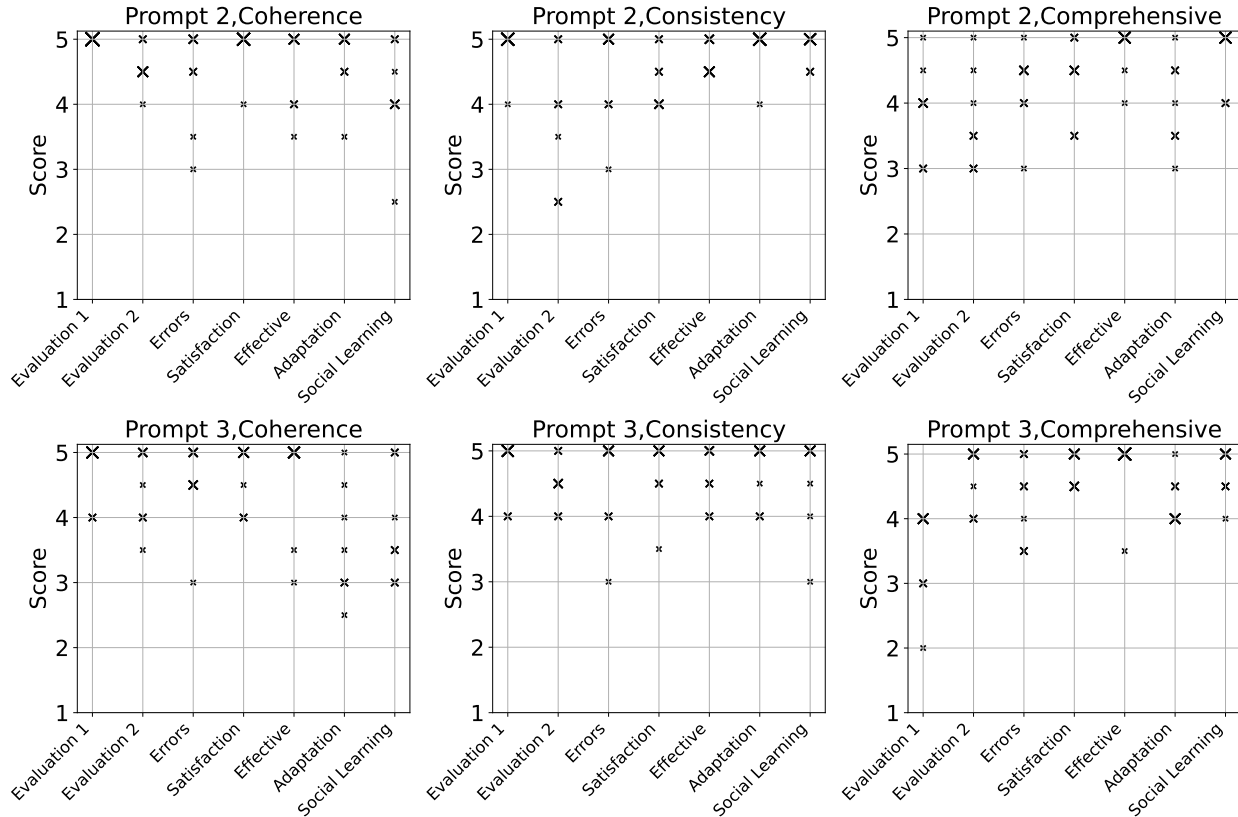


Figure 5: Evaluation of results using LLaMA Prompt 2 (summary) and Prompt 3 (theme).

question, we asked the model the following prompt: “The responses below are responses from different students regarding digital notes. Create a summary paragraph with 5 sentences that contains all of the main ideas from the responses below.” This prompt was chosen to provide more context for the large language model and to provide a specific format for the LLM to output data. The response was then compared with the summaries through the manual codings. Additionally, to gather themes, we asked the model to “Describe 10 themes from the data below”. The purpose of this prompt was to compare the themes generated by LLaMA with the themes generated by the grounded theory approach.

