A Novel Method for Thematically Analyzing Student Responses to Open-ended Case Scenarios

Umair Shakir

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Engineering Education

Andrew S. Katz, Chair
David B. Knight
Jacob R. Grohs
Justin L. Hess

November 10, 2023

Blacksburg, VA

Keywords: natural language processing, open-ended assessments, engineering case studies, automatic short answer grading, computerized qualitative data analysis

A Novel Method for Thematically Analyzing Student Responses to Open-ended Case Scenario

Umair Shakir

ABSTRACT (Academic)

My dissertation is about how engineering educators can use natural language processing (NLP) in implementing open-ended assessments in undergraduate engineering degree programs. Engineering students need to develop an ability to exercise judgment about better and worse outcomes of their decisions. One important consideration for improving engineering students' judgment involves creating sound educational assessments. Currently, engineering educators face a trad-off in selecting between open- and closed-ended assessments. Closed-ended assessments are easy to administer and score but are limited in what they measure given students are required, in many instances, to choose from a priori list. Conversely, open-ended assessments allow students to write their answers in any way they choose in their own words. However, open-ended assessments are likely to take more personal hours and lack consistency for both inter-grader and intra-grader grading. The solution to this challenge is the use of NLP. The working principles of the existing NLP models is the tallying of words, keyword matching, or syntactic similarity of words, which have often proved too brittle in capturing the language diversity that students could write. Therefore, the problem that motivated the present study is how to assess student responses based on underlying concepts and meanings instead of morphological characteristics or grammatical structure in sentences. Some of this problem can be addressed by developing NLP-assisted grading tools based on transformer-based large language models (TLLMs) such as BERT, MPNet, GPT-4. This is because TLLMs are trained on billions of words and have billions of parameters, thereby providing capacity to capture richer semantic representations of input text. Given the availability of TLLMs in the last five years, there is a significant lack of research related to integrating TLLMs in the assessment of open-ended engineering case studies. My dissertation study aims to fill this research gap.

I developed and evaluated four NLP approaches based on TLLMs for thematic analysis of student responses to eight question prompts of engineering ethics and systems thinking case scenarios. The study's research design comprised the following steps. First, I developed an example bank for each question prompt with two procedures: (a) human-in-the-loop natural language processing (HILNLP) and (b) traditional qualitative coding.

Second, I assigned labels using the example banks to unlabeled student responses with the two NLP techniques: (i) k-Nearest Neighbors (kNN), and (ii) Zero-Shot Classification (ZSC). Further, I utilized the following configurations of these NLP techniques: (i) kNN (when k=1), (ii) kNN (when k=3), (iii) ZSC (multi-labels=false), and (iv) ZSC (multi-labels=true). The kNN approach took input of both sentences and their labels from the example banks. On the other hand, the ZSC approach only took input of labels from the example bank. Third, I read each sentence or phrase along with the model's suggested label(s) to evaluate whether the assigned label represented the idea described in the sentence and assigned the following numerical ratings: accurate (1), neutral (0), and inaccurate (-1). Lastly, I used those numerical evaluation ratings to calculate accuracy of the NLP approaches. The results of my study showed moderate accuracy in thematically analyzing students' open-ended responses to two different engineering case scenarios. This is because no single method among the four NLP methods performed consistently better than the other methods across all question prompts. The highest accuracy rate varied between 53% and 92%, depending upon the question prompts and NLP methods. Despite these mixed results, this study accomplishes multiple goals.

My dissertation demonstrates to community members that TLLMs have potential for positive impacts on improving classroom practices in engineering education. In doing so, my dissertation study takes up one aspect of instructional design: assessment of students' learning outcomes in engineering ethics and systems thinking skills. Further, my study derived important implications for practice in engineering education. First, I gave important lessons and guidelines for educators interested in incorporating NLP into their educational assessment. Second, the open-source code is uploaded to a GitHub repository, thereby making it more accessible to a larger group of users. Third, I gave suggestions for qualitative researchers on conducting NLP-assisted qualitative analysis of textual data. Overall, my study introduced state-of-the-art TLLM-based NLP approaches to a research field where it holds potential yet remains underutilized. This study can encourage engineering education researchers to utilize these NLP methods that may be helpful in analyzing the vast textual data generated in engineering education, thereby reducing the number of missed opportunities to glean information for actors and agents in engineering education.

A Novel Method for Thematically Analyzing Student Responses to Open-ended Case Scenario

Umair Shakir

GENERAL AUDIENCE ABSTRACT

My dissertation is about how engineering educators can use natural language processing (NLP) in implementing open-ended assessments in undergraduate engineering degree programs. Engineering students need to develop an ability to exercise judgment about better and worse outcomes of their decisions. One important consideration for improving engineering students' judgment involves creating sound educational assessments. Currently, engineering educators face a trade-off in selecting between open- and closed-ended assessments. Closed-ended assessments are easy to administer and score but are limited in what they measure given students are required, in many instances, to choose from *a priori* list. Conversely, open-ended assessments allow students to write their answers in any way they choose in their own words. However, open-ended assessments are likely to take more personal hours and lack consistency for both inter-grader and intra-grader grading. The solution to this challenge is the use of NLP. The working principles of the existing NLP models are the tallying of words, keyword matching, or syntactic similarity of words, which have often proved too brittle in capturing the language diversity that students could write. Therefore, the problem that motivated the present study is how to assess student responses based on underlying concepts and meanings instead of morphological characteristics or grammatical structure in sentences. Some of this problem can be addressed by developing NLP-assisted grading tools based on transformer-based large language models (TLLMs). This is because TLLMs are trained on billions of words and have billions of parameters, thereby providing capacity to capture richer semantic representations of input text. Given the availability of TLLMs in the last five years, there is a significant lack of research related to integrating TLLMs in the assessment of open-ended engineering case studies. My dissertation study aims to fill this research gap.

The results of my study showed moderate accuracy in thematically analyzing students' open-ended responses to two different engineering case scenarios. My dissertation demonstrates to community members that TLLMs have potential for positive impacts on improving classroom practices in engineering education. This study can encourage engineering education researchers to utilize these NLP methods that may be helpful in analyzing the vast textual data generated in engineering education, thereby

reducing the number of missed opportunities to glean information for actors and agents in engineering education.

Dedication

To all who dream, strive, and execute enduring reforms in the existing unbalanced socio-economic structure of the world for material, physical, and spiritual empowerment of people that are deprived of their *free life* because they are situated in marginalized socioeconomic class, race, gender, religion, or geographical location.

Acknowledgements

I acknowledge all of those who encouraged or helped me during my life to stand up after each failure when I mentally, physically, or spiritually refused to do so. With this, I want to particularly call out a few names.

Dr. Andrew Katz, my advisor, is foremost in completing my doctoral journey. You have imprinted on me how I see, observe, or perceive this world. You were the first to introduce me to the fun part, where we predict, model, or explain complex human behavior. You are my motivation to start my day with reading, writing, or thinking. You have patiently listened to all of my raw ideas and naive questions. You showed me how to weave threads into discrete parts to form a holistic piece, whether it be a sentence, paragraph, or presentation. You taught me how to translate my abstract thinking into tangible artifacts, so I can communicate or debate with colleagues. Without you, Dr. Katz, I would never have been able to complete my PhD.

Next, I pay tribute and give complements to my committee members: Drs. Knight, Grohs and Hess for their guidance and probing, constructive, comments or questions when I lost track. Particularly, your guidance was helpful when I shifted my dissertation ideas. Thank you, you all remained with me during this bumpy ride. Dr. Knight, you always helped me scope my work. Dr. Grohs, you have always helped me think outside the box and connect my findings to practice or educators. Dr. Hess, I am indebted to you for your comments to contextualize my work and acknowledge critics within my research space. Particularly, I appreciate the flexibility and understanding of my committee members when I missed my milestone deadlines.

Among my graduate degree fellows in the Department of Engineering Education, VT; Malle, you are my milestone partner. I am proud of this partnership. You have given me undeniable help when you loudly spoke 40,000 words of my dissertation write-up during our review sessions. Being able to listen to my writing is the privilege in drawing flaws in my arguments. The IDEEAS lab members, I am thankful for your feedback along various milestones of my PhD.

Then, I want to acknowledge my friends: Muhammad Ali Naveed, Muhammad Imran Akbar, Muhammad Faheem Haider, and Shahzad Kamran. Ali Naveed, you are my emotional and financial props in the US. You have listened patiently to my spiritual rands, my career paths, and even hereafter beliefs and skepticisms in our long walks and drives. You ask me real-tough questions with humor and sarcasm. Imran Akbar, you are who stand behind where I am in my life now. You are the true support for me in life decision-making. Faheem Haider, you are such a

perfect listener for all of my life! Shahzad Kamran, you are the true person to which I rely most for executing my dreams.

Then, all of my teachers from high school to university to professional career. I have acquired wisdom from various people who came into my life at different stages. Mirza Aitqe-ur-Rehman you are my motivation to work hard for perfection.

Lastly but not least, my wife Nadia, and children: Khadija, Taha, Abdullah and Zainab. Naida Shakir, my wife, you do suffer a lot during my PhD and I hope you deserve more than that. To my parents who provide me with the safe space with their hard work, so, I could physically, intellectually grow.

In last, all of those people in past, present, and future I share smile, wealth, or wisdom for their empowerment without any return transaction for myself. Those moments are my source of energy, happiness, and hope.

AI Statement

I used Generative Artificial Intelligence (GAI) tools—ChatGPT and Claude—in preparing my dissertation material. Next, I explain why and how I used the GAI tools.

1 To identify mechincal mistakes in my write-up. To achieve this, I used the following prompt as input to the GAI tools:

- "You are an expert editor for academic dissertation. You are given text for the dissertation. You need to correct any grammatical mistake. The other strict instructions are that you should not change the order, content, and voice of the given text. You should not define acronyms. Your response should be rectified text. The text is as follows:"
- "Please give me list of proposed changes"

2 To improve my R and Python code lines for generating graphs given in Ch 4: Results. An example of prompt is:

- "You are an expert R programmer. I have given you code lines for generating bar chart using ggplot in R. Please suggest code lines to change color palette that is friendly for reader with color blindness."

3 To get alternative words in headings, sub-headings and sometimes in sentences. An example of prompt is:

- "Act as if you are a very creative writer and helpful research assistant. I need your help. I am a researcher using natural language processing in educational settings. I have given you a section heading, "Handling Lexical Diversity", please suggest alternative words for " Handling" in the given heading"

I have reviewed, and edited as per need, all the outputs of the GAI tools before final inclusion in this dissertation material. I take full responsibility for the material presented in this dissertations' manuscript.

Table of Contents

Dedication……..vi
Acknowledgement …….....vii
AI Statement ……. ……..…..ix
List of Tables……...xiii
List of Figures ……..xiv
List of Abbreviations ……xv
CHAPTER 1: INTRODUCTION..... 1
1.1. RESEARCH MOTIVATION..... 1
1.2. THE PROBLEM..... 2
1.3. RESEARCH PURPOSE AND RESEARCH QUESTION..... 4
1.3.1 Sub-Research Questions..... 4
1.4. CASE SCENARIOS..... 5
1.5. RESEARCH DESIGN..... 6
1.6. LIMITATIONS OF THE STUDY..... 8
1.7. CONTEMPORARY RELEVANCE AND RESEARCH CONTRIBUTIONS OF THE STUDY..... 9
1.8. THE STUDY'S IMPLICATIONS FOR PRACTICE..... 10
1.9. THE STUDY'S KEYWORD DEFINITION..... 11
1.9.1 Natural Language Processing..... 11
1.9.2 Machine Learning..... 11
1.9.3 Supervised Machine Learning..... 11
1.9.4 K nearest neighbor (kNN)..... 11
1.9.5 Zero-shot Classification (ZSC)..... 11
1.9.6 Ethics..... 11
1.9.7 Systems Thinking..... 11
1.9.8 Confusion Matrix..... 12
1.9.9 Recall..... 12
1.9.10 Precision..... 12
1.9.11 Accuracy..... 12
1.9.12 F1-Score..... 12
1.9.13 Quadratic Weighted Kappa (QWK)..... 12
CHAPTER 2:LITERATURE REVIEW..... 13
2.1. CHAPTER OVERVIEW..... 13
2.2. HISTORICAL EVOLUTION OF NLP METHODS..... 13
2.3. USE OF NLP IN OPEN-ENDED EDUCATIONAL ASSESSMENTS..... 15
2.3.1 Reference-based Approaches..... 16
2.3.2 Response-based Approaches..... 18
2.3.2.1. Use of Lexical Features in Supervised ML..... 18
2.3.2.2. Use of Syntactic Features in Supervised ML..... 19
2.3.2.3. Use of Unsupervised ML and its Combination with Supervised ML..... 21
2.3.2.4. Use of Semantic Features in Supervised ML..... 24
2.4. SEMANTIC FEATURES-BASED NLP APPROACHES..... 24
2.4.1 Knowledge-based Methods..... 24
2.4.2 Corpus-based Methods (Fixed Word Embeddings..... 25
2.4.2.5. Use of Fixed Word Embeddings in Evaluation of Student Responses..... 26
2.5. CONTEXTUALIZED WORD EMBEDDINGS..... 30

2.6. USE OF TLLMS IN EVALUATION OF STUDENT OPEN-ENDED RESPONSES 30
2.6.1 Feature-based Approaches 31
2.6.2 Fine-Tuning-based Approaches 32
2.7. USE OF NLP APPROACHES IN ENGINEERING EDUCATION 34
2.7.1 Use of Lexical Features in Supervised ML 35
2.7.2 Use of Syntactic Features in Supervised ML 36
2.7.3 Use of Unsupervised ML 37
2.7.4 Use of Semantic Features extracted with TLLMs 38
2.8. TLLMS-SPECIFIC LIMITATIONS AND THEIR ENVIRONMENTAL, FINANCIAL, AND SOCIAL IMPACTS 40
2.9. EXISTING ETHICS ASSESSMENT METHODS 42
2.9.1 Ways of Categorizing Psychometric Instruments 42
2.9.2 Ways of Applying and Accessing Case Studies 44
2.10. EXISTING SYSTEMS THINKING ASSESSMENT METHODS 45
2.11. CHAPTER SUMMARY 48
CHAPTER 3:RESEARCH METHODS 50
3.1. CHAPTER OVERVIEW 50
3.2. CASES SCENARIOS 50
3.2.1 Big Belly Trash Can Ethics Case Scenario 50
3.2.2 Abeesee Village Systems Thinking Case Scenario 51
3.3. DIVIDING STUDENT RESPONSES INTO TRAINING AND TESTING DATASETS 51
3.4. SELECTING QUESTION PROMPTS FROM CASE SCENARIOS 52
3.5. DATA ANALYSIS 54
3.5.1 Pre-processing of Text 55
3.5.2 Developing Example Bank 57
3.5.2.1. Example Bank via the HILNLP Approach 59
3.5.2.2. Example Bank via Traditional Qualitative Coding 61
3.5.2.3. Comparing the Example Banks developed with HILNLP and TraditionalQualitative Coding 61
3.5.3 Labeling the Unlabeled Student Responses 62
3.5.3.1. k Nearest Neighbors (kNN) 62
3.5.3.2. Zero-shot classification (ZSC) 65
3.6. SUMMARY OF DATA ANALYSIS 67
3.7. ACCURACY EVALUATION PROCEDURE 68
3.8. STUDY METHOD-SPECIFIC LIMITATIONS 69
3.8.1 kNN Methodology 69
3.8.2 Graphical or Mathematical Representations 70
3.8.3 Sentence- or Phrase-level Analysis 70
3.8.4 Subjective Choices Between HILNLP and Traditional Qualitative Coding 70
3.8.5 Generative AI and the Study's Method 70
3.8.6 Lack of Standardized Metrics and Standardized Data Set 71
CHAPTER 4: RESULTS 72
4.1. CHAPTER OVERVIEW 72
4.2. DIFFERENCES BETWEEN INPUT-UNLABELED SENTENCES AND OUTPUT-ASSIGNED LABELS 73
4.2.1 kNN Methods 74
4.2.2 ZSC Method (multi-label=true) 76

4.3. SUB-RQ1 78
4.3.1 Big Belly Trash Can Ethics Case Scenario 78
4.3.1.1. ethics_q1 78
4.3.1.2. ethics_q4 80
4.4. ABEESEE VILLAGE SYSTEMS THINKING CASE SCENARIO 82
4.4.1 sys_q3 82
4.4.2 sys_q4 84
4.4.3 sys_q5 86
4.4.4 sys_q6 88
4.4.5 sys_q7 89
4.4.6 sys_q11 91
4.5. SUB-RQ2 AND SUB-RQ3 93
CHAPTER 5:DISCUSSION AND CONCLUSION 97
5.1. CHAPTER OVERVIEW 97
5.2. DISCUSSION OF THE RESULTS 97
5.2.7 Limitations of the Results 97
5.2.8 Interpretation of the Results 99
5.2.9 Connecting the Results to the Literature 102
5.3. RESEARCH QUALITY MEASURES AND TRANSFERABILITY OF THE NLP APPROACH 104
5.4. LESSON LEARNED AND GUIDANCE ON THE USE OF MY NLP APPROACH 104
5.4.1 Question Phrasing 105
5.4.2 Data Pre-Processing 105
5.4.3 Summarization Techniques or Keyword Weighing 105
5.4.4 Co-References 105
5.4.5 Handling Off-Topic Statements 106
5.4.6 Intermediate Step for Automatic Grading 106
5.4.7 Matching Methods: ZSC versus kNN 107
5.4.8 Growing Example Bank 107
5.4.9 Comparing Human Scoring with Automatic Grading System Scoring 107
5.4.10 Handling Lexical Diversity 108
5.4.11 Not a Train-once-and-forever Solution 108
5.5. ADDRESSING COMMON CRITIQUES OF THE USE OF NLP IN TEACHING AND LEARNING 108
5.5.1 Educators and their CS Disciplinary Expertise 109
5.5.2 Conflict between Traditional Qualitative Research and NLP 109
5.6. CONTRIBUTION OF THE STUDY 110
5.6.1 A Method for Automatic Analysis of Open-ended Responses 111
5.6.2 A Scalable Method for Qualitative Data Analysis and Promoting Mixed Methods 111
5.6.3 A Timely Inquiry 112
5.7. DIRECTIONS FOR FUTURE RESEARCH 112
5.8. FINAL THOUGHTS 114
REFERENCES 115
APPENDIX A 144
APPENDIX B. 147

LIST OF TABLES

Table 1.1: Counts of Students' Responses for Case Scenarios……..5
Table 2.1: Summary of Studies that Used Lexical And Syntactic Features in Reference- or Response-Based NLP Approaches……... 23
Table 2.2: Summary of Studies that Used Fixed Word Embeddings……29
Table 2.3: Summary of Studies that Used Transformer-Based Large Language Models……34
Table 2.4: Summary of Studies that Used NLP from the Engineering Education Literature…..40
Table 3.1: Counts of Student Responses for Each Case Scenario, Training, and Testing Samples…….……52
Table 3.2: The Selected Question Prompts from Case Scenarios and their Abbreviations……..53
Table 3.3: Pre-Processing Methods Used for Question Prompts……..56
Table 3.4: Preliminary Codebook for a Question Prompt* of Systems Thinking Scenario……..57
Table 3.5: Counts of (Parsed) Student Responses in Example Bank and With-Held Samples….59
Table 3.6: Counts of Input and Output Sentences for NLP Approaches……..……..…65
Table 3.7: Examples for a Question Prompt* of Systems Thinking Case Scenario With ZSC (Multi-Label=True)…….……..67
Table 3.8: Evaluation Rating Example for a Question Prompt of Ethics Case Scenario……..…68
Table 4.1: Counts of Input Sentences and Labels Assigned (or not-Assigned) by KNN Approaches ……...……75
Table 4.2: Counts and Proportion for Input Sentences and Labels Assigned by ZSC Approaches……77
Table 4.3: Evaluation Ratings of Assigned Labels by NLP Approaches for ethics_q1…..……79
Table 4.4: Evaluation Ratings of Assigned Labels by NLP Approaches for ethics_q4…..……81
Table 4.5: Evaluation Ratings of Assigned Labels by NLP Approaches for sys_q3…….83
Table 4.6: Evaluation Ratings of Assigned Labels by NLP Approaches for sys_q3…...……..85
Table 4.7: Evaluation Ratings of Assigned Labels by NLP Approaches for sys_q5……..…...…87
Table 4.8: Evaluation Ratings of Assigned Labels by NLP Approaches for sys_q6…….88
Table 4.9: Evaluation Ratings of Assigned Labels by NLP Approaches for sys_q7……..……...90
Table 4.10: Evaluation Ratings of Assigned Labels by NLP Approaches for sys_q11……..…...92

LIST OF FIGURES

Figure 1.1: Overview of the Study's Research Designs.....7
Figure 2.1: Process Flow Diagram for the Literature Review on the Use of NLP in Open-Ended Educational Assessments.....16
Figure 3.1: Overview of Data AnalysisSteps.....54
Figure 3.2: Options for Pre-Processing of Raw Text Corpus.....56
Figure 3.3: Methods for Developing Example Bank.....58
Figure 3.4: Using KNN Process for Labeling Unlabeled Responses.....63
Figure 3.5: Using ZSC Process for Labeling Unlabeled Responses.....66
Figure 4.1: Proportion for Input Sentences that Were Not Included in Accuracy Evaluation by KNN Approaches.....76
Figure 4.2: Proportion of Input Sentences that Got Assigned More Than One Label in the ZSC (Multi -Label=True).....77
Figure 4.3: Counts for Evaluation Ratings of Assigned Labels by NLP Approaches for ethics_q1.....79
Figure 4.4: Counts for Evaluation Ratings of Assigned Labels by NLP Approaches for ethics_q4.....81
Figure 4.5: Counts for Evaluation Ratings of Assigned Labels by NLP Approaches for sys_q3.....83
Figure 4.6: Counts for Evaluation Ratings of Assigned Labels by NLP Approaches for sys_q4.....85
Figure 4.7: Counts for Evaluation Ratings of Assigned Labels by NLP Approaches for sys_q5.....87
Figure 4.8: Counts for Evaluation Ratings of Assigned Labels by NLP Approaches for sys_q6.....89
Figure 4.9: Counts for Evaluation Ratings of Assigned Labels by NLP Approaches for sys_q7.....91
Figure 4.10: Counts for Evaluation Ratings of Assigned Labels by NLP Approaches for sys_q11.....93
Figure 4.11: Distribution of True Positive Rates Across Question Prompts and Case Scenarios.....94
Figure 4.12: Distribution of Neutral Ratings Rates Across Question Prompts and Case Scenarios.....95
Figure 4.13: Distribution of False Positive Rates Across Question Prompts and Case Scenarios.....96

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
ML	Machine Learning
TLLMS	Transformer-Based Large Language Models
ASAG	Automatic Short Answer Grading
AEG	Automatic Essay Grading
kNN	k Nearest Neighbor
ZSC	Zero Shot Classification
HILNLP	Human-in-the-Loop Natural Language Processing
GAI	Generative Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
MPNET	Masked and Permuted Pre-training for Language Understanding
LIWC	Linguistic Inquiry and Word Count
sys_q	Question Prompts of Systems Thinking Case Scenario
ethics_q	Question Prompts of Ethics Case Scenario

Chapter 1: Introduction

This dissertation is about how engineering education researchers and practitioners can use natural language processing (NLP) in implementing open-ended assessments in undergraduate engineering degree programs. In doing so, I will answer the following research question (RQ): *How can we apply NLP approaches that use transformer-based large language models to thematically analyze students' responses to open-ended question prompts of case scenarios?* In sections 1.1 and 1.2, I give rationale for why I choose this critical problem space and RQ by identifying research gaps in the engineering education community's understanding related to NLP, and the need to integrate NLP into education assessments.

1.1. Research Motivation

Textual data (such as publications, student responses to open-ended assignments, teaching evaluation statements, and interview transcripts) form an important means of dialogue between the various agents and actors of the engineering education ecosystem. Some challenging issues with textual data are that it is complex and idiosyncratic, and its analysis incurs huge cost in terms of person-time and expert training of researchers and practitioners. Due to this cost, many researchers and practitioners do not analyze textual data. Therefore, textual data presents a missed opportunity for our community members to glean information for effective decision-making in engineering education ecosystem. While there is a lot to be gained through manual analysis of textual data, we could benefit from a computer-assisted approach which could quickly and accurately reveal trends and patterns in given textual datasets. This need presents the following challenge: how can we handle text data on a large scale in resource efficient ways? The solution I propose to this challenge is the use of NLP—a set of techniques at the intersection of computer science, statistics, and linguistics. Natural language processing tools enable computers to understand and generate natural languages which evolve through human use over time, for example, English, French or German (Hirschberg & Manning, 2015; Terrace et al., 1981).

Before recent model releases like GPT-4, NLP tools were dictionary-based and deterministic as they relied on a restricted set of rules and vocabulary encoded into algorithms to detect patterns and relationships in text corpora. Unfortunately, such a working principle often renders those dictionary-based NLP tools inflexible to respond to variations in words for describing the same idea (Kalyan et al., 2021; Mikolov, Sutskever, et al., 2013). For example, consider the following two sentences: (i) how do locals think about the heating problem, or (ii) what is the residents' perspective on the electric power issue. These two sentences express the same idea of the resident's opinion in different words. The dictionary-based NLP tools would not be able to identify those sentences as statements expressing the same idea.

However, engineering education researchers now have methods to resolve this inflexibility by developing NLP tools based on recent advances such as transformer-based large language models (TLLMs). Prominent examples are Facebook's RoBERTa (Y. Liu et al., 2019), Google's BERT (Devlin et al., 2019), Microsoft MPNet (Song et al., 2020). The working principle underlying TLLMs is called distributional semantics and is exemplified by the quote: "You shall know a word by the company it keeps" (Firth, 1968, p.179). These TLLMs are designed to learn dependencies in sequences of words (sentences and paragraphs) to model sequential and hierarchical structures in human language. Given how new these TLLMs are, the NLP approaches based on them have been underexplored for applications in assessment practices of student learning outcomes in engineering education.

1.2. The Problem

Given the nature of engineering work and its potential impact on community stakeholders, engineering students need to develop an ability to reason through complex scenarios and exercise judgment about better and worse potential outcomes of their engineering decisions. Part of that development can come in the form of teaching of systems thinking and ethics competencies in undergraduate engineering education. As part of that education, it is helpful to assess students' development in their ability to recognize broader economical, legal, cultural, and ethical factors, and their impacts on various stakeholders (Bielefeldt et al., 2018; M. Davis & Riley, 2008; Stephan, 1999). Moreover, in the interest of preparing students for the engineering challenges in an increasingly complex, interconnected world, it has been well-recognized that engineering educators of today need to address not only technical skills, but also professional skills, including engineering ethics and systems thinking. This raises the question of how to demonstrate students have developed this ability in their undergraduate engineering programs.

Popular approaches to ethics and systems thinking assessment include closed-ended items in the form of validated instruments and case studies with open-ended questions (Camelia et al., 2018; Castelle & Jaradat, 2016; Finelli et al., 2012; Zoltowski et al., 2013). Currently, engineering educators face a tradeoff in selecting between open-ended and closed-ended assessments. While closed-ended assessments may be limited in what they assess and the information they provide, they do have some positive aspects. For example, they are relatively quick and easy to administer and score, and they provide grading consistency across students. Engineering educators use these methods to produce information about students' outcomes efficiently. A drawback of closed-ended assessments is that students are forced to recognize the answer from an *a priori* list for a given problem rather than construct their own response. This approach can be suboptimal for a more authentic assessment of student knowledge and understanding—that is what teachers strive to

achieve. For example, what if someone wanted to assess ethical sensitivity to stakeholder identification using multiple-choice items? That format would require listing potential stakeholders, but the ability that instructor most likely wants to assess is whether students can identify those stakeholders themselves without such priming. Instead, one would want to allow students to write their own responses. Conversely, in the case of open-ended assessments, question prompts are open to being answered in any way in students' own words. Open-ended assessments therefore can be conduits to engage students in higher-order thinking, reasoning, and judgment (Anderson, 2016; Brookhart, 2010). In this sense, students can demonstrate their understanding (or lack of understanding) in creative and informative ways. However, open-ended assessments have their own downsides: they are likely to take more resources, in terms of personal hours and time, and lack consistency for both inter-grader and intra-grader grading. Inter-grader consistency refers to consistent grading between multiple graders, while intra-grader consistency refers to consistent grading by a single grader across students over time. Furthermore, drawbacks that characterize human scoring of open-ended assessments include grading fatigue, disparity in training and background knowledge of graders, and the inherent subjectivity associated with interpretation of open-ended responses (Nehm & Haertig, 2012).

To address drawbacks of human scoring of open-ended responses, educators and researchers have incorporated NLP techniques in their assessment practices of open-ended responses such as automatic short answers grading (ASAG) in diverse subject contexts like Biology, Chemistry, and Physics (Bai & Stede, 2022; Blessing et al., 2021; Caratozzolo et al., 2022; Zhai et al., 2021). Those existing NLP techniques incorporate a variety of methods, for example, words frequencies, keyword matching, or syntactic similarity of words, which often proved too brittle in capturing lexical diversity in which students could write. Here, I give two distinct examples. The first is the Winograd Schema Challenge, in which the problem is identifying the referent of a pronoun. For example, "The trophy did not fit in the suitcase because it was too small" and "The trophy did not fit in the suitcase because it was too big" (Levesque et al., 2012; Winograd, 1972). The problem is in identifying what "it" refers to. The second example is negation handling. For instance, "The students did not like the class and the instructor" and "The students did like the class and not the instructor". The problem is identifying the correct scope of the negation: whether it applies to "the class", "the instructor", or both. These are some non-trivial limitations of previous NLP techniques. Automatic assessment tools established on those NLP models achieve reduced accuracy and scalability when applied in classrooms (Burrows et al., 2015; Haller et al., 2022; Shah & Pareek, 2022). In contrast, modern TLLMs can mitigate some of those challenges because they use the attention mechanism to selectively focus on relevant parts of the input text, thereby capturing relations between words of a sentence over longer distances and

providing richer semantic representations of text (Devlin et al., 2019; Reimers & Gurevych, 2019a; J. Wei et al., 2022).

Given that TLLMs have been available in last five years, there is a significant lack of research related to integrating TLLMs in the evaluation of open-ended case scenarios, broadly within the field of education and particularly within engineering education. To assess students using open-ended scenarios and written responses, at least two processes need to happen from the instructor's side. First, one must identify themes in students' responses. Second, one must then apply the relevant rubrics. This study focuses on the first process. To the best of my knowledge, there is no research study published in engineering education that has explored the application of TLLMs in assessment of student responses to engineering ethics and systems thinking case scenarios. My dissertation study aims to fill this research gap.

1.3. Research Purpose and Research Question

The purpose of this study is to apply and evaluate performance of NLP approaches using TLLMs for thematic analysis of students' responses to open-ended case scenarios in the engineering education context. Accordingly, the overarching research question (RQ) guiding this study is:

> *How can we apply natural language processing approaches that use transformer-based large language models to thematically analyze student responses to open-ended question prompts of case scenarios?*

1.3.1 *Sub-Research Questions*

The sub-research questions (Sub-RQs) guiding this inquiry are:

Sub-RQ1: How well do different NLP processes (i.e., k nearest neighbors, zero-shot classification) label responses?

Sub-RQ2: Does the answer to sub-RQ1 vary by question prompts in a case scenario?

Sub-RQ3: Does the answer to sub-RQ1 vary by case scenarios (i.e., a systems thinking scenario vs an ethics case scenario)?

In the remaining sections of Chapter 1:, I first provide details of the research site and the two different case scenarios I used to answer the aforementioned sub-RQs. Then I provide an overview of the technical setup of the study's NLP approach to illustrate how the process functions.

Finally, I conclude with a discussion of the current iteration of my NLP approaches: its limitations and its implication for researchers and practitioners.

1.4. Case Scenarios

To illustrate my NLP approach, I applied and evaluated its accuracy in thematically labeling written responses of students to the following two case scenarios: (i) the Big Belly Trash Can Ethics Case Scenario, and (ii) Abeesee Village Systems Thinking Case Scenario. I used data collected from multiple engineering courses at Virginia Tech. The original data for both case scenarios was collected as students' assignments but not with the explicit purpose of being used in research related to NLP, like this dissertation study. However, because of their relevance, availability, and amount, I leveraged those student assignments here for demonstration purposes. In my dissertation study, I used 755 student responses for the ethics case scenario and 424 responses for the systems thinking case, as shown in Table 1.1.

Table 1.1: Counts of Students' Responses for Case Scenarios

Case Scenario	Response Count
(i) Big Belly Solar Trash Cans	755
(ii) Abeesee Village Systems Thinking	424

In case scenario (i), the first-year engineering program (FYE) in the Department of Engineering Education at Virginia Tech teaches students an ethics module, that comprises a case-based instructional design of 2 hours in a semester. In the FYE, instructors use a variety of case scenarios, though the most popular one is the Big Belly Trash Can ethics case. After discussing the case and its related material in class sessions with peers and teachers, students are required to submit their written responses to question prompts about: (a) recognition of an ethical issue, (b) identification of a stakeholder, (c) possible decision choices according to various ethical decision-making theories, and (d) consequences of those decisions on the chosen stakeholder. Students are also given the grading rubric to follow for writing their responses. The case study and question prompts are given in Appendix A.

In case scenario (ii), Grohs et al. (2018) developed a case scenario to assess systems thinking competencies. The scenario is framed in a community setting, the fictitious town of

Abeesee (pronounced like A.B.C.), facing heating issues in harsh winters. The respondents' reasoning process is captured through their written responses to the question prompts, which are distributed across the following three phases: (1) processing, (2) response, and (3) critique. First, in the processing phase, the question prompts are about the identification of the problem, stakeholders, and respondents' decision-making process and their goals. Second, in the response phase, the question prompts ask respondents to (a) outline a plan addressing the identified problem, (b) anticipate challenges in implementing their proposed plan, and (c) list potential measures of successful outcomes. Lastly, in the critique phase, someone else's solution is given to respondents, and they (a) interpret its goals, (b) predict its unintended consequences, and (c) judge the adequacy of resources if the given solution is implemented. The case scenario and question prompts are given in Appendix B.

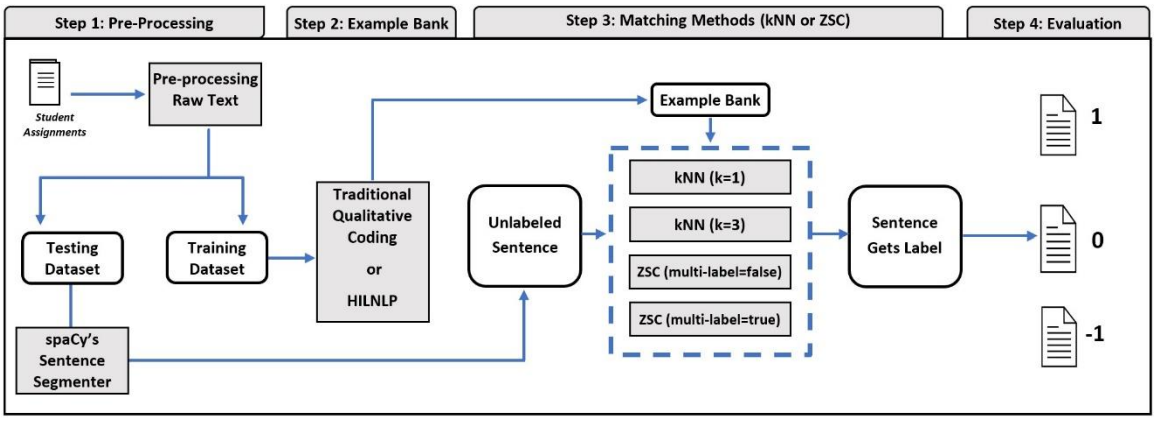
Notably, some of the question prompts of both case scenarios were phrased in a suboptimal manner for the NLP approaches. I have chosen only eight among thirteen question prompts from both case scenarios (i.e., two from (i) ethics case scenario and six from (ii) systems thinking case scenario) because students' responses to those question prompts tend to be more structured and focused on one idea at a time (at least in a parseable manner). The NLP approaches work best when the respondent focuses on one idea at a time. Therefore, the chosen question prompts present the best opportunity to demonstrate how my NLP approach could work for the thematic analysis of student responses.

1.5. Research Design

The study's research design comprises four steps as shown in Figure 1.1. As the first step, I pre-processed the raw text data before passing it to the NLP workflow. As the second step, I developed an example bank for each question prompt with two procedures: (a) human-in-the-loop natural language processing (HILNLP) and (b) traditional qualitative coding. The purpose of an example bank, that comprises student responses and their assigned labels, is to develop a saturated space that covers all possible aspects of an answer to a question prompt. The HILNLP workflow that I used receives raw texts and produces suggested groupings of those texts to which a human user could ascribe labels. The technical implementation of the HILNLP workflow is described in Chapter 3. For developing an example bank with the traditional qualitative coding method, I used a thematic analysis method to label student responses of a question prompt from the ethics case scenario (Clarke & Braun, 2017). As the third step in Figure 1.1, I assigned labels using the example banks to unlabeled student responses with following two NLP techniques: (i) k-Nearest Neighbors (kNN), and (ii) Zero-shot Classification (ZSC). The kNN approach took input of both sentences and their labels from the example banks. On the other hand, the ZSC approach only took

input of labels from the example bank. The technical implementation of (i) and (ii) is provided in Chapter 3.

Figure 1.1: Overview of the Study's Research Designs



** White represents data (e.g., student responses)*
*** Gray represents process*

As the fourth step in Figure 1.1, after assigning labels to student responses, I read each sentence or phrase to evaluate whether the assigned code represents the idea described in the sentence. If yes, then I assigned it a rating of an accurate label as 1. On the other hand, if not, then I assigned it a rating of an inaccurate label as -1. Between those extreme ratings, I had a third neutral category as 0. I assigned this category in instances of ambiguity or partial credit; for example, a sentence could be about more than one idea, or the sentence itself might be ambiguous. Lastly, I used those numerical evaluation ratings to calculate the total number (and proportions) of labeled instances that were (a) accurate (1), (b) inaccurate (-1), and (c) neutral (0).

The aforementioned quantitative evaluation procedure allowed me to answer my sub-RQs in the following way. First, to answer sub-RQ1 (How well do different NLP processes (e.g., k nearest neighbors, zero-shot) label responses?), I selected a question prompt and compared its evaluation ratings across four NLP approaches—(i) kNN (k=1), (ii) kNN (k=3), (iii) ZSC (multi-labels = false), and (iv) ZSC (multi-labels = true). Second, to answer sub-RQ2 (Does the answer to sub-RQ1 vary by question prompts in a case scenario?), I selected a case scenario and compared evaluation ratings across its question prompts for each of the NLP approaches from (i) to (iv). Third, to answer sub-RQ3 (Does the answer to sub-RQ1 vary by case scenarios (e.g., Abeesee system thinking scenario vs ethics case scenario)?), I looked across all question prompts from both of the case scenarios. Then, I compared the evaluation ratings for each of the NLP approaches from (i) to (iv). Lastly, I summarized those evaluation ratings and used this summary to develop the best practices for performing the thematic analysis of student responses through the

investigated NLP processes. Those best practices answered my overarching RQ: *How can we apply the NLP approaches based on transformer-based large language models to thematically analyze students' responses to open-ended question prompts of case scenarios*?

For the technical implementation of the data analysis steps shown in Figure 1.1, I used Google Colab notebooks, written using a combination of the R and python programming languages. All code is available in the github repository at: https://github.com/andrewskatz.

1.6. Limitations of the Study

I classified the limitations of my study into two categories: (i) study design-specific, and (ii) TLLM-specific. Related to (i), I identified limitations due to the manner in which the dataset was collected and how the example bank was developed. Related to (ii), I summarized limitations of TLLMs, as reported in literature, to endorse its negative social and environmental impacts on society.

Regarding study design-specific limitations, the first limitation arose from using assignments that had already been given and collected, but not with the explicit intent of being used in a methodological study like this dissertation study. Although students were often given clear formatting directions, the directions were not always followed. For example, responses to two sequential question prompts may have been written as a single response. This tended to create issues with the data pre-processing that were not always caught. This could lead to sub-par performance of the study's NLP approaches.

In the current implementation, the second limitation is that I selected individual sentences as the labeling unit rather than using paragraphs or whole responses. This is because the embedding model used in this study has a character limit in the range of 300–500 characters. By deciding to split responses at the sentence level, it is possible to lose context when a student develops an argument in more than one sentence. This challenge of eliciting multiple pieces of information at once due to the question phrasing, response format, and character limit of the study's embedding model is an inherent limitation of my dissertation study.

The third limitation is related to the kNN methodology. First, it requires one to predefine a value for k (i.e., the number of nearest neighbors required for matching). However, determining the optimal value of k often involves a process of trial and error for a given dataset (Hechenbichler & Schliep, 2004; Tan, 2005). Second, the kNN approach does not label all input sentences due to (a) the values chosen in this study for k and (b) the similarity score threshold. This limitation could be minimized in future by having a larger example bank.

The fourth limitation is that I did not use standardized datasets (e.g., SciESt, Beetle) and calculate common accuracy metrics (e.g., F1 score, recall, and precision). Notably, I give definitions of these metrics in section 1.9. These metrics are typically reported in literature on NLP tools in computer science or ASAG in education. I acknowledge this is a non-trivial limitation of my NLP approaches for engineering education community to benchmark performance of my NLP tools compared to other NLP approaches reported in the literature.

Regarding TLLM-specific limitations, it is important to acknowledge concerns about environmental and financial costs of TLLMs (Rillig et al., 2023). These costs could potentially limit communities and languages that can contribute to or benefit from these advanced technologies. In addition, TLLMs operate on statistical relationships of word co-occurrences rather than actual comprehension of the world. This may lead to various social impacts that could range from increased misinformation and privacy threats to the perpetuation of biases and stereotypes (Bender et al., 2021; Johri et al., 2023). Therefore, when TLLMs are used in downstream tasks, they may inadvertently contribute to biased decisions. For example, the problematic association of "doctor" with "man" and "nurse" with "woman" exists in TLLMs (Gonen & Goldberg, 2019; Ullmann, 2022). I accounted for bias in my study in two ways: (a) during the pre-processing, I de-identified instructors' names from student assignments and created integer identification numbers for students' names to protect privacy, (b) given the gender-neutral nature of responses used in this study, I maintain gender bias would likely not manifest in the results.