The result was analyzed in a similar manner as the LLaMA results from the topic analysis. Each response was given a 3-pronged score by seven referees with respect to cohesion, accuracy, and comprehensiveness. For prompt 2 which we used to gather summaries, the details are presented in Figure 5. The average scores of the referees for coherence ranged from 4.14 (“Social Learning”) to 5.0 (“Evaluation1”) depending on the question. The median of the scores for coherence was 4.0. The average scores of the referees for consistency range from 3.786 (“Evaluation2”) to 4.86 (“Evaluation 1”, “Adaptation”, “Social Learning”) across the questions. The median of the scores for consistency is 4.71. The average scores of the referees for comprehensiveness range from 3.786 (“Evaluation2”) to 4.79 (“Effective”) depending on the question. The median of the score for comprehensiveness was 4.21. For prompt 3 which we used to gather themes, the detailed scores are shown in Figure 5. The average scores of the referees for coherence range from 3.64 (“Adaptation”) to 4.71 (“Evaluation1”) depending on the question. The median of the scores for

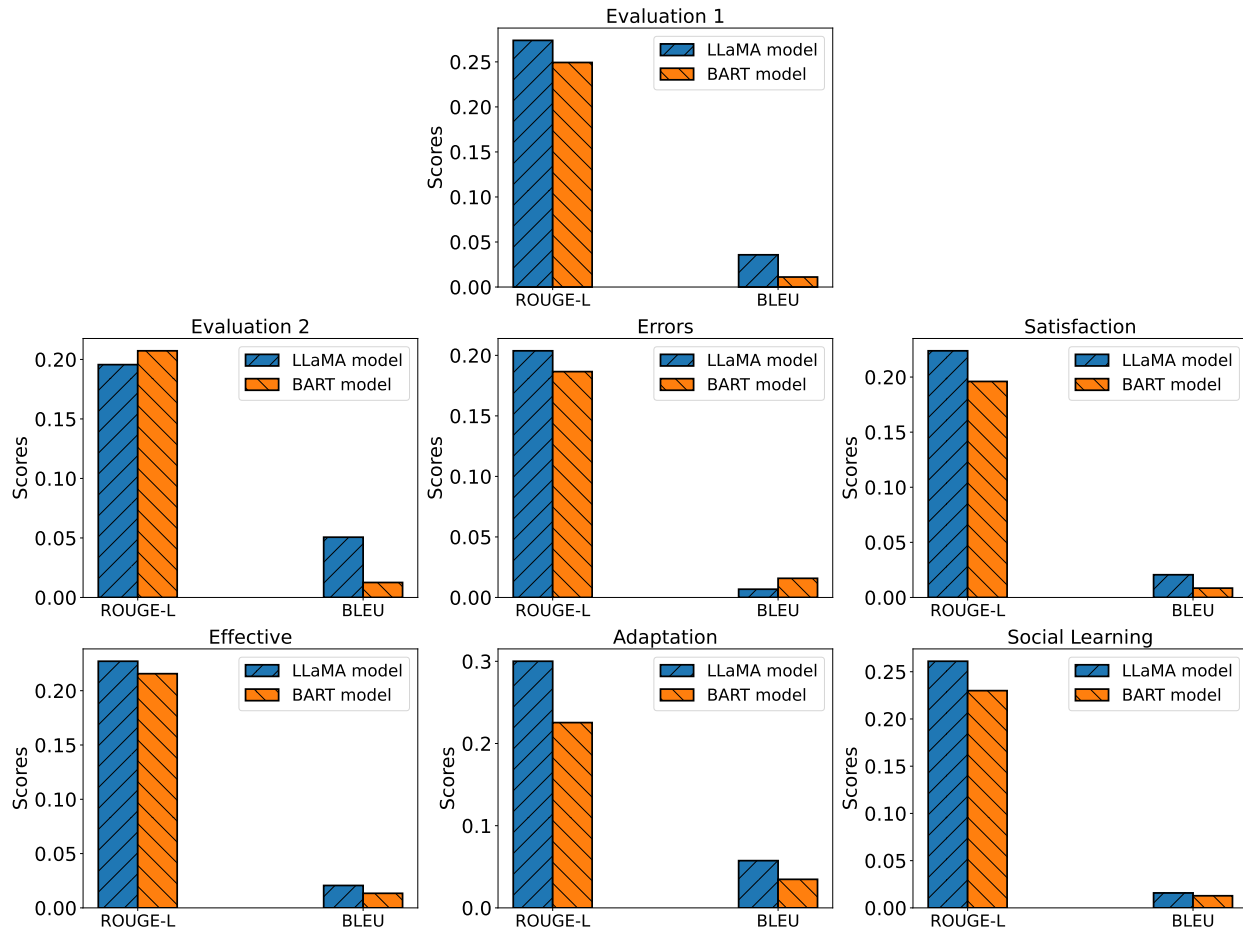


Figure 6: Rouge and Bleu score for all the questions.

coherence was 4.5. The average scores of the referees for consistency range from 3.79 (“Evaluation2”) to 4.86 (“Evaluation1”, “Adaptation”, “Social Learning”) across the questions. The median of the scores for consistency was 4.71. The average scores of the referees for comprehensiveness ranged from 3.43 (“Evaluation1”) to 4.79 (“Satisfaction”, “Effective”) depending on the question. The median of the score for comprehensiveness was 4.64. From these statistics, we saw that for different questions combined with different prompts, the results from the LLaMA model could vary significantly. For example, for the question “Social Learning”, the LLaMA model performed very well according to prompt 2 and the consistency measure, while it did not perform as well for prompt 1.

We analyzed the results and reflections of the referees who gave the highest score for a specific analysis and those who gave the lowest score. Appendix 1 demonstrates an example of 5.0 scored answers by any referee for coherence, consistency, and comprehensiveness respectively. In addition, we investigated what reflections the referees talked about during the scoring, when they gave a low score for an answer and included them in the appendix as well.

As in previous publications that evaluated the similarity of the answers with the manually coded reference text using automatic score metrics, we computed the ROUGE-L 10 and BLEU scores

for the summary answers of the two LLMs BART and LLaMA (Figure 6). The scores are generally low but we noticed LLaMA in general scored higher than BART which is consistent with the scores by the referees. The resulting low automatic scores are not surprising for they are very similar to what Tang et. al. reported for their data. It was reported in their article that the automatic scores do not correlate well with the quality of the summaries. Once again, it demonstrates the importance for human researchers to check the result and interpretation of a qualitative study.

Conclusion

We found the natural language model BERT outperformed (87% overall accuracy) the traditional SVM and RF ML methods (79% overall accuracy) regarding accuracy and gave more robust results than SVM and RF when the data has an unbalanced number of labels.

For the effectiveness of the methods we investigated in performing topic modeling, we found the LDA-generated list of topics coupled with the interpretation by our researcher matched very well with the manually identified topics. In addition to LDA, we used the state of the art large language model LLaMA to simulate topic generation with prompt1. The results have mixed qualities depending on the open-ended questions. For some questions, all three measures scored higher than 4.5 with 4.0 as “satisfied with the result”.

LLMs have been found to be suitable for summarization tasks. In our experiments, we used two LLM models, BART and LLaMA. We found the LLaMA v2 7B model outperformed BART and performed considerably well in some of the questions. For summaries from LLaMA, all but two questions received scores more than 4.1 for all three measures. For theme finding, all but three questions received scores of more than 4.2 for all three measures. For the question “Evaluation2”, the LLaMA gave helpful suggestions in addition to the themes.

In the process of comparing the results from LLM models and our manual coding based on the grounded theory, we indeed observed the commonly known issue of “Hallucination” of such AI methods, which seems to be dependent on the responses of a specific question.

There is very little manual work required to apply the BERT model for sentiment analysis, while more work is needed for the researcher to run the preprocessing for the SVM or RF methods for such free text analysis. For topic modeling, again the LLM models need little preprocessing but the LDA topic modeling requires manual work in the interpretation of generated topic lists and responses with high probabilities. The downside of LLaMA for topic modeling is that it does not always give associated responses for a particular topic. For the summarization of responses or theme finding tasks, LLM models need minimal limited preprocessing except preparing for the prompts and writing scripts to automatically generate queries. Although the BART model performed less well than the LLaMA model, it has the advantage that it allows users to select whether the summary includes as many words as possible from the original source.

In summary, we suggest machine learning based analysis/codings can be considered as valuable alternative coding methods to triangulate with partial manual analysis. We have seen better results in sentiment analysis using BERT instead of the traditional SVM and RF methods. The new LLM model LLaMA performed reasonably well in summary generation. The techniques involved in such methods are very attainable for the larger community of researchers.