1.7. Contemporary Relevance and Research Contributions of the Study

Recent headlines have featured questions about how ChatGPT, a TLLM, could impact the ways we teach. I suggest my dissertation is contemporary to the engineering education community because I demonstrate how TLLMs would have potential positive impacts on improving instructional practices in engineering education. My dissertation study focused on the following aspect of instructional design: assessment of students' learning outcomes related to engineering ethics and systems thinking competencies in engineering courses. Next, I describe the outcomes and contributions of my dissertation study for researchers and practitioners.

The first contribution is that I developed and evaluated a method using TLLMs for thematically analyzing student descriptive answers. This is an impactful contribution since it lays the foundation for automatic assessment tools that incorporate state-of-the-art TLLMs for open-ended case scenarios in engineering classrooms. The second contribution is that I introduced and provided a use case for the HILNLP approach for qualitative analysis of textual data at scale. Manual analysis of textual data is resource-intensive even with small samples and presents

challenges related to inter- and intra-coder reliability (Creswell & Poth, 2016; Tashakkori & Teddlie, 2009). On the other hand, the HILNLP approach could resolve these challenges by taking the first pass of identifying similar sentences. Moreover, the HILNLP approach could enable qualitative researcher to uncover novel themes and patterns in textual data by analyzing in its entirety, rather than in a sequential manner as in manual analysis.

Natural language processing is not widely used in engineering education (Berdanier et al., 2018; Bhaduri, 2018; Johri et al., 2023; Qadir, 2022). To address this gap, my dissertation study lies at the intersection of NLP and engineering education, and advocates the use of NLP in engineering education. Further, this research also addresses the call from engineering education community members for developing novel approaches to tackle emerging challenges in the field (Borrego & Bernhard, 2011; J. M. Case & Light, 2011). In addition to research contributions, the results of my research have implications for practice, which I describe below.

1.8. The Study's Implications for Practice

First, in section 5.4, I have given important lessons learned from my study for educators interested in incorporating NLP into their open-ended assessment practices. In addition to those lessons, I have also given guidelines about how to navigate those lessons in the future implementation. Second, our project team uploaded open-source code to a GitHub repository (referenced in section 1.5), thereby making it more accessible to a larger group of users. Third, in this research study I curated a list of identified codes related to engineering ethics and systems thinking constructs in students' written responses. I will share that list with the first-year ENGE instructional team. That list could help the team to determine to what extent students are consistently missing a particular perspective in responding to case scenarios, e.g., a minority group stakeholder who is being negatively affected by an engineering project. The ENGE instructional team may adapt their pedagogy to nudge students to that missing perspective in their classroom teaching.

1.9. The Study's Keyword Definition

1.9.1 *Natural Language Processing*

The principle aim of NLP is to gather information on how humans understand and use language through the development of computer programs intended to process and understand language in a manner similar to humans ((Jurafsky & Martin, 2008, Crossley, 2013 Manning & Schütze, 1999).

1.9.2 *Machine Learning*

Machine learning commonly refers to a broader set of algorithms with the ability to adapt its parameters to given data automatically (Bishop, 2006)

1.9.3 *Supervised Machine Learning*

In supervised ML, human-labeled data are used to train the machine in order to generate a model based on a set of attributes extracted from the data (Jordan and Mitchell 2015).

1.9.4 *K nearest neighbor (kNN)*

kNN is a supervised learning technique used for classification and regression tasks. The kNN algorithm searches the training dataset for the "k" nearest neighbors and returns the output label as the majority class (for classification) or average (for regression) of these "k" neighbors (Altman, 1992; Cover & Hart, 1967)

1.9.5 *Zero-shot Classification (ZSC)*

ZSC refers to the task of classifying objects into classes that were not seen during training. It relies on transferring knowledge from seen classes to unseen classes, often through auxiliary attributes (Pushp & Srivastava, 2017; Yin et al., 2019).

1.9.6 *Ethics*

"Ethics is based on well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues." (Velasquez et al., 1987, p. 1)

1.9.7 *Systems Thinking*

"Systems thinking as a system of synergistic analytic skills used to improve the capability of identifying and understanding systems, predicting their behaviors, and devising modifications to them in order to produce desired effects."(Arnold & Wade, 2017, p.1).

1.9.8 *Confusion Matrix*

This is adapted from (Silva et al, 2019)

	Predicted No	Predicted Yes
Actual NO	True Negative (TN)	False Positive (FP)
Actual YES	False Negative (FN)	True Positive (TP)

1.9.9 *Recall*

Recall measures the fraction of actual positives that were correctly identified (Bulut et al., 2022; Yik et al., 2021).

Recall = TP /(TP + FN)

1.9.10 *Precision*

Precision measures the fraction of identified positives that are actually positive (Bulut et al., 2022; Yik et al., 2021).

Precision = TP /(TP + FP)

1.9.11 *Accuracy*

It is defined as the ratio of the number of correct predictions to the total number of inputs (Kerkhof, 2020b)

Accuracy = (TP + TN)/ (TP + FP + TN + FN)

1.9.12 *F1-Score*

The F-1 Score is the harmonic mean of precision and recall to provide a balance between the two when their values differ significantly (Kerkhof, 2020b)

$F1 - Score = 2 \times (Precision \times Recall)/(Precision + Recall)$

1.9.13 *Quadratic Weighted Kappa (QWK)*

QWK measures the agreement between the human graded score and the predicted score. (Putnikovic & Jovanovic, 2023)

Chapter 2: Literature Review

2.1. Chapter Overview

In Chapter 1:, I presented the overarching RQ for this study: How can we apply NLP approaches that use transformer-based large language models to thematically analyze students' responses to open-ended question prompts in case scenarios? To understand the body of knowledge related to the RQ, I review literature both from STEM education and engineering education. I first provide the historical evolution of the NLP discipline. Next, I review studies investigating the use of NLP in the analysis (and/or grading) of student-generated descriptive text. In these sections, there is variation along the following dimensions: (a) what text features are used, (b) how these text features are extracted, and (c) what machine learning (ML) models are trained with extracted text features. Notably, I give distilled definitions of computer science (CS) terminologies at relevant points for readers' understanding. Following this, I discuss limitations of TLLMs related to their social, environmental, and financial impacts on society. Lastly, I review the existing assessment methods for engineering ethics and systems thinking competencies because the use cases of my NLP approach are related to these.

2.2. Historical Evolution of NLP Methods

Natural language processing (NLP) is collection of approaches used to analyze natural languages that evolve through human use (Hirschberg & Manning, 2015; Terrace et al., 1981). During the 1950s-1980s, NLP tools were rule-based and deterministic, primarily using a relatively small set of predefined rules and limited vocabulary, encoded by researchers into computer algorithms. This process was time-consuming and often proved incapable because those hand-crafted rules would only cover a portion of the extensive diversity of natural languages. During the 1980-2000, an increase in computational capacities led to the emergence of statistics-based NLP approaches. These approaches used frequency-based lexical features to identify patterns and relationships in text corpora, and statistical inferences for performing NLP tasks. The heavy reliance of these statistics-based NLP tools on frequency-based lexical features limited their ability to handle structures and content beyond specific textual dataset (i.e., similar to training data), that was a significant limitation (Ahmad et al., 2022; Burrows et al., 2015; Zhou et al., 2023).

In the 2000-2010 period, NLP tools began to incorporate dependencies of words in a sentence, named as syntactic features, in text analysis. For instance, dependency n-grams were developed by grouping subjects and verbs. These syntactic-feature based NLP tools analyzed text with the underlying assumption: similarly structured sentences were more likely to have a similar meaning. However, these approaches had limited flexibility in capturing the semantic meanings of

words (Putnikovic & Jovanovic, 2023; Ramnarain-Seetohul et al., 2022). In that era, proliferation of data on the Internet and further leaps in computational capacities allowed the use of large digital text corpora for developing NLP tools. Consequently, NLP tools began to incorporate and train ML algorithms with human-annotated data. Examples of those ML algorithms include regression, support vector machines (SVM), and naïve Bayesian classifier (C. D. Manning & Schütze, 1999; Sebastiani, 2002). From 2010 to the present, a major milestone in the NLP field was the development of word embedding models. Some examples of these models are Word2Vec, Global Vectors for Word Representation (GloVe), and Embeddings for Language Models (ELMo) (Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2018). These models map words or phrases into higher dimensional vector representations, thereby capturing both semantic and syntactic information. To remind readers, the former refers to the meaning of words whereas the latter corresponds to words' structural role in sentences. Word embedding models use neural network architectures that were introduced in the 1980s and have become central to the NLP field. This means that understanding modern NLP requires understanding the scope of common neural network model architectures.

Existing neural network models in the NLP field can be classified into sequential and attention-based models. Examples of sequential models include Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs) neural networks. Sequential models are designed to learn dependencies in sequences of words (sentences and paragraphs) to model sequential and hierarchical structures in language. However, RNNs tend to focus more on short-term context information while not being able to robustly capture longer-range dependencies of words in sentences. To address this issue, LSTM neural networks were developed and proved more effective in capturing relations between words of a sentence, thereby providing richer sentence and paragraph representations (Hochreiter & Schmidhuber, 1997; Sundermeyer et al., 2012). However, LSTMs encounter difficulties with longer passages due to their sequential processing mechanism. In contrast, attention-based models, were introduced in 2018 and are commonly known as TLLMs, process input text in parallel and utilize the attention mechanism to selectively focus on relevant parts of the input text, enabling them to better capture long-range dependencies (Vaswani et al., 2017). Prominent examples of TLLMs include GPT-4, BERT, XLNet, and MPNet (Devlin et al., 2019; Radford et al., 2019; Song et al., 2020; Z. Yang et al., 2019). These TLLMs have demonstrated state-of-the-art performance across a wide range of NLP tasks such as machine translation, text classification, and sentiment analysis (Gao et al., 2023; Mu et al., 2023; W. Zhang et al., 2023). My dissertation study is about how those TLLMs can be used in thematically analyzing open-ended responses of engineering students to two engineering case scenarios.

In sum, the NLP field has evolved from rule-based to statistics-based to modern TLLMs. This development has enabled researchers to develop computer applications that can process, understand, and generate human language in more flexible and comprehensive ways. This chronological development in the NLP field coincides with how NLP tools have been incorporated into open-ended educational assessment, which I explore in the next section.

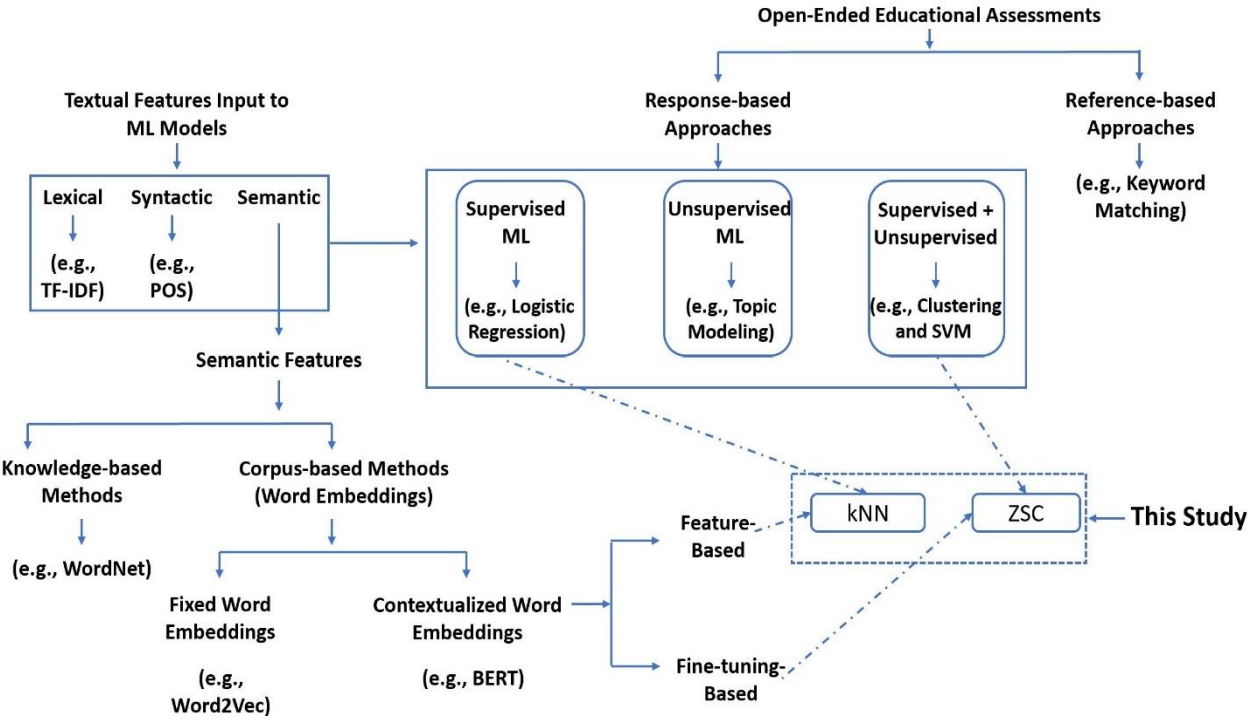
2.3. Use of NLP in Open-ended Educational Assessments

Literature on the use of NLP in open-ended educational assessments can be categorized into two areas: Automatic Short Answer Grading (ASAG) and Automatic Essay Grading (AEG). These two categories can be differentiated as follows: 1) the length of short answers ranges from one phrase to one to four paragraphs, whereas essays extend from one paragraph to several pages; 2) ASAG focuses on the content of the answers, while AEG emphasizes grammar or writing style (Burrows et al., 2015; Ramnarain-Seetohul et al., 2022; Shah & Pareek, 2022). Given these differences in ASAG and AEG, I scope my literature review to focus on ASAG and similar literature (such as student descriptive answers to survey questions) because my identified problem and RQ are relevant to that research area.

The NLP approaches used to analyze or grade student open-ended responses could be categorized based on how the NLP approach uses human-scored answers. The categories are as follows: reference-based or response-based (Galhardi & Brancher, 2018; Putnikovic & Jovanovic, 2023). In the reference-based approach, student responses are scored based on their similarity to expert-provided reference answers, or by choosing the highest scoring answer. On the other hand, the response-based approach uses human-scored answers to train ML models. These models are then used to assign scores to new responses.

Another categorization variable for NLP approaches is the types of textual features extracted from the input text. These features are as follows: lexical, syntactic, or semantic features. Lexical features are derived from the words and vocabulary present in the text (e.g., counts of words). Syntactic features are derived from the grammar and sentence structure in the text such as parts of speech (POS) tags. Semantic features represent the underlying meaning and concepts within the text rather than only vocabulary or grammar. I utilize a combination of the following two categorization variables—(a) how the NLP approach used human-scored answers, and (b) the types of textual features extracted from the input text—to organize the sections from 2.3 to 2.6 of the literature review. In addition, I demonstrate this literature review as a process flow diagram in Figure 2.1. In this figure, I particularly visualize where my dissertation study situates within the existing body of literature.

Figure 2.1: Process Flow Diagram for the Literature Review on the Use of NLP in Open-ended Educational Assessments



In the following sections, I first review studies that used reference-based approaches for ASAG. Second, I provide an overview of studies about response-based approaches that utilized lexical and syntactic features to train ML algorithms. Third, I give examples of unsupervised ML in NLP. Fourth, I move to discuss initial word embedding models developed to extract semantic feature in text and how those were used in the assessment of student open-ended responses. Lastly, I examine the application of TLLMs in educational assessments. The purpose of this extensive literature review is to outline prior work in this space of using NLP to assess students, where that work has fallen short, and identify the area where my work will contribute to the body of knowledge.

2.3.1 *Reference-based Approaches*

In reference-based approaches, the fundamental mechanism involved in assessing student open-ended responses is the comparison of a candidate response with reference (or graded) answers. The distinction between approaches lies in how they accomplish this mechanism.

A rudimentary NLP approach searches for similar keywords between reference and candidate text. Notable examples of this work are Open Mark and Indus Marker grading systems

(S. Jordan, 2009). Building off of similar keywords, a more sophisticated NLP approach called pattern matching was developed (Butcher & Jordan, 2010). In this method, a lexicon of words is compiled, with the elements of that lexicon being referred to as n-grams. In this context, n-gram refers to a string of n number of words (Jurafsky & Martin, 2009a). For example, a bigram is a two-word sequence while a trigram is a three-word sequence. Continuing with this example, given the phrase "many students are taking exams," a bigram would produce: "many students," "students are," "are taking," "taking exams." A special case of n-grams is when n equals to 1; this is known as a Bag-of-Words (BOW). N-gram-based assessment methods measure the degree of overlap of n-grams between a reference text and a student response text. For instance, one might consider the simple binary presence or absence of n-grams that exist both in the reference text and the student response.

Likewise, another common lexical feature is the frequency of n-grams in a document. This is called the term frequency (TF). However, TF alone could present challenges when comparing verbose responses with terse ones. To address this challenge, the TF can be multiplied by a proportional weight to penalize general words (e.g., prepositions, stop words, or articles), resulting in a metric called the term frequency-inverse document frequency (TF-IDF). Notably, Galhardi & Brancher (2018) found in their literature review of 44 ASAG-related papers that the most popular lexical feature used in their reviewed papers is n-grams and its TF-IDF, which was present in 30 (70%) of their reviewed papers. Another lexical feature often employed includes text statistics such as the length of the response, count of unique words, verb counts, and similar lexical statistics. However, algorithms based on keywords matching or n-gram frequency lack knowledge of grammar or syntax, and are highly prone to failure from cases such as different words with identical meanings.

To overcome these limitations to some extent, hand-crafted grammar or syntactic rules were embedded in NLP approaches. Haudek et al. (2012) and Nehm & Haertig (2012) built dictionaries and lexical feature extraction rules in the SPSS Text Analysis (SPSSTA) software. In their hand-crafted dictionaries, they identified keywords and phrases used by 812 undergraduate students' responses to a biology exam. Their NLP approach involved identifying biology terms, locating synonyms of such terms, and identifying distribution patterns of those terms in responses. Although the model was able to identify some information correctly, the authors cautioned about the paid subscription of SPSS and its limited ability to automate scoring of new student answers. Developing lexical feature libraries requires significant human time, effort, and expertise. Further, these libraries typically cannot be used for automated grading processes. Therefore, to automate

grading of new student answers, lexical and syntactic features are integrated with supervised or unsupervised ML algorithms in response-based approaches for ASAG.

2.3.2 *Response-based Approaches*

In response-based approaches, student responses are used to train supervised or unsupervised ML models. A supervised ML algorithms trained on a dataset that has human-assigned scores combined with textual information extracted from the responses, such as lexical, syntactic, or semantic features. During training, the model learns regression or classification rules based on input textual features. After training, the model applies the rules to assign scores (or labels) to new student responses. For example, an answer could be correct, partially correct, or incorrect. Some common classifiers include logistic regression, SVM, random forest, or naive Bayes classifiers. Additionally, some approaches use ensemble methods that combine predictions from multiple different classifiers. On the other hand, unsupervised ML models are designed to learn from and make score or correctness predictions based on unlabeled data. Examples of unsupervised ML include clustering (e.g., K-means clustering and hierarchical clustering) and dimension reduction (e.g., principal component analysis). Next, I review published papers to examine how response-based approaches and textual features (lexical, syntactic, and semantic features) have been used in educational assessment.

2.3.2.1. Use of Lexical Features in Supervised ML

As an example of an approach that used both a lexical feature complemented by a supervised ML algorithm, Ha et al. (2011) and Nehm et al. (2012) developed a classification-based ML software called the Summarization Integrated Development Environment (SIDE). The SIDE program was designed to score undergraduate science students' written explanations by using a BOW model. The model marked the presence or absence of each element of the BOW matrix and counted the frequency of the present elements in students' response. These features were subsequently fed into an SVM model for classification and scoring purposes. An SVM identifies decision boundaries (imagine separating two teams on a field with the most effective line) that differentiate various classes of data (known as support vectors) with the largest margin. Once the decision boundary is established, SVM classifies new data points by determining on which side of the line the points fall. Applying a similar mechanism (i.e., BOW used as the input textual feature for the SVM model) of the SIDE software, Moharreri et al. (2014) developed an automatic assessment tool for evolution theory and Yik et al. (2021) developed for Lewis's acid model. As another example of supervised ML model used for classifying student responses, Wilson et al. (2022) used the TF-IDF method to vectorize students' responses and used these vector representations to develop two classifiers: k-nearest neighbors and logistic regression. The data

came from students at the University of Colorado Boulder who completed the Physics Measurement Questionnaire (PMQ)—an assessment tool for studying student reasoning around measurement uncertainty. Authors found that their logistic regression classifier yielded better classifications than their k-nearest neighbor approach and with the same level of agreement as that between two humans categorizing the data.

In the past decade, end-to-end neural networks have replaced classical supervised ML and dominated most areas of NLP-related research, and ASAG is no exception (Bai & Stede, 2022; Haller et al., 2022). Unlike in non-neural network-based supervised ML approaches, neural models learn a dense, non-interpretable vector representation of the input text(s) and feed it to an output classification or regression layer. As recent examples of the ASAG approach that used lexical feature and neural networks, Zhang et al. (2022) and Zhai et al. (2022) combined three distinct lexical features into a latent feature and then fed it into an RNN for training purposes. These lexical features included: (i) difference in word count between student and reference answers; (ii) maximum IDF value of matched words between student and reference answers; and (iii) cosine similarity between TF-IDF vectors of student and reference answers. Through this process, (Zhai et al., 2022; Zhang et al., 2022) demonstrated that RNNs can extract latent features that are more representative (or informative) of input text than the original lexical features. Despite these positive findings, lexical feature-based supervised ML approaches are considerably limited in their ability to capture the meaning of text. For example, words are treated as independent from one another, and therefore syntactic relationships between words are ignored. For instance, “students are studying” and “studying are students” are considered equivalent in lexical based NLP approaches. To address these limitations, syntactic features are used to train supervised ML algorithms.

2.3.2.2. Use of Syntactic Features in Supervised ML

Syntactic features are defined as roles and relationships of words within a sentence. To capture syntactic relationships, dependency n-grams are developed by POS tagging. In this process, each word in the n-gram corpus is labeled with its corresponding POS tag (such as noun, verb, adjective, or adverb). Dependency n-gram models typically use a triple format containing two words and the dependency relationship between them. For instance, the Stanford Parser (Klein & Manning, 2003) is a commonly used tool for generating POS tags from text. In ASAG, the underlying philosophy in the syntactic-feature based NLP approach is that if two answers share numerous POS tags, they possess a similar syntactic structure and are more likely to convey the same meaning.

Pulman & Sukkarieh (2005) developed a prototype of the c-rater system that automatically grades short answers. Their system used a Hidden Markov Model (HMM) POS tagger to extract syntactic features from the responses—an HMM is a statistical model used to represent data with underlying hidden states. These syntactic features were then used to train a naive Bayes ML classifier to assign grades. The c-rater has been commercialized by the Educational Testing Service as c-rater-ML (Heilman & Madnani, 2013a, 2013b). For grading purposes, the c-rater-ML extracts several types of lexical and syntactic features: (a) concepts expressed in words, sequences of words, and sequences of characters; (b) syntactic relationships between these concept words; and (c) the response length. The c-rater-ML engine uses SVM, wherein weights of all features are used for scoring. As an example of application of c-rater-ML in educational assessment, Lee et al. (2019) used the c-rater-ML to analyze secondary school student responses about scientific argumentation.

Expanding beyond a single, supervised ML model, multiple different supervised ML models can be trained and combined to make predictions, thereby making better predictions than an individual model. This process is called ensemble learning. As an example, Roy et al. (2016) proposed an NLP approach based on an ensemble of two classifiers. The first classifier used a TF-IDF representation of BOW. The second classifier used five similarity measures covering lexical, semantic, and vector-space dimensions between reference and student answer. Ultimately, the classifiers were combined in a weighted fashion to form an ensemble used to predict the final score (label). In another example, Jescovitch et al. (2021) trained an ML model based on an 8-classification algorithm ensemble. In their work, text features of each document were extracted as n-grams and used as input in the ensemble algorithm to predict whether each given document belongs to each class. Using ensemble models can give better predictions, but ensemble models also have downsides. First, it takes more resources to train and use ensemble models. Second, it can be difficult to understand why the ensemble model makes certain predictions because many models are combined. After describing the downsides of ensemble models, in next paragraph, I discuss the limitation of supervised ML models in NLP.

Supervised ML requires human-annotated data, and NLP researchers caution about the inherent subjectivity when humans annotate data to establish desired output from the model (Bender & Friedman, 2018; T. Sun et al., 2019; Wilson et al., 2022). In contrast with supervised ML, unsupervised ML often operates without human annotations and can detect patterns in a text corpus that may differ from what humans would expect. This discrepancy can occur due to several reasons such as (a) humans tend to label data in an ordered manner, whereas unsupervised ML can analyze all data simultaneously; (b) there are inter- or intra-coder consistency issues when

manually labeling data for training purposes, while this issue is not relevant to unsupervised ML. Although these issues for supervised ML may be resolved with one-time, upfront fixed costs during labeling the original training data, they can sometimes be prohibitively expensive in practice. This raises the question about alternatives to supervised ML, i.e., unsupervised models, which I discuss next.

2.3.2.3. Use of Unsupervised ML and its Combination with Supervised ML

A widely used unsupervised ML approach in the NLP field is topic modeling. The most popular topic modeling algorithm is Latent Dirichlet Allocation (LDA). The LDA algorithm is a probabilistic approach that presumes documents are comprised of various topics, and each topic is characterized by a distinct distribution of words (Blei et al., 2003). The LDA approach estimates these probable topics and corresponding word distributions based on word co-occurrence statistics within a set of documents. The technique has been used to model the distribution of topics in a variety of contexts (Chauhan & Shah, 2022). In education, researchers have used the LDA algorithm for thematic analysis of student responses, among other things. For instance, in the physics education field, Geiger et al. (2022) employed LDA to identify distinct ideas in student written responses to open-ended questions about electric circuit design. In the math education field, Cronin et al. (2019) identified key themes using the LDA approach in 21,313 feedback data entries from student consultation sessions at two mathematics support centers.

Since the traditional LDA approach relies on word co-occurrences, such topic models struggle to identify latent topics in sparse texts (e.g., short answers) (Li et al., 2017). Limitations such as these have led researchers to extend the traditional LDA model, leading to another topic modeling algorithm based on the Dirichlet Multinomial Mixture (DMM) distribution. The DMM assumes that each document can be represented by just one latent topic; in contrast, LDA allows documents to have multiple topics (Blei et al., 2003; Li et al., 2017). Vadapally et al. (2022) applied both DMM and LDA for analyzing students' responses to the minute papers in an undergraduate software engineering course. In their minute papers, students provide short answers to two questions: what they learned and what they did not learn in each class. Vadapally et al. concluded that DMM performed better than LDA for generating latent topics in those short texts. Despite topic models can find useful topics (themes) in texts, but they are limited. I gave following two reasons: First, topic models solely rely on word co-occurrences and cannot capture the underlying meanings of words. Second, interpretation of derived topics from topic models is not universal. Two different analysts might interpret the derived topics differently based on their knowledge and experience. After describing the challenges of topic models, in the next paragraph, I discuss the limitation of unsupervised ML models in NLP.

Aldea et al. (2020) cautioned that unsupervised ML methods alone might not completely automate grading processes due to lack of reference-responses which encompass all possible ways to answer an open-ended question. Therefore, researchers have combined both supervised and unsupervised ML algorithms for assessment of student answers. As an example of this approach, Rosenberg & Krist (2021) analyzed 845 middle school students' responses about science model explanations using a sequence of unsupervised and supervised ML algorithms. First, they developed a document-term matrix by tokenization with unigram and converted frequencies of those unigram on a log-scale. Then, they used a combined hierarchical agglomerative and k-means clustering technique to identify similar students' responses. Subsequently, they performed descriptive coding to identify categories in the students' responses. Lastly, they trained three supervised ML algorithms—naïve Bayes, SVM, and sequential neural network—to classify held-out students' responses. The authors found that the agreement between ML-assigned codes and manual coding ranged from 0.62 for naïve Bayes to 0.66 for SVM.

However, the combination of supervised and unsupervised ML models in NLP is not without challenges. In the case of supervised ML, one inherent limitation is the trade-off between optimization for a specific task and generalizability across different contexts. Training a supervised ML model with specific labeled data may not generalize well to unseen data from other contexts. Additionally, unsupervised ML models can cluster data that should be separated, introducing noise that could propagate if the unsupervised model is combined with a supervised ML model downstream. Next, for readers' overview, in Table 2.1, I have summarized the research studies along with their respective NLP methods and limitations that I cited in the sections 2.3.2. Table 2.1 shows that NLP researchers have used a variety of lexical and syntactic text features, combined with both supervised and unsupervised ML approaches to assess students' written responses. Despite their useful performance in some contexts, lexical features can still fail to capture the meanings of sentences, and syntactic features can only do so to a limited degree. On the other hand, capturing how students' responses and reference responses are connected not by their words or sentence structure but by their meaning and concepts, can significantly enhance the performance of automatic grading of student written responses (Ahmad et al., 2022; Haller et al., 2022; Magliano & Graesser, 2012; Putnikovic & Jovanovic, 2023). To clarify, semantic features represent the deeper meaning and concepts conveyed in the text, beyond just the vocabulary and grammar captured by lexical and syntactic features. Over the past 10 years, researchers have proposed several semantic feature-based NLP approaches to build on this concept, which I review next.

Table 2.1: Summary of Studies that used Lexical and Syntactic Features in Reference- or Response-Based NLP Approaches

Papers	NLP Method(s)	Limitations
(Jordan, 2009; Butcher and Jordan, 2010)	Matching keywords	Methods depends upon morphological characteristics of words and resource-intensive for training supervised ML models. Further, these models are task-dependent
(Nehm and Haertig 2012; Haudek et al., 2011)	TF-IDF	
(Ha et al., 2011; Nehm et al. ,2012; Moharreri et al. 2014; Yik et al., 2021)	BOW model + Support Vector Machine-based classification	
(Pulman & Sukkarieh,2005; Heilman & Madnani, 2013a, 2013b; Lee et al. 2019)	POS tagging and decision tree learning and Naive Bayesian machine learning algorithms	
(Roy et al., 2016; Jescovitch et al. 2021)	Ensemble ML	
(Zhang et al., 2022; Zhai et al., 2022)	Lexical Feature + RNN for supervised training	
(Rosenberg & Krist, 2021)	TF-IDF+ hierarchical agglomerative and k-means clustering	
(Wilson et al., 2022)	TF-IDF- KNN + logistic regression	
(Geiger et al., 2022; Cronin et al. 2019)	LDA	Methods depends upon words co-occurrence and interpretation of outputs significantly depends upon analyst
(Vadapally et al., 2022)	Dirichlet Multinomial Mixture (DMM)	

2.3.2.4. Use of Semantic Features in Supervised ML

It is noteworthy that I have given semantic-feature-based NLP approaches a separate level two heading, instead of a level four heading under section 2.3.2, due to its significance in my dissertation.

2.4. Semantic Features-based NLP Approaches

Natural language processing techniques for extracting semantic features of text are commonly categorized as follows: (a) knowledge-based and (b) corpus-based (Galhardi & Brancher, 2018; Kerkhof, 2020a; Shah & Pareek, 2022). The knowledge-based techniques use a knowledge database that already stored hand-coded semantic relationships between words and reflect the way the analyst (NLP developer) perceives those semantic relationships. This external database is used to calculate the binary semantic similarity between words. For example, a word pair could be designated as synonyms (having the same or similar meaning) or antonyms (having opposite meanings).

In contrast to knowledge-based methods, corpus-based methods typically use large public corpora—like Wikipedia or digital libraries—to automatically establish semantic relationships between words as vectors in a higher-dimensional space based on their usage in those corpora. These corpus-based semantic models are commonly called word embeddings. Word embeddings calculate semantic proximity between words or phrases as real number rather than binary (synonym or antonym). The important idea to note about embedding models is their ability to take raw text and generate a representation for that text in high-dimensional vectors to perform subsequent mathematical operations. It is not without limits, but this functionality opens many possibilities.

2.4.1 *Knowledge-based Methods*

One example of the knowledge-based method is WordNet, a semantic lexical database that groups English words into sets called synsets, like {car, auto, automobile, machine, motorcar}. Each synset represents a distinct concept; the aforementioned example is about vehicles (Fellbaum, 2010; Miller, 1995). These synsets are interlinked by semantic relationships such as synonym, antonym, and hypernym-hyponym with other synsets. To clarify, hypernym-hyponym represents one word in a more general form and the other in a more specific instance of the same concept. An example would be as follows: hypernym: vehicle and hyponym: car. Galhardi and Brancher (2018) found that WordNet-based semantic features were used in 11 (25%) of the 44 papers included in their systematic reviews of ASAG literature.

As an example, Pribadi et al. (2017) employed WordNet to find semantic synonyms of words used by students in their short answers about computer architecture related questions. In their study, after sentence tokenization, student answers and graded responses were converted into the WordNet synsets. Sentence tokenization is the task of segmenting a text into individual sentences using punctuation marks like periods (.), exclamation marks (!), and question marks (?) as delimiter to identify sentence boundaries. The Dice Coefficient is then used as a semantic similarity measure. This coefficient is calculated as twice the number of common synsets between two responses, divided by the total number of distinct synsets in both responses. A coefficient value close to 1 indicates high semantic similarity, while a coefficient value close to 0 indicates low semantic similarity. The author chose a threshold value for the coefficient that was considered to select semantically similar pairs for student and reference answers. These selected pairs were then given the same score as the corresponding reference answer. The authors found that their ASAG system did not yield the best result. This is because their system produced low similarity values when measuring two sentences of different lengths despite having the number of overlapping words by almost 80%.

To improve knowledge-based ASAG system, researchers have combined knowledge-based measures with ensemble-based supervised ML models. For example, Sahu & Bhowmick (2020) developed an ensemble-based supervised ML model using a combination of knowledge-based measures and a stacked regression approach. In this regression task, the student's answer was categorized as "correct", "partially correct incomplete", "contradictory", "irrelevant", or "non-domain" based on its semantic similarity with the corresponding model answer. The authors reported that their proposed stacked regression based ensemble model showed a huge improvement in F1 scores on standardized ASAG datasets: ScientBank and Beetle.

Despite their utility as demonstrated in the aforementioned study, knowledge-based methods require significant human efforts to build domain-specific knowledge thesauruses that encode the lexical, syntactic, and semantic understanding of human language. To overcome the limitations of manual efforts, NLP researchers have developed corpus-based methods, commonly known as fixed word embeddings. Fixed word embeddings automatically identify semantic relationships between words by using large public corpora like Wikipedia and the ways in which those words appear in sentences alongside other words in those corpora.

2.4.2 *Corpus-based Methods (Fixed Word Embeddings*

Some examples of fixed word embedding models are Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). Word2Vec is a pre-trained word embedding model on a

portion of the Google News dataset (approximately 100 billion words). This model comprises 300-dimension vectors for 3 million words. The GloVe word embedding model is trained on the Common Crawl and consists of 300-dimension vectors for 2.2 million words. These word embeddings represent words as vectors in a high-dimensional space, where each dimension holds semantic or syntactic features of the words. The intuition behind word embeddings is explained by the distributional hypothesis in linguistics: words appearing in similar contexts often share similar meanings. Alternatively, this hypothesis is exemplified in linguist John Firth's quote, "You shall know a word by the company it keeps" (Firth, 1968, p.179). For example, the word vectors of 'fantasy' and 'imagination' are close in high-dimensional space since their semantic meaning is similar, in part because they often appear in similar contexts.

In evaluation of student responses, word embeddings are used to address the challenge that motived the present study: how to compare student responses to relevant reference responses based on the concepts and meanings that are behind words instead of the words themselves or their grammatical structure in sentences.

2.4.2.5. Use of Fixed Word Embeddings in Evaluation of Student Responses

To address the challenge mentioned in the preceding paragraph, NLP researchers have developed the following two-step method: first, both candidate and reference answers are converted into a high dimensional vector space using a text embedding model. Second, various distance measures are used to compare distance (semantic similarity) between the vectors of candidate and reference answers. Some popular distance measures are cosine distance, Euclidian distance, and Manhattan distance. The cosine distance is considered favorable as this does not include the length of text (magnitude of text vectors), which is regarded as irrelevant for measuring semantic similarity between two text statements (Kerkhof, 2020; Qiao & Hu, 2023). Similar to the Dice Coefficient, those distance measures are expressed using a real number between 0 ~ 1, where 0 represents no semantic similarity and 1 represents an exact match in meaning (Kerkhof, 2020a).

In the first scoping review on the use of word embeddings in ASAG, Putnikovic & Jovanovic (2023) found that a total of 17 research studies have been published that used word embeddings in ASAG. Further, they found most of the articles used word embeddings, mainly to estimate the similarity of student and model answers using the cosine similarity measure. As an example, Magooda et al. (2016) used Word2Vec and GloVe for the vector representation of student and reference answers. They then calculated the Manhattan distance between those two text categories—Manhattan distance is the total distance between two points if one follows a grid-based path between them. Those measurements were then used to train an SVM ML model for

automatic grading purposes. The authors found that their system's performance was comparable to that of a lexical and syntactic feature-based system like c-rater.

In addition to non-neural network ML models, semantic features extracted from word embeddings have also been used in training of modern neural networks for automatic grading or automatic classification systems. For instance, Ariely et al. (2023) and Jiang et al. (2020) used word embeddings followed by a LSTM model for scoring student responses in science education. They concluded that extracting semantic features using word embeddings and subsequently feeding them into neural sequential network models like LSTMs improved the performance of automated scoring systems.

Likewise, example studies that use word embeddings in automatic text classification systems are (Sun et al., 2019; Capuano et al., 2021). In the former example, Sun et al. (2019) used Word2Vec to feed an LSTM neural network for identifying binary classification of urgent posts in massive open online courses (MOOC). In the latter example, Capuano et al. (2021) developed an experimental setup to classify MOOC forum posts with a combination of word embeddings to train an attention-based hierarchical RNN. The data corpus contained 29,604 learner forum posts from 11 Stanford University public online classes within the Humanities, Medicine, and Education fields. The classification of MOOC forum posts was performed for the following attributes: subject area, domain topics, sentiment polarity, level of confusion, and level of urgency of a forum post. The authors concluded that their experimental setup was able to successfully detect subject, domain, sentiment, confusion, and urgency of forum posts, achieving an accuracy between 74% and 88%.

While some of these models illustrate the utility of applying embedding models in education settings, fixed word embedding models like Word2Vec, Doc2Vec, and GloVe were developed in the early 2010s, rendering them outdated. Other models have built on these original embedding models. For example, recently, Forsyth & Mavridis (2021) applied the newly developed at the time Spacy "en_vectors_web_lg" embedding model in their automatic grading system for high school students' short answers related to computer science concepts. de Araujo et al. (2023) and Almazova et al. (2021) used Wiki40b-lm-multilingual (Guo et al., 2020), and USE-multilingual2 (Yang et al., 2019) to feed into an LSTM neural network-based classifier, respectively. The reason is likely because USE-multilingual2 was pre-trained on questions-answers pairs from web forums, which have more informal communication, and the latter Wiki40b-lm-multilingual1 on Wikipedia articles, which are written as formal encyclopedia entries.

The purpose of the classifier was to categorize students' think-aloud protocols in an online environment.

In addition to being slightly outdated, word embedding models also may strip the full context of a document into individual words because these models are limited to converting individual words into vectors. To address this limitation, two strategies are often used. First, sentence embeddings can be created by summing or averaging individual word embeddings. Second, researchers can use purpose-built models that are designed to generate embeddings for larger units of text, such as entire sentences or documents. For example, Le & Mikolov (2014) developed the Doc2Vec models. The Doc2Vec model considers the order and context of words in a document during the vectorization process, which allows the model to preserve semantic relationships between different pieces of text in documents.

Bulut et al. (2022) utilized the Doc2Vec model in the qualitative coding of medical students' responses for the Situational Judgment Test, a common open-ended assessment for medical school admissions in North America. First, the authors developed codes based on a theoretical framework related to professionalism. Second, they used the Doc2Vec model to embed both the code list and student documents. Lastly, they employed a cosine-similarity measure between the centroid of embeddings for each document and predefined codes to determine semantic similarity and assign labels to response documents.

Using Doc2Vec, Romero-Gómez & Orjuela-Cañon (2022) compared it with TF-IDF and Word2Vec methods for the thematic analysis of biomedical engineering students' responses. These responses related to what they understood about Bioinformatics before and after taking the undergraduate Advanced Bioinformatics course. The vector representations of responses, before and after taking the course, belonging to the same student were compared using cosine distance. The authors concluded that TF-IDF showed fewer dissimilarities than Word2Vec and Doc2Vec between two documents belonging to the same students. Therefore, authors concluded that the TF-IDF text vectorization method might not be appropriate for analyzing the learning gains of students because the TF-IDF model cannot detect such differences.

Table 2.2: Summary of Studies that used Fixed Word Embeddings

Paper(s)	NLP Methods	Limitations
(Pribadi et al., 2017; Shaukat et al. 2021, Sahu and Bhowmick, 2020)	WordNet	Wordnet is a pre-built knowledge thesaurus, which a resource-intensive and task dependent
(Guerrero and Wiley, 2019; Magooda et al. 2016)	Word2Vec and Glove	All of the mentioned methods produce fixed embedding for words. For example, the word "port" could refer to a shipping port or a USB port, but fixed word embedding models would generate the same vector to represent both meanings.
(Ariely et al. 2023; Jiang et al. 2020; Sun et al. 2019)	Word2Vec + LSTM	
(Forsyth and Mavridis, 2021)	Spacy "en_vectors_web_lg"	
(de Araujo et al., 2023; Almazova et al. 2021)	Wiki40b-lm-multilingual1	
(Bulut et al. 2022)	Doc2 Vec	
(Romero-Gómez & Orjuela-Cañon 2022)	TFIDF, Word2Vec, and Doc2Vec	
(Capuano et al., 2021)	Word + Sentence embeddings	

The aforementioned studies have demonstrated that fixed word embeddings are useful, to some extent, in comparing student responses to reference responses based on the concepts and meanings. I summarized the research studies that I cited in the literature review section 2.4, along with their respective NLP methods and limitations, in Table 2.2. However, fixed word embeddings like Word2Vec and GloVe have a major limitation, to which I turn in the below paragraph.