Discussion

We presented a first study that applied and evaluated traditional machine learning and the new generative-AI large language model approaches in an inductive analysis of free response survey questions in a real engineering education research setting.

We found many meaningful insights through the sentiment analysis task. BERT had the highest accuracy across all of the models used for sentiment analysis (Fig. 2). These results depict how BERT performed better than the traditional machine learning models, indicating that natural language modeling methods might perform better than the traditional machine learning methods when analyzing the tone of qualitative open-ended responses. Additionally, we found that within the conventional machine learning methods, the TF-IDF preprocessing method performed better than the GloVe word embedding for both the SVM and Random Forest models.

For the summarization task, both LLaMA and BART had promising results. LLaMA performed well overall for generating summaries and identifying key themes across all three measures: accuracy, consistency, and comprehension. Some responses, such as evaluation 1 prompt 2's coherence, had a 5.0/5.0 score across all reviewers. In some cases, the model even gave suggestions to the specific complaints that students mentioned. The BART results indicate that it had average performance compared to LLaMA, and the results were very varied as most questions had a wide range of scores. However, this model was more convenient because it does not require a prompt, and it runs faster compared to other models, including LLaMA.

Our results found that together the models we investigated can contribute to a sentiment analysis and provide key themes and summaries, but these results are not reliable on their own for analysis. We recommend using machine learning methods and natural language processing when analyzing large amounts of data, but it's necessary for a human expert to verify the results to ensure that the results are accurate. We do not recommend using models to analyze data in instances where misinformation could result in harm, such as the summarization of medical texts, because models often generate incorrect interpretations and hallucinations [6].

The LLM models have the advantage that they can efficiently output multiple results from multiple runs. On our local PC without any GPU or other advanced hardware, BART required approximately 1-2 minutes to generate a complete summary. LLaMA, on the other hand, required 3-5 minutes to generate each response locally. In compliance with the grounded theory, in our manual grading/evaluation of these results in terms of coherence, consistency, and comprehensiveness, the LLM results were compared with manual coding as well as the raw data. However, it took a significant amount of time for several members to complete the task. From the evaluation we performed on LLMs, we recommend that human analysts familiarize themselves with the raw data, select a portion of the raw data to perform a manual coding task, and compare results with those from LLMs to make discrete judgments. Researchers need to consider different data types, contents, lengths, and numbers of responses, as well as the differences in prompts for they all influence the yielded reports. Therefore, we suggest a prudent approach is for researchers to complete a representative sample of the data analysis manually and use their own judgment to evaluate the results of LLM in a qualitative analysis.

One of the complex issues in analyzing qualitative data is interpretation for it is highly dependent on context. Possible differences in qualitative researchers' knowledge and theoretical

frameworks, literature awareness, and expertise all factor in the interpretation process [1]. Consequently, the iterative processes of qualitative data analysis, usually completed by multiple coders to check interpretations, make multiple discussions and negotiations a norm in analyzing qualitative data [14]. As Maxwell (2005) stated, “any qualitative study requires decisions about how the analysis will be done, and these decisions should inform, and be informed by, the rest of the [study] design” (p. 95). In the entire process of analyzing our data, we find threads in the data using both manual coding and ML including the LLM models to generate a holistic study process. We found as a comparative methodological project, both similarities and differences are potentially valuable. For example, differences between the manual coding and LLM/ML responses could be used as another valuable source of information and could provide unique perspectives for the manual coders as well as machine learning designers. Additionally, similarities between the models and manual coding responses would reaffirm the importance of key themes or sentences when creating categories and summaries.

Limitations

This study has the following limitations. First, the data size is relatively small (n=129). We speculate that small data size might make certain distinctive responses hard to group into major themes. Whether a larger data size results in more dependable responses will likely depend on the specific LLaMA models chosen, and their characteristics (e.g. context, maximum token length, and training), and is out of the scope of this study.

Secondly, the models we employed exhibit certain limitations. Specifically, for the BART model, its maximum input token count is 1024. This necessitates the segmentation of extensive survey responses into smaller clusters, with BART then tasked with generating a summary for each cluster before merging them. This approach means that the model can only summarize a portion of the responses at any one time, leading to the potential for repetitive information in the amalgamated summary. For instance, if each cluster contained a similar theme, the final merged summary may repetitively feature sentences related to this theme several times. Moreover, BART’s inability to comprehend prompts further restricts its application scope. For LLaMA, the selection of prompts is subjective, which does not guarantee the identification of the most suitable prompt. Furthermore, although LLaMA is capable of generating a specific number of themes, it lacks the ability to map responses to certain themes.

Third, typical LLM hallucination [31] appeared in our analysis. LLM hallucination has multiple causes such as source-reference divergence in training data, dependence on incomplete or contradictory data entries, stretching or twisting the logic, or adding logical relationships that are not supported by source data [32, 33]. It is so prevalent to such an extent as for us to rethink the reliance of using “coherence” as a standard for overall evaluation of LLM models. For instance, if qualitative data to questions have contradicting or multiple dimensions, possibly in contradiction with one another, then, the high coherence rates achieved on such data become questionable. The human rating of coherence might unduly overrate surface coherence. In fact, if opposing and contradicting data were used, it is possible that a lower rate of coherence means a higher rate of validity.

Because we wanted other educators to replicate our results, we chose to run the models on more typical faculty computing environments that don’t require expensive or rare hardware choices.

Further, new LLM models are constantly being published. Thus, our limited selection of LLMs, and the LLMs that we evaluated are unlikely to be as accurate as models that are now available to the readers of this paper. Similarly, our training set was limited because we used survey results with short student responses, and this could have limited the models' abilities to analyze the sentiment of key themes in data. A future study could also investigate more prompts and use prompt engineering to create more refined prompts optimized for each large language model [34].

Future work

In the future, we could investigate the application of the LLM models LLaMA, BART and other ML methods such as LDA, BERT on different raw data with different types, length and prompt engineering that are needed. Protocols for reporting the convergence rates or other validation methods between two types of coding might be necessary. Software tools that can help streamline the process will be a great asset for education researchers to explore the new AI based technologies.

Appendix

Here we present the sample summaries generated by LLM and human coders. For the complete coding results, please visit links below:

Manual coding: <https://uofi.box.com/s/exlkaqysevfw2c34vdcfcd7bpjs0oh3>

LLM generated summary: <https://uofi.box.com/s/t6qgcg256lgv6k5e5pkq92v19ccu9kxk>

Manual-Written Summary for "Evaluation 2" question

Summary: Overall, there was a mixed bag of student sentiments within the responses. On one hand, complaints could be coded with several ideas. For example, the theme of comprehensibility could include how students struggled with going through large blocks of text or interpreting handwriting and formulas. Complaints surrounding the level of detail could be both a lack or an overwhelming amount of detail. Students also wished to have digital notes become more of supplemental tool that could build upon lectures, with features like practice problems or notes from course staff. Otherwise, some pointed that it could be redundant. And others wanted the notes to be accessible for more lectures. Complaints on formatting could include the presentation of the notes or the amount of visuals. These were all areas that students pointed to as areas of improvement. On the other hand, positive feedback cited the effectiveness of the digital notes as a study tool, as well as it's ease of use and convenience.

Excellent LLM-Generated Summary for "Evaluation 2" question

Summary: Based on your feedback, here are some possible improvements that could be made to the digital notes: 1. More concise and organized format: Some students found the digital notes to be too detailed or overwhelming. Consider breaking up the notes into more concise sections or using headings and subheadings to organize the content better. 2. Interactive elements: Adding interactive elements such as quizzes, practice problems, or hyperlinks to relevant resources could make the digital notes more engaging and useful for students. 3. Improved readability: Some

students mentioned that the formatting in certain areas of the digital notes was difficult to read. Consider using a cleaner and more consistent format to improve readability. 4. Handwritten notes integration: While some students prefer handwritten notes, others may find it easier to make their own notes on digital notes. Consider providing an option for both digital and handwritten notes to cater to different learning styles. 5. Practice problems: As mentioned by a few students, having a “practice problems” section in the digital notes could be helpful for working through homework problems. 6. Consistency in formatting: To improve readability, it’s important to maintain consistency in formatting throughout the digital notes. This includes using headings, subheadings, bullet points, and other formatting tools consistently. 7. Improved accuracy of transcripts: While the digital notes are helpful for referring to when rewatching lectures or going over material that has been covered before, some students mentioned that the transcripts are not always accurate. Ensure that the transcripts are carefully reviewed and corrected if necessary. 8. Hyperlinking to relevant resources: Providing hyperlinks to relevant resources such as textbook problems or other online materials could help students better understand the concepts covered in class. 9. More concise digital notes: Some students mentioned that they would prefer more concise digital notes, rather than lengthy slides with a lot of information. Consider breaking up longer lectures into shorter sections or using summaries to condense the information. 10. Improved readability for handwritten notes: To make handwritten notes easier to understand, consider using a consistent font and formatting throughout the notes. This can help students who are trying to review the material on their own.