Fixed word embeddings map each word to one context-insensitive vector. This is problematic when dealing with polysemy—which refers to the capacity of a word or phrase to have multiple meanings depending upon its context. For example, the word "port" could refer to a shipping port or a USB port, but fixed word embedding models would generate the same vector to represent both meanings. To address this limitation of fixed word embeddings, NLP researchers have recently developed contextualized word embeddings that can generate vector representation of a word based on its context (i.e., surrounding words) (Peters et al., 2018; Vaswani et al., 2017). Next, I turn to describing contextualized word embeddings.

2.5. Contextualized Word Embeddings

The contextualized word embedding models can be classified as: (i) sequence-based and (ii) attention-based (Ahmad et al., 2020; Haller et al., 2022). Sequence-based word embeddings process input text in an ordered manner and tend to focus more on short-term context information. An example of sequence-based word embedding is ELMo. As a result, they struggle to robustly capture longer-range dependencies in text. In contrast to sequential models, attention-based models, commonly known as TLLMs, process input text in parallel and use the attention mechanism to selectively focus on relevant parts of the input, enabling them to better capture long-range dependencies. Some examples of TLLMs are BERT, MPENT, XLNET, and GPT-3.5. Next, I explore how TLLMs have been used in analyzing open-ended responses.

2.6. Use of TLLMs in Evaluation of Student Open-ended Responses

There are two approaches to incorporate TLLMs for automated analysis (or scoring) of open-ended responses: (a) feature-based approaches or (b) fine-tuning-based approaches. In the feature-based approaches, the pre-trained TLLM is frozen and used only to generate contextualized vector representations of the input data. These representations are then extracted out of the TLLM and fed into a separate downstream task-specific ML model (Shaik et al., 2022). On the other hand, in the case of fine-tuned-based approaches, TLLMs can be fine-tuned for downstream tasks in a specific domain to further improve the accuracy of pre-trained TLLMs, a process called transfer learning (Emerson et al., 2023; Radford et al., 2021; Raffel et al., 2020). Transfer learning provides the possibility for NLP researchers to fine-tune TLLMs without incurring huge training costs. This is because NLP researchers can exploit existing features that TLLMs have learned from large corpora of text data during their initial development and training phase—e.g., BERT is trained on Wikipedia and digital book corpora (Devlin et al., 2019). Next, I review research studies where researchers have used TLLMs in feature-based approaches.

2.6.1 *Feature-based Approaches*

Researchers have not only utilized pre-trained TLLMs but also compared them with fixed word embeddings or lexical-feature based text analysis methods. Here I present two example studies that demonstrate that current TLLMs outperformed previous lexical-feature based or fixed word embedding based methods across NLP classification tasks.

In the first example, Liu et al. (2022) compared a three-way classifier model comprised of a BERT embedding that was then fed into a CNN (BERT-CNN) with other six baseline models to classify across several dimensions of cognitive and emotional engagement. The data corpus contained MOOC discussion form posts from 8,867 participants registered in an Introduction to Psychology course. The three lexical-feature based methods used in the study were: (i) TF-IDF fed into a KNN classifier, (ii) TF-IDF fed a Naive Bayesian classifier, and (iii) TF-IDF fed a Random forest classifier. The remaining three baseline methods were: (iv) the Word2Vec-based word embeddings fed into CNN, (v) the Word2Vec-based word embeddings fed into RNN, and (vi) the Word2Vec-based word embeddings fed into attention-based-LSTM. Compared with these six baseline methods, the BERT-CNN model improved the F1 score for emotional and cognitive engagement recognition tasks by 10% and 8%, respectively. In the second example, Riordan et al. (2020) compared (a) lexical-features fed into SVR algorithm, (b) GloVe embeddings fed into RNN algorithm, and (c) BERT for automatic grading of science explanations written by K-12 students. Their study found that even base BERT, a TLLM without fine-tuning, outperformed the other two conventional methods.

Further, Shah & Pareek (2022) conducted a literature review of NLP for automatic evaluation of short answers. They concluded that studies which used NLP approaches based on BERT reported more than 90% precision in their classification tasks. Besides TLLMs being used in supervised ML approaches, TLLMs are also combined with unsupervised ML approaches (such as clustering and dimension reduction) in NLP workflows. This is done to achieve dual purpose: minimize human input to reduce bias, and reduce resource costs in terms of number of analysts and person-hours (Haller et al., 2022; Kerkhof, 2020a). The following paragraph gives two such examples.

As the first example, Chang et al. (2021) applied the NLP approach comprised of WordNet, BERT, and clustering for thematic analysis of 5,000 responses collected from healthcare workers about working during the COVID-19 pandemic. First, the authors used WordNet to normalize linguistic variations in responses. For example, the phrases “video visit”, “video call” and “virtual meeting” were normalized to “video call”. Second, those comments were vectorized

using the BERT model. Third, they used the t-SNE dimension reduction algorithm to identify groups of similar comments. Lastly, the authors read those similar comments for thematic analysis. In their paper, the authors corroborated the utility and efficiency of their NLP method. It allowed their team to complete thematic analysis of 5,000 text responses in two days rather than the two months their team would have needed without NLP. As the second example, from the physics education field, Wulf and their team developed an NLP workflow comprised of the following three steps: first, sentence embedding through BERT; second, dimension reduction through UMAP; and third, clustering via HBDSCAN (Wulff et al., 2021, 2022a, 2023). Their data contained reflection responses of pre-service physics teachers while watching a video vignette. Wulff et al. concluded that their NLP workflow was successful in classifying teachers' written reflections according to the reflection-supporting theoretical model.

In addition to feature-based approach for TLLMs, in another study, Wulff and colleagues fine-tuned BERT for reflection classification tasks using the same data (reflection responses of preservice physics teachers and found improved performance compared to base BERT (Wulff et al., 2022b). Next, I provide examples of how researchers have used fine-tuning of TLLMs for ASAG.

2.6.2 *Fine-Tuning-based Approaches*

As the first example of use of TLLMs in fine-tuning based approaches, Sung et al. (2019) fine-tuned the base BERT model by augmenting data from psychology-specific resources for ASAG. On two psychology related ASAG datasets, the authors demonstrated an improvement of 6% to 10% in accuracy as compared to the base BERT. As the second example, Camus and Filighera (2020) experimented with fine-tuning different TLLMs such as BERT, RoBERTa, AlBERT, XLM, and XLMRoBERT through the process of knowledge distillation. They reported an improvement of up to 13% in macro-average-F1 score over SciEntsBank. Through knowledge distillation, the authors found that the BERT and its variant could be fine-tuned using only a few human-annotated examples on NLP-assisted text classification tasks. These fine-tuned BERT models could approximate BERT's benchmark performance at a fraction of the computational cost. Although effective, one should note that fine-tuning TLLMs on a specific domain and task may limit its cross-domain generalization (Ahmad et al., 2022a).

Another limitation of the aforementioned TLLMs is that they generate word-level contextualized embedding. To address this limitation, researchers have developed sentence-level TLLMs to extract the semantics of longer text segments such as sentences or paragraphs for tasks like ASAG. An example of sentence-level TLLM is the Sentence BERT (SBERT) model,

introduced by (Reimers & Gurevych, 2019). Condor et al. (2021) compared the SBERT with Word2Vec and BOW models for ASAG dataset. The dataset was collected in the 2019 field test of a Critical Reasoning for College Readiness assessment by the Berkeley Evaluation and Assessment Research Center. Authors found that overall SBERT performed better than the other two models investigated in their study. Although, their results are not promising for the generalizability of auto grading models to unseen questions.

Therefore, the sentence-level TLLMs are also fine-tuned to improve performance. For example, Ahmed et al. (2022b) fine-tuned SBERT for ASAG Mohler's dataset from the computer science field. Ahmed et al. found that their fine-tuned SBERT model outperformed conventional BERT, SBERT, and GloVe embedding models for automatic grading of Mohler's dataset. The authors recommended that combining sentence-level embedding from TLLMs with fine-tuning approaches can enable NLP workflow to extract nuanced semantic information, thereby substantially improving the accuracy of automatic evaluation methods for student writing.

In Table 2.3, I list the publications with their NLP methods those I cited in section 2.6. Moreover, given these recent findings, the research community has been concluding that TLLMs offer substantial advantages over prior NLP-based approaches to ASAG. However, in a recent literature review about deep learning approaches in ASAG, Haller et al. (2022) concluded that the best-performing models in ASAG tasks are those that combine hand-engineered features with TLLMs. As an example the biology field, TLLMs can capture general semantic understanding of student answers but might not capture specific nuances of a correct biology answer. This is where hand-engineered features are helpful to capture the use of specific biological terms. Therefore, the authors suggest that by combining the TLLM's broad understanding of language with these specific hand-engineered features, ASAG can both understand semantics of the answer and pay attention to the specific technical terms that are crucial for domain specific answers.

Table 2.3: Summary of Studies that used Transformer-based Large Language Models

Paper(s)	NLP Method(s)
(Liu et al., 2022)	BERT into CNN; TF-IDF fed into: (i) a KNN classifier, (ii) a Naive Bayesian classifier, and (iii) a Random forest (RF). The other Word2Vec-based models were (iv) Convolutional neural networks, (v) RNN, and (vi) attention-based-LSTM.
(Chang et al., 2021)	WordNet, BERT, and clustering
(Wulff et al., 2021, 2022a, 2023)	first, sentence embedding through BERT; second, dimension reduction through UMAP; and third, clustering via HBDSCAN
(Wulff et al. 2022b)	fine-tuned BERT for classification task of reflection responses of preservice physics teachers
(Sung et al., 2019)	BERT model was augmented with data from psychology-specific resources for ASAG
(Camus and Filighera, 2020)	fine-tuning BERT, RoBERTa, AlBERT, XLM, and XLMRoBERT
(Condor et al., 2021; Ndukwe et al., 2022; Ahmad et al., 2022)	SBERT model

2.7. Use of NLP Approaches in Engineering Education

In this section, I review examples in the existing literature about how the NLP approaches reviewed above have been used in engineering education field. I organize this section as follows: first, I discuss lexical features-based NLP methods. Second, I review syntactic feature-based NLP methods. Third, I summarize research studies utilizing topic modeling techniques. Last, I provide an overview of engineering education literature related to the use of semantic feature-based NLP methods including TLLMs.

2.7.1 *Use of Lexical Features in Supervised ML*

Lexical feature-based NLP approaches use frequency-based (words counts) and dictionary-based (word matching) techniques to extract lexical features of textual data. These features are combined with supervised ML algorithms to automate text analysis tasks. Next, I provide examples of both frequency- and dictionary-based techniques in the following paragraphs.

As an example of frequency-based technique, Soledad et al. (2017) analyzed undergraduate engineering students' responses on the Student Perception of Teaching Survey at Virginia Tech. First, the authors manually labeled those open-ended responses according to the theoretical constructs of the Success and Caring components in the MUSIC Model of Academic Motivation. Second, they used those annotated responses to train a TF-IDF model to automate the classification process. To remind the reader, the TF-IDF model is a method for text vectorization in which the textual features in the document are represented by weighted frequencies of individual words.

Another NLP scholar from Virginia Tech, Bhaduri & Roy (2017) used the TF-IDF approach to analyze the mission statements of 59 engineering colleges in the U.S.: 29 public, and 30 private. First, authors extracted 713 unique unigram tokens present in the corpus after the stop-words (e.g., articles) were removed. Those pre-processed documents with unigram tokens were passed through the TF-IDF feature extractor. Each document is encoded in a 713-dimensional feature space and the output of the TF-IDF feature extractor resulted in a 59x713 dimensional matrix. Their study found that there were indeed differences in the vocabulary of words used in mission statements of public versus private engineering colleges.

In her second study, Bhaduri (2018) collected responses from 152 first-year engineering students at Virginia Tech in a metacognition class intervention. First, she manually coded student responses into a three-way metacognition level: high, medium, and low. After developing the annotated dataset, she extracted the following lexical features: number of sentences, number of tokens, and number of POS tags in each student response. Those features were then used to train the following three supervised ML classifiers: SVM, logistic regression, and random forest. She concluded that the random forest classifier was more accurate than the other two classifiers.

One more example of a study using the TF-IDF model is (Verleger, 2014). The author classified engineering student team performance on Model-Eliciting Activities (MEAs). The MEAs are open-ended engineering problems where teams of students produce a written document describing the steps to solve a given engineering problem. The authors used word frequencies and corresponding grading rubric's category labels to train a decision tree ML model. Subsequently,

that ML model was used to automatically classify new student responses into different categories of the MEA rubric.

As an example of dictionary-based lexical features, Berdanier et al. (2020) used these features to analyze a corpus of 54 interview transcripts about graduate engineering student career preparation. The authors created two codebooks about the theoretical framework of Community of Practice: expert-hand-curated and machine-generated. The first dictionary comprised 69 words and phrases. The second dictionary was formed by mining the theoretical framework sections of 14 journal articles that primarily employed the Community of Practice theory. For analysis purposes, every interview transcript was vectorized and every value in a vector indicated how many times the corresponding dictionary word appeared in that transcript. After vectorization, the data was clustered and represented in high-dimensional space using PCA visualization and pairwise-distance plots.

In another study, Berdanier et al.(2018) demonstrated two examples of the applications of NLP in engineering education. In the first example, the authors combined both frequency- and dictionary-based features for discourse analysis of 500 engineering résumés according to descriptions of engineering competencies developed by the American Association of Engineering Societies. First, they handcrafted a dictionary of labels by coding 100 engineering résumés. Second, the authors developed an NLP workflow in Python computer language that tallied up the frequencies of those code words in the remaining 400 engineering résumés. In the second example, Berdanier and her colleagues developed a classifier for genre analysis of research articles according to Swales rhetorical moves. The data corpus comprised literature review sections of papers published in 2017-2018 in the *Journal of Propulsion and Power*, a journal published through the American Society of Mechanical Engineers. First, frequency-based lexical features were extracted by POS counts and then those were fed into a LSTM classifier for genre analysis.

Following lexical feature-based methods, in the next section, I provide example studies from engineering education literature that used syntactic-feature-based NLP approaches.

2.7.2 *Use of Syntactic Features in Supervised ML*

Syntactic feature-based NLP methods rely on grammar or structural roles of words in sentences. In engineering education, researchers have combined syntactic features of text with text vectorization methods. For instance, Jayakodi et al. (2015) used WordNet similarity to classify engineering course exam questions according to Bloom's taxonomy. First, verbs (POS tags) in exam questions were extracted, and then those along with Bloom's taxonomy verb lists were

vectorized using WordNet. The cosine similarity score between verbs of exam questions and Bloom's taxonomy were used for exam question classification according to Bloom's taxonomy.

Using WordNet, Arbogast & Montfort (2016) calculated lexical diversity index in semi-structured interviews of undergraduate engineering students to understand their mental processes during engineering problem-solving. First, authors completed POS tagging in transcripts with the Stanford POS tagger. Second, they used WordNet to group those POS tags based on similarity in meanings. Third, they calculated a lexical diversity index that incorporated those grouped POS tags. Authors found that a large amount of engineering jargon was used by engineering students in interviews.

Additionally, the syntactic-features have been incorporated in mobile educational applications. An example is CourseMIRROR that has been used in engineering classrooms to analyze and generate summaries of in-situ student reflections (Butt et al., 2022; Fan et al., 2015, 2017). The CourseMIRROR functions in three steps. First, it uses POS tagging to segregate noun phrases in students' reflections. Second, these noun phrases are clustered based on semantic similarity via Latent Semantic Analysis and the K-Medoids algorithm. Third, representative phrases in each cluster are chosen via the LexRank model for instructor consideration to achieve the following purpose: phrases mentioned by more students should attract more attention from the instructor.

Next, I provide an example study on topic modeling, an unsupervised ML method, from the engineering education literature.

2.7.3 *Use of Unsupervised ML*

The LDA algorithm is a probabilistic approach that presumes documents are comprised of various topics, and each topic is characterized by a distinct distribution of words (Blei et al, 2003). An example study of the LDA algorithm in engineering education is (Nanda et al., 2022). They utilized LDA to examine themes in peer-to-peer comments of first-year engineering students in their engineering foundation courses. Students were directed to provide constructive feedback in writing to themselves and teammates on their teamwork behaviors via the Comprehensive Assessment of Team-Member Effectiveness (CATME) interface. This tool is commonly used to manage undergraduate engineering teams in U.S. colleges. As of 2019, CATME had over 7,000 active instructor accounts across more than 2,000 institutions worldwide (Wang et al., 2019). Given the richness of data collected through this platform, engineering education researchers have also used the CATME data set in semantic feature-based NLP methods, the methods to which I turn next.

2.7.4 *Use of Semantic Features extracted with TLLMs*

Continuing the example of the CATME dataset, Wei et al. (2020) introduced an NLP-based pipeline tool for de-identifying peer-to-peer comments. This task, similar to POS tagging, was performed at the word level in a sentence. The authors used a combination of GloVe—a fixed word embedding model—and neural network to identify and replace names with pseudonyms. Another example study related to the CATME dataset is (Wang et al., 2019). The authors converted peer-to-peer comments from the CATME survey into a numerical scale using combination of word embeddings from RoBERTa and neural network classifier. First, the authors manually rated CATME survey responses on a scale of 1-5 and embedded those rated responses using RoBERTa. Second, those embeddings were fed into neural network-based classifier for training purposes. This workflow achieved an F1 score of 0.67 on the testing dataset.

Continuing a prior example of the CourseMIRROR mobile application, Magooda et al. (2022) recently updated the application with TLLM. They integrated the DistilBERT—a distilled version of BERT— into the application followed by a reflection quality prediction module based on the SVM model. The purpose was to allow real-time feedback for students as they write and submit reflections. Butt et al. (2022) showed the efficacy of CourseMIRROR for enhancing students' engagement in engineering classrooms.

Moreover, engineering education researchers have compared semantic feature-based approaches with lexical and syntactic features-based approaches. For example, Becker et al. (2019) compared lexical- and BERT-based NLP approaches to evaluate misconceptions in electrical circuit design related short question answers. In the lexical-based approach, they created various rules that comprised a search for an ordered set of between two and four electrical circuit design related keywords. After establishing those rules, they were applied to each answer on a sentence-by-sentence basis for identifying misconceptions. On the other hand, the authors used BERT for predicting binary classification: each answer contained a sentence conveying a misconception (positive), or it did not (negative). The precision score for the BERT model was 0.90 as compared to 0.63 for lexical-based approach.

To further improve accuracy of TLLMs, engineering education researchers have also fine-tuned those pre-trained models according to engineering education contexts. For instance, Ganesh et al. (2022) fine-tuned RoBERTa—an optimized version of BERT—for three-way classification (positive, neutral, and negative) of industrial engineering students' responses about engineering identity and transformative experiences. The authors fine-tuned RoBERTa for sequence classification using the Michigan Electrical Engineering and Computer Science Targeted

Sentiment Analysis Dataset. They found that the fined-tuned RoBERTa model improved macro-F1 score by 6.7 from 48.4 to 55.1 from the base RoBERTa model.

In the engineering education community, Katz et al. (2021) have introduced a TLLMs-based human-in-the-loop-NLP (HILNLP) workflow for text analysis. This workflow consists of the following steps: (1) pre-processed data is embedded in a high-dimensional space (ranging from 768 to 1,024) using TLLMs, (2) this high-dimensional space is then reduced to a range of 5-10 dimensions through a combination of linear (PCA) and nonlinear (UMAP) dimension reduction processes, (3) clustering algorithms (HBDSAN) are used to produce/identify homogeneous clusters, and (4) these clusters are qualitatively coded by a researcher to identify potential themes.

Katz et al. (2021) suggested that their HILNLP allowed qualitative researchers to handle larger data volumes while decreasing the time and coordination efforts needed for team analysis. In (Katz et al., 2021), the HILNLP approach reduced the time from 12 hours to just 3 hours in analyzing over 3,000 student SPOT survey responses. The HILNLP approach has been utilized for analyzing various data corpuses, including: (a) how engineering faculty members define assessments (Chew et al., 2022); (b) students' semi-structured interviews about social justice issues (Shakir et al., 2022); (c) students' post-semester surveys about the Engineering Projects in Community Service program (Anakok et al., 2022); and (d) student self-reflections (Gamieldien, Case, et al., 2023, 2023; Gamieldien, McCord, et al., 2023). In Table 2.4, I list the publications I summarized in section 2.7 and note their NLP methods.

Table 2.4: Summary of Studies that use NLP from the Engineering Education Literature

Paper(s)	NLP Method(s)
(Soledad et al., 2017; Verleger, 2014)	BOW
(Bhaduri & Roy; 2017; Berdanier et al. 2018a)	TF-IDF
(Berdanier et al., 2018b)	SVM, logistic regression, and random forest POS + LSTM classifier TF+PCA visualization and pairwise-distance plots.
(Jayakodi et al. 2015; Arbogast and Montfort 2016)	POS tagging + WordNet
(Fan et al., 2015, 2017)	POS tagging+ LSA and the K-Medoids algorithm.
(Magooda et al., 2022)	DistilBERT—a distilled version of the BERT+ SVR machine learning algorithm
(Nanda et al., 2022)	LDA
(Wang et al. 2019)	RoBERTa+ neural network-based classifier
(Wei et al. 2020)	GloVe+ neural network
(Becker et al. 2019)	Keyword + BERT
(Ganesh et al. 2022)	Fine-tuned RoBERTa
(Katz et al., 2021; Chew et al. 2022; Shakir et al. 2022; Anakok et al., 2022)	MPNet, BERT,

2.8. TLLMs-specific Limitations and their Environmental, Financial, and Social Impacts

Even though TLLM-based models have shown promise in their education applications, they are not without their own drawbacks. I classify impacts of TLLMs into two categories: (a) environmental and financial impacts, and (b) social impacts. In this section, I summarize the literature on these impacts.

Regarding environmental and financial impacts, researchers have cautioned about impacts of TLLMs due to their huge resource consumption (Bender et al., 2021; Dodge et al., 2022; Rillig et al., 2023). For instance, Strubell et al. (2019) calculated the energy required to train the BERT model was equivalent to a flight from the US to England. This resource usage relies heavily on non-renewable energy sources and increases carbon emissions. In addition to carbon emissions, data centers can cause other environmental issues like high water usage and potential soil pollution (Bender et al., 2021; Dodge et al., 2022; Rillig et al., 2023). According to these studies, this resource consumption in developing TLLMs is more likely to disproportionately affect marginalized communities. Moreover, these communities often do not benefit from the resulting language technology. Another downside of the high financial costs of developing TLLMs is that it raises entry barriers, thereby limiting who can contribute to this research area and which languages can fully benefit from these technologies. For instance, more than 90% of the world's languages spoken by over a billion people have minimal to no support in terms of language technology (Bender et al., 2021). To mitigate the challenge of resource consumption and for transparency, some researchers have suggested that TLLM developers should report the resource costs. With that information, users of TLLMs can then consider the trade-offs between resource consumption and TLLMs' performance (Bender et al., 2021; Dodge et al., 2022; Rillig et al., 2023).

Regarding social impacts, researchers have discussed impacts of TLLMs such as biases, privacy threats, and increased misinformation. First, TLLMs have the potential ability to propagate social biases and stereotypes when used in downstream NLP tasks (Bartl et al., 2020; Gonen & Goldberg, 2019; Ullmann, 2022). TLLMs are trained on textual data which primarily originates from the internet. The textual data available on the internet has social biases and stereotypes embedded in it. Therefore, TLLMs trained on this data also inherit those biases and stereotypes. As an example of gender bias, in these kinds of models, an artifact from training on biased data on the internet is the model's learned association between "doctor" "man" and a second association between "nurse" and "woman." Moreover, some researchers claim that TLLMs encode more bias against identities marginalized along multiple dimensions (Caliskan, 2021; Crenshaw, 1990). Although researchers have proposed de-biasing methods, complete removal of social biases and stereotypes from TLLMs is undesirable (Bender & Friedman, 2018; Gonen & Goldberg, 2019). In TLLM research community, there is a debate on whether completely removing biases from TLLMs should even be the goal. This is because a completely bias-free TLLM might be just an inaccurate representation of our lives and societies. Second, ill-motivated users could potentially access personal information from TLLMs because they are trained on public internet data (Bender et al., 2021). Third, high-quality output from TLLMs can appear truthful and such outputs can

easily be mistaken for expert opinions.This is because current TLLMs do not have true natural language understanding and they simply operate on prediction based on patterns in the training data (Bender et al., 2021; Johri et al., 2023).

I endorse the aforementioned environmental, financial, and social impacts and researchers should consider those impacts when using TLLMs in their work. However, I believe TLLMs can be used to glean nuanced insights from textual data at a large scale; but they must still be deployed judiciously

Next, I describe the existing assessments methods for ethics and systems thinking—the use cases in this dissertation study.

2.9. Existing Ethics Assessment Methods

Recently, Kim & Bairaktarova (2023) conducted a literature review of existing engineering ethics instruments and found that most focus on measuring individual students' abilities at the individual level. Furthermore, I categorized these ethics instruments into two main types: psychometric instruments and case studies. Psychometric instruments tend to be quantitative in nature, while case studies are typically qualitative (Hess et al., 2023).

2.9.1 *Ways of Categorizing Psychometric Instruments*

Psychometric ethics assessment instruments can be categorized across several dimensions: (i) measured ethics construct, (ii) format, (iii) theoretical framework, and (iv) original target population.

Categorizing these ethics assessments by their measured constructs produces several groupings. For example, many instruments focus on ethical sensitivity, defined as an ability "to identify and recognize relevant ethical issues emerging from a situation" (Borenstein et al., 2008, p.13). This is also similar to ethical issue recognition, as used in the Test for Ethical Sensitivity in Construction (TESC) (Sands II et al., 2020; Sands & Simmons, 2014). Another construct measured is ethical knowledge, as in the study of ethics and curricular experiences from Finelli et al. (2012). This can be defined as knowledge of ethical principles. A third common construct measured, appearing in EERI, ESIT, SER, and DIT, is ethical reasoning. Ethical reasoning is defined as an ability to apply moral theories or logical arguments to reason through ethical dilemmas. With a slightly different view towards ethics related constructs, one could measure interest in ethics, perceptions of the value of ethics education, feelings of autonomy in the classroom activities related to ethics, feelings of connection or relatedness with their classmates, perceptions of one's own competence when it comes to ethical issues, and understanding of systems thinking as it

relates to ethical issues. These were measured in the Survey of Ethical Reasoning (SER) (Lewis et al., 2019). Bloom's taxonomy provides an alternative schema for organizing ethics assessments as some instruments measure understanding of ethical theories or professional engineering codes of ethics while others measure higher order constructs such as an ability to apply ethical theories in ethical dilemmas and evaluation of ethical decision-making (Junaid et al., 2021). Finally, one can try to measure ethical behavior, focusing on realized or hypothetical responses to ethical issues. Relevant issues pertaining to ethical behavior for students may include cheating, volunteerism, or boundaries around shared work on course assignments (Finelli et al., 2012).

Beside categorizing by constructs measured, I categorize by the formats of ethics assessment instruments, which can include closed-ended items, open written response questions, and even oral responses. For example, in the EERI, participants are given six separate case scenarios related to ethics dilemmas such as issues of safety, design standards, and constraints of cultural norms. A participant reads a scenario and is then presented with twelve unique items that may bear on an ethical decision about the dilemma (Odom & Zoltowski, 2019). The participant is then asked to rate each item on an ordinal scale in reference to how significant they think the item is when considering how to respond to the dilemma, after which they are asked to rank the top four items which they reasoned to be most important. In a similar format, in the EDM, participants receive twelve scenarios relevant to their field research setting (i.e., health, social, or biological science). Each scenario provides contextual and background information for six ethical (and technical) events that followed. For each event, six to eight action items are provided to participants as a potential response to resolve the issue. They are then asked to select two different responses that they felt would most likely resolve the problem. Other classic closed-ended formats ask respondents to rank a series of statements in terms of their perceived importance of ethical reasoning and their confidence in applying the ethical reasoning process as on the SER (Lewis et al., 2019). For more open-ended formats, participants can receive case scenarios related to industry and respond to a prompt to "reflect on the situation, and write down at least 3 issues you are concerned with and/or questions you may have about the situation, and please be as descriptive as possible" (Sands II et al., 2020, p.9). For a second example, in the Moral Judgment Interview (MJI), three hypothetical moral dilemmas are presented to participants in an oral interview (Colby et al., 1983). Each dilemma is then followed by 9-12 standardized probe questions designed to elicit participant's moral judgment process.

Of course, most of these assessments build upon various theoretical frameworks. Many assessments refer to a neo-Kohlbergian moral development theory (Colby et al., 1983), e.g., DIT, EERI, SER, SRM, and ESIT. The ethical sensitivity and moral imagination framework developed

by Johnson and Werhane (Johnson, 1994) is a second example and is used as the foundation for TESSE. A third is Terenzini and Reason's College Impact Model (Terenzini & Reason, 2005), as applied in SEED. The TESC uses Rest's four component model of cognitive moral development (Rest, 1986). Finally, some reference taxonomies such as the taxonomy of ethical behavior from (Helton-Fauth et al., 2003) or even Bloom's taxonomy, as mentioned above.

Finally, one can categorize ethics assessment methods by their original target populations. For example, the EERI, SEED, TESSE, and ESIT have each been developed (or used) with undergraduate engineering students (Borenstein et al., 2008, 2010 Finelli et al., 2012 Odom & Zoltowski, 2019). The DTEC on the other hand has been used with undergraduate students working in design teams (Oakes et al., 2011). On the other end of the higher education spectrum, the SkillSET has been used with graduate and professional students (Berry et al., 2013). Some instruments have been used across that spectrum, such as the PMEAR (Rudnicka et al., 2013). Finally, others such as the SER and the ESIT have been applied beyond engineering to students across STEM disciplines.

Despite the availability of the aforementioned ethics assessment methods, case studies are one of the most common methods to teach and assess ethics in engineering programs. Illustrating this point, in a literature review study of ethics education interventions, Hess and Fore (2018) found that 80% of reviewed papers in their study of ethics interventions incorporated case studies as the way to teach and to assess engineering ethics content. Therefore, due to common teaching and assessment method and availability of student responses, I selected an engineering ethics case study as a use case for demonstrating my NLP approaches.

2.9.2 *Ways of Applying and Accessing Case Studies*

(Hess and Fore, 2018; Hess et al., 2021) and (Martin et al., 2021) have categorized case studies used in engineering ethics in the US and Irish engineering education contexts, respectively. The authors used the following categorizing variables: (a) historical versus hypothetical, (b) thick information versus thin information, (c) evaluative versus participative, and (d) featuring macro issues versus micro issues. As an example, (Ermer, 2004) is a hypothetical, thin information, participative case study featuring a micro issue about a catalyst used in a product. Engineering ethics cases typically include individualistic, hypothetical, and historical scenarios (Hess and Fore, 2018; Hess et al., 2021; Martin et al., 2021). Martin et al. (2021) concluded that their study faculty participants highlighted the need to switch from hypothetical scenarios towards more realistic case settings.

2.10. Existing Systems Thinking Assessment Methods

Systems thinking educators use a battery of measurement methods to assess students' learning outcomes in systems thinking-related courses. Demonstrating this point, Dugan et al. (2021) conducted a systematic literature review of 27 systems thinking assessments in engineering and found that 16 of the 27 assessments targeted professionals and/or postsecondary students in engineering. In this section, I categorize (and summarize)—similar to ethics assessments in the previous section—the systems thinking assessments used in the engineering education context. I have categorized systems thinking assessments across the following dimensions: (i) theoretical constructs measured, (ii) measurement methodology, (iii) format, and (iv) medium used for student responses submission.

The first categorizing variable I used is theoretical constructs. Commonly measured constructs in systems thinking assessment instruments are (a) ability to identify elements of systems and (b) ability to identify relationships at various levels for constructing a problem space with specific boundaries between the elements of systems. However, the terminology used to refer to these two constructs varies across assessment instruments. For instance, (a) is termed as individual objects and processes in the Systems Thinking Assessment Rubric (STAR) (Lavi et al., 2020, 2021) and concepts in the Systemic Synthesis Questions (SsynQs) (Hrin et al., 2016). Other instruments used the following terminologies: components, system structure, key variables, and terms (Brandstädter et al., 2012; Hrin et al., 2016; Keynan et al., 2014; Lavi et al., 2020, 2021; Meilinda et al., 2018; Rehmann et al., 2011). Related to (b), instruments not only recognize relationships between system elements but also levels of relationships. These levels of relationships are key to characterizing complexity of identified systems or problem spaces. For example, the Climate Change Systems Thinking Instrument (CCSTI) includes both identifying relationships within one level of organization and analyzing those relations across different levels of organization (Meilinda et al., 2018). Likewise, the STAR instrument asked participants to examine structural and procedural relations within individual objects and processes. Additionally, participants were instructed to include the refinement of those relations into hierarchical functions (Lavi et al., 2020, 2021). Grohs et al. (2018) included a separate section in their scoring rubric that checks if participant responses are aligned across different aspects of systems thinking skills in their response. Taylor et al. (2020) considered identifying roles/purposes for each system element as part of the relationship construct. Lastly, feedback loops represent advanced forms of relationships between system elements. Identifying those feedback loops was explicitly included in the evaluation rubrics of (Davis et al., 2020; Hu & Shealy, 2018; Meilinda et al., 2018; Sweeney & Sterman, 2000).

In addition to constructs (a) and (b), another theoretical construct measured in systems thinking assessments is the identification of factors that influence a given system or problem space. Many assessment methods go beyond the traditional technical factors and include social, economic, environmental, political, and legal aspects of a given problem space. Frank (2010) referred to these factors as engineering and non-engineering consequences. Jaradat (2014) labeled them as non-technical issues, while (Camelia et al., 2018; Camelia & Ferris, 2018) described them as political, social, and environmental responsibilities. Notably, both Grohs et al. (2018) and Hu & Shealy (2018) incorporated contextual aspects throughout their scoring rubric, and explicitly included the identification of stakeholders in rubrics. In the case of Hu and Shealy (2018), stakeholder considerations were one of several dimensions influencing a holistic score. Meanwhile, Grohs et al. (2018) evaluated awareness of stakeholders as a distinct construct.

Lastly, another construct measured in assessment methods was temporal awareness, which is an ability to account for dynamic behavior of elements, relationships, and influencing factors. For example, Grohs et al. (2018) included an assessment question about identifying both short-term and long-term goals and consequences. Another assessment that emphasized time was (Keynan et al., 2014) that involved thinking temporally as an advanced systems thinking competency. Furthermore, STAR included temporary objects and decision nodes as a systems thinking attribute in its rubric (Lavi et al., 2020, 2021).

The second categorizing variable I used is measurement methods. The constructs of systems thinking have been measured in various ways, including behavior-based, preference-based, self-reported, and cognitive methods (Dungun et al., 2021). In their systematic literature review of systems thinking assessments in engineering, Dungun et al. (2021) concluded that 19 out of the 27 instruments included in their review paper were behavior-based. Behavior-based measurement methods evaluate a participant's knowledge or skill based on their performance in specific tasks such as drawing a concept map, answering open-ended questions, or completing a fill-in-the-blank activity. Examples of behavior-based systems thinking instruments include the CCSTI and the STAR; the former includes multiple-choice questions, while the latter encompasses concept-mapping (Meilinda et al., 2018; Lavi et al., 2020, 2021).

Second, preference-based assessments aim to characterize individuals' values and aptitudes towards systems thinking perspectives. These instruments require participants to indicate the extent to which a statement aligns with their values and interests on a scale (Camelia et al., 2018; Kordova & Frank, 2018). For instance, Castelle & Jaradat (2016) developed participants' systems thinking profiles based on their responses to 39 binary questions presented in a

cybersecurity case scenario. This method of developing a systems thinking profile is also integrated into the authors' virtual reality gaming-based assessment (Jaradat et al., 2019). Interestingly, Dungun et al. (2021) compared behavior- and preference-based assessment methods in engineering and found that preference-based assessments tended to push beyond a narrow technical focus more than behavior-based assessments. However, as mentioned earlier, the majority of systems thinking assessments are behavior-based. This suggests that engineering educators may undervalue the importance of considering broader contextual aspects of a problem when evaluating systems thinking skills in engineering students.

Third, self-reported assessments ask participants, rather than an external observer, to provide their own evaluations of their understanding and knowledge of systems thinking competencies. For example, Hadgraft et al. (2008) asked students to rate their learning of 14 systems thinking skills. Similarly, the Engineering Systems Thinking Survey, developed by Degen et al. (2018), was divided into two sections. The first section contains Likert-scale questions on self-efficacy regarding systems thinking skills, while the second section measures knowledge and skills through multiple-choice questions. However, Davis et al. (2023) compared performance of engineering students on a self-report assessment and a scenario-based assessment used in engineering education. Their findings indicated that solely using self-report assessments have limitations and suggested educators should incorporate other assessment formats. Fourth, cognitive-based assessments measure brain activity during solving systems thinking problems. An example of this approach is Hu and Shealy's Assessment (2018), where participants wore a functional near-infrared spectroscopy (fNIRS) cap to monitor brain activity during concept mapping activities.

The third categorizing variable I used is format. Instruments can also be categorized based on their format into either closed- or open-ended types. For closed-ended instruments, participants are presented questions with predefined answers or are asked to complete missing information. Keynan et al. (2014) is an example of the multiple-choice format, in which participants were presented with 15 terms related to an ecosystem. They were instructed to select three terms from the list, and among the three chosen terms, two needed to be similar to each other but different from the third term. Another example of a multiple-choice assessment is the CCSTI instrument (Meilinda et al., 2018). Example assessments of the fill-in-the-blank format are (Hrin et al., 2016; Sweeney & Sterman, 2000; Timofte & Popuș, 2019), where participants were provided a diagram or graph for completion.

The primary examples of open-ended assessments are scenario-based instruments. In these, participants are presented with a realistic or fictitious problem followed by a series of open-ended questions. Open-ended questions can vary in their complexity and the degree of scaffolding provided to elicit participant responses. For instance, the Systems Assessment Test (SysTest) used a single prompt where, after reading a customer needs statement, students were directed to describe the system (Tomko et al., 2017). Conversely, Grohs et al.'s (2018) assessment instrument consisted of multiple prompts designed to operationalize systems thinking constructs, using the scenario of heating problems in the fictitious town of Abeesee. Davis et al. (2020) developed an assessment tool which comprised several paragraphs describing the real-world shrinkage of Lake Urmia in northwest Iran. Apart from written responses, one part of the Brandstädter et al. (2012) instrument asked participants to draw non-directed concept maps after reading a scenario about the Blue Mussel in the sea ecosystem.

The fourth categorizing variable I used is medium used for response submission. One could categorize assessment methods based on the medium used to represent participant responses, which include textual, oral, visual, and simulation. First, (Davis et al., 2020; Grohs et al., 2018) are examples of instruments where participants conveyed their response in textual medium. Second, (Rehmann et al., 2011) asked students to submit their responses as oral presentations. Third, visual-based responses vary in the degree of structure provided to draw visuals. For example, the STAR instrument follows a highly structured conceptual model approach based on object-process methodology (Lavi et al., 2020, 2021). Conversely, (Vanasupaa et al., 2008) allowed participants to draw free-form rich visuals. Fourth, (Brandstädter et al., 2012) and (Hu & Shealy, 2018) blended both concept-mapping and written mediums. Lastly, Jaradat et al. (2019) used a virtual reality gaming environment where counts of touch controller characterized how students react to uncertain situations.

2.11. Chapter Summary

The NLP field has evolved from rule-based to statistics-based to modern TLLMs. This development has enabled researchers to develop computer applications that can process, understand, and generate human language in more flexible and comprehensive ways. This chronological development in the NLP field coincides with how NLP tools have been incorporated into open-ended educational assessment. In the last fifty years, NLP and education researchers have used a variety of lexical and syntactic text features, combined with both supervised and unsupervised ML approaches to assess students' written responses. Despite their useful performance in some contexts, lexical features can still fail to capture the meanings of sentences, and syntactic features can only do so to a limited degree. On the other hand, capturing how

students' responses and reference responses are connected not by their words or sentence structure but by their meaning and concepts, can significantly enhance the performance of automatic grading of student written response. To achieve this purpose, TLLMs have been used recently. Given that TLLMs have been available in the last five years, there is a significant lack of research related to integrating TLLMs in the evaluation of open-ended case scenarios, broadly within engineering education. My dissertation study aim to fill this research gap.

Chapter 3: Research Methods

3.1. Chapter Overview

In Chapter 3, I describe the methodology used to answer my RQ. First, I describe two use cases and provide my rationale for selecting question prompts from those to demonstrate application of my NLP approach. Next, I discuss steps involved in data analysis. These sequential steps are as follows: (i) pre-processing of text, (ii) developing example banks through two approaches: (ii-a) human in the loop natural language processing (HILNLP), and (ii-b) traditional qualitative coding, and (iii) assigning codes to unlabeled student responses through two NLP methods: (iii-a) k nearest neighbor, and (iii-b) zero-shot classification. Finally, I elaborate the procedure to evaluate accuracy of those assigned codes to answer my RQ. Notably, I document not only the NLP algorithms used to implement the aforementioned steps, but also provide a detailed justification of the choices I made when using those NLP algorithms in this study. This detailed memo-ing is helpful for researchers and practitioners to transfer or adapt my NLP approach in their own settings.

3.2. Cases Scenarios

In my dissertation, I applied and evaluated NLP approaches to student written responses to the following two case scenarios: (i) the Big Belly Trash Can Ethics Case Scenario, and (ii) Abeesee Village Systems Thinking Case Scenario. Details about these two case scenarios are provided below.

3.2.1 *Big Belly Trash Can Ethics Case Scenario*

The first-year engineering program (FYE) in the Department of Engineering Education at Virginia Tech teaches students an ethics module. The module comprises a case-based instructional design of two hours in a semester. For the assessment of the ethics module, the majority of the FYE instructors use an ethics case study. While there are several cases that instructors might use, the most popular one is the Big Belly Trash Can ethics case. After reading the case and its related material, the students are required to submit their written responses to question prompts about: (a) recognition of an ethical issue, (b) identification of a stakeholder, (c) possible decisions according to various ethical decision-making theories, and (d) consequences of those decisions on various stakeholders. Students are also given the grading rubric to follow for writing their responses. The case scenario and question prompts are provided in Appendix A.

To collect students' submissions, Dr. Katz as principal investigator (PI), and I as co-PI received a VT Institutional Review Board determination of "Non-human subject research" (IRB #22-092) for a project titled, "Assessing Ethical Decision Making in Engineering Education." Under this project, I collected students' submissions to the Big Belly Trash Can case scenario from

the FYE instructors for Spring 2021 and Spring 2022. I used only responses from students who consented to participate in research associated with the class, resulting in a total of 755 student responses in this study. Of these 755 responses, 550 were from students of Spring 2022 and 205 were from students of Spring 2021.

3.2.2 *Abeesee Village Systems Thinking Case Scenario*

Grohs et al. (2018) developed a case scenario to assess systems thinking competencies. The scenario is framed in a community setting, the fictitious town of Abeesee (pronounced like A.B.C.), facing heating issues in harsh winters. The respondents' reasoning process is captured through their written responses to the question prompts, which are distributed across the following three phases: (1) processing, (2) response, and (3) critique. First, in the processing phase, the question prompts are about the identification of the problem, stakeholders, and respondents' decision-making process and its goals. Second, in the response phase, the question prompts ask respondents to (a) outline a plan addressing the identified problem, (b) anticipate challenges in implementing their proposed plan, and (c) list potential measures of successful outcomes. Lastly, in the critique phase, someone else's solution is given to respondents, and they (i) interpret its goals; (ii) predict its unintended consequences, and (iii) judge the adequacy of resources in the given solution implementation. The case scenario and question prompts are given in Appendix B.

In my dissertation, I used 424 students' responses to the Abeesee case scenario that were collected by Grohs and his team in the project titled (IRB# 20-688), "Solving Complex Problems through Transdisciplinarity." These 424 students were from two settings: (a) 262 students from the Virginia Tech Rising Sophomore Abroad Program between 2017 and 2022, and (b) 162 students from statistics and human development related undergraduate courses at Virginia Tech in Spring 2021.

3.3. Dividing Student Responses into Training and Testing Datasets

In NLP studies, it is common practice to divide the total dataset into two samples: one for training NLP algorithm and the other for its testing. Following this practice, I divided the collected student responses in this study into these two samples. Notably, in this study, the training sample represents the dataset that was used to develop example banks, while the testing sample represents withheld student responses that were labeled through my NLP approach. I used the collection-setting-context as the demarcation for dividing the collected student responses into training and testing samples. For instance, in the Big Belly Trash Can case scenario, I used student responses from Spring 2022 for developing the example bank and from Spring 2021 for assigning codes using my NLP approaches, as shown in Table 3.1. I chose the collection-setting-context as

demarcation line for training and testing samples to demonstrate that the example bank developed using student responses in one classroom setting could be used to analyze student responses from other classroom settings, if both samples belong to the same case scenario.

Table 3.1: Counts of Students' Responses for Each Case Scenario, Training, and Testing Samples

Case Scenario	Total Student Responses	Used in Example Bank	Used for Testing
Big Belly Solar Trash Cans	755	550	205
Collection-Site-Context	Virgnia Tech	First-Year Engineering course in Spring 2022	First-Year Engineering course in Spring 2021
Abeesee Village Systems Thinking	424	262	162
Collection-Site-Context	Virginia	Rising Sophomore Abroad Program	Statistic and Human Development courses

3.4. Selecting Question Prompts from Case Scenarios

Given the response counts listed in Table 3.1 for each case scenario, the original data for both case scenarios were collected as student assignments but not with the explicit purpose of being used in research related to the NLP like this dissertation study. The NLP approaches that I tested work best when the respondent focuses on one idea at a time. As such, some of the question prompts were phrased in a suboptimal manner for the NLP approaches. This is because students might have described multiple ideas in a single short sentence, which could sometimes lead to noisy performance of the NLP approaches. This challenge of eliciting multiple pieces of information at once due to the question phrasing and response format is an inherent limitation of my dissertation study. Therefore, I selected question prompts from both case scenarios where student responses tended to be more structured and focused on one idea at a time (at least in a

parseable manner). These selected prompts provided the best opportunity to demonstrate how the NLP approaches could perform thematic analysis of student responses.

For the ethics case scenario, I used students' responses to the following two question prompts (out of a total of six): (i) identify an ethical dilemma, and (ii) identify a stakeholder and explain the impact on the stakeholder. These two question prompts including their abbreviations used to reference them throughout this study are listed in Table 3.2 and highlighted in Appendix A. This table will be a useful reference for the reader in Chapter 4 when analyzing the results.

Table 3.2: The Selected Question Prompts from Case Scenarios and their Abbreviations

Case Scenario	Question Prompt	Abbreviations
Big Belly Solar Trash Cans	Ethical dilemma is clearly and thoroughly identified and related to the case study.	ethics_q1
	A clear description is included of a stakeholder, including how they are related to the case study. Their relationship to the ethical issue/dilemma is clearly explained.	ethics_q4
Abeesee Village Systems Thinking	Given what you know from the scenario, please write a statement describing your perception of the problems and/or issues facing Abeesee.	sys_q3
	What additional information do you need before you could begin to develop a response in Abeesee? Consider both detail and context of the problems/issues you identified.	sys_q4
	What groups or stakeholders would you involve in planning a response to the problems/issues in Abeesee?	sys_q5
	Please briefly describe the process you would use planning a response to the problems/issues in Abeesee.	sys_q6
	What would you expect a successful plan to accomplish?	sys_q7
	What challenges do you see in implementing your plan? What are the limitations of your approach?	sys_q11

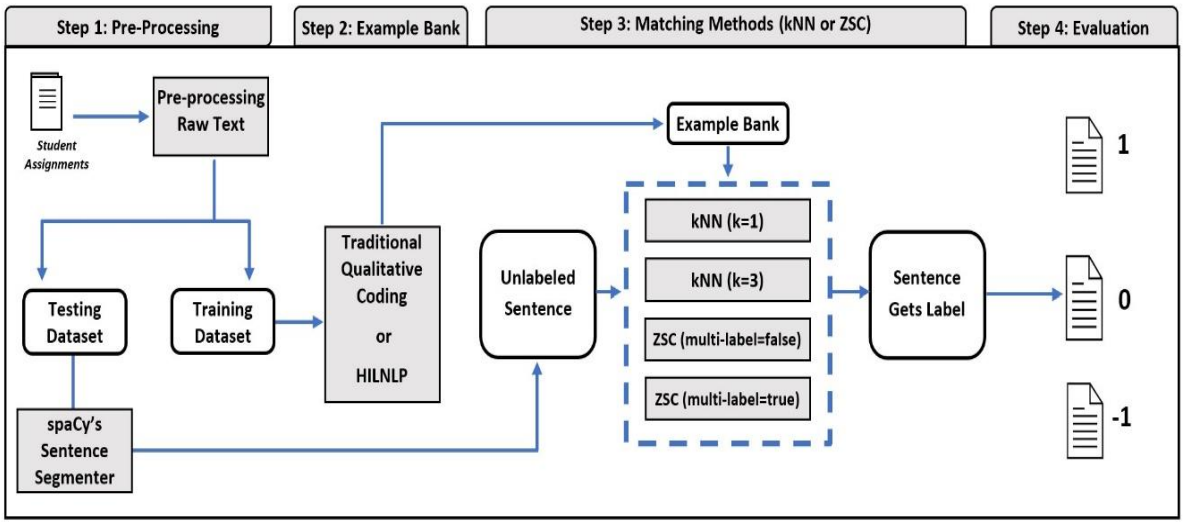
For the systems thinking case scenario, I used student responses to all question prompts except, (i) "Given what you know and a budget of $50,000, develop a plan that would address the Abeesee situation....Use a numbered, step-by-step guide, and recipe style to explain your response

plan" and (ii) critique phase question prompts (Grohs et al., 2018, p. 118). This means that I used only six question prompts (out of a total of 13) from the systems thinking case scenario. These question prompts with the abbreviations used to reference them throughout this study are also listed Table 3.2 and highlighted in Appendix B.

3.5. Data Analysis

To answer my study's sub-RQs, the data analysis comprised four processes as shown in Figure 3.1—previously captioned as Figure 1.1 in Chapter 1:. First, I pre-processed the raw text data before passing it to the NLP workflow. Second, I developed the example banks to identify themes in the students' responses with two procedures: (a) HILNLP and (b) traditional qualitative coding. In the third step, I assigned labels to unlabeled students' responses using the example bank with the following four NLP approaches: (i) k Nearest Neighbors (kNN) top score (k=1), (ii) kNN majority vote (k=3), (iii) Zero-shot Classification (ZSC) (multi-label=false), (iv) ZSC (multi-label=true). Notably, the kNN approaches took input of both sentences and their labels from the example bank. On the other hand, the ZSC approaches took only the input of labels from the example bank. Lastly, I read those (newly) labeled students' responses to evaluate whether the codes assigned to those responses through the kNN and ZSC approaches were accurate or not. Here, accuracy meant that the assigned code represented the idea expressed in student responses.

Figure 3.1: Overview of Data Analysis Steps



** White represents data (e.g., student responses)*
*** Gray represents process*

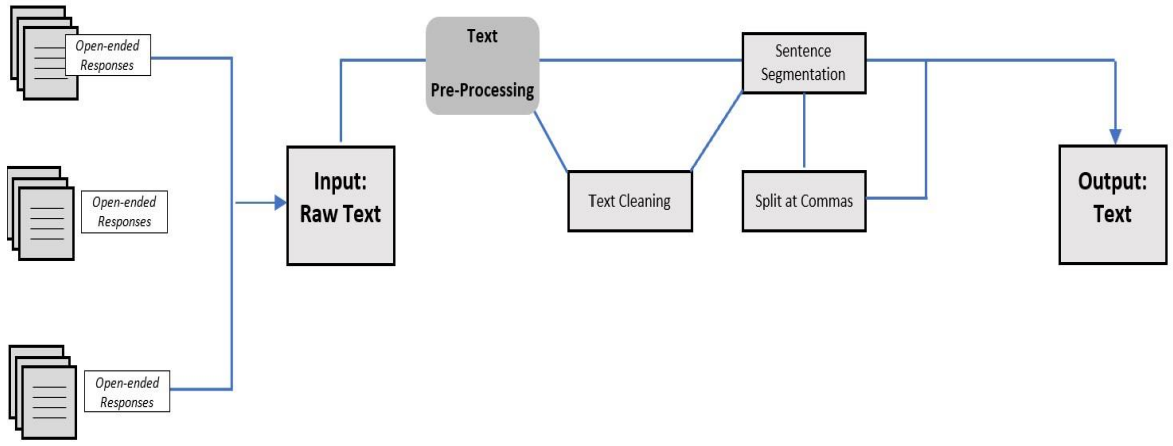
For the technical implementation of the data analysis steps shown in Figure 3.1, I used Google Colab notebooks, written using a combination of the R and Python programming languages. I describe the technical implementation of the four steps of data analysis below.

3.5.1 *Pre-processing of Text*

For both case scenarios, I collected students' assignments as pdf files from instructors. I skimmed the pdf files to see how consistent formatting was among students' submissions. In the case of the ethics scenario, there was a wide variation in the formatting that would have been problematic for downstream NLP tasks. I decided to resolve those formatting issues in students' submissions. I converted the original pdf files into txt files and then manually added unique symbol patterns, both at the start and end of each question's answer, to be able to separate students' responses to the question prompts. With the help of those symbol patterns, I extracted each student's answer to each of the two question prompts. These excerpts were then cleaned to remove Arabic numerical symbols such as "[1], "[2]", etc., or legend symbols. In the case of the systems thinking scenario, students' responses had fairly consistent formatting for each question prompt.

Here, a noteworthy step in data analysis was that students' responses were typically a single block of text which may or may not have consisted of multiple sentences. Before passing these blocks of text to the NLP processes, I had three options to choose from: (a) no split of a single contiguous block of text, (b) split at sentence level via spaCy's sentence segmenter (Honnibal & Montani, 2017), or (c) split at phrase level within a sentence via punctuation of a comma, as shown in Figure 3.2. The purpose of splitting was to optimize the performance of the NLP approaches because they worked best when the input text focused on one idea at a time. A block of text (e.g., a paragraph or a sentence) might have expressed multiple topics at a time, but the vast majority of single phrases or some sentences might have expressed fewer (i.e., only one) topics. From (a)-(c), which option worked for what type of question prompt to thematically analyze students' responses was an open question that I addressed on a question-by-question basis, as provided in Table 3.3.

Figure 3.2: Options for Pre-processing of Raw Text Corpus



As given in Table 3.3, for all question prompts from both case scenarios, except sys_q4 and sys_q5, I segmented the responses at the sentence level using spaCy's sentence segmenter. However, for sys_q4 and sys_q5, I segmented the student responses at phrase level within a sentence using commas as delimiters. I took this approach because these questions asked students to list additional information or stakeholders, which students typically listed as (proper) nouns separated by commas within single sentences. Conversely, for all other question prompts, splitting at the phrase level might disrupt the reasoning students developed across sentence(s).

Table 3.3: Pre-processing Methods used for Question Prompts

Pre-processing Method	Question Prompt
Split at sentence level via spaCy's sentence segmenter	ethics_q1 ethics_4 sys_q3 sys_q6 sys_q7 sys_q11
Split at phrase level within a sentence via punctuation of a comma	sys_q4 sys_q5

3.5.2 *Developing Example Bank*

To label student responses through the NLP approaches, first, I developed an example bank for each of the question prompts that I selected for my dissertation study. The purpose was to develop a saturated space (that covered all possible aspects) of responses with assigned labels to a question prompt. For example, in the systems thinking case scenario, there was one question related to stakeholders: "What groups or stakeholders would you involve in planning a response to the problems/issues in Abeesee?" The example bank of this question prompt (an abridged version is shown in Table 3.4) included sentences related to the following labels: builders, businesses, charities, citizens, consultants, donors, energy companies, engineers, environmental groups, etc. After developing the example bank, unlabeled responses were matched to those example responses in the example bank for assigning labels. I described those matching processes in the section 3.5.3 on labeling the unlabeled student responses.

Table 3.4: Preliminary Codebook for a Question Prompt* of Systems Thinking Scenario

Example Bank Sentences	Example Labels
people who run businesses in that part of the town	Businesses
a professional charity society that helps these people like Mercy medical.	Charities
I would involve the Abeesee People	Citizens
outsiders who can be brought in to consult about solutions	Consultants

*What groups or stakeholders would you involve in planning a response to the problems/issues in Abeesee?

To develop example banks, I utilized the following two approaches, as depicted in Figure 3.3, depending on the question prompt: (i) HILNLP approach, and (ii) traditional qualitative coding. For all question prompts considered in this study except ethics_q1, the HILNLP was used to develop example banks, as provided in Table 3.5. For ethics_q1, I developed the example bank with traditional qualitative coding procedure. I give details of (i) in section 3.5.2.1 and (ii) in section 3.5.2.2.

Figure 3.3: Methods for Developing Example Bank

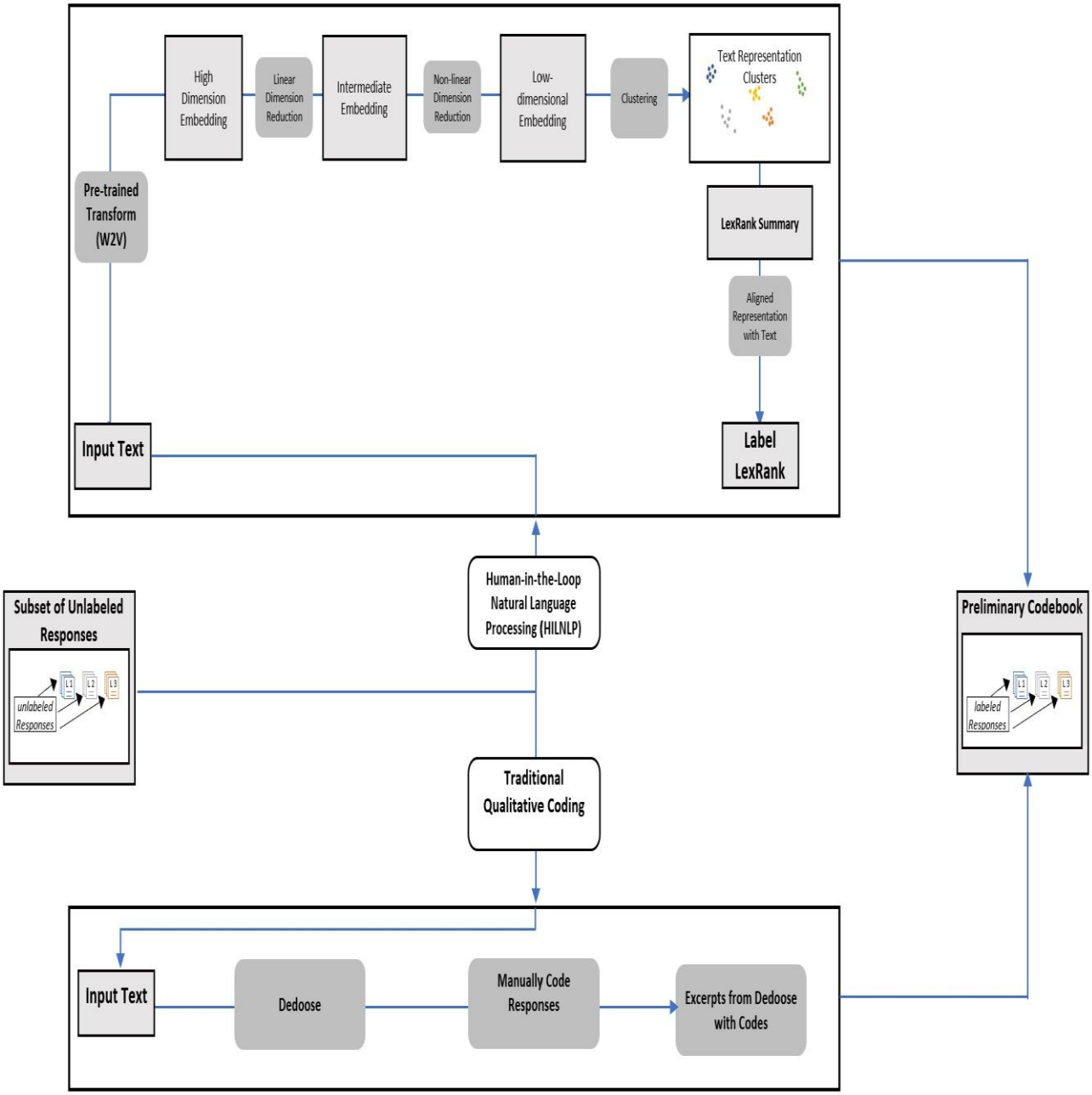


Table 3.5: Counts of (Parsed) Student Responses in Example Bank and With-held Samples

Question Prompt	Method Used for Developing Example Bank	Counts of Labeled Statements in Example Bank	Counts of Unlabeled Statements in With-held Sample
ethics_q1	Traditional Qualitative Coding	743	911
ethics_q4	HILNLP	2185	1036
sys_q3		401	247
sys_q4		1182	440
sys_q5		900	396
sys_q6		913	336
sys_q7		478	188
sys_q11		712	282

3.5.2.1. Example Bank via the HILNLP Approach

From the dataset of each case scenario, I sampled students' responses to develop an example bank through the HILNLP workflow as I have described in Section 3.3. For example, in the Big Belly Trash Can case scenario, I used student responses from Spring 2022 for developing the example bank using the HILNLP approach. Next, I explained the technical implementation of the HILNLP.

Technical Implementation of the HILNLP workflow.

The technical implementation of the HILNLP workflow was completed through the following six sequential steps: (a) sentence embedding, (b) dimension reduction, (c) clustering, (d) LexRank summary, (e) LexRank summary labeling, (f) augmentation.

Sentence Embedding. I embedded the pre-processed text—sentences of students' written responses—into a 768-dimensional vector space using pre-trained Microsoft's Masked and Permuted Pre-training for Language Understanding (MPNet) (Song et al., 2020). The pre-training meant that these models had been trained on large corpora of text (e.g., all of Wikipedia) to

generate embeddings and therefore allowed easy off-the-shelf use. These embeddings were intended to be high-dimensional abstract representations of text in a vector space. A less mathematical way of stating the preceding sentence was that we wanted to try and represent each response with a long array of numbers. What each of those numbers meant by itself was not particularly important. The key was in how each of those arrays of numbers (i.e., numerical representations of the sentences) related to each other. The advantage of this sentence embedding into vector space was that it enabled mathematical operations (e.g., clustering, cosine similarity for semantic similarity) on the subsequent numerical representations of the text.

Dimension Reduction. Theoretically, the clustering could have taken place in this original embedding space. Historically, clustering algorithms suffer in higher (768) or intermediate dimensional spaces (in the range of 50-100) since every point (i.e., text embedding) is far from every other point—the phenomenon defined as the curse of dimensionality (Assent, 2012; Bellman, 2017). Therefore, this reduction process aimed to project the original text embeddings into a lower dimensional space where clustering could meaningfully occur. First, I used Principal Components Analysis (PCA) to reduce the original embedding space (d=768-dimensional space) into an intermediate embedding space (in the range of 65-80 dimensions) (Jolliffe, 2002; Sankaran, 2022). Second, I used the Uniform Manifold Approximation and Projection (UMAP) to reduce the intermediate embedding space (d=768-dimensional space) to a lower dimensional space (d=5) because UMAP worked well only up to fewer than 100 dimensions (Allaoui et al., 2020; McInnes et al., 2020). The combination of PCA and UMAP rendered minimum loss of information during dimension reduction. After those dimensionality reduction steps, the data was ready for clustering.

Clustering. I used a Hierarchical Agglomerative clustering algorithm with Ward linkage to discover semantic groupings in the students' responses(Murtagh & Legendre, 2014). The resultant clusters theoretically had semantically homogeneous texts within a cluster without necessarily relying on syntactic similarity. For example, in our paper (Shakir et al., 2022), we explored the perspectives of undergraduate students about the roles of social class, gender, and race in shaping their educational experiences. We used the HILNLP workflow to analyze students' interview transcripts. With the workflow I used for my dissertation and for the aforementioned paper, Student A's response, "I think income level is kind of related to what you are getting now in terms of financial resources", and Student B's response "Because like, money indeed buys you opportunities." were clustered together. This was because the combination of steps including word embedding to dimension reduction to clustering could identify those statements as discussing the same theme, even though the two responses shared no words in common. In short, the HILNLP workflow that I used for my prior paper and this dissertation study received raw texts and produced

suggested groupings of those texts to which a human user could ascribe specific meanings. I then described the process of identifying the themes in each cluster.

LexRank Summary and its Coding. With the clusters now formed, I needed to summarize the themes discussed. To do this, one had several options belonging to either abstractive or extractive summarization. I used LexRank, an extractive text summarization process that produced document summaries without paraphrasing information. The LexRank algorithm does this by identifying and sub-setting the most salient sentences in the original document(s) (Erkan & Radev, 2004). Using the LexRank approach, I identified representative student responses from the clustering results. The ultimate output from this process was a subset of students' responses to which I assigned labels and made an initial example bank. I did this by reading each student's response to assign it a label. My advisor performed member checking for the quality of that coding process(Creswell & Poth, 2016; Lincoln & Guba, 1985). I counted the number of response sentences by each label in the example bank. It is noteworthy here that I decided to have at least five instances of response sentences in the example bank for each label. When there were fewer than five instances for a label, I augmented the example bank with my written sentences that should have expressed the same idea in other words. I added special identification digits to those sentences to separate them from student responses. Next, I describe how I also developed an example bank with the traditional qualitative coding method.

3.5.2.2. Example Bank via Traditional Qualitative Coding

To consider whether the codebook generation process impacted the labeling outcomes, I also used a traditional qualitative coding approach to label 550 student responses for the ethics_q1 question prompt. I suggest this number sufficed to develop a saturated example bank that covered all aspects of responses to the question prompt. I uploaded the data in Dedoose for traditional thematic analysis (Clarke & Braun, 2017). I read students' responses, defined and assigned codes, and selected excerpts. I iterated on these processes to refine and merge codes. When I observed that new codes were not emerging from the data, I downloaded all codes and their excerpts from Dedoose as an initial example bank. It is noteworthy to contrast example banks developed with both the HILNLP and traditional qualitative coding method, to which I turned next.

3.5.2.3. Comparing the Example Banks developed with HILNLP and Traditional Qualitative Coding

In traditional qualitative coding, a qualitative code (or thematic) unit can be a phrase, a sentence, or multiple sentences, depending on researcher judgment. In contrast, in the HILNLP approach the text prior to coding was preprocessed and segmented either at the sentence or comma

level. Therefore, assigning a single code to multiple sentences simultaneously was not an option. In this dissertation study, the primary difference between example banks developed using the traditional qualitative coding method and the HILNLP method is as follows: in the former, labeled example responses might contain one or more sentences, whereas in the latter, labeled example responses include a phrase or a single sentence. The final counts of example responses included in example banks are listed in Table 3.5. Using these example responses, next I describe the NLP approaches used to code unlabeled (or new) student responses.

3.5.3 *Labeling the Unlabeled Student Responses*

First, the withheld student responses were pre-processed with the method described in Section 3.5.1—similar to student responses used in example banks. The counts of these pre-processed unlabeled response sentences are given in Table 3.5. Next, these unlabeled (or new) responses were matched with example responses using the following two NLP methods: (a) kNN (Hechenbichler & Schliep, 2004) and (b) ZSC (Pushp & Srivastava, 2017; Yin et al., 2019). When one compared (a) and (b), it is worth noting that (a) needed an example bank with both sentences and labels. On the other hand, (b) needed just the labels themselves—and this made it more resource efficient than (a). This was because, in the case of the ZSC approach, one could start the thematic analysis with just a preliminary list of candidate labels and not need example sentences. Moreover, in the kNN approach, there was no guarantee that all responses would get a label. At least with the ZSC approach, everything would get a label—though there was an expression of uncertainty about the label's accuracy. This can be a tradeoff between the two approaches because it may not always be desirable to assign a label, such as in instances where there is ambiguity about the topics in an unlabeled response. I now describe the technical implementation of the kNN and ZSC approaches.

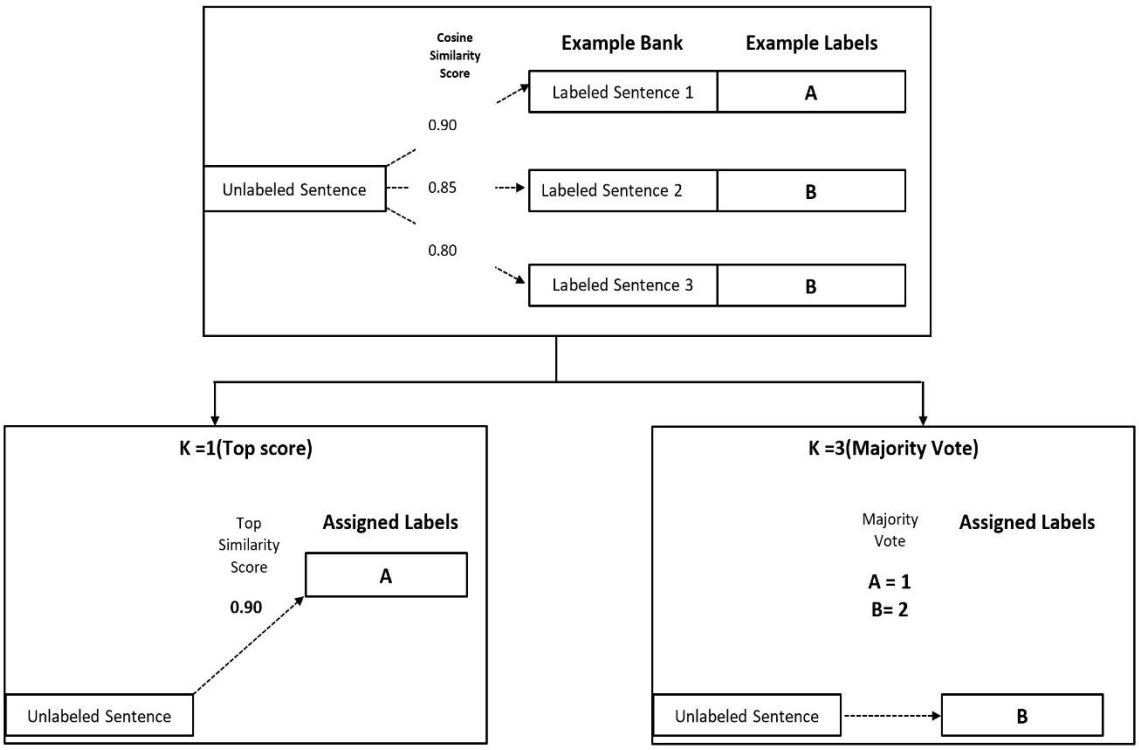
3.5.3.1. k Nearest Neighbors (kNN)

For the kNN methods for matching, the technical implementation included the following three steps: (a) sentence embedding, (b) calculating cosine similarity, and (c) assigning labels by either identifying the top score (k=1) or a majority vote (k=3). Notably, these methods applied one label to each student statement.

Sentence Embedding. Similar to the HILNLP workflow, I embedded text—sentences or phrases from the unlabeled dataset and the example bank—into a 768-dimensional vector space using the MPNet embedding model (Song et al., 2020). After embedding both data sets, I determined the semantic similarity between unlabeled sentences and labeled sentences using the cosine distance measure.

Cosine Similarity Score. I used the cosine distance measure between embedding vectors of labeled and unlabeled sentences. I chose this for two reasons: First, it is the most used distance measure among others (e.g., Euclidian distance, and Manhattan distance) in word embeddings to estimate the similarity of student and reference answers in ASAG (Putnikovic & Jovanovic, 2023). Second, tshe cosine distance is considered favorable as this does not include the length of text (magnitude of text vectors), which is regarded as irrelevant for measuring semantic similarity between two text statements (Kerkhof, 2020; Qiao & Hu, 2023). Theoretically, the similarity score ranged from 0 to 1. The maximum value, a similarity score of 1, represented the exact match between unlabeled and labeled sentences. As the similarity score between two sentences decreased from 1 to 0, we inferred that those two sentences did not match each other—in other words, they were less likely to be about the same topic. Each unlabeled sentence had a similarity score from its comparison with each sentence from the example bank as shown in Figure 3.4. The question was which label of an example bank sentence should be assigned to an unlabeled sentence. To achieve this purpose, I used the following two kNN processes: (a) top score (k = 1) and (b) majority vote (k =3).

Figure 3.4: Using kNN Process for Labeling Unlabeled Responses



Assigning labels Top Score (k =1). The unlabeled sentence was assigned the code of an example bank sentence for which it had the maximum similarity score. Likewise, in Figure 3.4, the unlabeled sentence was assigned label A because the unlabeled sentence had a maximum similarity score of 0.90 with labeled sentence 1. Theoretically, all unlabeled sentences in top score (k =1) method should match a response in the example bank, though the top similarity score can range from 1 to 0. Notably, I selected 0.70 as the threshold value of the top match similarity score to assign a code to the unlabeled sentences. Any unlabeled sentences with similarity scores below 0.7 to example responses were not incorporated for subsequent analysis in this dissertation study. The rationale for this methodological decision was that the likelihood of an inaccurate label below this similarity score increases significantly based on my experience and judgment. After applying a threshold value of 0.70 similarity score in the kNN topscore method, the final number of (newly) labeled response sentences through the NLP method was lower than the number of input sentences for all question prompts investigated in this study. These counts are listed in Table 3.6.

Assigning labels Majority Vote (k=3). When k =3, first, the NLP method selected the three example bank sentences with the highest similarity scores with an unlabeled sentence. Among labels of those three example sentences, any label with a majority vote (2 or more) was assigned to the unlabeled sentence. For example, in Figure 3.4, the unlabeled sentence was assigned label B because it had the majority of 2. If any label did not have the majority vote (2 or more), the unlabeled sentence remained unlabeled. Let's assume sentence 3 of the example bank in Figure 3.4 had the label C rather than B. In this case, the unlabeled sentence was not assigned any label among the three (A, B, C) because none of the labels had a majority. Similar to the top score kNN method, the number of (newly) labeled response sentences through kNN majority vote (k=3) was lower than the number of input sentences. These counts are listed in Table 3.6.

Table 3.6: Counts of Input and Output Sentences for NLP Approaches

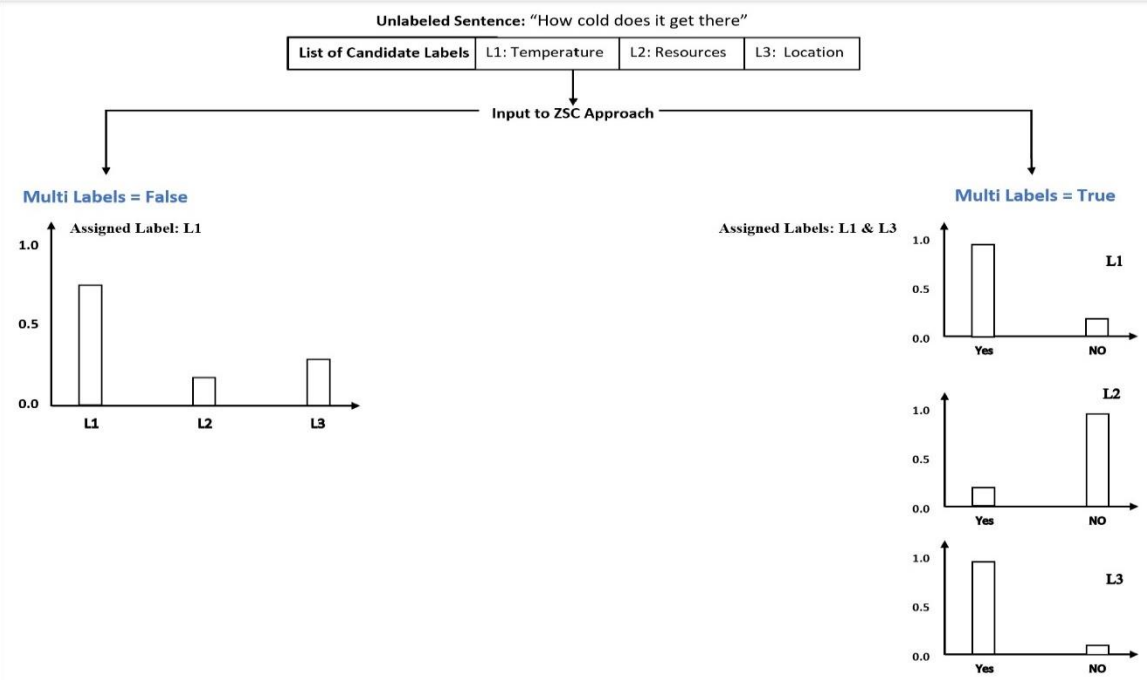
Question prompt	Total Input (Parsed) Sentences	Output-Assigned Labels			
		Top Score (k=1)	Majority Vote (k=3)	ZSC (multi= False)	ZSC (multi= True)
ethics_q1	911	602	476	911	990
ethics_q4	1036	659	899	1036	1053
sys_q3	247	199	219	247	384
sys_q4	440	259	381	440	716
sys_q5	396	267	358	396	553
sys_q6	336	213	198	336	1254
sys_q7	188	152	152	188	302
sys_q11	282	116	159	282	294

3.5.3.2. Zero-shot classification (ZSC)

Without task-specific training, the ZSC approach aims to associate an appropriate label to a text snippet using ready-made, pre-trained natural language inference (NLI) models (Yin et al., 2019). In my dissertation study, this approach took the following inputs: (a) the students' sentences that I was interested in labeling and (b) the list of candidate labels from the example banks. As shown in Figure 3.5, consider the following example unlabeled text , "how cold does it get there?", and candidate labels "temperature" (L1), "resources" (L2), and "location" (L3). The ZSC model embedded the unlabeled text and candidate labels into the same vector space. Then, first, the ZSC approach posed the unlabeled sentence (how cold does it get there?) as the premise. Second, the approach turned each candidate label (i.e., L1, L2, L3) into a hypothesis to determine whether the hypothesis was true or false given the premise. In the case of the example given in Figure 3.5, the ZSC approach constructed the following three hypotheses:

Is the text "how cold does it get there?" about temperature?
Is the text "how cold does it get there?" about resources?
Is the text "how cold does it get there?" about location?

Figure 3.5: Using ZSC Process for Labeling Unlabeled Responses



Using the NLI models, the ZSC approach fitted discrete probability distributions to determine probabilities for labels (or the posed hypotheses). The NLI models had the following two options to determine probabilities: (i) multi-labels=false and (ii) multi-labels=true. Notably, (i) applied one label to each student statement, while (ii) applied more than one label to each statement.

Assigning labels through ZSC (multi-labels=false). In this case, the NLI models fitted a single discrete probability distribution among all three candidates, as depicted on the left side of Figure 3.5. This meant that the sum of probabilities for L1, L2, and L3 was equal to one. I chose the single label with the highest probability (i.e., L1) to assign to the unlabeled sentence. Therefore, all input response sentences for each question prompts investigated in this study received one label through ZSC (multi-labels = false), as I listed these counts in Table 3.6.

Assigning labels through ZSC (multi-labels= true). In this case, the NLI models fitted separately the Yes-No probability distribution to each of the candidate labels, as shown on the right side of Figure 3.5. I chose a threshold value of 0.90 for label probabilities to assign them to unlabeled sentences. This was because, according to my experience with the ZSC approach (multi labels = true), labels with probabilities above the threshold value of 0.9 were more relevant to unlabeled sentences. Likewise, for the example given in Figure 3.5, L1 and L3 were assigned to

the unlabeled sentence. To further illustrate the phenomenon of assigning more than one label to a single response sentence through ZSC (multi-labels=true), I gave two example response sentences for sys_q3 with their assigned labels in

Table **3.7**. Notably, ZSC (multi-labels = true) is the only NLP method among all four considered in this dissertation study where the number of output labeled sentences exceeds the initial input numbers as listed in Table 3.6.This is because ZSC (multi-labels = true) is only method among four NLP methods investigated in this study that can assign more than one label to a sentence.

Table 3.7: Examples for a Question Prompt* of Systems Thinking Case Scenario with ZSC (multi-label=true)

(Parsed) Student Responses	Assigned Label
They need a safe inexpensive way to heat their homes	Affordability
They need a safe inexpensive way to heat their homes	Heat Availability
The problem Abeesee people face is the isolation of their living spaces and the lack of wealth in the area.	Poverty
The problem Abeesee people face is the isolation of their living spaces and the lack of wealth in the area.	Remote Location

**Given what you know from the scenario, please write a statement describing your perception of the problems and/or issues facing Abeesee.*

3.6. Summary of Data Analysis

In sum, for data analysis in my dissertation study I first pre-processed students' responses on a question-by-question basis. Second, I developed preliminary codebooks for question prompts that I selected for my dissertation study of ethics and systems thinking case scenarios. To develop example banks, I used traditional qualitative coding for the following one question prompt of the ethics case scenario: (i) identifying an ethical dilemma (ethics_q1). For all other question prompts, I used the HILNLP approach to develop example banks. Third, I used the kNN and the ZSC approaches to thematically analyze student responses. For the kNN approach, I used these two processes: (i) k=1 and (ii) k =3. For the ZSC approach, I used: (iii) multi-labels=false and (iv) multi-labels=true. I deployed (i)-(iv) for a total of eight question prompts (two for the ethics case scenario and six for the systems thinking scenario)—that I selected for my dissertation study and given in Table 3.2. Next, I describe the evaluation procedure to compare the accuracy of (i)-(iv)

for thematic analysis of student responses to each of the eight question prompts for the engineering case scenarios.

3.7. Accuracy Evaluation Procedure

After assigning labels to (parsed) student responses, I read each sentence or phrase to evaluate whether the assigned code represented the idea described in the sentence. If yes, then I assigned it a rating of an accurate label as 1. On the other hand, if not, then I assigned it a rating of an inaccurate label as -1. In between those extreme ratings, I had a third category of neutral as 0. I used this category in instances of ambiguity or partial credit; for example, a sentence could have been about more than one idea, or the sentence itself might have been ambiguous. To demonstrate this evaluation procedure, Table 3.8. provides three example response sentences for ethics_q1 coded using the kNN majority vote (k=3) method along with their assigned accuracy ratings. For instance, in Table 3.8 the third example sentence discussed Wi-Fi and privacy issues, but it was assigned the "access to income" label. Therefore, I rated this sentence as inaccurately labeled (-1) in my evaluation. Lastly, I used those numerical evaluation ratings to calculate the total number (and proportions) of sentences that were labeled (a) accurately, (b) inaccurately, and (c) neutral by the NLP approaches.

Table 3.8: Evaluation Rating Example for a Question Prompt* of Ethics Case Scenario

Response Sentences	Assigned Label	Evaluation Rating		
I believe that the biggest ethical dilemma that is presented into this case study is the rejection of a source of income for the homeless that surround the urban areas around the bay.	Access to Income	Accurately Labeled	True Positive	1
An ethical dilemma would definitely be seen the most within the homeless population.	Access to Income	Partial Credit or Ambiguous Sentence	Neutral	0
The wifi hotspot feature in these trash cans is the ethical dilemma from this case study that I am going to study.	Access to Income	Inaccurately Labeled	False Positive	-1

** Identify Ethical dilemma related to the case study.*

The aforementioned quantitative evaluation procedure allowed me to answer my research questions in the following way. First, to answer sub-RQ1(How well do different NLP processes

(e.g., k nearest neighbors, zero-shot) label responses?), I selected a question prompt and compared its evaluation ratings across NLP approaches (i) kNN(k=1), (ii) kNN (k=3), (iii) ZSC (multi-labels = false), and (iv) ZSC (multi labels = true). Second, to answer sub-RQ2 (Does the answer to sub-RQ1 vary by assessment questions in a case scenario?), I selected a case scenario and compared evaluation ratings across its question prompts for each of the NLP approaches from (i) to (iv). Third, to answer sub-RQ3 (Does the answer to sub-RQ1 vary by case scenarios? (e.g., Abeesee system thinking scenario vs ethics case scenario), I looked across all questions from both of the case scenarios. Then, I compared the evaluation ratings for each of the NLP approaches from (i) to (iv). Lastly, I summarized those evaluation ratings. I used this summary to develop the best practices to perform the thematic analysis of students' responses through the investigated NLP approaches. Those best practices answered my dissertation's RQ: How can we apply the NLP approaches based on transformer-based language models to thematically analyze students' responses to open-ended question prompts of case scenarios? Given details of my study's research design, I now mention limitations specific to the study's design that could qualify my results in Chapter 4.