Comments:

1. Even gave suggestions to the specific complaints that students mentioned
2. Included key points and was somewhat cohesive, but it was a list rather than a cohesive summary.

Relatively Poor LLM-Generated Summary for “Evaluation 2” question

Summary: Digital notes offer several advantages, including ease of accessibility, organization, interactivity, time-saving, searchability, and consistent formatting. They can also provide interactive elements such as quizzes or practice problems to help students engage more deeply with the material. However, some students may find it challenging to understand the content of digital notes due to the format or way the information is presented, lack face-to-face interaction and immediate feedback, and have limited content coverage, visual aids, and interactivity compared to traditional classroom settings. Additionally, some students may find it difficult to annotate or highlight important information within digital notes due to the format or lack of tools for doing so. Despite these cons, many students find digital notes to be a helpful tool in their learning journey, providing a convenient and organized way to review course material and engage with course content.

Comments:

1. Paragraph was very cohesive. However, it made a pros and cons list instead of providing the main complaints. Was also inaccurate because it interpreted the suggestions as a pro of classtranscribe.
2. Why did it mention the benefit of the notes? The question is for the complaints.

Acknowledgements

This work was funded in part by the Institute for Inclusion, Diversity, Equity, and Access in the Grainger College of Engineering, University of Illinois Urbana-Champaign (Grant #: GIANT2021-03 and GIANT2022-08), and Microsoft Grant for developing UDL practices in college campuses.

References

- [1] P. Bazeley, *Qualitative Data Analysis: Practical Strategies*. Thousand Oaks, CA: Sage, 2013.
- [2] J. Rouder *et al.*, “What to Do with All Those Open-Ended Responses? Data Visualization Techniques for Survey Researchers,” *Survey Practice*, Aug 2021. [Online]. Available: www.surveypractice.org/article/25699-what-to-do-with-all-those-open-ended-responses-data-visualization-techniques-for-survey-researchers
- [3] C. Vanover *et al.*, *Analyzing and Interpreting Qualitative Research: After the Interview*. Sage Publications, 2022.
- [4] R. Nawaz, Q. Sun, M. Shardlow, G. Kontonatsios, N. R. Aljohani, A. Visvizi, and S.-U. Hassan, “Leveraging AI and Machine Learning for National Student Survey: Actionable Insights from Textual Feedback to Enhance Quality of Teaching and Learning in UK’s Higher Education,” *Applied Sciences*, vol. 12, no. 1, p. 514, 2022. [Online]. Available: <https://doi-org.proxy2.library.illinois.edu/10.3390/app12010514>
- [5] G. Nanda, K. A. Douglas, D. R. Waller, H. E. Merzdorf, and D. Goldwasser, “Analyzing Large Collections of Open-Ended Feedback from MOOC Learners Using LDA Topic Modeling and Qualitative Analysis,” *IEEE Transactions on Learning Technologies*, vol. 14, no. 2, pp. 146–160, 2021. [Online]. Available: <https://doi-org.proxy2.library.illinois.edu/10.1109/TLT.2021.3064798>
- [6] L. Tang, Z. Sun, B. Idnay, J. G. Nestor, A. Soroush, P. A. Elias, Z. Xu, Y. Ding, G. Durrett, J. F. Rousseau, C. Weng, and Y. Peng, “Evaluating large language models on medical evidence summarization,” *Nature Digital Medicine*, vol. 158, 2023.
- [7] B. Zhang, H. Yang, and T. Zhou, “Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models,” *ICAIF ’23*, November 2023.
- [8] X. Ding, K. Sun, Z. Xiao, S. Varadhan, J. Li, N. Gersich, A. Agarwal, M. Gopannagari, C. Vogiatzis, D. Dalpiaz, J. Amos, L. Angrave, and H. Liu, “Evaluating the Low-Stakes Assessment Performance, Student Perceived Accessibility, Belongingness, and Self-Efficacy in Connection to the Use of Digital Notes in Engineering and Computing Courses,” in *ASEE Conference*, 2023.
- [9] H. Liu, L. Angrave, J. Amos, D. Dalpiaz, C. Vogiatzis, Z. Xiao, S. Varadhan, and J. Louie, “A Digital Book-Based Pedagogy to Improve Course Content Accessibility for Students with and without Disabilities in Engineering and other STEM courses,” in *ASEE Conference*, 2022.
- [10] Y. C. Tie, M. Birks, and K. Francis, “Grounded theory research: A design framework for novice researchers,” *Sage Open Med*, vol. 7, p. 2050312118822927, 2019.
- [11] B. Glaser and A. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Mill Valley, CA: Sociology Press, 1967.
- [12] K. Charmaz and R. Thornberg, “The Pursuit of Quality in Grounded Theory,” *Qualitative Research in Psychology*, vol. 18, no. 3, pp. 305–327, 2021.
- [13] L. M. Stough and S. Lee, “Grounded Theory Approaches Used in Educational Research Journals,” *International Journal of Qualitative Methods*, vol. 20, pp. 1–13, 2021.