3.8. Study Method-specific Limitations

This section includes the discussion about the study's design specific limitations. These are related to the following: (a) kNN method, (b) graphical or mathematical representations, (c) sentence or phrase level analysis, (d) subjective choices for the HILNLP and traditional qualitative coding procedures, and (e) comparison metrics and standardized data set.

3.8.1 *kNN Methodology*

The kNN approach does not label all sentences. This discrepancy can be attributed to the unbalanced example bank, which might show bias in terms of the frequency of example sentences for a qualitative label. For example, the label of "privacy concerns" in the first question prompt of ethics case scenario (ethics_q1), most of the example sentences related the label of "access to income." The absence of example sentences may cause testing responses to remain unlabeled during the matching process using the kNN method. The kNN methods are ill-suited for rare labels because they require several example sentences to establish for matching tasks. To control this limitation, I developed at least five example responses for a label in the example bank. Another limitation with the kNN method is the need to pre-define a value for nearest neighbors (k) that can be influenced by data characteristics. If we are confident that unlabeled responses have a similar matching sentence in the example bank, then the k=1 could work well. Conversely, if there is uncertainty in the example responses and each label has a smaller number of instances, then one

should opt for a higher value of k. However, determining the optimal value of k is dependent on researcher judgment and a process of trial-and-error.

3.8.2 *Graphical or Mathematical Representations*

Student responses to open-ended question prompts may include mathematical or graphical representations, particularly in engineering courses. An instructor manually parsing through students' written responses can easily interpret those mathematical or graphical representations. However, my NLP approach could not include those for grading.

3.8.3 *Sentence- or Phrase-level Analysis*

In the current implementation, I selected individual sentences because the data were structured in such a way that responses to multiple sub-questions were contained within the same paragraph and would thus confound the labeling. By splitting the sentences, I might lose context and reasoning when a student develops an argument in more than one sentence.

3.8.4 *Subjective Choices Between HILNLP and Traditional Qualitative Coding*

I subjectively chose to perform traditional qualitative coding of student responses to ethics_q1 and HILNLP for all question prompts investigated in this study. The choice of the HILNLP and traditional coding also led to length differences in thematic units in the HILNLP and traditional qualitative analysis. In traditional qualitative analysis, a thematic unit could range from a word, phrase, sentence, or even an entire paragraph that I perceived as completely capturing a qualitative label. Unfortunately, this variability in the scope of a thematic unit is not possible in the example bank generated through the HILNLP due to the preprocessing of text.

3.8.5 *Generative Artificial Intelligence and the Study's Method*

ChatGPT, a GAI tool, a has made several things possible related to the study's method. First, I used MPNet for vectorization of student responses, which has a bandwidth of 300-500 characters. To help with the bandwidth limitations, I split student responses using Spacy's segmenter. However, GPT-4, the model behind ChatGPT, can vectorize much larger pieces of text without losing as much information. Therefore, the pre-processing steps I took may not be necessary if an instructor uses ChatGPT.

Second, as a conversational GAI tool, an instructor would not need to use programming code for NLP tasks. So, the codebase I developed may not be necessary now. For example, (i) an instructor could directly input student responses into the ChatGPT interface and ask what themes are mentioned in the responses, or (ii) provide both the codebook (themes) and student responses as inputs, then ask ChatGPT about patterns of themes in the responses. However, despite the

benefits of (i) and (ii), there are still issues of deployment, student privacy, and cost when using ChatGPT that would need to be considered when using ChatGPT in academic settings.

3.8.6 *Lack of Standardized Metrics and Standardized Data Set*

First, my dissertation study lies at the intersection of computer science and education fields, where research studies report metrics for performance of the investigated NLP methods. The common accuracy metrics are recall, precision, and F1 metrics. I provided their definitions in section 1.9. These metrics are calculated by binary classification of results as true positives and true negatives on standardized datasets. Following this tradition of reporting metrics, my study has a limit, as I did not report these metrics. I justified this choice by using a three-way classification system (true positive, neutral, and true negative) rather than a binary in my accuracy check procedure. I included a third category, the neutral rating, to account for two situations. First, the pre-processing of text can generate a few meaningless phrases (like isolated transition adverbs). Second, students may pack more than one idea into short sentences, necessitating multiple qualitative codes rather than a single one. I argued that these two situations do not deserve a true negative tag in a binary classification context and do not justify penalizing the NLP approaches in accuracy calculations. This new category makes it technically unfeasible to calculate and report the standardized metrics of precision, recall, and F1 for my NLP approaches. This is a non-trivial limitation for the contemporary comparison of my NLP approaches, which others reported in the literature.

Second, ASAG has established itself as a distinct body of literature within the field of education. Research studies related to ASAG employ their NLP tools on standardized public datasets (e.g., SciESt, Beetle), whereas my NLP approach is applied to a restricted dataset. However, my datasets are often considered to be the intellectual property of the instructors and are not publicly available. The unique composition of the testing dataset prevents the comparison of the performance of my NLP approach with other NLP approaches reported in ASAG studies.

Despite the aforementioned study-method specific limitations, I suggest the novelty of my NLP approaches contributes significantly to the existing body of knowledge by (a) introducing a new accuracy measure, and (b) providing a novel testing.

Chapter 4: Results

4.1. Chapter Overview

Chapter 4 presents the results of NLP approaches that are described in Chapter 3 and used to achieve the study's purpose: to apply NLP approaches that use TLLMs to thematically analyze student responses to open-ended question prompts of case scenarios. In doing so, I explored the following sub-RQs:

Sub-RQ1: How well do different NLP processes (i.e., k-nearest neighbors, zero-shot classification) label responses?

Sub-RQ2: Does the answer to sub-RQ1 vary by assessment questions in a case scenario?

Sub-RQ3: Does the answer to sub-RQ1 vary by case scenarios (i.e., a systems thinking scenario vs an ethics case scenario)?

To answer the above sub-RQs, I present results of the manual accuracy evaluation of assigned codes by the kNN and ZSC approaches to student responses to eight open-ended question prompts from two case scenarios - two are from the ethics case scenario, and six are from the systems thinking case scenario. For the kNN approach, I used these two processes: (i) top score (k=1) and (ii) majority vote (k=3). For the ZSC approach, I used the following two configurations: (iii) multi-labels=false and (iv) multi-labels=true. In total, four NLP approaches were used across eight question prompts, making 32 cases for manual accuracy evaluation.

To demonstrate how I present results in this chapter, I consider here one case of kNN (k =1) for ethics_q1 among those 32 cases. For reporting results, I provide counts and proportions of three evaluation ratings: true positive (accurately labeled), false positive (inaccurately labeled) and neutral (partial credit or ambiguous response). For each evaluation rating, I first tallied counts of labels assigned in that rating. Second, I normalized these counts and calculated their proportions by dividing them with the total number of output labels assigned by the kNN (when k =1) for ethics_q1. Following this, I reported these proportions as the results of my study.

I adopted the abovementioned results reporting methodology because output labels assigned vary across NLP approaches. To some extent, this normalization procedure could aid in comparing accuracy evaluation ratings across NLP approaches. However, this normalization procedure may be just oversimplification in a few cases because in those few cases there is a large difference (~500 sentences) in the number of output labels assign between kNN and ZSC approaches.

After calculating counts and proportions of the evaluation ratings for the four NLP approaches, I use tables and bar charts to juxtapose those on a question-by-question basis. The reader should note two important ideas here.

(i) The y-axis in bar charts shows counts of sentences and I changed limits of the y-axis by question prompt for readability.

(ii) Pre-processing of raw data (student responses) often resulted in meaningless, standalone phrases or fragments, for example, introductory transition phrases such as “therefore”, “hence”, “but”, etc. I termed those as “noise.” I separately listed counts of “noise” in tables and these were not included in calculations of proportions of the accuracy evaluation ratings. I took this approach to avoid that true positive ratings of NLP approaches were not unduly penalized due to assigning the wrong label to “noise”.

Along with presenting tables and bar charts for evaluation ratings for each question prompt, I also provide a descriptive synthesis of these evaluation ratings’ counts and proportions across four NLP approaches. In the next section, I explain the difference (or absence of difference) between the total number of input sentences for the NLP approaches and the output assigned labels that were manually evaluated for accuracy check.

4.2. Differences Between Input-Unlabeled Sentences and Output-Assigned Labels

To remind readers, I divided the collected student responses into two samples: one used for developing example bank and the other withheld for testing the labeling with the NLP approaches using the example bank. The example bank contained response sentences with assigned codes. The withheld student responses were split at phrase or sentence level. These unlabeled phrases or sentences for labeling purposes were then input into four NLP approaches: kNN topscore (k=1), kNN majority vote (k=3), ZSC (multi-label=false), and ZSC (multi-label=true). Each of the input sentences were assigned labels, but I used different threshold values of semantic similarity measure for various NLP methods to shortlist labeled sentences from those input sentences for subsequent manual accuracy evaluation.

For the kNN topscore (k=1), I used 0.70 as the threshold value for cosine distance. The rationale for this methodological decision was that the likelihood of an inaccurate label below this similarity score increases significantly based on my past experience with using the kNN methodology and k = 1 (i.e., a nearest match). For the kNN majority vote (k=3), only those labels were assigned to input unlabeled sentences that had a majority vote of two; otherwise, sentences

remained unlabeled. Therefore, in this study, the number of output labeled sentences was less than the number of input unlabeled sentences for both configurations of the kNN methods.

Next, for the ZSC (multi-label=false), a single label with the highest probability from the ZSC model was assigned to input unlabeled sentences. This is the only method in this study for which the number of output labeled sentences were equal to the number of input unlabeled sentences. Lastly, the ZSC (multi-label=true), I chose a probability value of 0.90, so labels with a probability value 0.90 and above from the ZSC model were assigned to input sentences. This approach was taken because, according to my experience with the ZSC model when multi labels = true in multiple contexts and datasets, labels with probabilities above the threshold value of 0.9 were more relevant to the input unlabeled sentences compared to suggested labels that fell below the 0.9 probability threshold. Finally, ZSC (multi-label=true) was the only method among the four NLP methods investigated in this study where a sentence could receive more than one label.

The combination of these observations means that the number of input unlabeled sentences and output labeled sentences used for the subsequent accuracy check for each of the NLP methods varied on a question-by-question basis. I reported those details by each question prompt in section 4.3 to section 4.4. Next, I describe (a) input sentences that were not labeled in the kNN methods and (b) input sentences that were assigned more than one label in ZSC (multi-label=true).

4.2.1 *kNN Methods*

Related to (a), the sys_q11 had the maximum proportion of input sentences that were not assigned labels by kNN approaches in this study. The kNN topscore (k=1) method did not assign labels to 166 out of 282 (59%) input sentences, whereas the kNN majority vote (k=3) method did not assign labels to 123 out of 282 (44%) input sentences. The reason approximately half of the input sentences in sys_q11 were not assigned labels in the kNN was that I did not have labeled sentences in the example bank to which those unlabeled input sentences could be matched (i.e., have semantic similarity). This is because the example bank for sys_q11 is deficient and I discuss this further in Chapter 5. The proportion of input sentences that were not assigned labels across various question prompts was almost always higher for the kNN topscore (k=1) method compared to the kNN majority vote (k=3) method, as depicted in Table 4.1and Figure 4.1.

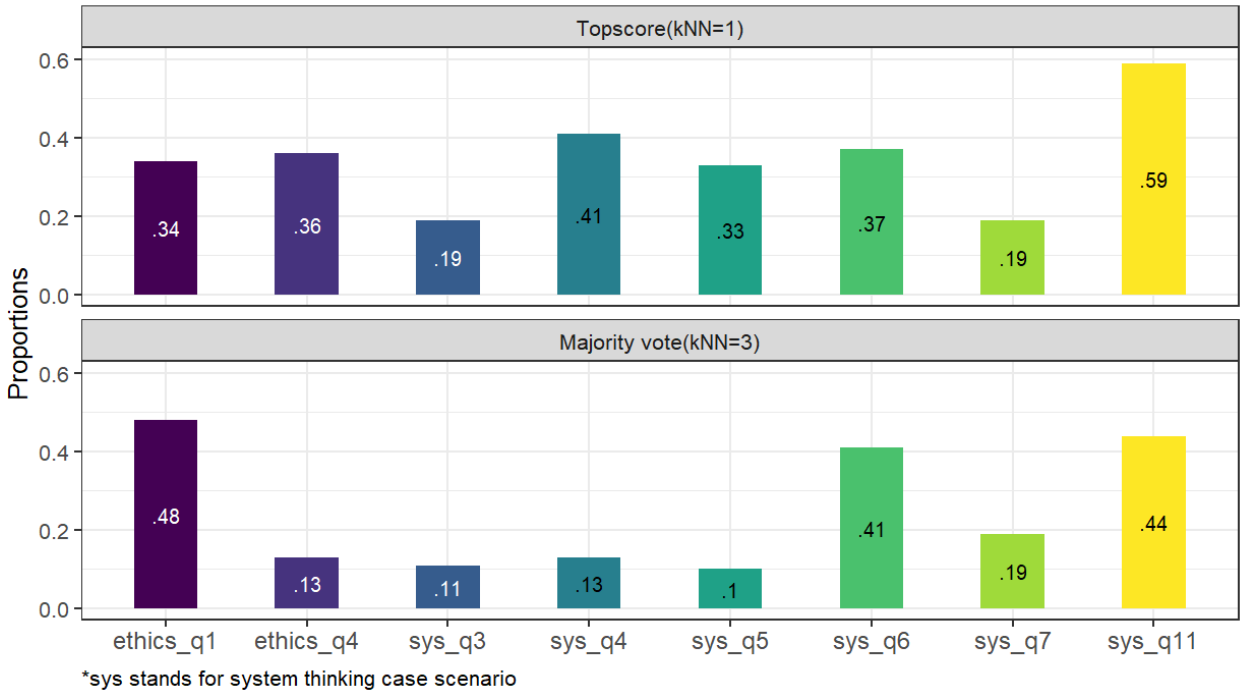
For example, with the kNN topscore (k=1) method, the proportions of input sentences that were not assigned labels were 36% (377 out of 1036) for ethics_q4 and 41% (181 out of 440) for sys_q4. In contrast, with the kNN majority vote (k=3) method, these proportions decreased to 13% (377 out of 1036) for ethics_q4 and (59 out of 440) for sys_q4. This discrepancy arises because the kNN topscore (k=1) method labels an input sentence based on a single similar example

sentence. If there are no highly similar example sentences in the example bank, the input sentence would not exceed the similarity threshold (i.e., 0.70) and would remain unlabeled. Conversely, the kNN majority vote method (k=3) labels an input sentence based on three nearest neighbors instead of one, providing more chances for the input sentence to match with similar example sentences.

Table 4.1: Counts of Input Sentences and Labels Assigned (or not-Assigned) by kNN Approaches

Case Scenario	Question	Counts of (Parsed) Sentences				
		Input	Assigned Labels		Not-Assigned Labels	
			(kNN =1)	*(kNN =3)*	*(kNN =1)*	*(kNN =3)*
ethics	q1	911	602	478	309 (34%)	433(48%)
ethics	q4	1036	659	899	377(36%)	137(13%)
sys	q3	247	199	219	48(19%)	28(11%)
sys	q4	440	259	381	181(41%)	59(13%)
sys	q5	396	267	358	129(33%)	38(10%)
sys	q6	336	213	198	123(37%)	138(41%)
sys	q7	188	151	152	36(19%)	36(19%)
sys	q11	282	116	159	166(59%)	123(44%)

Figure 4.1: Proportion for Input Sentences that were not included in Accuracy Evaluation by kNN Approaches



4.2.2 *ZSC Method (multi-label=true)*

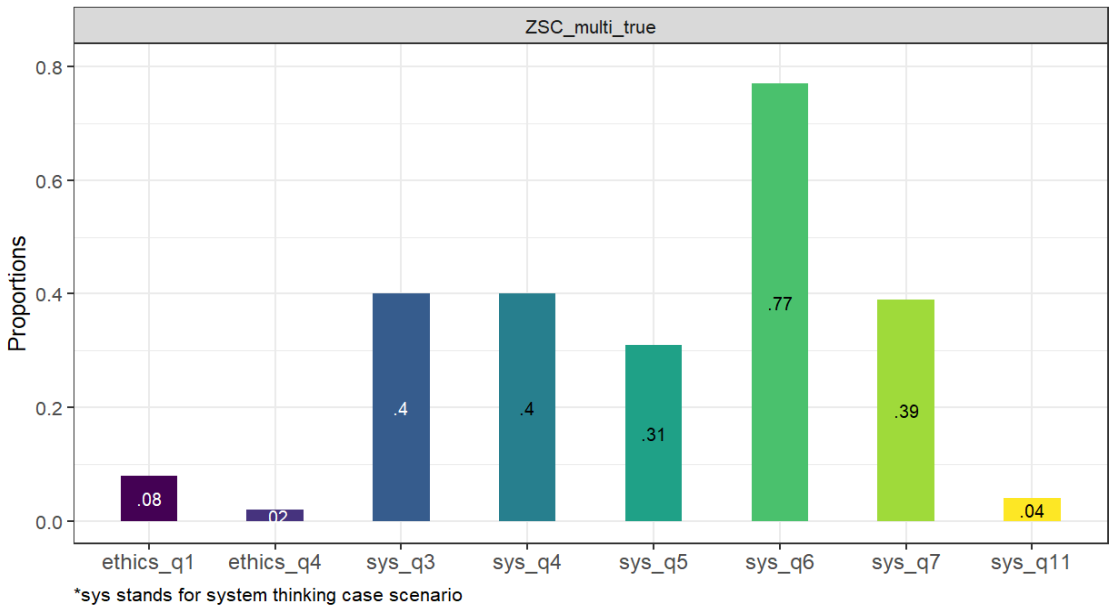
Related to ZSC (multi-label=true), the sys_q6 among all question prompts had the highest proportion of input sentences that had one or more label when using the ZSC (multi-label=true) method, i.e., 71% (259 out of 336), as shown in Table 4.2 and Figure 4.2. This was because multiple interrelated ideas were present in the unlabeled sentence and that labels input to the ZSC model were also closely related. I gave one example: "meet with the residents to hear their concerns and then discuss this among the government of Abeesee to decide a course of action." This example sentence received the following labels: "interact with government", and "interact with locals". Conversely, for ethics_q4 and sys_q11, the proportion of input sentences receiving more than one label was less than 5%.

For example, in ethics_q4, only 16 out of 1036 (2%) input sentences received more than one label using the ZSC (multi-label=true) method. This was because very different stakeholders were described in the sentences, and ten different labels were input into the ZSC model. Some of the input labels were homeless population, students, waste management workers, and wildlife. Besides sys_q6, ethics_q4, and sys_q11, the remaining question prompts had proportions of input sentences that received more than one label ranged between 30% and 40%, as shown in Table 4.2 and Figure 4.2. Next, I provide the results for my Sub-RQ1.

Table 4.2: Counts and Proportion for Input Senetences and Labels Assigned by ZSC Appraoches

Case Scenario	Question	Counts of (Parsed) Sentences			
		Input	Labels Assigned		Gets more than one label
			ZSC (multi= False)	ZSC (multi=True)	ZSC (multi=True)
ethics	q1	911	911	990	71(8%)
ethics	q4	1036	1036	1053	16 (2%)
sys	q3	247	247	384	99 (40%)
sys	q4	440	440	716	175(40%)
sys	q5	396	396	553	123(31%)
sys	q6	336	336	1254	259(71%)
sys	q7	188	188	302	73(39%)
sys	q11	282	282	294	11(4%)

Figure 4.2: Proportion of Input Sentences that got Assigned more than one Label in the ZSC (multi-label=true)



4.3. Sub-RQ1

To answer my Sub-RQ1, I provide results of the NLP approaches by each engineering case scenario and its each question prompt investigated in this dissertation study.

4.3.1 *Big Belly Trash Can Ethics Case Scenario*

For the ethics case scenario, I collected a total of 755 student responses. Of these 755, I used 550 to develop the example bank, while the remaining 205 were withheld for labeling using the NLP approaches. Now, I turn to report results for ethics_q1.

4.3.1.1. ethics_q1

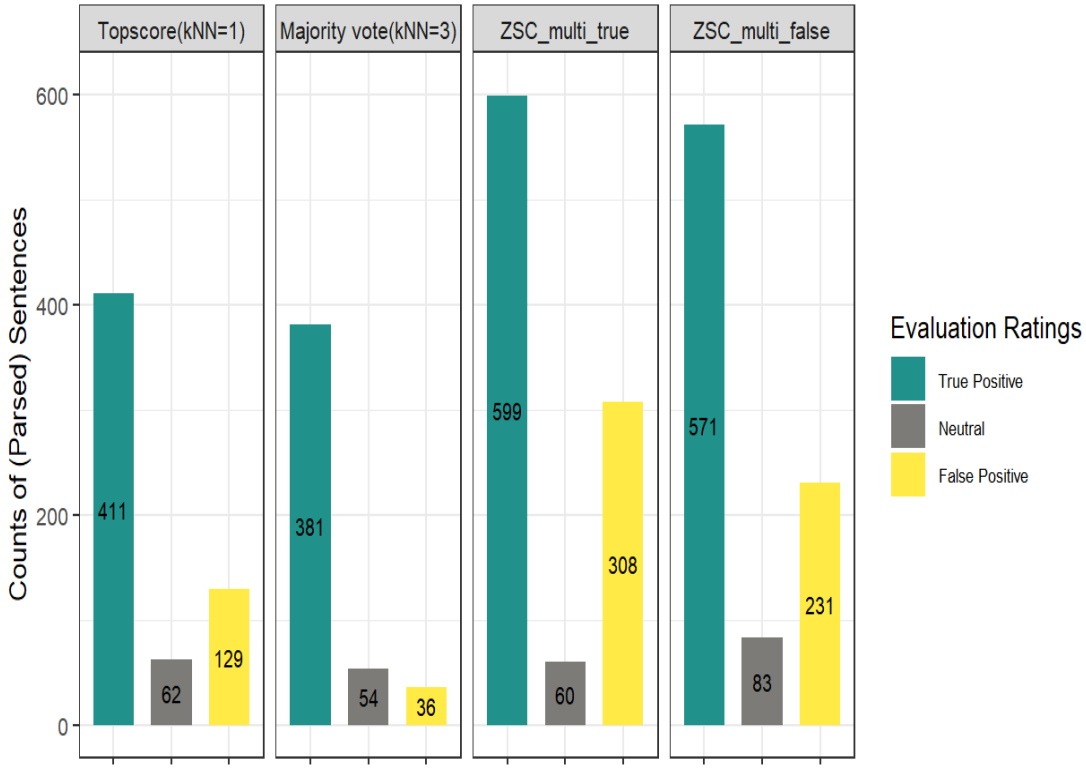
After traditional qualitative coding of 550 student responses, the example bank contained 743 excerpts with assigned codes. The withheld 205 student responses were split at the sentence level using Spacy's segmenter, resulting in 911 unlabeled sentences. These 911 sentences along with the 743 excerpts were input into the kNN methods. Those 911 sentences were also input into the ZSC model along with thirteen labels. After applying threshold values described in section 4.2, the number of sentences for evaluation were 476 for kNN majority vote (k=3) and 990 for ZSC (multi-label=true), as tabulated in Table 4.3.

For ethics_q1, kNN majority vote (k=3) yielded the highest accuracy rate with 80% (381 out of 476) being assigned an accurate label as shown in Table 4.3 and Figure 4.3. On the other hand, ZSC had labeled every input sentence but with lower accuracy rates between 61% to 63%. The ZSC (multi-label=true) had the highest false positive rate at 31%, meaning 231 out of 990 labels assigned were inaccurate codes. However, when comparing accuracy rates between kNN majority vote (k=3) and ZSC (multi-label=true), it is noteworthy that the kNN approach had labeled 476 sentences with at most one label per sentence compared to ZSC (multi-label=true) wherein the 911 input sentences could receive more than one label (990 in this instance). Neutral ratings were comparable at 9-11% across the four methods. Notably, ethics_q1 is the only question prompt among eight question prompts considered in this study for which I used traditional qualitative coding to develop an example bank, as discussed in Chapter 3. I argue that complete excerpts, which may have single or multiple sentences (e.g., "The case study mentions that their source of income by recycling bottles was also hindered by the trash collectors, but their source of food is also hindered.") , as examples used for labeling might be a reason for the better performance of the kNN approach than ZSC, which only required labels. Next, I describe the results for the ethics_q4.

Table 4.3: Evaluation Ratings of Assigned labels by NLP approaches for ethics_q1

Methods	Input Sentences	Total # of labels Assigned	True Positive	Neutral	True Negative	Noise
Topscore (kNN =1)	911	602	411 (68%)	62 (10%)	129 (21%)	0
Majority vote (kNN =3)		476	381 (80%)	54 (11%)	36 (8%)	5
ZSC (multi = True)		990	599 (61%)	60 (6%)	308 (31%)	23
ZSC (multi = False)		911	571 (63%)	83 (9%)	231 (29%)	26

Figure 4.3: Counts for Evaluation Ratings of Assigned labels by NLP approaches for ethics_q1



4.3.1.2. ethics_q4

The example bank for ethics_q4 was developed using the HILNLP method, and it contained 2,993 labeled response sentences. On the other hand, the test set contained 1,043 unlabeled sentences, which were obtained by splitting the withheld 205 student responses at the sentence level using Spacy's segmenter. These 1,043 sentences were input into the four NLP methods for evaluation purposes. The number of labels for evaluation were between 658 to 1,053 for different NLP methods, as listed in Table 4.4.

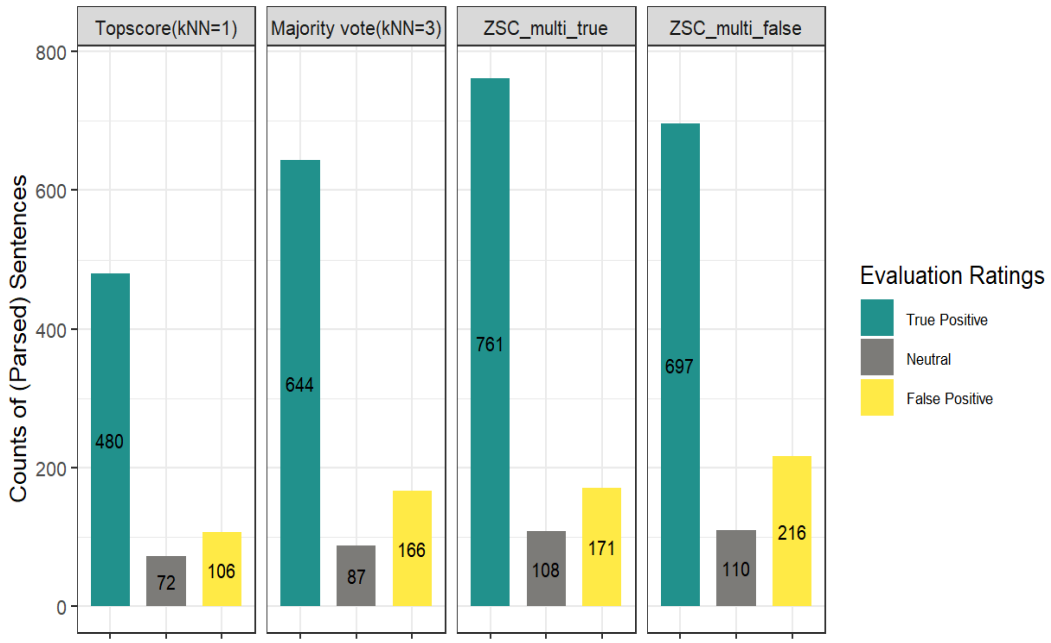
For ethics_q4, the true positive rate was approximately 72% across the following three NLP approaches: kNN top score (k=1), kNN majority vote (k=3), and ZSC (multi-label = true), as shown in Table 4.4 and Figure 4.4. However, among these three approaches, kNN top score (k=1) assigned accurate labels to 480 out of 658 (73%) total labeled sentences. On the other hand, ZSC (multi-label = true) accurately labeled more sentences, with 761 out of 1,053 (72%) total labels. The percentages of inaccurately labeled sentences were 16 to 21% across NLP approaches and neutral ratings were between 10 to 11%. For the approximately similar performance of kNN and ZSC approaches, I reasoned that ethics_q4 asked students to identify a stakeholder and the impact of the ethical dilemma on them. Students in their responses primarily included nouns e.g., homeless people, sanitary workers, government. Therefore, due to the prevalent mention of nouns in student responses in ethics_q4, the complete example responses in the kNN approach (which was mostly lists of nouns) and the only label for the ZSC approach (which was only noun phrases by design) became similar during the labeling of unlabeled sentences.

After presenting results for the ethics case scenario, I turn to describe accuracy results for the systems thinking case scenario.

Table 4.4: Evaluation Ratings of Assigned labels by NLP approaches for ethics_q4

Method	Input Sentences	Total # of labels Assigned	True Positive	Neutral	True Negative	Noise
Topscore (kNN =1)	1036	658	480 (73%)	72 (11%)	106 (16%)	0
Majority vote (kNN =3)		899	644 (72%)	87 (10%)	166 (18%)	2
ZSC (multi = True)		1053	761 (72%)	108 (10%)	171 (16%)	13
ZSC (multi = False)		1036	697 (67%)	110 (11%)	216 (21%)	13

Figure 4.4: Counts for Evaluation Ratings of Assigned labels by NLP approaches for ethics_q4



4.4. Abeesee Village Systems Thinking Case Scenario

For the systems thinking case scenario, I collected a total of 424 student responses. Of these 424, I used 262 to develop the example bank using the HILNLP procedure described in section 3.5.2.1. I used the remaining 162 student responses for testing the NLP approaches. Next, I describe the accuracy evaluation results of the NLP approaches for each question prompt of the system thinking case scenario.

4.4.1 *sys_q3*

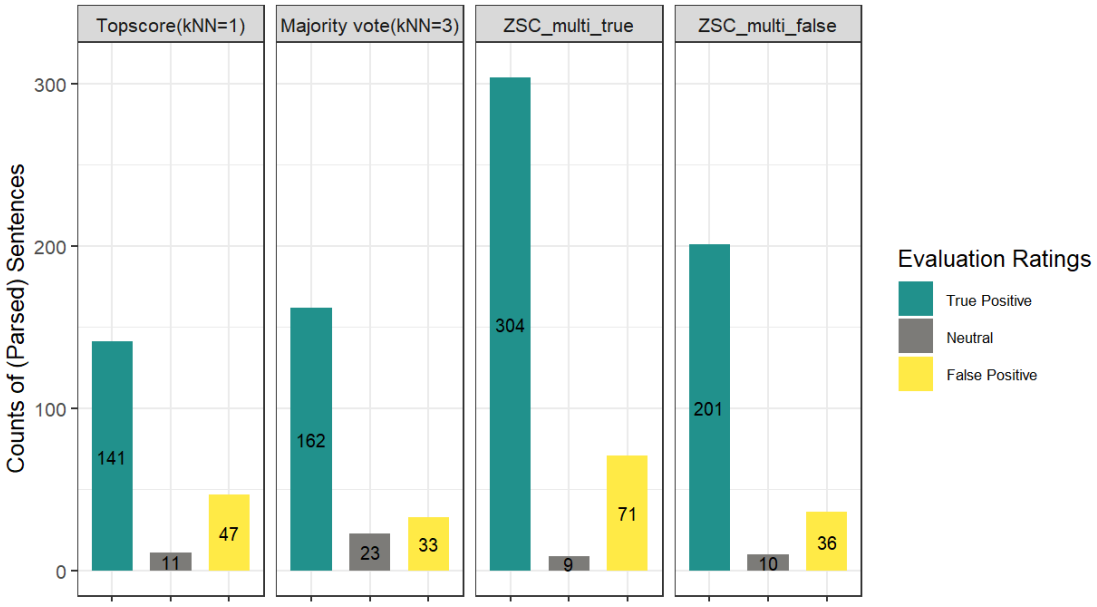
The example bank for the sys_q3 consisted of 401 labeled sentences, while the number of input unlabeled sentences was 247 for the NLP approaches. After the matching process of these 247 unlabeled sentence with 401 labeled sentences, the numbers of labels assigned were 218 and 384 for the kNN majority vote (k=3) and ZSC (multi-label =true), respectively.

For sys_q3, both configurations of the ZSC approach, in comparison to the two kNN methods, not only achieved higher true positive rates but also labeled more sentences. In the case of the ZSC approach, the accuracy rates were 81% and 79 % for (multi-label=false) and (multi-label=true), respectively. On the other hand, those accuracy rates for kNN majority vote (k=3) and kNN topscore (k=1) were 74% and 71%, respectively, as shown in Table 4.5 and Figure 4.5. When comparing the number of accurately labeled sentences, ZSC (multi-label= true) accurately coded 304 out of 384 (79%) labeled sentences, while kNN topscore (k=1) coded accurately 141 out of 199 (71%) labeled sentences. In sys_q3, the ZSC approaches had higher true positive rates than the kNN methods and for this I give two reasons. First, only nine labels (e.g., alternative heat sources, bad economy and poverty, deaths, harsh winter etc.) were used in the ZSC approach for sys_q3. Second, most of these input labels were mentioned in causal attribution in student responses. For example, deaths of townspeople were attributed to harsh winter and lack of heating resources due to prevalent poverty in the town. Therefore, fewer labels and most of them mentioned in student responses enabled the ZSC approach to accurately classify most of the total assigned labels.

Table 4.5: Evaluation Ratings of Assigned labels by NLP approaches for sys_q3

Method	Input Sentences	Total # of labels assigned	True Positive	Neutral	True Negative	Noise
Topscore (kNN =1)	247	199	141 (71%)	11 (6%)	47 (24%)	0
Majority vote (kNN =3)		218	162 (74%)	23 (11%)	33 (15%)	0
ZSC (multi = True)		384	304 (79%)	9 (2%)	71 (18%)	0
ZSC (multi = False)		247	201 (81%)	10 (4%)	36 (15%)	0

Figure 4.5: Counts for Evaluation Ratings of Assigned labels by NLP approaches for sys_q3



4.4.2 *sys_q4*

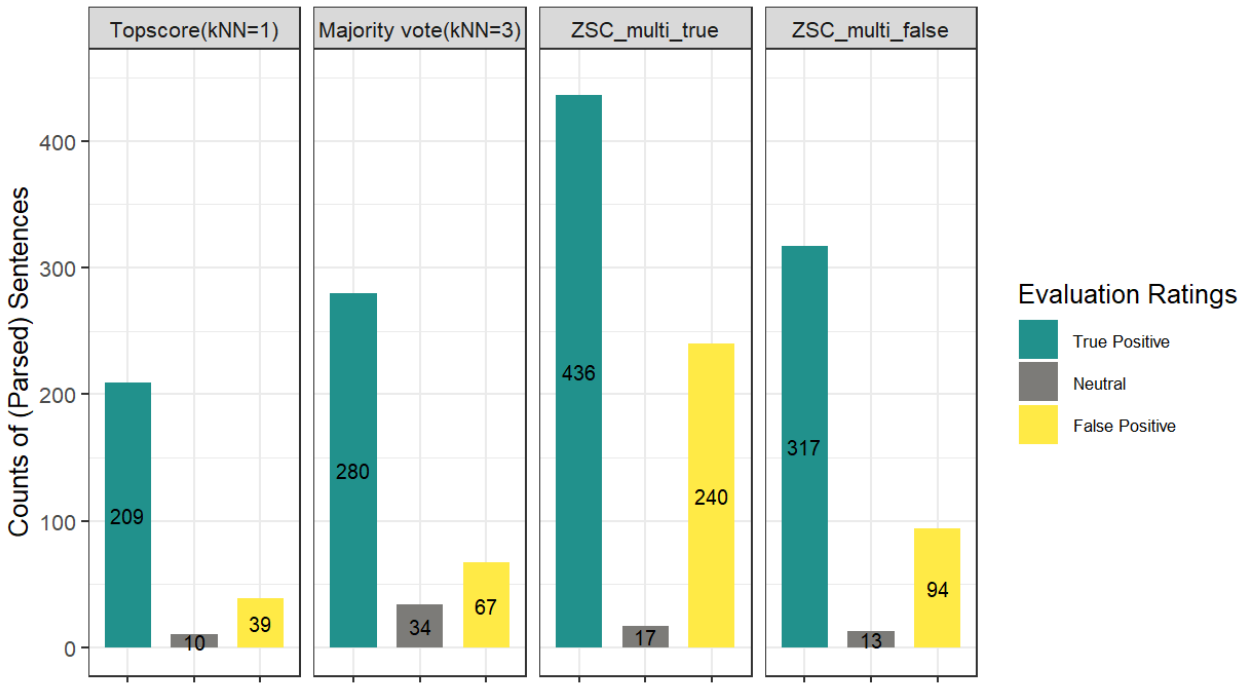
The sys_q4 was one of two question prompts in this dissertation study for which student responses were segmented at the phrase level using comma as a delimiter rather than the sentence level. Using the HILNLP, I developed the example bank for sys_q4 that had 1,182 labeled response phrases. These labeled phrases were then matched with 440 unlabeled response phrases from a withheld dataset of 162 students. For instance, in the kNN top score (k=1) approach, all 440 unlabeled response phrases were matched with one of the 1,182 labeled response phrases. Though, only 258 from 440 unlabeled response phrases were assigned labels as shown in Table 4.6 because not all of the matches were about the threshold similarity score of 0.7. Notably, sys_q4 has the highest number of response phrases across the eight question prompts that are marked as noise. This is because segmenting student responses at the comma level often resulted in transition words being isolated as standalone, meaningless phrases.

For sys_q4, the ZSC approaches had higher false positive rates than the kNN approaches as shown in Table 4.6 and Figure 4.6. For instance, ZSC (multi-label=true) inaccurately labeled 240 out of 716 (34%) assigned codes. On the other side, the kNN topscore (k=1) had inaccurately labeled only 15% (39 out of 258) of total assigned codes. In the ZSC approaches, the higher false positive rates might have been attributed to similar phrasing present in input labels. Some examples of potentially problematic input labels were resource cost, resource availability, root cause, and price rise cause. The former two labels shared the word “resource” and the latter two shared the word “cause”. However, these labels had underlying distinct definitions in the context of sys_q4. Since the ZSC method relied solely on the phrasing of labels, the presence of the shared term in labels might have led to potential errors in label assignment.

Table 4.6: Evaluation Ratings of Assigned labels by NLP approaches for sys_q4

Method	Input Sentences	Total # of labels assigned	True Positive	Neutral	True Negative	Noise
Topscore (kNN =1)	440	258	209 (81%)	10 (4%)	39 (15%)	0
Majority vote (kNN =3)		381	280 (73%)	34 (9%)	67 (18%)	0
ZSC (multi = True)		716	436 (61%)	17 (2%)	240 (34%)	23
ZSC (multi = False)		440	317 (72%)	13 (3%)	94 (21%)	16

Figure 4.6: Counts for Evaluation Ratings of Assigned labels by NLP approaches for sys_q4



4.4.3 *sys_q5*

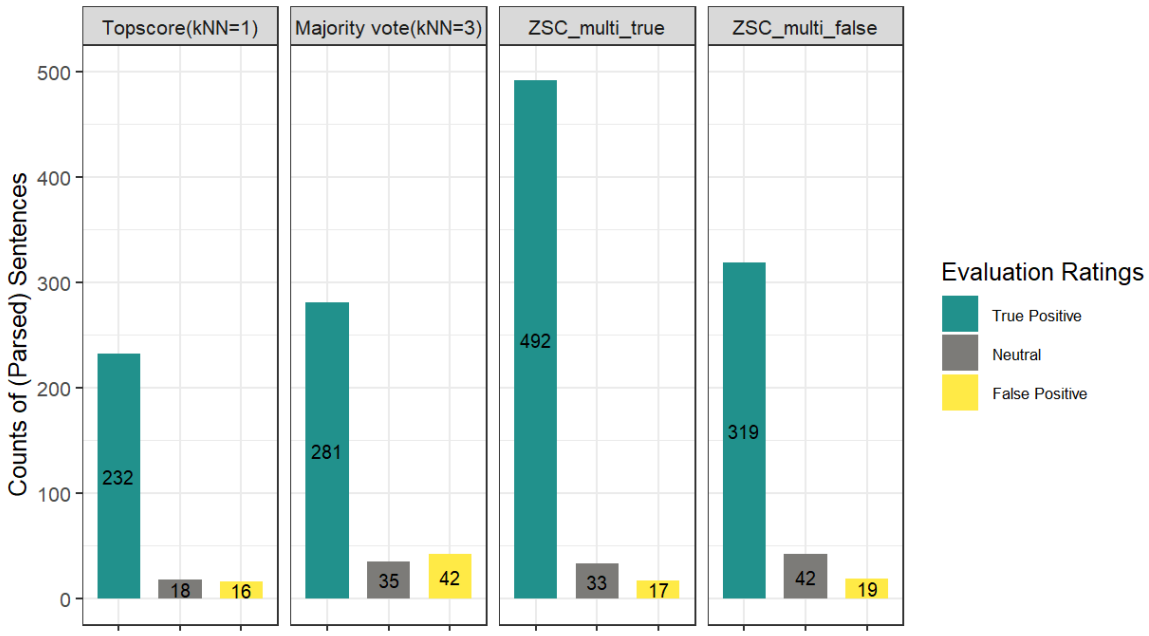
The sys_q5 is the second of two question prompts in this study where I segmented student responses at phrase level. The example bank of sys_q5 had 1,108 response phrases with assigned labels. The withheld dataset of 162 student responses was split into 396 unlabeled response phrases. When 1,108 labeled response phrases were matched to 396 unlabeled response phrases in the kNN methods, top score (k=1) had 276 assigned codes and majority vote (k=3) had 358 assigned codes, as listed in Table 4.7.

For sys_q5, all four of the NLP methods had accurately coded 78% or above of their respective total labeled response phrases. The sys_q5 is the only question prompt across both scenarios where all four NLP methods achieved a true positive rate exceeding 78% as shown in Figure 4.7. In terms of counts of labels assigned, ZSC (multi-label=true) had accurately coded 492 out of 553 (89%) labeled sentences, whereas kNN topscore accurately coded 281 instances out of 358 (78%) labeled instances. Similar to the stakeholder related question prompt in the ethics case scenario (ethics_q4), sys_q5 asked students to list stakeholders they would have involved in planning a response to the problems described in the systems thinking case scenario. However, the true positive rates of the four NLP methods for sys_q5 were higher than for ethics_q4. I give the following justification. The ethics_q4 asked two parts: first, to list a stakeholder, and then to state the impact of the dilemma on them. In doing so, students might have combined a stakeholder and the impact in a single sentence. Third, phrase level splitting may have also contributed to the issue. In contrast, sys_q5 only asked students to list stakeholders. So, students primarily listed nouns in their responses, such as locals, engineers, heating companies, government officials, or non-profit organizations. Therefore, due to the prevalent mention of nouns in student responses, all NLP methods in sys_q5 achieved higher true positive rates than all other question prompts investigated in this study.

Table 4.7: Evaluation Ratings of Assigned labels by NLP approaches for sys_q5

Method	Input Sentences	Total # of labels assigned	True Positive	Neutral	True Negative	Noise
Topscore (kNN=1)	396	266	232 (87%)	18 (7%)	16 (6%)	0
Majority vote (kNN=3)		358	281 (78%)	35 (10%)	42 (12%)	0
ZSC (multi=True)		553	492 (89%)	33 (6%)	17 (3%)	11
ZSC (multi=False)		391	319 (82%)	42 (11%)	19 (5%)	11

Figure 4.7: Counts for Evaluation Ratings of Assigned labels by NLP approaches for sys_q5



4.4.4 *sys_q6*

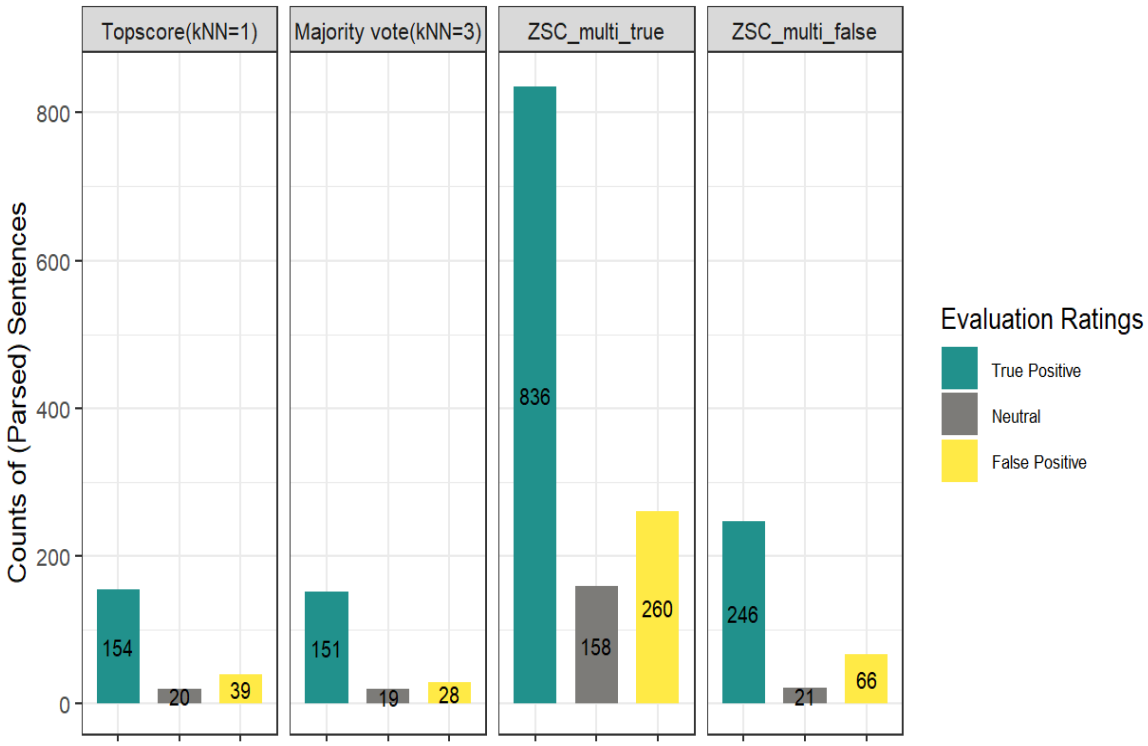
The example bank of sys_q6 had 913 labeled response sentences with 23 unique labels, while the number of unlabeled sentences was 336. When 23 unique labels were input with 336 unlabeled response sentences in the ZSC methods, its configuration of (multi-label = true) assigned 259 out of 336 (71%) sentences more than one label, with a total of 1,254 assigned, as listed in Table 4.8.

Opposite to sys_q5, the true positive rates for all four of the NLP approaches were less than 78% as shown in Table 4.8 and Figure 4.8. The kNN majority vote (k=3) had a maximum accuracy rate of 76%, and it assigned accurate labels to 151 out of 198 instances. When comparing the NLP methods for the sys_q6, it should be noted that the number of total labeled sentences significantly varied across NLP methods. For example, the ZSC (multi-label= true) had accurately coded 836 out of 1,254 (67%) total assigned labels, while kNN approach had labeled 154 out of 213 (72%) total assigned labels. Notably, sys_q6 was the only question prompt among the eight question prompts investigated in this dissertation study for which 259 sentences out of 336 (71%) input sentences received more than one label in the ZSC (multi-label= true) approach. This was because students had combined multiple closely aligned ideas in a short text. I gave one example sentence: "Then I would have contacted local officials." This example sentence received the following accurate codes in the ZSC (multi-label= true) approach: "collaborate with stakeholders", "feedback to solutions", "interact with government", and "interact with locals".

Table 4.8: Evaluation Ratings of Assigned labels by NLP approaches for sys_q6

Methods	Input Sentences	Total # of labels assigned	True Positive	Neutral	True Negative	Noise
Topscore (kNN =1)	336	213	154 (72%)	20 (9%)	39 (18%)	0
Majority vote (kNN =3)		198	151 (76%)	19 (10%)	28 (14%)	0
ZSC (multi = True)		1254	836 (67%)	158 (13%)	260 (21%)	0
ZSC (multi = False)		336	246 (73%)	21 (6%)	66 (20%)	3

Figure 4.8: Counts for Evaluation Ratings of Assigned labels by NLP approaches for sys_q6



4.4.5 *sys_q7*

The example bank for sys_q7 consisted of 478 labeled sentences, while the number of input unlabeled sentences for the NLP methods was 188. After the matching process of these 188 unlabeled sentences with 478 labeled sentences, the number of labels assigned in kNN topscore (k=1) and kNN majority vote (k=3) were 151 and 152, respectively, as mentioned in Table 4.9. Conversely, the number of labels assigned in the ZSC (multi-label=true) and ZSC (multi-label=false) were 302 and 188, respectively.

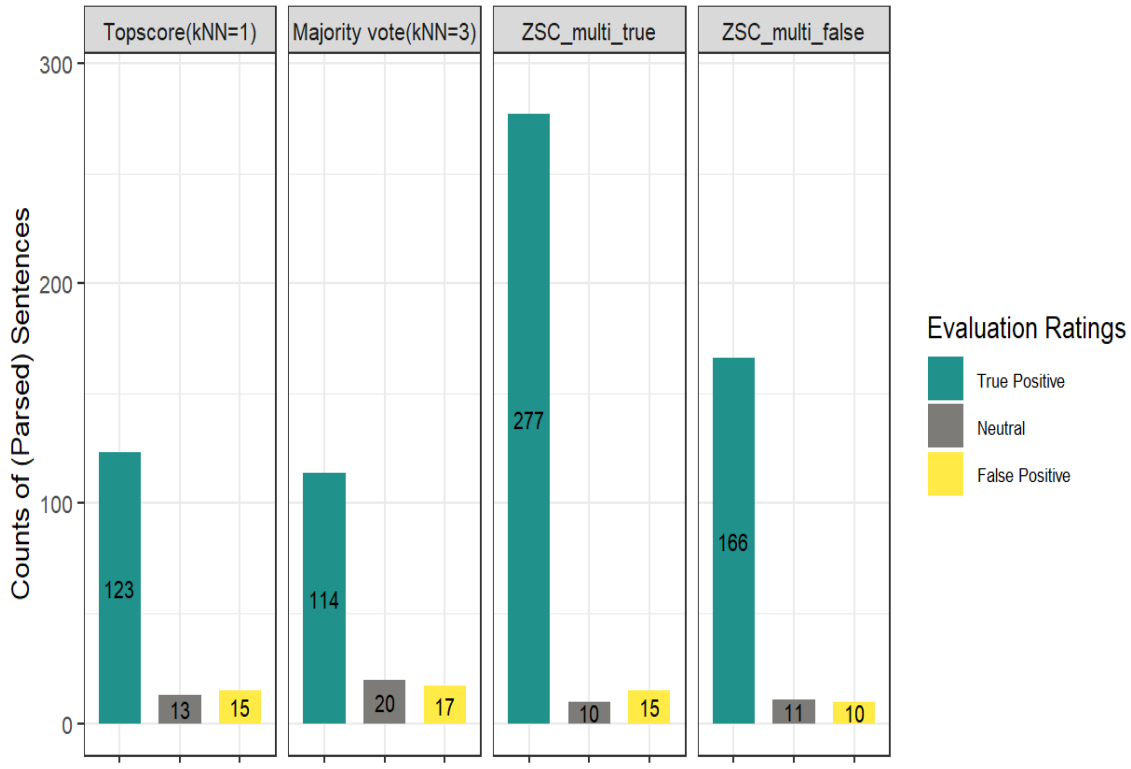
Similar to sys_q3, the ZSC methods in sys_q7 had higher true positive rates across total labeled sentences than the kNN methods. For instance, the ZSC (multi-label=true) and ZSC (multi-label=false) had accurately labeled 92% and 88% of their respective total labeled instances, as shown in Table 4.9 and Figure 4. **9**. Conversely, the kNN topscore (k=1) and kNN majority vote (k=3) had accurately labeled 81% and 75% of their respective total labeled instances. Moreover, the ZSC (multi-label=true) in the sys_q7 had achieved the highest accuracy rate (92%) among all four NLP methods used across the eight question prompts in this dissertation study. For the highest performance of ZSC (multi-label=true) in sys_q7, I give following two reasons. First, only ten

labels were used as input labels. Some examples of input labels for sys_q7 were “heating access”, “safe operation”, “reduce deaths” etc. Second, the underlying meaning of these input labels were mostly mentioned in student responses. For example, “no deaths due to hypothermia, all homes have a safe form of heat”. In this example sentence, the following labels were accurately assigned to the given example sentence: “heating access”, “safe operation”, “reduce deaths”. Therefore, having fewer input labels compared to other question prompts and the fact that those labels mirror language that students used enabled the ZSC approach to achieve highest accuracy rate (92%).

Table 4.9: Evaluation Ratings of Assigned labels by NLP approaches for sys_q7

Method	Input Sentences	Total # of labels assigned	True Positive	Neutral	True Negative	Noise
Topscore (kNN =1)	188	151	123 (81%)	13 (9%)	15 (10%)	0
Majority vote (kNN =3)		152	114 (75%)	20 (13%)	17 (11%)	1
ZSC (multi = True)		302	277 (92%)	10 (3%)	15 (5%)	0
ZSC (multi = False)		188	166 (88%)	11 (6%)	10 (5%)	1

Figure 4. 9: Counts for Evaluation Ratings of Assigned labels by NLP approaches for sys_q7



4.4.6 *sys_q11*

For sys_q11, the number of example response sentences in the example bank was 712 with 18 unique labels. The number of unlabeled response sentences was 282. After the matching process of these 282 unlabeled sentences with 712 labeled sentences, the number of labels assigned in kNN topscore (k=1) and kNN majority vote (k=3) were 116 (41%) and 159 (56%), respectively. Conversely, when 282 unlabeled sentences with 18 unique labels were input into the ZSC model, the number of assigned labels were 294 for the ZSC (multi-label=true), whereas every sentence of 282 unlabeled sentences was assigned one of 18 labels in the ZSC (multi-label=false).

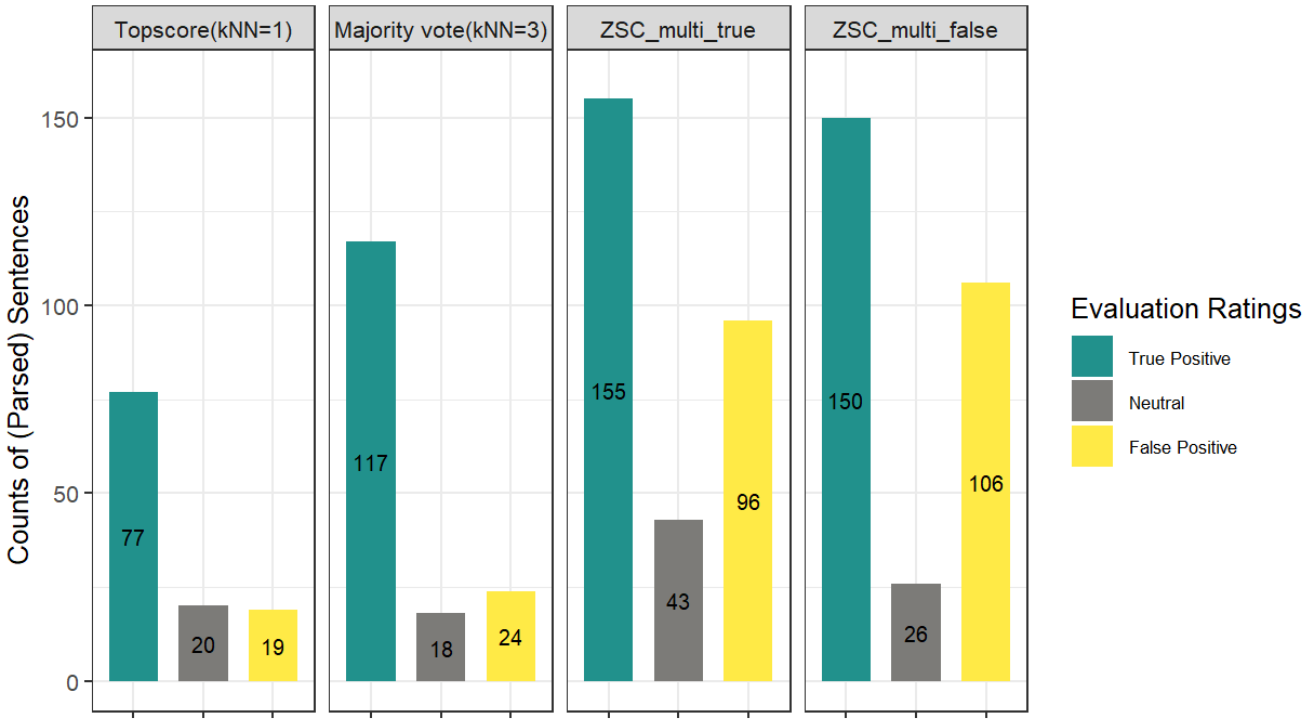
For sys_q11, both configurations of the ZSC method accurately labeled only 53% of their total assigned labels as listed in Table 4.10 and Figure 4.10. On the other hand, kNN majority vote (k=3) and kNN topscore (k=1) had accurately labeled 74% and 66% of their total labeled sentences, respectively. In addition, the ZSC approaches had the highest false positive rate for all four NLP methods across eight question prompts investigated in this dissertation study. According to Table 4.10, the ZSC (multi-label= true) and the ZSC (multi-label= false) had inaccurately assigned labels to 96 out of 294 (33%) and 106 out of 282 (38%) of total labeled sentences,

respectively. The difficulty of ZSC approaches in assigning accurate labels might have been due to multiple ideas being packed in sys_q11 question (i.e., “What challenges did you see in implementing your plan? What were the limitations of your approach?”). In their responses, students combined different ideas, which was difficult for the ZSC approaches to accurately classify based on only phrasing of input labels.

Table 4.10: Evaluation Ratings of Assigned labels by NLP approaches for sys_q11

Methods	Input Sentences	Total # of labels Assigned	True Positive	Neutral	True Negative	Noise
Topscore (kNN =1)	282	116	77 (66%)	20 (17%)	19 (16%)	0
Majority vote (kNN =3)		159	117 (74%)	18 (11%)	24 (15%)	0
ZSC (multi = True)		294	155 (53%)	43 (15%)	96 (33%)	0
ZSC (multi = False)		282	150 (53%)	26 (9%)	106 (38%)	0

Figure 4.10: Counts for Evaluation Ratings of Assigned labels by NLP approaches for sys_q11



4.5. Sub-RQ2 and Sub-RQ3

To answer my Sub-RQ2 and Sub-RQ3, I summarize the distributions of proportions for the following: accurately labeled sentences (true positives), sentences labeled for partial credit or as ambiguous (neutral), and inaccurately labeled sentences (false positives) out of total assigned instances by the four NLP methods.

When comparing true positive rates, the kNN majority vote (k=3) achieved similar true positive rates across the question prompts in both case studies as shown in Figure 4.11**:**. In six question prompts of the systems thinking case scenario, the true positive rates ranged from 74% to 78%, while the true positive rate was 80% for ethics_q1 and 72% for ethics_q4. Conversely, the ZSC (multi-label=true) method exhibited a wide variation in true positive rates across the question prompts in the case studies, as depicted in Figure 4.11**:**. For instance, for the sys_q11, the true positive rate was 53%, while it was 92% for the sys_q7. For the ethics case scenario, true positive rates ranged from 61% to 72%. Likewise, when comparing the proportion of neutral ratings proportions out of total labels assigned, the kNN majority vote (k=3) achieved comparable proportion of neutral rating across the question prompts in both case studies that ranged from 9-13% as shown in Figure 4.12. Conversely, the ZSC (multi-label=true) method showed a wide

variation in the proportion of neutral ratings out of the total assigned labels across the question prompts in both case studies, that ranged from 2% to 15% as depicted in Figure 4.12. This is because the kNN majority vote (k=3) and ZSC (multi-label=true) differ in their working principles. The kNN method considers k nearest neighboring example sentences in assigning labels which could help in capturing the context. In contrast, the ZSC method focuses on the specific phrasing of input labels, making it more susceptible to discrepancies if phrasing is suboptimal—for this I gave examples the below paragraph when comparing false positives rates across the NLP approaches.

Figure 4.11: Distribution of True Positive Rates Across Question Prompts and Case Scenarios

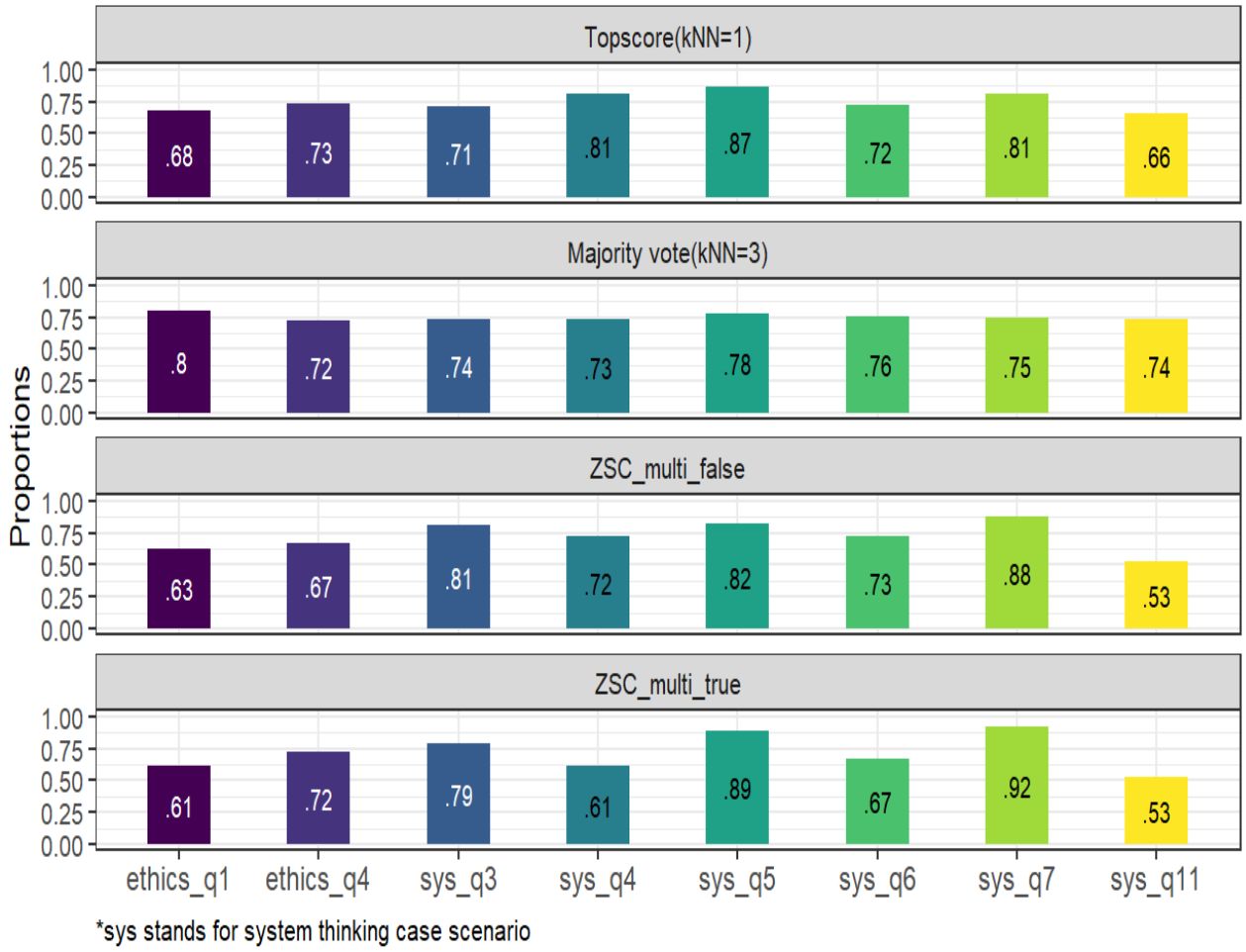
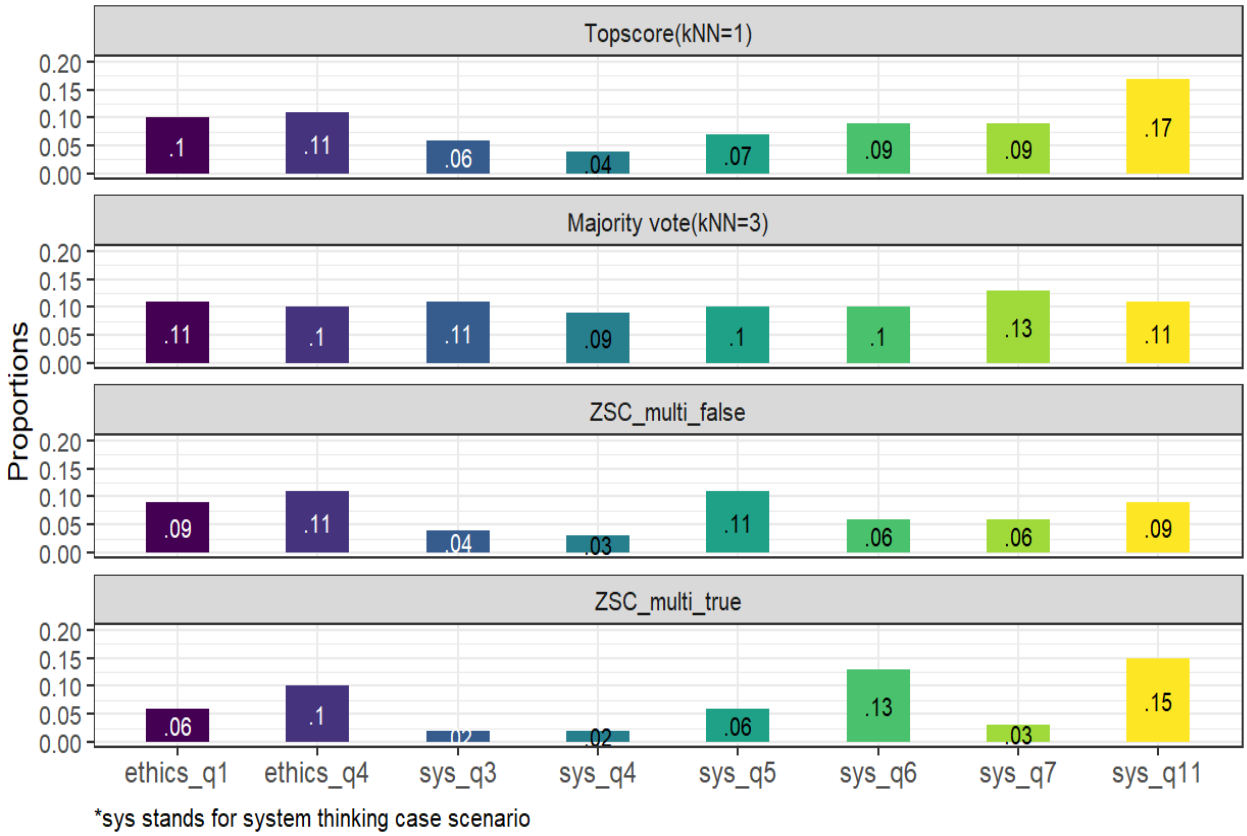
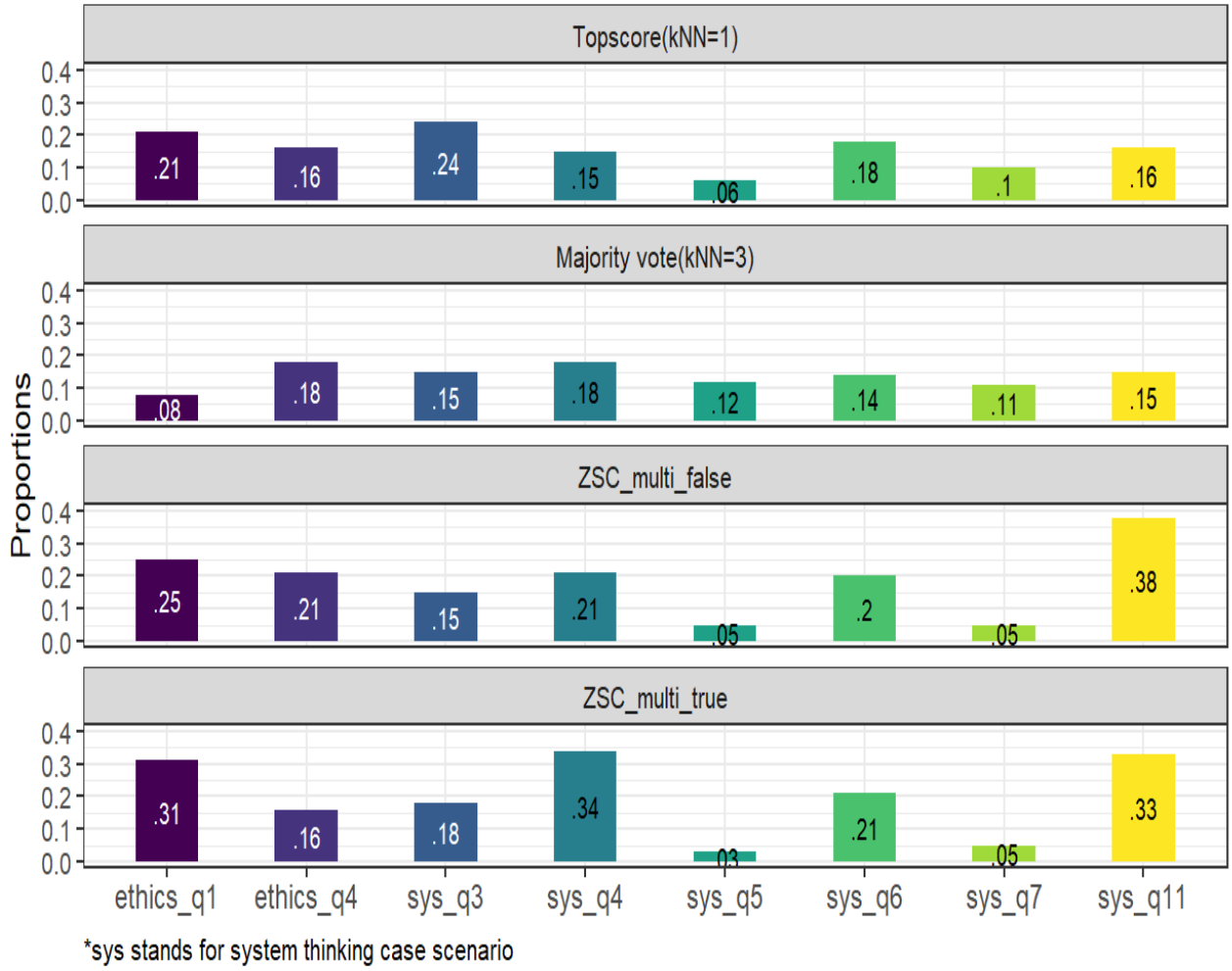


Figure 4.12: Distribution of Neutral Ratings Rates Across Question Prompts and Case Scenarios



When comparing false positive rates, ZSC (multi-label=false) and ZSC (multi-label=true) had the highest false positive rates for sys_q11, i.e., 38% and 33% respectively. On the other hand, both ZSC methods showed the lowest false positives rates for sys_q5 and sys_q7 question prompts, ranging between 3% and 5% as shown in Figure 4.13. This discrepancy is due to the ZSC model predominantly relying on the phrasing of input labels for classification. In the case of sys_q5, it only asked students to list stakeholders, so students primarily listed nouns in their responses. Examples of input labels in the ZSC approaches were locals, engineers, heating companies, government officials, and non-profit organizations. Since students responded predominantly with nouns that clearly matched the input labels, the ZSC methods were able to classify the text with the correct labels in this case. Conversely, sys_q11 asked students to articulate challenges and limitations in a single question. This led to students mixing different ideas in responses. In addition, the input labels had close lexical similarity, like “product affordability”, “product efficiency”, and “product durability”, which conflated the accuracy of ZSC approaches.

Figure 4.13: Distribution of False Positive Rates Across Question Prompts and Case Scenarios



Chapter 5: Discussion and Conclusion

5.1. Chapter Overview

As presented in Chapter 4, the results of my study showed moderate accuracy in thematically analyzing students' open-ended responses to two different engineering case scenarios. In Chapter 5, I synthesize information from my research that would help to address the following challenge: how can we evaluate student written responses to case studies at scale? In doing so, first, I provide discussion about my study's results. Second, I discuss the research quality measures I adopted in my study to enable transferability of my results to other contexts. Third, I provide guidelines for educators to develop automated analysis or grading systems for open-ended responses. Fourth, I acknowledge and offer points to mitigate skepticism about automatic grading or qualitative analysis of student open-ended responses. Fifth, I present how my study contributes to ongoing conversations in the engineering education field. Sixth, I provide directions for future research. Lastly, I conclude this manuscript with my concluding remarks.

5.2. Discussion of the Results

In this section, first I provide limitations of my results. Then, I analyze and interpret my results. Lastly, I connected my results with the existing literature.

5.2.7 *Limitations of the Results*

I acknowledge there are numerous limitations associated with my results. These limitations may qualify the claims about implications of my results for research and practice in engineering education, which I present in section 5.4.

First, not all of the student responses I have collected in this study are assigned labels by my NLP approaches. For instance, for ethics_q3, 48% of total student responses collected in the study are not assigned labels by the kNN majority vote (k=3) because the example bank is deficient and did not include semantically similar example sentences. Some examples of those unlabeled sentences are : (i) "Also, prospective students are more likely to have a good impression of a clean campus, which may lead to an increased acceptance rate of the college.", (ii) "General students benefit because the campus is cleaner." I suggest the main reason why some responses are not labeled is the absence of semantically similar sentences in the example bank. I vectorized both unlabeled responses and example responses using a TLLM. If those responses are not semantically closer, the vectors representing them would be far apart in a higher-dimensional space, resulting in a low cosine similarity score. Therefore, when I employed a threshold value of cosine similarity to shortlist responses for labeling purposes, any sentences below the threshold value remain unlabeled. The solution may be to include more relevant labeled examples in the example bank.

Second, there are non-trivial proportions of inaccurately assigned labels among the total assigned labels by my NLP methods. For example, the proportion of inaccurately assigned labels ranges between 3% and 38% of total assigned labels across question prompts investigated in this study. In practice, this may mean student responses would be inaccurately assessed. There are the following two driving factors for inaccurate labeling of NLP methods. First, the pre-processing of student responses can result in introductory sentences being left as standalone sentences without context. The model then matches these introductory sentences with well-thought-out sentences, to which the introductory sentences share a few portions, and mislabels them. For example, the following three sentences about the ethical dilemma of identity (ethics_q3) are mislabeled as "access to income": (i) "Between 1000 and 2000 people in Berkeley experience homelessness a year," (ii) "At the University of Berkeley, big belly trash cans are installed," (iii) "One ethical dilemma/issue from this case study is more applicable now than ever." Second, if ideas are closely aligned, the model may classify a sentence based on minor details rather than the main idea. For example, the following sentences about what additional information is required (sys_q4) are mislabeled as "house structure": (i) "How many families per household?" (ii) "How many people are in each household?" (iii) "How many people live in one home?" However, the statements in example bank described physical infrastructure of houses rather than social composition.

Third, not a single NLP method among the four NLP methods—kNN topscore (k=1), kNN majority vote (k=3), ZSC (multi-label=false), ZSC (multi-label=false)—consistently performs best across the eight question prompts of two case scenarios. The nature of case scenarios did not contribute to performance of NLP methods. Therefore, it is difficult to claim which NLP method may consistently perform better than other NLP methods in different contexts.

Fourth, the proportion of accurately assigned labels out of the total labels assigned by NLP methods is less than 90% and ranges between 60% to 89% across the question prompts. The exception is ZSC (multi-label=true) for sys_q7, which accurately labeled 92% of the total assigned labels. Note that the true positive and false positive rates do not necessarily add to 100%. This is because of the neutral (i.e., 0) rating, which was applied when a assigned label was neither clearly correct nor incorrect. This wide variation in the proportion of accurately assigned labels among NLP methods might create challenges in predicting which labels would be (in)accurately assigned using the NLP approaches. If we did have such an ability to predict inaccuracy in the NLP methods, one could develop an automatic flagging system within the NLP approach to set aside potentially inaccurate labels for manual analysis.

Fifth, my example banks that contained example responses with labels were limited. Theoretically, more example responses with labels could be added to the example banks for assigning codes to unlabeled student responses. This expansion of the example banks may help in the following two ways: first, it may decrease the number of unlabeled student responses; second, it may increase the proportion of accurately assigned labels out of the total assigned labels.

Sixth, my study did not calculate the number of false negatives, a number commonly used in accuracy matrices reported for NLP classification tasks, such as the F-1 score, Recall, or Precision (Galhardi & Brancher, 2018; Kerkhof, 2020a; Shah & Pareek, 2022). The ZSC (multi-label=true) is the only method among the four NLP methods that assigned more than one label to any given student response. For example, 71% (259 out of 336) of total input sentences for sys_q6 received more than one label, making a total of 1258 output labels assigned with ZSC (multi-label=true). Among those 1258, 836 (67%) were accurately assigned labels. Theoretically, the other three approaches investigated in this study should have assigned those labels. If not, those missed (not assigned) labels should be calculated as false negatives.

Lastly, numerous TLLMs are available for use in NLP tasks, but I have only used MPNet in my study. There is a possibility that other TLLMs may perform better than MPNet, which is an open question for future research.

In sum, given the aforementioned limitations of my results, I acknowledge that my NLP approaches showed moderate accuracy in the thematic analysis of student responses to engineering case scenarios. Despite these limitations, there are important observations from my results for automatic analysis (or grading) of student open-ended responses, which I provide next.

5.2.8 *Interpretation of the Results*

The first observation is that student responses can be pre-processed in a multitude of ways, and the pre-processing step impacts the accuracy rate of the NLP approaches. For example, the accuracy rate of the kNN topscore (k=1) for sys_4 and sys_5, where I split student responses at the phrase level, was higher than the accuracy rate of the kNN topscore (k=1) for the other six question prompts, where I split student responses at the sentence level using Spacy's segmenter. Conversely, splitting student responses at the phrase level in sys_q4 did not improve the accuracy rate of the remaining three NLP approaches compared to their accuracy rates for other question prompts. Therefore, it is important to make appropriate choices for pre-processing techniques based on the characteristics of the dataset and NLP approach (Berdanier et al., 2018).

The second observation is that when choosing between the kNN and the ZSC approaches, there is a trade-off between the number of assigned labels and the accuracy of those labels. The kNN approaches always assign fewer labels than the ZSC approaches, though there is often higher inaccuracy in the labels assigned by the ZSC approaches. As an example, for sys_q11, the number of labels assigned by the kNN methods was 116 and 159, and the accuracy was 66% and 74% respectively. On the other hand, for the ZSC approaches, the number of assigned labels was 282 and 294, and the accuracy was 53% for both. Future research can address this trade-off as follows. Related to the kNN approaches, one may expand the example bank by adding new example sentences with labels to increase the number of labels assigned by the kNN approaches. Particularly, if example sentences expressing any specific idea are missing in the example bank, new responses about that idea could not achieve above-threshold semantic similarity with existing example sentences in the bank. Therefore, the addition of alternative phrasing or new example sentences to the example bank will help increase the likelihood of matching unlabeled sentences with labeled sentences. Related to the ZSC approaches, the accuracy of the ZSC approaches may depend upon the phrasing of labels. One should have a minimum number of words as a label. For example, "affordability" as a single label is better than "affordability of heating material". In the case of closely aligned ideas, a more generic label would be better than a more fine-resolution label.

The third observation is that a smaller number of input labels in the ZSC approaches helps to increase the proportion of accurately assigned labels out of the total assigned labels. For example, the number of input labels for sys_q3 is ten; the proportions of accurate labels were 88% and 92% for ZSC (multi-label=false) and ZSC (multi-label=true) respectively. Conversely, when the number of input labels is 18 in sys_q11, the proportion of accurate labels for both configurations of ZSC was 53%. Therefore, one should create fewer input labels with fewer words in each label to achieve higher accuracy for the ZSC approaches.

The fourth observation is related to something mentioned in the limitation section: no single method among the four NLP methods performed consistently better than the other methods across all question prompts investigated in this study. The highest accuracy rate varies by question prompt and NLP method. For example, the kNN majority vote (k=3) performed best in ethics_q1 (the question prompt is "Identify ethical dilemma related to the given case study") with 80% accuracy. On the other hand, the ZSC (multi-label=true) achieved the highest accuracy rate of 92% in sys_q7 (the question prompt is "What would you expect a successful plan to accomplish?"). This raises the question: which method works best in what situation, to which I turn next.

(1) **kNN (k=1):** One may use kNN topscore (k=1) in the following conditions: (a) example bank is saturated, which means it includes all possible ways to answer a question prompt, (b) example sentences in the example bank focus on one idea at a time, and (c) unlabeled sentences closely aligned (have semantic similarity and describe same idea) with example sentences.

(2) **kNN (k=3):** One may opt for the kNN majority vote (k=3) method when: (a) one aims to assign more than one label and (b) unlabeled sentences are not closely aligned with sentences in the example bank. In this instance, the kNN majority vote (k=3) would help to match unlabeled sentences with multiple close neighbors and assign multiple labels.

(3) **ZSC (multi-label=false):** One may choose the ZSC (multi-label=false) approach in the following scenario: An instructor wants to assess whether students have mentioned a key concept in their answers and the instructor has curated a list of those concepts. Notably, this approach may be considered similar to keyword matching-based ASAG systems. However, these systems are criticized for their inability to capture the same keywords expressed in other words. I suggest the ZSC models are an advancement because the ZSC models use contextualized word embeddings and can capture alternative phrasings of keywords.

(4) **ZSC (multi-label=true):** One may choose the ZSC (multi-label=true) approach in the following scenario: an instructor wants to assess multiple concepts in student answers. For this instance, I suggest ZSC (multi-label=true) since it has the ability to apply multiple labels to a single statement.

Lastly, for instructors who may be interested in assessment of systems thinking or engineering ethics learning objectives using these NLP methods, I give the following two suggestions.

(1) One may use ZSC methods for learning goals related to ethical sensitivity or awareness (e.g., stakeholder identification). In question prompts related to these learning objectives, students generally mentioned nouns. The instructor can assess these responses by inputting a list of concepts expressed as nouns into the ZSC model.

(2) Conversely, the kNN method may be helpful for assessing learning goals associated with ethical reasoning. (e.g., identifying ethical dilemmas or problem spaces, and the

impact of dilemmas or problems on stakeholders). Students in their responses to these types of learning goals' question prompts tend to develop their reasoning over multiple sentences. The kNN method may be better for assessment of that reasoning because it enables matching those student passages with similar example sentences or paragraphs from the example bank.

5.2.9 *Connecting the Results to the Literature*

After this comparison, I connected my findings with literature about data pre-processing, splitting dataset for ML algorithms, and the difference with supervised ML approach. After connecting my results with NLP literature, I connected my study with use case literature. The literature is about ethics and system thinking. Lastly, I connected my findings with existing engineering education studies that have used NLP techniques.