- [14] J. A. Maxwell, *Qualitative Research Design: An Interactive Approach*. Thousand Oaks, CA: Sage, 2005.
- [15] Y. Meng, J. Huang, G. Wang, Z. Wang, C. Zhang, Y. Zhang, and J. Han, “Discriminative Topic Mining via Category-Name Guided Text Discriminative Topic Mining via Category-Name Guided Text Embedding,” *arXiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2003.12237>
- [16] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, “A Survey on Evaluation of Large Language Models,” 2023.
- [17] Z. Xiao, X. Yuan, Q. Liao, R. Abdelghani, and P. Oudeyer, “Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding,” in *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, ser. IUI ’23 Companion. Association for Computing Machinery, 2023, pp. 75–78. [Online]. Available: <https://doi.org/10.1145/3581754.3584136>
- [18] H. Rose and A. Meyer, *Teaching Every Student in the Digital Age: Universal Design for Learning*. ERIC, 2002.
- [19] “The UDL Guidelines,” Sep 2022. [Online]. Available: <https://udlguidelines.cast.org/>
- [20] K. Charmaz, *Constructing Grounded Theory*. New York, NY: Sage, 2014.
- [21] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach Learn*, vol. 20, pp. 273–297, 1995. [Online]. Available: <https://doi.org/10.1007/BF00994018>
- [22] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001. [Online]. Available: <https://doi.org/10.1023/A:1010933404324>
- [23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *ArXiv*, 2018. [Online]. Available: <https://arxiv.org/abs/1810.04805>
- [24] C. Sammut and G. Webb, *TF-IDF*. Encyclopedia of Machine Learning, 2011. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_832
- [25] J. Pennington, R. Socher, and C. Manning, “GloVe: Global Vectors for Word Representation,” 2014.
- [26] S. Syed and M. Spruit, “Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation,” in *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2017. [Online]. Available: <https://doi.org/10.1109/DSAA.2017.61>
- [27] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet Allocation,” *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
- [28] Ollama, “ollama,” 2023. [Online]. Available: <https://github.com/ollama>
- [29] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.
- [30] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “LLaMA: Open and Efficient Foundation Language Models,” 2023. [Online]. Available: <https://arxiv.org/abs/2302.13971>
- [31] Y. Liu *et al.*, “Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models,” *Metrad*, 2023. [Online]. Available: <https://doi-org.proxy2.library.illinois.edu/10.1016/j.metrad.2023.100017>

- [32] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2401.11817>
- [33] J.-Y. Yao, K.-P. Ning, Z.-H. Liu, M.-N. Ning, and L. Yuan, "LLM lies: Hallucinations are not bugs, but features as adversarial examples," 2023. [Online]. Available: <https://arxiv.org/abs/2310.01469>
- [34] J. Zamfirescu-Pereira, R. Wong, B. Hartmann, and Q. Yang, "Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 1–21. [Online]. Available: <https://doi.org/10.1145/3544548.3581388>