There are notable differences between my study and the studies reported in the ASAG literature. First, ASAG-related studies typically use public datasets such as SciEntsBank, Beetle, Texas 2011, and ASAP-SAS, while I used private educational datasets (student assignments) available at Virgina Tech. Second, ASAG-related studies evaluate the performance of their automatic grading systems on tasks like 2-way, 3-way, or 5-way classification of student answers from those public datasets. As an example, a 3-way classification task includes labeling student answers into correct, partially correct, or incorrect categories (Bai & Stede, 2022). On the other hand, the NLP classification approaches I used in my study had many more labels than two, three, or even five (i.e., 18 labels for sys_q11). Third, ASAG-related studies report their systems' performance as common metrics like accuracy, precision, recall, F-1 score, and Quadratic Weighted Kappa (QWK). For definitions of these metrics, readers may refer to section 1.9. However, I did not use these common metrics to quantify the performance of my NLP approaches. This is because I introduced the neutral rating, besides true positive or false positive, which was employed when the assigned label by my NLP approach was neither correct nor incorrect. This neutral rating made the calculation of those common metrics technically infeasible.

Given the aforementioned differences, I attempt to provide a qualitative comparison of my NLP approaches' performance to ASAG systems reported in the literature. The F1 scores achieved using BERT and its variants on the SciEntsBank dataset range from 79-96% (Camus & Filighera, 2020; Forsyth & Mavridis, 2021). In the first scoping review of embedding in ASAG (Putnikovic & Jovanovic_2023), the accuracy of reviewed studies ranged between 42% and 82% (see table XI, p. 9), which are comparable to my result (53% to 92%). This equivalent accuracy shown in the qualitative comparison demonstrate the utility of my NLP approaches. Next, I compare my data pre-processing technique to commonly used in ASAG-studies. The equivalent accuracy shown in

the above qualitative comparison demonstrate the utility of my NLP approaches. Next, I compare my data pre-processing technique to commonly used in ASAG-studies.

My study did not use common pre-processing techniques such as stemming, lemmatization, and removal of punctuations or stop words (Jurafsky & Martin, 2009; Kerkhof, 2020; Manning et al., 2014). Recent works have questioned the necessity of such techniques (Crossley et al., 2019; Kumar & Boulanger, 2021). Despite the fact that I had not used those techniques in my NLP approaches, I achieved comparable performance with ASAG systems that used common pre-processing techniques. Therefore, my findings may support the recent works which claim that pre-processing techniques do not impact the accuracy of NLP-assisted grading systems, although this would require further testing beyond the scope of this dissertation.

ASAG systems based on supervised ML depend on the size and distribution of labels in the training dataset. Labels that are underrepresented or overrepresented can lead to bias in those ASAG systems (Blodgett et al., 2020; Shermis, 2014). Since I did not train any ML model from scratch in my tested approaches, the volume of data needed to use such approaches is smaller than the amount needed when training a brand new supervised model. This is noteworthy because labeling datasets in supervised ML is resource-intensive and task-dependent. My study aligns with recent works aiming to reduce reliance on labeled datasets in ASAG systems.

Researchers have cautioned that students may attempt to game ASAG systems, particularly based on keyword matching mechanism, by stringing together incoherent keywords in their responses (Almazova et al., 2021; H.-S. Lee et al., 2019). My NLP approaches are less prone to such system gaming by students because these are based on TLLMs and use the underlying meaning of words rather than the morphological matches or grammatical structure.

I categorize the engineering education literature related to NLP into two categories. The first category is studies that use frequency or dictionary-based feature extraction methods such as (Berdanier et al., 2018, 2020; Bhaduri, 2018; Bhaduri & Roy, 2017; Soledad et al., 2017). My approach has demonstrated notable improvements compared to these studies, as indicated by approximate comparisons between reported accuracy scores. The second category is research studies published by my research group at the department of Engineering Education, Virginia Tech (Anakok et al., 2022; Chew et al., 2022; Gamieldien, 2023; Gamieldien, Case, et al., 2023; Gamieldien, McCord, et al., 2023; Shakir et al., 2022). These studies use state-of-the-art TLLMs in their NLP approaches. I contributed to these approaches by incorporating (i) a different context (focused on students' ethics and system thinking assignments), (ii) kNN approach for matching methods , and (c) test the utility of the ZSC approach in different context.

Next, I compare my NLP approaches with other NLP techniques that have been used in the engineering ethics education field. Taraban et al have employed various frequency-based NLP approaches in ethics education (Taraban et al., 2018, 2022, 2017, 2019; Taraban, Marcy, et al., 2020; Taraban, Robledo, et al., 2020). My NLP approaches are based on recent TLLMs; hence, they are different from the frequency-based approaches used by Taraban et al. For example, in (Taraban, Marcy, et al., 2020), the authors used IBM Watson Natural Language Classifier (Watson-NLC) and a naïve Bayes classifier for distinguishing sentences related to ethics from other sentences in student responses. They reported that Watson-NLC achieved an 81% accuracy rate. In other studies (Taraban et al., 2022, 2019), the authors used the Linguistic Inquiry and Word Count (LIWC) tool to analyze student reflections on ethics. I agreed with the authors that LIWC has limitations in the context of ethics education as its dictionaries may not reflect all categories pertinent to ethics assignments. As per limited knowledge and review of published literature, I suggests Taraban et al. is the only research group that has published related to the use of NLP in engineering ethics. However, their NLP methods are dictionary-based and different than my work completed in this study. This is because I have used TLLMs rather than hand-coded dictionaries to vectorize student engineering ethics responses.

5.3. Research Quality Measures and Transferability of the NLP Approach

Related to quality measure that I employed in my research, my advisor performed member checking in the following two stages of the research process. First, when I was developing the example banks (or coding dictionaries), I solicited his feedback iteratively at various points to improve labels and their phrasing (Miles et al., 2014; Saldaña, 2014, 2021). Second, when I was manually evaluating each assigned label by my NLP approaches, my advisor audited my evaluation ratings. Having my advisor engaged as a member-checking auditor at various stages in my evaluation procedure lends credibility to the numbers and proportions of evaluation ratings (i.e., accurately labeled, inaccurately labeled, and neutral ratings) reported in my results.

Next, I provide important lessons learned from my study for practitioners and educators interested in incorporating NLP into their classroom open-ended assessments.

5.4. Lesson Learned and Guidance on the use of my NLP Approach

In this section, I provide 11 important lessons learned from my study, followed by guidance on navigating these lessons to aid future implementation of my NLP approaches in classrooms (Katz et al. 2023, forthcoming).

5.4.1 *Question Phrasing*

The phrasing of the question significantly affects the quality of student responses and its NLP-assisted analysis. Unclear and multipart questions can be problematic (Chin & Osborne, 2008). For example, a single question prompt asks students to identify multiple stakeholders and explain why they choose those stakeholders in the same question prompt. In answering this the question prompt, student could pack chosen multiple stakeholders and the explanations for choosing them into a single paragraph. During data pre-processing, it would be difficult to segment each stakeholder and its explanation, which could affect the accuracy of NLP-assisted data analysis.

Guidance 1: Compose targeted questions. Do not pack too many elements into a single question. Divide a complex question into several shorter questions. These fairly scoped question phrasing are preferable for better performance of NLP methods investigated in this study. However, , I suggest GAI may mitigate this limitation in the future.

5.4.2 *Data Pre-Processing*

Splitting responses at sentence-level for NLP-assisted analysis works well when students provide independent, self-contained information in each sentence (Jurafsky & Martin, 2009; Kerkhof, 2020; Manning et al., 2014). However, when sentences depend significantly on each other, conducting NLP-assisted analysis at sentence-level can render it less accurate.

Guidance 2: Avoid sentence-splitting for questions that require several sentences to answer. Instead, split responses by paragraph or consider analyzing at whole-response level.

5.4.3 *Summarization Techniques or Keyword Weighing*

One may not be interested in splitting students' responses at sentence- or paragraph-level. Instead, one may aim to analyze at whole-response level (Ahmed et al., 2022; Condor et al., 2021; Reimers & Gurevych, 2019). In this case, I have the following recommendation.

Guidance 3: One could employ text summarization techniques at the data pre-processing stage to eliminate redundant content. Alternatively, since all keywords or concepts do not have equal importance in evaluating student response, weighing keywords according to pre-established criteria can help to eliminate redundant content.

5.4.4 *Co-References*

When a pronoun has an ambiguous referent, there are challenges with resolving co-references. As an example, consider the following sentences: "The students read the case studies.

They liked them." In these sentence, resolving referents for the pronouns "they" and "them" can be difficult (Lee et al., 2017; Ng & Cardie, 2002; Soon et al., 2001).

Guidance 4: Employ a coreference resolution technique in data pre-processing stage. Coreference resolution is a method for explicitly replacing pronouns with their referent. Neural coreference resolution by Lai et al. (2022) is a viable option. In the aforementioned example, the coreference resolution would yield the following sentences: "The students read the case studies. The students liked the case studies."

5.4.5 *Handling Off-Topic Statements*

Any NLP system for analysis of student-written responses must be flexible enough to accommodate students' tendency to diverge from the main topic (Higgins et al., 2004, 2006). Occasionally, students may write three or four sentences that are tangentially related to the main topic. For example, in the case of ethics case scenario, a question asks student to compare the effects of two different ethical frameworks on chosen stakeholder. Students often begin their answers by giving definitions of ethical frameworks. These definitions are related to the question, thus not entirely non sequiturs. However, these definitions do not help instructors to assess whether students can understand the differences in impacts on the stakeholder when applying the two different ethical frameworks.

Guidance 5: Instructor should deploy a mechanism to overlook the parts of student response that do not significantly contribute to evaluating student understanding of the construct as intended by the question prompt. For example, if the goal is to evaluate whether students can distinguish between impacts on stakeholder when applying two different ethical frameworks, it would be beneficial to ask separate questions, such as one for definitions of the ethical frameworks and another for how applying those ethical frameworks could lead to different impacts on the stakeholder.

5.4.6 *Intermediate Step for Automatic Grading*

The current implementation of my NLP approach only performs thematic analysis of student responses. In practice, the thematic analysis of student responses serves as an intermediate step for automatic grading (Arbogast & Montfort, 2016; Odden et al., 2021). I suggest the following recommendation in addition to the steps I developed in this dissertation study.

Guidance 6: As the first step, one may manually append scores to labeled responses in the example banks as per the grading rubric. Next, using my NLP approaches, one may match new student responses to example responses for scoring. The philosophy behind Guidance 6 is: first, to

determine which of the scored responses in example bank is most similar to the new response, and then assign the score of matched scored response to the new response.

5.4.7 *Matching Methods: ZSC versus kNN*

The kNN method requires an example bank that comprised sentences and their respective labels. In contrast, the ZSC approach uses only labels. One could consider the following when choosing between these two methods.

Guidance 7: If one does not already have example sentences that are collected and labeled, they could start analyzing student responses using the ZSC approach with an *a priori* list of labels only. Otherwise, kNN is preferable because the accuracy of the ZSC approach depends on phrasing of input labels, which can vary. For example, the labels for the topic of "access to heat" can be: "power access", "heating problems", or "energy issues." This variation in phrasing of input labels for expressing the same topic might impact performance of the ZSC approach.

5.4.8 *Growing Example Bank*

Open-ended questions expect students to express their reasoning and understanding in different ways. This implies that there is no gold standard for a reference answer, so multiple reference answers can be correct. Therefore, one should expand the example bank to include all possible ways to answer an open-ended question. A larger example bank would increase the likelihood of NLP approaches to match new student responses with similar scored responses.

Guidance 8: One can expand the example bank by asking students or subject matter experts to develop alternative phrasings for each pre-existing response in the example bank. Alternatively, one could use generative artificial intelligence (GAI) to create alternate phrasings, which a human can audit.

5.4.9 *Comparing Human Scoring with Automatic Grading System Scoring*

A question persistently posed about ASAG systems is whether they can score written responses as accurately as a human grader (Wiser et al., 2016; Zhai et al., 2020, 2020). Therefore, one can develop fidelity of ASAG systems as follows.

Guidance 9: One can determine the level of agreement between scores assigned by a ASAG system and a human grader. This can be achieved by quantifying the percentage of exact or adjacent matches of scores assigned by a ASAG system and a human grader.

5.4.10 *Handling Lexical Diversity*

Students can express an idea in different words. For example, a student can write the following statements about selecting stakeholder: (i) “I chose city council”, (ii) “I selected governing body of city”, (iii) “I selected public administration office”. All of these statements express the same idea: the student chose government as their stakeholder. Therefore, it is impossible to predict all permutations of how a student might express an idea and encode those permutations in dictionaries (Zupanc & Bosnić, 2017).

Guidance 10: Employ an NLP system that does not rely on hand-coded rules. Use state-of-the-art TLLMs in the NLP system because these models can capture the semantics of student responses without any hand-coded rules.

5.4.11 *Not a Train-once-and-forever Solution*

In courses where content and assessment questions do not change over time, setting up an ASAG system can be a one-time task (Roy et al., 2016). Conversely, in many courses where content and questions change regularly, ASAG system cannot be a train-once-and-forever solution.

Guidance 11: Instructors will need to create or update an ASAG system’s example bank for every new assessment.

After offering guidelines on the use of NLP in open-ended assessments, I address common critiques about the acceptance of these NLP tools in teaching and research.

5.5. Addressing Common Critiques of the Use of NLP in Teaching and Learning

I have divided the discussion about critiques of the use of NLP in teaching and learning into the following two sub-sections: section 5.5.1: Educators and their CS Disciplinary Expertise, and section 5.5.2: Conflict between Traditional Qualitative Research and NLP

In section 5.5.1, I first explain challenges for educators in designing or using NLP tools in their classroom practices. Second, from educators’ perspectives, I describe a critique for NLP tools developed solely by an NLP expert. Third, I advocate for collaboration between educators and NLP experts. Lastly, given the recent release of the state-of-the-art NLP technology, I suggest an approach that could be adopted by educators to develop CS expertise.

In section 5.5.2, I give two viewpoints to demonstrate the conflict between traditional qualitative coding and the working principles of NLP tools based on supervised ML. Then, I advocate for the HILNLP approach, based on TLLMs and unsupervised ML algorithms, that was developed by Katz et al. (2021) and used in this study.

5.5.1 *Educators and their CS Disciplinary Expertise*

Generally, many educators may possess limited CS knowledge and skills to design NLP tools for analyzing text generated in their classrooms. Developing that CS expertise often presents a substantial learning curve for educators (Akgun & Greenhow, 2022; X. Chen et al., 2022; Ludvigsen et al., 2018; Shaik et al., 2022; Zawacki-Richter et al., 2019). However, there has been a recent paradigm shift for learning CS skills and the learning curve for educators has been reduced. Now, educators can use chatbots like ChatGPT or Claude for assistance with the CS coding process. This is exemplified in Andrej Karpathy's—former Director of AI at Tesla and OpenAI—quote, "English is the hottest new programming language" (Karpathy, 2023).

Besides designing their own NLP tools, educators can utilize commercially available, user-friendly NLP tools, such as the LIWC (Pennebaker et al., 2007). However, I argue that the use of these NLP tools by educators could be problematic. Existing NLP tools come with standard options optimized for diverse but limited datasets (more limited than those used in typical TLLM training); employing these standard options might result in sub-par performance for a specific educational dataset(Jordan & Mitchell, 2015; Swithenby, 2006; Whitelock & Brasher, 2006). Consequently, when an individual lacks a sufficient understanding of the computerized system they are using, they might be more easily influenced by the results the system presents, whether those results are accurate or inaccurate (Gin, 2023). Notably, this critique could also apply to TLLM-based NLP systems so, implementing these systems do not resolve the issue of required CS disciplinary expertise.

At the same time, NLP experts may develop an automatic grading system, but they may also lack expertise in teaching and learning. Therefore, educators may not trust a system developed solely by NLP experts. Therefore, I would advocate for collaboration between NLP experts and educators to develop NLP tools that will be helpful for accurately and quickly gleaning information from textual data generated in education ecosystem.

5.5.2 *Conflict between Traditional Qualitative Research and NLP*

I suggest there are inherent tensions between traditional qualitative coding and the working principles of NLP tools based on supervised ML. To support this proposition, I give three viewpoints.

The first is about definitions of qualitative codes. When qualitative researchers start qualitative coding, they may lack a clear definition for some constructs of interest, especially when employing certain types of coding such as *in vivo* coding (Charmaz, 2006; Saldaña, 2014; Strauss & Corbin, 1998). Instead, constructs' definitions gradually emerge as data analysis progresses.

Even in the case of *a priori* coding, the definition of codes may evolve as more data is explored (Saldaña, 2021). On the other hand, in the case of NLP tools based on supervised ML, we need predefined codes and a substantial amount of corresponding labeled data. To address this conflict, I suggest the HILNLP, where contributions come from both the machine and a human researcher. Such a system can help to identify codes in the text by effectively running a first pass on the text to group similar responses together. Another benefit of the HILNLP approach is this can reduce time required for text analysis. The human user can utilize these groupings to perform further analysis to fine tune and identify meanings in ways that only a human could (Gin, 2023; Katz et al., 2021b; Kumar & Boulanger, 2021).

The second viewpoint is about the significance of the quantity of available data. In the case of traditional qualitative coding, if new patterns during textual analysis are not emerging, qualitative researchers may determine that "saturation" has been reached (Charmaz, 2006; Lincoln & Guba, 1985). Therefore, a qualitative researcher would not analyze more data since it would not contribute significantly to the results, so limited data availability may not present a problem. On the other hand, in the case of NLP tools based on supervised ML, labeling enough data to train a robust classifier is necessary. This means that more available data is better than limited data. Additionally, in ML, data points that appear infrequently may be considered noise, but from the perspectives of traditional qualitative analysis, the least mentioned ideas might be more interesting (Chang et al., 2021). This is an on-going, inherent dilemma between traditional qualitive coding and NLP tools based on supervise ML due to their conflicting underlying philosophies.

The third viewpoint is about whether NLP can be helpful for analysis of longer interview transcripts. I suggest employing NLP for analyzing interview transcripts does not give the desired results as a qualitative researcher, i.e., a thick description of the phenomenon under investigation. I gave the following two reasons: (i) during the pre-processing, it is difficult to distinguish between when the participant is describing the core phenomenon and when the participant deviates from the topic in interview transcripts, (ii) the analysis of interview transcripts requires an understanding of the implicit meanings of text. In contrast, NLP tools primarily operate on explicit textual meaning.

Next, I describe the contributions of my dissertation study to engineering education research and practice.

5.6. Contribution of the Study

The following are contributions of my dissertation study: it (i) demonstrates automatic analysis of student responses using TLLMs, (ii) provides a scaled method for analyzing large

volumes of textual data, and (iii) offers timely inquiry, given that TLLM-based tools like ChatGPT have recently gained traction in many sectors because of their potential use cases.

5.6.1 *A Method for Automatic Analysis of Open-ended Responses*

Assessments provide evidence for student learning and should inform curriculum design such as course content and instructional method (Pellegrino & Chudowsky, 2001; Pellegrino et al., 2016; Wiggins & McTighe, 2005). This backward curriculum design encourages educators to think like an assessor before developing their courses. The availability of open-ended assessments and the ease of their implementation can influence what knowledge and skills could be covered in a course. Open-ended assessments are time intensive to grade though, so many researchers have proposed automatic grading tools to expedite the grading process (Haller et al., 2022; Kerkhof, 2020). My dissertation study builds off this previous work by using recent TLLMs, which analyze student answers based on their meanings rather than just their words or grammatical structure. Therefore, my dissertation study is a step towards automatic scoring of open-ended assessments, thereby promoting the use of these assessments in engineering courses—a widely recommended assessment practice but one that is less frequently used in engineering classrooms (Brookhart, 2010; Chew et al., 2022).

5.6.2 *A Scalable Method for Qualitative Data Analysis and Promoting Mixed Methods*

Manual qualitative data analysis is a resource intensive process, even for textual data from only a few individuals (Crowston et al., 2012; Tashakkori & Teddlie, 2009). To address this challenge, my dissertation study helps characterize the performance of a HILNLP approach. Such an analysis can help qualitative researchers to work with larger textual data volumes than is currently feasible. Quantitative research methods are considered favorable for their ease to collect information at scale. Though, quantitative information is criticized because it mostly lies in low-dimensional space. On the other hand, qualitative research methods can collect multidimensional information but to collect and analyze qualitative information is resource-intensive. This can be mitigated through using NLP. As an example of the use of NLP in traditional thematic analysis, a researcher would read a portion of the data, identify common themes, and collect excerpts of themes into a codebook. Then, the codebook including excerpts and their labels can be helping classifying themes in the remaining dataset using NLP approaches.

Moreover, I view the HILNLP as a tool for promoting mixed-method research in engineering education. By providing feasible ways for analyzing qualitative data at scale, the HILNLP could identify systematic patterns (quantitative information) in textual data, allowing researchers to integrate both qualitative and quantitative information in their research. For example, a researcher using this approach could quickly identify differences between the number

of themes mentioned by male and female students in their teaching evaluation statements, and make statistical inferences (Chang et al., 2021; Chen et al., 2018).

5.6.3 *A Timely Inquiry*

Recent headlines have featured how GAI tools—most of which have been developed on TLLMs that have the ability to generate language like humans—could have better or worse impacts for teaching and learning (Choi et al., 2023; Dwivedi et al., 2023; Eke, 2023; Kasneci et al., 2023). Given how contemporary this debate is, my dissertation is a timely investigation aimed at informing the engineering education community how the use of TLLMs in classrooms can be beneficial for assessment. My dissertation study investigated one aspect of instructional design: the assessment of students' learning outcomes. I accomplished this through demonstrating how TLLMs can be useful for automatic analysis of student responses based on their underlying meanings rather than words or sentence structure. Next, I describe future work.

5.7. Directions for Future Research

In this section, I give potential avenues for future research that one could pursue by using work completed in my dissertation study.

The first direction for future work is to append scores as per the grading rubric to response labeled by my NLP approaches. When assessing students' learning outcomes with open-ended case scenarios and written responses, instructors typically need to complete two processes. In the first process, the instructor identifies themes in student responses. In the second process, the instructor applies the relevant grading rubrics. My dissertation study was focused on the first process. Therefore, in future work, one can complete the second process: assign scores as per the grading rubric to the labeled responses by my NLP approach.

Second, my work is a proof of concept. Currently, one cannot implement my NLP approach in the FYE classes at Virginia Tech in the upcoming semester because there were limitations due to inconsistent accuracy. However, future work could address these limitations. To help facilitate that process, the labeled responses of my dissertation study can serve as an example bank for labeling (or scoring) new student assignments in the engineering ethics or systems thinking modules. I iterate here the underlying mechanism for scoring student assignment with my NLP approach: once the space of possible answers has been saturated and representatives of each response type are included in an example bank, one can check new responses for semantic similarity using my NLP approach against labeled (scored) sentences in the example bank. Using threshold value of semantic similarity measure, the new response can be matched to an example

response and get assigned a label/score. As an additional future step, following implementation of my NLP approach in classrooms, one can then calculate agreement between scores assigned by the NLP approach and manual grader as an evaluation step. Doing so would be helpful for the engineering education community to know how consistent the NLP-based grading system is compared to manual scoring.

Third, one could study faculty or student perspective about the effectiveness of NLP-based grading tools in future research. Faculty members or students can be asked to rank on a Likert scale the usefulness of the grading tool.

Fourth, one could use the NLP approach in other contexts of teaching and learning, e.g., feedback to students from teachers. With the rise in student enrollment leading to larger classrooms, university instructors face the challenge of providing timely and meaningful feedback (Edalati, 2020; Hirschberg & Manning, 2015; Shaik et al., 2022). For instance, if students only identify a subset of potentially impacted stakeholders in the stakeholder identification question, the NLP-based system could provide a prompt highlighting other stakeholders for them to consider.

Fifth, I was unable to systematically assess the NLP approach's handling of text from students for whom English is a second language. The current dataset contained minimal examples of non-native English text, preventing a thorough evaluation of this aspect in my study. Further research with more diverse language speakers and datasets is needed to evaluate the approach's effectiveness for grading responses from non-native English speakers.

Sixth, the use and growth of NLP in engineering education may be impeded by the availability of public datasets. Conducting experiments with standard public datasets enhances NLP models' comparability, allowing them to be improved. Some examples of standard public datasets in ASAG are Mohler, SemEval, and Beetle (Kerkhof, 2020). However, I acknowledge that education datasets are often restricted in their distribution due to FERPA laws and IRB regulations (Case et al., 2022; Paretti et al., 2023). As such, these types of datasets are often not publicly available. In the future, we should develop a benchmark dataset specific to automatic grading in engineering education. Consequently, tasks such as extending existing datasets, data augmentation, and creating synthetic data through GAI models may be future avenues for NLP-related research in engineering education.

Lastly, future work may focus on developing a user-friendly application programming interface (API). Currently, my NLP approaches require manual adjustments directly in the

codebase, rather than operating through a convenient API. A user-friendly API could enable the broader research community to utilize this NLP approach without CS expertise.

In conclusion, I suggest the abovementioned directions for future research will help provide the engineering research community with accessible tools to assist their sensemaking of large-scale qualitative data.

5.8. Final Thoughts

The aim of this study was to introduce an NLP-based approach to a research field where it holds potential yet remains underutilized. This study can encourage engineering education researchers to utilize these NLP methods that may be helpful in analyzing the vast textual data generated in engineering education, thereby reducing the number of missed opportunities to glean information. The approach I used in this dissertation is only the tip of what is possible with newer GAI models such as GPT-4 from Open AI, Claude from Anthropogenic, and Bard from Google. As such, I encourage scholars to explore potential applications of GAI in pedagogical, ethical, and economic dimensions of education practice and research. First, an example of pedagogical dimension is how teacher can use GAI in providing feedback to students on their assessments. Second, an example of ethical dimension is who is best equipped to educate and develop the individual student in the current digital age—multinational corporation or a public education system. There are concerns that significant changes related to GAI will arrive, from outside the formal public education system, through organizations such as LinkedIn, lynda.com, Amazon, or Coursera (Chiu et al., 2023; Xu & Ouyang, 2022). Third, an example of economic dimension is potential success of GAI in reducing some of the financial cost of teaching and learning but at the expense of other unanticipated consequences for public schools, colleges, and universities.

References

Ahmad, M., Junus, K., & Santoso, H. B. (2022). Automatic content analysis of asynchronous discussion forum transcripts: A systematic literature review. *Education and Information Technologies, 27*(8), 11355–11410. https://doi.org/10.1007/s10639-022-11065-w

Ahmed, A., Joorabchi, A., & Hayes, M. J. (2022). On the Application of Sentence Transformers to Automatic Short Answer Grading in Blended Assessment. *2022 33rd Irish Signals and Systems Conference (ISSC)*, 1–6. https://doi.org/10.1109/ISSC55427.2022.9826194

Akgun, S., & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics, 2*(3), 431–440. https://doi.org/10.1007/s43681-021-00096-7

Aldea, A. I., Haller, S. M., & Luttikhuis, M. G. (2020). *Towards Grading Automation of Open Questions Using Machine Learning*. 584–593.

Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020). Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In A. El Moataz, D. Mammass, A. Mansouri, & F. Nouboud (Eds.), *Image and Signal Processing* (pp. 317–325). Springer International Publishing. https://doi.org/10.1007/978-3-030-51935-3_34

Almazova, N. A., Hallstrom, J. O., Fowler, M., Hollingsworth, J. E., Sitaraman, M., Kraemer, E. T., & Washington, G. (2021). Automated Analysis of Student Verbalizations in Online Learning Environments. *52nd ACM Technical Symposium on Computer Science Education, SIGCSE 2021, March 13, 2021 - March 20, 2021*, 1272. https://doi.org/10.1145/3408877.3439660

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, *46*(3), 175–185. https://doi.org/10.2307/2685209

Anakok, I., Woods, J., Huerta, M., Schoepf, J., Murzi, H., & Katz, A. (2022). Students' Feedback About Their Experiences in EPICS Using Natural Language Processing. *2022 IEEE Frontiers in Education Conference (FIE)*, 1–9. https://doi.org/10.1109/FIE56618.2022.9962557

Anderson, M. A. (2016). Pedagogical Support for Responsible Conduct of Research Training. *Hastings Center Report, 46*(1), 18–25. https://doi.org/10.1002/hast.533

Arbogast, C. A., & Montfort, D. (2016, June 26). *Applying Natural Language Processing Techniques to an Assessment of Student Conceptual Understanding*. 2016 ASEE Annual Conference & Exposition. https://peer.asee.org/applying-natural-language-processing-techniques-to-an-assessment-of-student-conceptual-understanding

Ariely, M., Nazaretsky, T., & Alexandron, G. (2023). Machine Learning and Hebrew NLP for Automated Assessment of Open-Ended Questions in Biology. *International Journal of Artificial Intelligence in Education*, *33*(1), 1–34. https://doi.org/10.1007/s40593-021-00283-x

Arnold, R. D., & Wade, J. P. (2017). A Complete Set of Systems Thinking Skills. *INSIGHT*, *20*(3), 9–17. https://doi.org/10.1002/inst.12159

Assent, I. (2012). Clustering high dimensional data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(4), 340–350.

Bai, X., & Stede, M. (2022). A Survey of Current Machine Learning Approaches to Student Free-Text Evaluation for Intelligent Tutoring. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-022-00323-0

Bartl, M., Nissim, M., & Gatt, A. (2020). *Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias* (arXiv:2010.14534). arXiv. https://doi.org/10.48550/arXiv.2010.14534

Becker, J., Sior, E., Hoy, J., & Kahanda, I. (2019). Board 11: Predicting At-Risk Students in a Circuit Analysis Course Using Supervised Machine Learning. *2019 ASEE Annual Conference & Exposition Proceedings*, 32185. https://doi.org/10.18260/1-2--32185

Bellman, R. (2017). *High-dimensional Outlier Detection: The Subspace Method*. https://www.semanticscholar.org/paper/High-dimensional-Outlier-Detection%3A-the-Subspace-Bellman/68875270c89eac81d3b9fc47095e62337634bb49

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 🦜. 610–623.

Berdanier, C. G. P., Baker, E., Wang, W., & McComb, C. (2018). Opportunities for Natural Language Processing in Qualitative Engineering Education Research: Two Examples. *2018 IEEE Frontiers in Education Conference (FIE)*, 1–6. https://doi.org/10.1109/FIE.2018.8658747

Berdanier, C. G. P., McComb, C. M., & Zhu, W. (2020). Natural Language Processing for Theoretical Framework Selection in Engineering Education Research. *2020 IEEE Frontiers in Education Conference (FIE)*, 1–7. https://doi.org/10.1109/FIE44824.2020.9274115

Berry, R. M., Borenstein, J., & Butera, R. J. (2013). Contentious Problems in Bioscience and Biotechnology: A Pilot Study of an Approach to Ethics Education. *Science and Engineering Ethics*, *19*(2), 653–668. https://doi.org/10.1007/s11948-012-9359-6

Bhaduri, S. (2018). *NLP in Engineering Education-Demonstrating the use of Natural Language Processing Techniques for Use in Engineering Education Classrooms and Research*. Virginia Tech.

Bhaduri, S., & Roy, T. (2017, June 24). *Demonstrating Use of Natural Language Processing to Compare College of Engineering Mission Statements*. 2017 ASEE Annual Conference & Exposition. https://peer.asee.org/demonstrating-use-of-natural-language-processing-to-compare-college-of-engineering-mission-statements

Bielefeldt, A. R., Polmear, M., Knight, D., Swan, C., & Canney, N. (2018). Intersections between Engineering Ethics and Diversity Issues in Engineering Education. *Journal of Professional Issues in Engineering Education and Practice*, *144*(2), 04017017. https://doi.org/10.1061/(ASCE)EI.1943-5541.0000360

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer; WorldCat.org. http://catdir.loc.gov/catdir/enhancements/fy0818/2006922522-t.html

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Blessing, G., Azeta, A., Misra, S., Chigozie, F., & Ahuja, R. (2021). A Machine Learning Prediction of Automatic Text Based Assessment for Open and Distance Learning: A Review. In A. Abraham, M. Panda, S. Pradhan, L. Garcia-Hernandez, & K. Ma (Eds.), *Innovations in Bio-Inspired Computing and Applications* (pp. 369–380). Springer International Publishing. https://doi.org/10.1007/978-3-030-49339-4_38

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). *Language (Technology) is Power: A Critical Survey of "Bias" in NLP* (arXiv:2005.14050). arXiv. https://doi.org/10.48550/arXiv.2005.14050

Borenstein, J., Drake, M., Kirkman, R., & Swann, J. (2008). *The Test of Ethical Sensitivity in Science and Engineering (TESSE): A Discipline Specific Assessment Tool for Awareness of Ethical Issues*. 13.1270.1-13.1270.10. https://peer.asee.org/the-test-of-ethical-sensitivity-in-science-and-engineering-tesse-a-discipline-specific-assessment-tool-for-awareness-of-ethical-issues

Borrego, M., & Bernhard, J. (2011). The Emergence of Engineering Education Research as an Internationally Connected Field of Inquiry. *Journal of Engineering Education*, *100*(1), 14–47. https://doi.org/10.1002/j.2168-9830.2011.tb00003.x

Brandstädter, K., Harms, U., & Großschedl, J. (2012). Assessing System Thinking Through Different Concept-Mapping Practices. *International Journal of Science Education*, *34*(14), 2147–2170. https://doi.org/10.1080/09500693.2012.716549

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. Ascd.

Bulut, O., MacIntosh, A., & Walsh, C. (2022). Leveraging Natural Language Processing for Quality Assurance of a Situational Judgement Test. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium* (pp. 84–88). Springer International Publishing. https://doi.org/10.1007/978-3-031-11647-6_14

Burrows, S., Gurevych, I., & Stein, B. (2015). The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Butcher, P. G., & Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. *Computers & Education*, *55*(2), 489–499. https://doi.org/10.1016/j.compedu.2010.02.012

Butt, A. A., Anwar, S., & Menekse, M. (2022, August 23). *WIP: Role of digital nudging strategies on STEM students' application engagement*. 2022 ASEE Annual Conference &

Exposition. https://peer.asee.org/wip-role-of-digital-nudging-strategies-on-stem-students-application-engagement

Caliskan, A. (2021, May 10). Detecting and mitigating bias in natural language processing. *Brookings*. https://www.brookings.edu/research/detecting-and-mitigating-bias-in-natural-language-processing/

Camelia, F., & Ferris, T. L. J. (2018). Validation Studies of a Questionnaire Developed to Measure Students' Engagement With Systems Thinking. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *48*(4), 574–585. https://doi.org/10.1109/TSMC.2016.2607224

Camelia, F., Ferris, T. L. J., & Cropley, D. H. (2018). Development and Initial Validation of an Instrument to Measure Students' Learning About Systems Thinking: The Affective Domain. *IEEE Systems Journal*, *12*(1), 115–124. https://doi.org/10.1109/JSYST.2015.2488022

Camus, L., & Filighera, A. (2020). Investigating Transformers for Automatic Short Answer Grading. *Artificial Intelligence in Education*, *12164*, 43–48. https://doi.org/10.1007/978-3-030-52240-7_8

Capuano, N., Caballé, S., Conesa, J., & Greco, A. (2021). Attention-based hierarchical recurrent neural networks for MOOC forum posts analysis. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 9977–9989. https://doi.org/10.1007/s12652-020-02747-9

Caratozzolo, P., Rodriguez-Ruiz, J., & Alvarez-Delgado, A. (2022). Natural Language Processing for Learning Assessment in STEM. *2022 IEEE Global Engineering Education Conference (EDUCON)*, 1549–1554. https://doi.org/10.1109/EDUCON52537.2022.9766717

Case, J. M., & Light, G. (2011). Emerging Research Methodologies in Engineering Education Research. *Journal of Engineering Education*, *100*(1), 186–210. https://doi.org/10.1002/j.2168-9830.2011.tb00008.x

Case, J., Matusovich, H., Paretti, M., Sochacka, N., & Walther, J. (2022, August 23). *Changing the Paradigm: Developing a Framework for Secondary Analysis of EER Qualitative Datasets*. 2022 ASEE Annual Conference & Exposition. https://peer.asee.org/changing-the-paradigm-developing-a-framework-for-secondary-analysis-of-eer-qualitative-datasets

Castelle, K. M., & Jaradat, R. M. (2016). Development of an Instrument to Assess Capacity for Systems Thinking. *Procedia Computer Science*, *95*, 80–86. https://doi.org/10.1016/j.procs.2016.09.296

Chang, T., DeJonckheere, M., Vydiswaran, V. G. V., Li, J., Buis, L. R., & Guetterman, T. C. (2021). Accelerating Mixed Methods Research With Natural Language Processing of Big Text Data. *Journal of Mixed Methods Research*, *15*(3), 398–412. https://doi.org/10.1177/15586898211021196

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

Chauhan, U., & Shah, A. (2022). Topic Modeling Using Latent Dirichlet allocation: A Survey. *ACM Computing Surveys*, *54*(7), 1–35. https://doi.org/10.1145/3462478

Chen, N.-C., Drouhard, M., Kocielnik, R., Suh, J., & Aragon, C. R. (2018). Using Machine Learning to Support Qualitative Coding in Social Science: Shifting the Focus to Ambiguity. *ACM Transactions on Interactive Intelligent Systems*, *8*(2), 9:1-9:20. https://doi.org/10.1145/3185515

Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C. (2022). Two Decades of Artificial Intelligence in Education: Contributors, Collaborations, Research Topics, Challenges, and Future Directions. *Educational Technology & Society*, *25*(1), 28–47.

Chew, K. J., Ross, A., Katz, A., & Matusovich, H. M. (2022). Defining Assessment: Foundation Knowledge Toward Exploring Engineering Faculty's Assessment Mental Models. *2022 IEEE Frontiers in Education Conference, FIE 2022, October 8, 2022 - October 11, 2022*, *2022-October*. https://doi.org/10.1109/FIE56618.2022.9962667

Chin, C., & Osborne, J. (2008). Students' questions: A potential resource for teaching and learning science. *Studies in Science Education*, *44*(1), 1–39. https://doi.org/10.1080/03057260701828101

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *4*, 100118. https://doi.org/10.1016/j.caeai.2022.100118

Choi, E. P. H., Lee, J. J., Ho, M.-H., Kwok, J. Y. Y., & Lok, K. Y. W. (2023). Chatting or cheating? The impacts of ChatGPT and other artificial intelligence language models on nurse education. *Nurse Education Today, 125*, 105796. https://doi.org/10.1016/j.nedt.2023.105796

Clarke, V., & Braun, V. (2017). Thematic analysis. *The Journal of Positive Psychology, 12*(3), 297–298. https://doi.org/10.1080/17439760.2016.1262613

Colby, A., Kohlberg, L., Gibbs, J., Lieberman, M., Fischer, K., & Saltzstein, H. D. (1983). A Longitudinal Study of Moral Judgment. *Monographs of the Society for Research in Child Development*, *48*(1/2), 1–124. https://doi.org/10.2307/1165935

Condor, A., Litster, M., & Pardos, Z. (2021). Automatic Short Answer Grading with SBERT on Out-of-Sample Questions. In *International Educational Data Mining Society*. International Educational Data Mining Society. https://eric.ed.gov/?id=ED615495

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.*, *43*, 1241.

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Cronin, A., Intepe, G., Shearman, D., & Sneyd, A. (2019). Analysis using natural language processing of feedback data from two mathematics support centres. *International Journal of Mathematical Education in Science and Technology*, *50*(7), 1087–1103. https://doi.org/10.1080/0020739X.2019.1656831

Crossley, S. A., Bradfield, F., & Bustamante, A. (2019). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research*, *11*(2), 251–270.

Crowston, K., Allen, E. E., & Heckman, R. (2012). Using natural language processing technology for qualitative data analysis. *International Journal of Social Research Methodology*, *15*(6), 523–543. https://doi.org/10.1080/13645579.2011.625764

Davis, K. A., Grote, D., Mahmoudi, H., Perry, L., Ghaffarzadegan, N., Grohs, J., Hosseinichimeh, N., Knight, D. B., & Triantis, K. (2023). Comparing Self-Report Assessments

and Scenario-Based Assessments of Systems Thinking Competence. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-023-10027-2

Davis, K., Ghaffarzadegan, N., Grohs, J., Grote, D., Hosseinichimeh, N., Knight, D., Mahmoudi, H., & Triantis, K. (2020). The Lake Urmia vignette: A tool to assess understanding of complexity in socio-environmental systems. *System Dynamics Review*, *36*(2), 191–222. https://doi.org/10.1002/sdr.1659

Davis, M., & Riley, K. (2008). Ethics across the graduate engineering curriculum: An experiment in teaching and assessment. *Teaching Ethics*, *9*(1), 25–42. https://doi.org/10.5840/tej20089115

de Araujo, A., Papadopoulos, P. M., McKenney, S., & de Jong, T. (2023). Automated coding of student chats, a trans-topic and language approach. *Computers and Education: Artificial Intelligence, 4*, 100123. https://doi.org/10.1016/j.caeai.2023.100123

Degen, C. M., Muci-Küchler, K. H., Bedillion, M. D., Huang, S., & Ellingsen, M. (2018). *Measuring the Impact of a New Mechanical Engineering Sophomore Design Course on Students' Systems Thinking Skills*. ASME 2018 International Mechanical Engineering Congress and Exposition. https://doi.org/10.1115/IMECE2018-87624

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., Luccioni, A. S., Smith, N. A., DeCario, N., & Buchanan, W. (2022). *Measuring the carbon intensity of ai in cloud instances*. 1877–1894.

Dugan, K. E., Mosyjowski, E. A., Daly, S. R., & Lattuca, L. R. (2021, July 26). *Systems Thinking Assessments: Approaches That Examine Engagement in Systems Thinking*. 2021 ASEE Virtual Annual Conference Content Access. https://peer.asee.org/systems-thinking-assessments-approaches-that-examine-engagement-in-systems-thinking

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and

implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Edalati, M. (2020). *The Potential of Machine Learning and NLP for Handling Students' Feedback (A Short Survey)* (arXiv:2011.05806). arXiv. https://doi.org/10.48550/arXiv.2011.05806

Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology*, *13*, 100060. https://doi.org/10.1016/j.jrt.2023.100060

Emerson, A., Min, W., Rowe, J., Azevedo, R., & Lester, J. (2023). Multimodal Predictive Student Modeling with Multi-Task Transfer Learning. *13th International Conference on Learning Analytics and Knowledge: Towards Trustworthy Learning Analytics, LAK 2023, March 13, 2023 - March 17, 2023*, 333–344. https://doi.org/10.1145/3576050.3576101

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479. https://doi.org/10.1613/jair.1523

Ermer, G. E. (2004). Using case studies to teach engineering ethics and professionalism. *Teaching Ethics*, *4*(2), 33–40.

Fan, X., Luo, W., Menekse, M., Litman, D., & Wang, J. (2015). CourseMIRROR: Enhancing Large Classroom Instructor-Student Interactions via Mobile Interfaces and Natural Language Processing. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 1473–1478. https://doi.org/10.1145/2702613.2732853

Fan, X., Luo, W., Menekse, M., Litman, D., & Wang, J. (2017). *Scaling reflection prompts in large classrooms via mobile interfaces and natural language processing*. 363–374.

Fellbaum, C. (2010). WordNet. In *Theory and applications of ontology: Computer applications* (pp. 231–243). Springer.

Finelli, C. J., Holsapple, M. A., Ra, E., Bielby, R. M., Burt, B. A., Carpenter, D. D., Harding, T. S., & Sutkus, J. A. (2012). An Assessment of Engineering Students' Curricular and Co-Curricular Experiences and Their Ethical Development. *Journal of Engineering Education*, *101*(3), 469–494. https://doi.org/10.1002/j.2168-9830.2012.tb00058.x

Firth, J. (1968). Descriptive linguistics and the study of English. *World Englishes: Critical Concepts in Linguistics. Ed. by K. Bolton and B. Kachru*, *3*, 203–217.

Forsyth, S., & Mavridis, N. (2021). Short Answer Marking Agent for GCSE Computer Science. *2021 IEEE World Conference on Engineering Education (EDUNINE)*, 1–6. https://doi.org/10.1109/EDUNINE51952.2021.9429163

Frank, M. (2010). Assessing the interest for systems engineering positions and other engineering positions' required capacity for engineering systems thinking (CEST). *Systems Engineering*, *13*(2), 161–174. https://doi.org/10.1002/sys.20140

Galhardi, L. B., & Brancher, J. D. (2018). Machine Learning Approach for Automatic Short Answer Grading: A Systematic Review. In G. R. Simari, E. Fermé, F. Gutiérrez Segura, & J. A. Rodríguez Melquiades (Eds.), *Advances in Artificial Intelligence—IBERAMIA 2018* (pp. 380–391). Springer International Publishing. https://doi.org/10.1007/978-3-030-03928-8_31

Gamieldien, Y. (2023). *Innovating the Study of Self-Regulated Learning: An Exploration through NLP, Generative AI, and LLMs*. https://vtechworks.lib.vt.edu/handle/10919/116274

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

Gamieldien, Y., McCord, R., & Katz, A. (2023). *Utilizing Natural Language Processing to Examine Self-Reflections in Self-Regulated Learning* (SSRN Scholarly Paper 4487795). https://doi.org/10.2139/ssrn.4487795

Ganesh, A., Scribner, H., Singh, J., Goodman, K., Hertzberg, J., & Kann, K. (2022). Response Construct Tagging: NLP-Aided Assessment for Engineering Education. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 250–261. https://doi.org/10.18653/v1/2022.bea-1.29

Gao, F., Jiang, H., Blum, M., Lu, J., Liu, D., Jiang, Y., & Li, I. (2023). *Large Language Models on Wikipedia-Style Survey Generation: An Evaluation in NLP Concepts* (arXiv:2308.10410). arXiv. https://doi.org/10.48550/arXiv.2308.10410

Geiger, J. M., Goodhew, L. M., & Odden, T. O. (2022). *Developing a natural language processing approach for analyzing student ideas in calculus-based introductory physics*. 2022 Physics Education Research Conference Proceedings.

Gin, B. C. (2023). Evolving natural language processing towards a subjectivist inductive paradigm. *Medical Education*, *n/a*(n/a), 1–3. https://doi.org/10.1111/medu.15024

Gonen, H., & Goldberg, Y. (2019). *Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them* (arXiv:1903.03862). arXiv. https://doi.org/10.48550/arXiv.1903.03862

Grohs, J. R., Kirk, G. R., Soledad, M. M., & Knight, D. B. (2018a). Assessing systems thinking: A tool to measure complex reasoning through ill-structured problems. *Thinking Skills and Creativity*, *28*, 110–130. https://doi.org/10.1016/j.tsc.2018.03.003

Grohs, J. R., Kirk, G. R., Soledad, M. M., & Knight, D. B. (2018b). Assessing systems thinking: A tool to measure complex reasoning through ill-structured problems. *Thinking Skills and Creativity*, *28*, 110–130. https://doi.org/10.1016/j.tsc.2018.03.003

Guo, M., Dai, Z., Vrandečić, D., & Al-Rfou, R. (2020). *Wiki-40b: Multilingual language model dataset*. 2440–2452.

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: Prospects and limitations. *CBE Life Sciences Education*, *10*(4), 379–393. https://doi.org/10.1187/cbe.11-08-0081

Hadgraft, R. G., Carew, A. L., Therese, S. A., & Blundell, D. L. (2008). *Teaching and assessing systems thinking in engineering*. 230–235.

Haller, S., Aldea, A., Seifert, C., & Strisciuglio, N. (2022). *Survey on Automated Short Answer Grading with Deep Learning: From Word Embeddings to Transformers* (arXiv:2204.03503). arXiv. http://arxiv.org/abs/2204.03503

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. *CBE Life Sciences Education*, *11*(3), 283–293. https://doi.org/10.1187/cbe.11-08-0084

Hechenbichler, K., & Schliep, K. (2004). *Weighted k-nearest-neighbor techniques and ordinal classification*.

Heilman, M., & Madnani, N. (2013a). *ETS: Domain adaptation and stacking for short answer scoring*. 275–279.

Heilman, M., & Madnani, N. (2013b). *Henry-core: Domain adaptation and stacking for text similarity*. 96–102.

Helton-Fauth, W., Gaddis, B., Scott, G., Mumford, M., Devenport, L., Connelly, S., & Brown, R. (2003). A New Approach to Assessing Ethical Conduct in Scientific Work. *Accountability in Research*, *10*(4), 205–228. https://doi.org/10.1080/714906104

Hess, J. L., Kerr, A. J., Lin, A., & Chung, A. (2023). A Systematic Review of the 2016 National Academy of Engineering Exemplary Ethics Programs: Revisions to a Coding Framework. *Science and Engineering Ethics*, *29*(6), 36. https://doi.org/10.1007/s11948-023-00456-y

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, *12*(2), 145–159.

Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004). *Evaluating multiple aspects of coherence in student essays*. 185–192.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hrin, T. N., Fahmy, A. F. M., Segedinac, M. D., & Milenković, D. D. (2016). Systemic Synthesis Questions [SSynQs] as Tools to Help Students to Build Their Cognitive Structures in a Systemic Manner. *Research in Science Education*, *46*(4), 525–546. https://doi.org/10.1007/s11165-015-9470-1

Hu, M., & Shealy, T. (2018, June 23). *Methods for Measuring Systems Thinking: Differences Between Student Self-assessment, Concept Map Scores, and Cortical Activation During Tasks About Sustainability*. 2018 ASEE Annual Conference & Exposition.

https://peer.asee.org/methods-for-measuring-systems-thinking-differences-between-student-self-assessment-concept-map-scores-and-cortical-activation-during-tasks-about-sustainability

Jaradat, R. (2014). An Instrument to Assess Individual Capacity for System Thinking. *Engineering Management & Systems Engineering Theses & Dissertations*. https://doi.org/10.25777/wzh1-2563

Jaradat, R., Hamilton, M. A., Dayarathna, V. L., Karam, S., Jones, P., Wall, E. S., Amrani, S. E., & Hsu, G. S. E. (2019, June 15). *Measuring Individuals' Systems Thinking Skills through the Development of an Immersive Virtual Reality Complex System Scenarios*. 2019 ASEE Annual Conference & Exposition. https://peer.asee.org/measuring-individuals-systems-thinking-skills-through-the-development-of-an-immersive-virtual-reality-complex-system-scenarios

Jayakodi, K., Bandara, M., & Perera, I. (2015). An automatic classifier for exam questions in Engineering: A process for Bloom's taxonomy. *2015 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 195–202. https://doi.org/10.1109/TALE.2015.7386043

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J., Urban-Lurain, M., Doherty, J. H., & Haudek, K. C. (2021). Comparison of Machine Learning Performance Using Analytic and Holistic Coding Approaches Across Constructed Response Assessments Aligned to a Science Learning Progression. *Journal of Science Education and Technology*, *30*(2), 150–167. https://doi.org/10.1007/s10956-020-09858-0

Jiang, R., Gouvea, J., Hammer, D., Miller, E., & Aeron, S. (2020). *Automatic coding of students' writing via Contrastive Representation Learning in the Wasserstein space* (arXiv:2011.13384). arXiv. https://doi.org/10.48550/arXiv.2011.13384

Johnson, M. (1994). *Moral Imagination: Implications of Cognitive Science for Ethics*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/M/bo3684141.html

Johri, A., Katz, A. S., Qadir, J., & Hingle, A. (2023). Generative artificial intelligence and engineering education. *Journal of Engineering Education*, *n/a*(n/a). https://doi.org/10.1002/jee.20537

Jolliffe, I. T. (2002). *Principal component analysis for special types of data*. Springer.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Jordan, S. (2009). Assessment for Learning: Pushing the Boundaries of Computer-Based Assessment. *Practitioner Research in Higher Education, 3*(1), 11–19.

Junaid, S., Kovacs, H., Martin, D. A., & Serreau, Y. (2021). *What is the role of ethics in accreditation guidelines for engineering programmes in Europe? CONF*.

Jurafsky, D., & Martin, J. H. (2009a). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

Jurafsky, D., & Martin, J. H. (2009b). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Second edition). Pearson Prentice Hall. http://catdir.loc.gov/catdir/toc/ecip0812/2008010335.html

Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2021). *AMMUS: A Survey of Transformer-based Pretrained Models in Natural Language Processing* (arXiv:2108.05542). arXiv. https://doi.org/10.48550/arXiv.2108.05542

Karpathy, A. (2023). *English is the hottest new coding language—What is prompt engineering? | LinkedIn*. https://www.linkedin.com/pulse/english-hottest-new-coding-language-what-prompt-etienne-oosthuysen/

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Katz, A., Norris, M., Alsharif, A. M., Klopfer, M. D., Knight, D. B., & Grohs, J. R. (2021a). *Using Natural Language Processing to Facilitate Student Feedback Analysis*. 2021 ASEE Virtual Annual Conference Content Access.

Katz, A., Norris, M., Alsharif, A. M., Klopfer, M. D., Knight, D. B., & Grohs, J. R. (2021b, July 26). *Using Natural Language Processing to Facilitate Student Feedback Analysis*. 2021

ASEE Virtual Annual Conference Content Access. https://peer.asee.org/using-natural-language-processing-to-facilitate-student-feedback-analysis

Kerkhof, R. (2020a). *Natural Language Processing for scoring open-ended questions: A systematic review*.

Kerkhof, R. (2020b). *Natural language processing for scoring open-ended questions: A systematic review*. The 33rd Twente Student Conference on IT, Netherlands.

Keynan, A., Assaraf, O. B.-Z., & Goldman, D. (2014). The repertory grid as a tool for evaluating the development of students' ecological system thinking abilities. *STUDIES IN EDUCATIONAL EVALUATION*, *41*, 90–105. https://doi.org/10.1016/j.stueduc.2013.09.012

Kim, D., & Bairaktarova, D. (2023, June 25). *Assessment Instruments for Engineering Ethics Education: A Review and Opportunities*. 2023 ASEE Annual Conference & Exposition. https://peer.asee.org/assessment-instruments-for-engineering-ethics-education-a-review-and-opportunities

Klein, D., & Manning, C. D. (2003). *Accurate unlexicalized parsing*. 423–430.

Kordova, S., & Frank, M. (2018). Systems thinking as an engineering language. *American Journal of Systems Science*, *6*(1), 16–28.

Kumar, V. S., & Boulanger, D. (2021). Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined? *International Journal of Artificial Intelligence in Education*, *31*(3), 538–584. https://doi.org/10.1007/s40593-020-00211-5

Lai, T. M., Bui, T., & Kim, D. S. (2022). End-To-End Neural Coreference Resolution Revisited: A Simple Yet Effective Baseline. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8147–8151. https://doi.org/10.1109/ICASSP43922.2022.9746254

Lavi, R., Dori, Y. J., & Dori, D. (2021). Assessing Novelty and Systems Thinking in Conceptual Models of Technological Systems. *IEEE Transactions on Education*, *64*(2), 155–162. https://doi.org/10.1109/TE.2020.3022238

Lavi, R., Dori, Y. J., Wengrowicz, N., & Dori, D. (2020). Model-Based Systems Thinking: Assessing Engineering Student Teams. *IEEE Transactions on Education*, *63*(1), 39–47. https://doi.org/10.1109/TE.2019.2948807

Le, Q., & Mikolov, T. (2014). *Distributed representations of sentences and documents*. 1188–1196.

Lee, H.-S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education*, *103*(3), 590–622. https://doi.org/10.1002/sce.21504

Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end neural coreference resolution. *arXiv Preprint arXiv:1707.07045*.

Levesque, H. J., Davis, E., & Morgenstern, L. (2012). The Winograd schema challenge. *KR*, *2012*, 13th.

Lewis, E. J., Ludwig, P. M., Nagel, J., & Ames, A. (2019). Student ethical reasoning confidence pre/post an innovative makerspace course: A survey of ethical reasoning. *Nurse Education Today*, *75*, 75–79. https://doi.org/10.1016/j.nedt.2019.01.011

Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2017). Enhancing Topic Modeling for Short Texts with Auxiliary Word Embeddings. *ACM Transactions on Information Systems*, *36*(2), 11:1-11:30. https://doi.org/10.1145/3091108

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. sage.

Liu, S., Liu, S., Liu, Z., Peng, X., & Yang, Z. (2022). Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement. *Computers & Education*, *181*, 104461. https://doi.org/10.1016/j.compedu.2022.104461

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Roberta: A robustly optimized BERT pretraining approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Ludvigsen, S., Cress, U., Rosé, C. P., Law, N., & Stahl, G. (2018). Developing understanding beyond the given knowledge and new methodologies for analyses in CSCL. *International Journal of Computer-Supported Collaborative Learning*, *13*, 359–364.

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. *Behavior Research Methods*, *44*(3), 608–621. https://doi.org/10.3758/s13428-012-0211-3

Magooda, A. E., Zahran, M., Rashwan, M., Raafat, H., & Fayek, M. (2016). *Vector based techniques for short answer grading*. The twenty-ninth international flairs conference.

Magooda, A., Litman, D., Ashraf, A., & Menekse, M. (2022). *Improving the Quality of Students' Written Reflections Using Natural Language Processing: Model Design and Classroom Evaluation*. 519–525.

Manning, C. D., & Schütze, Hinrich. (1999). *Foundations of statistical natural language processing*. MIT Press; WorldCat.org. http://digitool.hbz-nrw.de:1801/webclient/DeliveryManager?pid=3277006

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. https://doi.org/10.3115/v1/P14-5010

Martin, D. A., Conlon, E., & Bowe, B. (2021). Using case studies in engineering ethics education: The case for immersive scenarios through stakeholder engagement and real life data. *Australasian Journal of Engineering Education*, *26*(1), 47–63. https://doi.org/10.1080/22054952.2021.1914297

McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (arXiv:1802.03426). arXiv. https://doi.org/10.48550/arXiv.1802.03426

Meilinda, M., Rustaman, N., Firman, H., & Tjasyono, B. (2018). Development and validation of climate change system thinking instrument (CCSTI) for measuring system thinking on climate change content. *Journal of Physics: Conference Series*, *1013*, 012046. https://doi.org/10.1088/1742-6596/1013/1/012046

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*.

https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative data analysis: A methods sourcebook.*

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Moharreri, K., Ha, M., & Nehm, R. H. (2014). EvoGrader: An online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, *7*(1), 15. https://doi.org/10.1186/s12052-014-0015-2

Mu, Y., Reheman, A., Cao, Z., Fan, Y., Li, B., Li, Y., Xiao, T., Zhang, C., & Zhu, J. (2023). *Augmenting Large Language Model Translators via Translation Memories* (arXiv:2305.17367). arXiv. https://doi.org/10.48550/arXiv.2305.17367

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification*, *31*, 274–295.

Nanda, G., Wei, S., Katz, A., Brinton, C., & Ohland, M. (2022, August 23). *Work-in-Progress: Using Latent Dirichlet Allocation to uncover themes in student comments from peer evaluations of teamwork.* 2022 ASEE Annual Conference & Exposition. https://peer.asee.org/work-in-progress-using-latent-dirichlet-allocation-to-uncover-themes-in-student-comments-from-peer-evaluations-of-teamwork

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming Biology Assessment with Machine Learning: Automated Scoring of Written Evolutionary Explanations. *Journal of Science Education and Technology*, *21*(1), 183–196. https://doi.org/10.1007/s10956-011-9300-9

Nehm, R. H., & Haertig, H. (2012). Human vs. Computer Diagnosis of Students' Natural Selection Knowledge: Testing the Efficacy of Text Analytic Software. *Journal of Science Education and Technology*, *21*(1), 56–73. https://doi.org/10.1007/s10956-011-9282-7

Ng, V., & Cardie, C. (2002). *Improving machine learning approaches to coreference resolution*. 104–111.

Oakes, W. C., Titus, C., Zoltowski, C. B., May, J. L., & Huyck, M. (2011). *The Creation of Tools for Assessing Ethical Awareness in Diverse Multi-Disciplinary Programs*. 22.1436.1-

22.1436.17. https://peer.asee.org/the-creation-of-tools-for-assessing-ethical-awareness-in-diverse-multi-disciplinary-programs

Odden, T. O. B., Marin, A., & Rudolph, J. L. (2021). How has Science Education changed over the last 100 years? An analysis using natural language processing. *Science Education*, *105*(4), 653–680. https://doi.org/10.1002/sce.21623

Odom, P. W., & Zoltowski, C. B. (2019, June 15). *Statistical Analysis and Report on Scale Validation Results for the Engineering Ethical Reasoning Instrument (EERI)*. 2019 ASEE Annual Conference & Exposition. https://peer.asee.org/statistical-analysis-and-report-on-scale-validation-results-for-the-engineering-ethical-reasoning-instrument-eeri

Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 604–624. https://doi.org/10.1109/TNNLS.2020.2979670

Paretti, M. C., Case, J. M., Benson, L., Delaine, D. A., Jordan, S., Kajfez, R. L., Lord, S. M., Matusovich, H. M., Young, E. T., & Zastavker, Y. V. (2023). Building capacity in engineering education research through collaborative secondary data analysis. *Australasian Journal of Engineering Education*, *28*(1), 8–16. https://doi.org/10.1080/22054952.2023.2214462

Pellegrino, J., & Chudowsky, N. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments. *Educational Psychologist*, *51*(1), 59–81. https://doi.org/10.1080/00461520.2016.1145550

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic inquiry and word count: LIWC [Computer software]. *Austin, TX: Liwc. Net*, *135*.

Pennington, J., Socher, R., & Manning, C. D. (2014). *Glove: Global vectors for word representation*. 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Pribadi, F. S., Adji, T. B., Permanasari, A. E., Mulwinda, A., & Utomo, A. B. (2017). *Automatic short answer scoring using words overlapping methods*. *1818*(1), 020042.

Pulman, S., & Sukkarieh, J. (2005). *Automatic short answer marking*. 9–16.

Pushp, P. K., & Srivastava, M. M. (2017). *Train Once, Test Anywhere: Zero-Shot Learning for Text Classification* (arXiv:1712.05972). arXiv. https://doi.org/10.48550/arXiv.1712.05972

Putnikovic, M., & Jovanovic, J. (2023). Embeddings for Automatic Short Answer Grading: A Scoping Review. *IEEE Transactions on Learning Technologies*, 1–13. https://doi.org/10.1109/TLT.2023.3253071

Qadir, J. (2022). *Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education*. TechRxiv. https://doi.org/10.36227/techrxiv.21789434.v1

Qiao, C., & Hu, X. (2023). Leveraging Semantic Facets for Automatic Assessment of Short Free Text Answers. *IEEE Transactions on Learning Technologies*, *16*(1), 26–39. https://doi.org/10.1109/TLT.2022.3199469

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J. (2021). *Learning transferable visual models from natural language supervision*. 8748–8763.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 140:5485-140:5551.

Ramnarain-Seetohul, V., Bassoo, V., & Rosunally, Y. (2022). Similarity measures in automated essay scoring systems: A ten-year review. *Education and Information Technologies*, *27*(4), 5573–5604. https://doi.org/10.1007/s10639-021-10838-z

Rehmann, C. R., Rover, D. T., Laingen, M., Mickelson, S. K., & Brumm, T. J. (2011). *Introducing Systems Thinking to the Engineer of 2020*. 22.961.1-22.961.16. https://peer.asee.org/introducing-systems-thinking-to-the-engineer-of-2020

Reimers, N., & Gurevych, I. (2019a). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. https://doi.org/10.18653/v1/D19-1410

Reimers, N., & Gurevych, I. (2019b). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (arXiv:1908.10084). arXiv. https://doi.org/10.48550/arXiv.1908.10084

Rest, J. R. (1986). *Moral development: Advances in research and theory*. Praeger; WorldCat.org. http://digitool.hbz-nrw.de:1801/webclient/DeliveryManager?pid=3657473

Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A., & Sauerland, U. (2023). Risks and Benefits of Large Language Models for the Environment. *Environmental Science & Technology*, *57*(9), 3464–3466. https://doi.org/10.1021/acs.est.3c01106

Riordan, B., Bichler, S., Bradford, A., King Chen, J., Wiley, K., Gerard, L., & C. Linn, M. (2020). An empirical investigation of neural methods for content scoring of science explanations. *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 135–144. https://doi.org/10.18653/v1/2020.bea-1.13

Romero-Gómez, A. F., & Orjuela-Cañon, A. D. (2022). Natural Language Processing Approach for Learning Process Analysis in a Bioinformatics Course. *2022 IEEE ANDESCON*, 1–5. https://doi.org/10.1109/ANDESCON56260.2022.9989686

Rosenberg, J. M., & Krist, C. (2021). Combining Machine Learning and Qualitative Methods to Elaborate Students' Ideas About the Generality of their Model-Based Explanations. *Journal of Science Education and Technology*, *30*(2), 255–267. https://doi.org/10.1007/s10956-020-09862-4

Roy, S., Bhatt, H. S., & Narahari, Y. (2016). *An Iterative Transfer Learning Based Ensemble Technique for Automatic Short Answer Grading* (arXiv:1609.04909). arXiv. https://doi.org/10.48550/arXiv.1609.04909

Rudnicka, E. A., Sacre, M. B., & Shuman, L. J. (2013). Development and evaluation of a model to assess engineering ethical reasoning and decision making. *The International Journal of Engineering Education*, *29*(4), 948–966.

Sahu, A., & Bhowmick, P. K. (2020). Feature Engineering and Ensemble-Based Approach for Improving Automatic Short-Answer Grading Performance. *IEEE Transactions on Learning Technologies*, *13*(1), 77–90. https://doi.org/10.1109/TLT.2019.2897997

Saldaña, J. (2014). *Coding and analysis strategies*.

Saldaña, J. (2021). *The coding manual for qualitative researchers* (4E.). SAGE; WorldCat.org.

Sands II, K. S., Pearce, A. R., Suh, M. J., Simmons, D. R., Fiori, C. M., & Mouras, and V. A. (2020). Toward a Technique of Evaluating Student Ethical Sensitivity to Professional Issues of the Construction Industry. *Journal of Construction Engineering and Project Management*, *10*(2), 1–20. https://doi.org/10.6106/JCEPM.2020.10.2.001

Sands, K. S., & Simmons, D. R. (2014). *Utilizing Think-Aloud Protocols to Assess the Usability of a Test for Ethical Sensitivity in Construction*. 24.1355.1-24.1355.16. https://peer.asee.org/utilizing-think-aloud-protocols-to-assess-the-usability-of-a-test-for-ethical-sensitivity-in-construction

Sankaran, K. (2022). *Statistical Data Visualization: PCA and UMAP Examples*. Statistical Data Visualization. https://krisrs1128.github.io/stat479/posts/2021-03-25-week10-5/

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Shah, N., & Pareek, J. (2022). Automatic Evaluation of Free Text Answers: A Review. In S. Rajagopal, P. Faruki, & K. Popat (Eds.), *Advancements in Smart Computing and Information Security* (pp. 232–249). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-23095-0_17

Shaik, T., Tao, X., Li, Y., Dann, C., McDonald, J., Redmond, P., & Galligan, L. (2022). A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis. *IEEE Access*, *10*, 56720–56739. https://doi.org/10.1109/ACCESS.2022.3177752

Shakir, U., Ovink, S., & Katz, A. (2022, August 23). *Using Natural Language Processing to Explore Undergraduate Students' Perspectives of Social Class, Gender, and Race*. 2022 ASEE

Annual Conference & Exposition. https://peer.asee.org/using-natural-language-processing-to-explore-undergraduate-students-perspectives-of-social-class-gender-and-race

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53–76.

Soledad, M., Grohs, J., Bhaduri, S., Doggett, J., Williams, J., & Culver, S. (2017). Leveraging institutional data to understand student perceptions of teaching in large engineering classes. *47th IEEE Frontiers in Education Conference, FIE 2017, October 18, 2017 - October 21, 2017, 2017-October*, 1–8. https://doi.org/10.1109/FIE.2017.8190608

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and Permuted Pre-training for Language Understanding. *Advances in Neural Information Processing Systems*, *33*, 16857–16867.
https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, *27*(4), 521–544.

Stephan, K. D. (1999). A Survey of Ethics-Related Instruction in U.S. Engineering Programs. *Journal of Engineering Education*, *88*(4), 459–464. https://doi.org/10.1002/j.2168-9830.1999.tb00474.x

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research techniques*.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). *Mitigating Gender Bias in Natural Language Processing: Literature Review* (arXiv:1906.08976). arXiv. https://doi.org/10.48550/arXiv.1906.08976

Sun, X., Guo, S., Gao, Y., Zhang, J., Xiao, X., & Feng, J. (2019). Identification of urgent posts in MOOC discussion forums using an improved RCNN. *2019 IEEE World Conference on Engineering Education (EDUNINE)*, 1–5. https://doi.org/10.1109/EDUNINE.2019.8875845

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). *LSTM neural networks for language modeling*. Thirteenth annual conference of the international speech communication association.

Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., & Arora, R. (2019). Pre-Training BERT on Domain Resources for Short Answer Grading. *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 6071–6075. https://doi.org/10.18653/v1/D19-1628

Sweeney, L. B., & Sterman, J. D. (2000). Bathtub dynamics: Initial results of a systems thinking inventory. *System Dynamics Review*, *16*(4), 249–286. https://doi.org/10.1002/sdr.198

Swithenby, S. J. (2006). *E-assessment for open learning*. European Association of Distance Teaching Universities (EADTU) Annual Conference.

Tan, S. (2005). Neighbor-weighted K-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications*, *28*(4), 667–671. https://doi.org/10.1016/j.eswa.2004.12.023

Taraban, R., LaCour, M. S., Marcy, W. M., & Burgess, R. A. (2017, June 24). *Developing Machine-Assisted Analysis of Engineering Students' Ethics Course Assignments*. 2017 ASEE Annual Conference & Exposition. https://peer.asee.org/developing-machine-assisted-analysis-of-engineering-students-ethics-course-assignments

Taraban, R., Marcy, W. M., & Koduru, L. (2018). *Tools to assist with collection and analysis of ethical reflections of engineering students*. *2018*.

Taraban, R., Marcy, W. M., Koduru, L., Schumacher, J. R., & Iserman, M. (2019, June 15). *Board 74: Using Machine Tools to Analyze Changes in Students' Ethical Thinking*. 2019 ASEE Annual Conference & Exposition. https://peer.asee.org/board-74-using-machine-tools-to-analyze-changes-in-students-ethical-thinking

Taraban, R., Marcy, W. M., LaCour, M. S., Koduru, L., Prasad HC, S., & Zasiekin, S. (2020). Using the Web to Develop Global Ethical Engineering Students. *Advances in Engineering Education*.

Taraban, R., Robledo, D., Donato, F. V., Campbell, R. C., Kim, J.-H., Na, C., & Reible, D. D. (2020, June 22). *Machine-assisted Analysis of Communication in Environmental Engineering*. 2020 ASEE Virtual Annual Conference Content Access. https://peer.asee.org/machine-assisted-analysis-of-communication-in-environmental-engineering

Taraban, R., Saraff, S., Zasiekin, S., & Biswal, R. (2022). A Psycholinguistic Analysis of Inter-Ethnic Views of Ethics. *East European Journal of Psycholinguistics*, *9*(1), Article 1. https://doi.org/10.29038/eejpl.2022.9.1.tar

Tashakkori, A., & Teddlie, C. (2009). Integrating qualitative and quantitative approaches to research. *The SAGE Handbook of Applied Social Research Methods*, *2*, 283–317.

Taylor, S., Calvo-Amodio, J., & Well, J. (2020). A Method for Measuring Systems Thinking Learning. *Systems*, *8*(2), Article 2. https://doi.org/10.3390/systems8020011

Terenzini, P. T., & Reason, R. D. (2005). *Parsing the first year of college: A conceptual framework for studying college impacts*. annual meeting of the Association for the Study of Higher Education, Philadelphia, PA.

Terrace, H. S., Petitto, L. A., Sanders, R. J., & Bever, T. G. (1981). Response: Ape Language. *Science*, *211*(4477), 87–88. https://doi.org/10.1126/science.7444455

Timofte, R. S., & Popuș, B. T. (2019). Assessment tasks to measure systems thinking and critical thinking in organic chemistry. *Acta Chemica Iasi*, *27*(2), 251–262.

Tomko, M., Nelson, J., Linsey, J., Bohm, M., & Nagel, R. (2017). Towards assessing student gains in systems thinking during engineering design. *DS 87-9 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 9: Design Education, Vancouver, Canada, 21-25.08.2017*, 179–188.

Ullmann, S. (2022). Gender Bias in Machine Translation Systems. In A. Hanemaayer (Ed.), *Artificial Intelligence and Its Discontents: Critiques from the Social Sciences and Humanities* (pp. 123–144). Springer International Publishing. https://doi.org/10.1007/978-3-030-88615-8_7

Vadapally, A., Dehbozorgi, N., & Chowdary Attota, D. (2022). Minute-Paper Dashboard: Identification of Learner's Misconceptions Using Topic Modeling on Formative Reflections. *2022 IEEE Frontiers in Education Conference, FIE 2022, October 8, 2022 - October 11, 2022, 2022-October*. https://doi.org/10.1109/FIE56618.2022.9962598

Vanasupaa, L., Rogers, E., & Chen, K. (2008). Work in progress: How do we teach and measure systems thinking? *2008 38th Annual Frontiers in Education Conference*, F3C-1-F3C-2. https://doi.org/10.1109/FIE.2008.4720378

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Velasquez, M., Andre, C., Shanks, T., & Meyer, M. J. (1987). Can ethics be taught. *Issues in Ethics*, *1*(1), 101–102.

Verleger, M. A. (2014). *Using Natural Language Processing Tools to Classify Student Responses to Open-Ended Engineering Problems in Large Classes*. 24.1338.1-24.1338.15. https://peer.asee.org/using-natural-language-processing-tools-to-classify-student-responses-to-open-ended-engineering-problems-in-large-classes

Wang, R., Wei, S., Ohland, M. W., & Ferguson, D. M. (2019). *Natural language processing system for selfreflection and peer-evaluation*. 229–238.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent Abilities of Large Language Models* (arXiv:2206.07682). arXiv. https://doi.org/10.48550/arXiv.2206.07682

Wei, S., Wang, R., Ohland, M. W., & Nanda, G. (2020, June 22). *Work in Progress: Automating Anonymous Processing of Peer Evaluation Comments*. 2020 ASEE Virtual Annual Conference Content Access. https://peer.asee.org/work-in-progress-automating-anonymous-processing-of-peer-evaluation-comments

Whitelock, D. M., & Brasher, A. (2006). *Developing a roadmap for e-assessment: Which way now?*

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design* (2nd ed.). Association for Supervision and Curriculum Development ASCDAscd.

Wilson, J., Pollard, B., Aiken, J. M., Caballero, M. D., & Lewandowski, H. J. (2022). Classification of Open-ended Responses to a Research-based Assessment Using Natural Language Processing. *Physical Review Physics Education Research*, *18*(1), 010141. https://doi.org/10.1103/PhysRevPhysEducRes.18.010141

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, *3*, 191–191. https://doi.org/10.1016/0010-0285(72)90002-3

Wiser, M. J., Mead, L. S., Smith, J. J., & Pennock, R. T. (2016). Comparing Human and Automated Evaluation of Open-Ended Student Responses to Questions of Evolution. *Proceedings of the Artificial Life Conference 2016*, 116–122. https://doi.org/10.7551/978-0-262-33936-0-ch025

Wulff, P., Buschhüter, D., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2022). Bridging the Gap Between Qualitative and Quantitative Assessment in Science Education Research with Machine Learning—A Case for Pretrained Language Models-Based Clustering. *Journal of Science Education and Technology*, *31*(4), 490–513. https://doi.org/10.1007/s10956-022-09969-w

Wulff, P., Buschhüter, D., Westphal, A., Nowak, A., Becker, L., Robalino, H., Stede, M., & Borowski, A. (2021). Computer-Based Classification of Preservice Physics Teachers' Written Reflections. *Journal of Science Education and Technology*, *30*(1), 1–15. https://doi.org/10.1007/s10956-020-09865-1

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2022). Utilizing a Pretrained Language Model (BERT) to Classify Preservice Physics Teachers' Written Reflections. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-022-00290-6

Wulff, P., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2023). *Enhancing writing analytics in science education research with machine learning and natural language processing–Formative assessment of science and non-science preservice teachers' written reflections*. *7*, 1027.

Xu, W., & Ouyang, F. (2022). A systematic review of AI role in the educational system based on a proposed conceptual framework. *Education and Information Technologies*, *27*(3), 4195–4223. https://doi.org/10.1007/s10639-021-10774-y

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., & Sung, Y.-H. (2019). Multilingual universal sentence encoder for semantic retrieval. *arXiv Preprint arXiv:1907.04307*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, *32*.

Yik, B. J., Dood, A. J., Arellano, D. C.-R. de, Fields, K. B., & Raker, J. R. (2021). Development of a machine learning-based tool to evaluate correct Lewis acid–base model use in written responses to open-ended formative assessment items. *Chemistry Education Research and Practice*, *22*(4), 866–885. https://doi.org/10.1039/D1RP00111F

Yin, W., Hay, J., & Roth, D. (2019). *Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach* (arXiv:1909.00161). arXiv. https://doi.org/10.48550/arXiv.1909.00161

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0

Zhai, X., Haudek, K. C., Stuhlsatz, M. A. M., & Wilson, C. (2020). Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment. *Studies in Educational Evaluation*, *67*, 100916. https://doi.org/10.1016/j.stueduc.2020.100916

Zhai, X., He, P., & Krajcik, J. (2022). Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching*, *59*(10), 1765–1794. https://doi.org/10.1002/tea.21773

Zhai, X., Krajcik, J., & Pellegrino, J. W. (2021). On the Validity of Machine Learning-based Next Generation Science Assessments: A Validity Inferential Network. *Journal of Science Education and Technology*, *30*(2), 298–312. https://doi.org/10.1007/s10956-020-09879-9

Zhang, L., Huang, Y., Yang, X., Yu, S., & Zhuang, F. (2022). An automatic short-answer grading model for semi-open-ended questions. *Interactive Learning Environments*, *30*(1), 177–190. https://doi.org/10.1080/10494820.2019.1648300

Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023). *Sentiment Analysis in the Era of Large Language Models: A Reality Check* (arXiv:2305.15005). arXiv. https://doi.org/10.48550/arXiv.2305.15005

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., & Sun, L. (2023). *A Comprehensive*

Survey on Pretrained Foundation Models: A History from BERT to ChatGPT (arXiv:2302.09419). arXiv. https://doi.org/10.48550/arXiv.2302.09419

Zoltowski, C. B., Buzzanell, P. M., Oakes, W. C., & Kenny, M. W. (2013). A qualitative study exploring students' engineering ethical reflections and their use in instrument validation. *2013 IEEE Frontiers in Education Conference (FIE)*, 1551–1553. https://doi.org/10.1109/FIE.2013.6685098

Zupanc, K., & Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems*, *120*, 118–132. https://doi.org/10.1016/j.knosys.2017.01.006

Appendix A

Big Belly Solar Case Study

Background

The problem of waste management in urban settings is a problem that cities have been working to tackle for a long time. Recently a number of new technologies, developed in part by engineers, have emerged to help combat common trash problems. The Big Belly Solar trash compactor system is one of the technologies that have been widely implemented, including on our own campus here at Virginia Tech. As with many new technologies, there is some controversy about whether these types of trash cans should be adopted widely, with arguments on either side. The cases that you will read about look at two perspectives of the Big Belly Solar roll out in the San Francisco Bay area--one in the City of San Francisco, and another across the bay at the University of California, Berkeley. The third article is more recent and elaborates on San Francisco's most recent efforts to prototype their own trash bin.

Before you read these three articles, it is important to understand some context that differs from what you may be familiar with. In contrast to the relatively rural setting that most of you are familiar with here at Virginia Tech, UC Berkeley is located in an urban environment. This means that some of the challenges found in the surrounding community, like homelessness, are more visible on their campus. It is also important to note that in both Berkeley and San Francisco, unlike Virginia, California offers a container deposit incentive such that someone can turn in used containers for 5 or 10 cents each. It is not uncommon for homeless people in the state to work to collect discarded bottles and make money from returning them as a source of income.

Case study articles

The following three articles present unique views of the implementation of the Big Belly solar trash cans, and cover some of the successes and challenges encountered. Read through the articles, and use this information to help you complete your ethics warm-up in class 10B, as well as the individual ethics report.

As you read through these articles, think about some of the stakeholders that are either directly or indirectly involved. If you have trouble identifying a stakeholder, you can pick something from the list at the end of this document. Also, think about the role that engineers might have taken on in each of the cases.

Berkeley Big Belly Solar Case

Big Belly Solar: Sproul’s New Waste Bins

San Francisco Big Belly Solar Case

Talkin' trash: Little appetite in San Francisco for Big Belly garbage bins

San Francisco Trash Bin Prototyping Efforts

Garbage odyssey: San Francisco’s bizarre, costly quest for the perfect trash can

Stakeholders you can consider

Berkeley Students

University administrators

Big Belly Solar engineers

City of San Francisco waste management workers

The San Francisco Department of Public Works administration

Homeless population

Other stakeholders you identify

Question Prompts Category and Grading Rubric

Category	Percent	Question Prompts
Q1: Identify Ethical Dilemma	**10%**	**Ethical dilemma is clearly and thoroughly identified and related to the given case study.**
Q2: Dilemma Explanation	15%	Explanation includes solid reasoning for why the chosen issue or dilemma was chosen. Answer is clearly related to ethical considerations and backed up by information included in or inferred from the provided case study.
Q3:Ethical Framework Description	20%	At least two ethical frameworks are described that could be related to the chosen issue/dilemma. These could be frameworks from the provided videos, or other well-developed frameworks researched from other sources.
Q4**:Stakeholder Description**	20%	**A clear description is included of a stakeholder, including how they are related to the case study. Their relationship to the ethical issue/dilemma is clearly explained.**
Q5: Framework Comparison	10%	A clear explanation is included that describes how the chosen stakeholder would be affected should the different ethical frameworks be applied in this situation.
Q6: Code of Ethics	15%	A fundamental canon from the NSPE code of ethics is identified. The response includes a well-thought-out connection between the canon and how it could relate to the situation presented in the case study.
Formatting and Writing Quality	10%	The responses to all questions are composed professionally with little to no errors. Any minor errors present do not distract the reader from the intended message. The format of the report follows the template or is organized in a way that is easy for the reader to understand.

Appendix B

System Thinking Case Scenario: Village of Abeesee Scenario.

A. I Vignette

The Village of Abeesee has about 50,000 people. Its harsh winters and remote location make heating a living space very expensive. The rising price of fossil fuels has been reflected in the heating expenses of Abeesee residents. Many residents are unable to afford heat for the entire winter (5 months). A University of Abeesee study shows that 38% of village residents have gone without heat for at least 30 winter days in the last 24 months. Last year, 27 Abeesee deaths were attributed to unheated homes. Most died from hypothermia/exposure (21), and the remainder died in fires or from carbon monoxide poisoning that resulted from improper use of alternative heat sources (e.g., burning trash in an unventilated space).

1. A.II Prompts

1. Processing Phase

1**. Given what you know from the scenario, please write a statement describing your perception of the problems and/or issues facing Abeesee.**

2. What additional information do you need before you could begin to develop a response in Abeesee? Consider both detail and context of the problems/issues you identified.

3. What groups or stakeholders would you involve in planning a response to the problems/issues in Abeesee?

4. Please briefly describe the process you would use to plan a response to the problems/issues in Abeesee.

5. What would you expect a successful plan to accomplish?

2) Response Phase

1. Given what you know and a budget of $50,000, develop a plan that would address the Abeesee situation maximizing the impact of your $50,000. Use a numbered, step-by-step guide, and recipe style to explain your response plan. For example, Step 1: Buy the noodles. Step 2: Boil water. Step 3: Add the noodles. Step 4: Drain the noodles.

2. **On the previous page, you developed a plan. Without specifically changing your plan, reflect on it. What challenges do you see to implementing your plan? What are the limitations of your approach?**

3) Critique Phase

Below, you will have been provided a plan for Abeesee that was developed by someone else.

Plan #46A

1. Develop an application process to allocate up to 100 grants of $500 (100 × $500 = $50,000) to low-income Abeesee residents.

2. Form a review committee comprised of 5 representatives from Abeesee stakeholder groups

3. Distribute $500 grants that can be used to make improvements to homes and residences to reduce exposure to low temperatures

and/or make heating sources safer. Do not allow residents to use grants to pay heating costs.

4. Request documentation of improvements

5. Track "days without heat" and "deaths attributed to unheated homes" to see if there is a decline.

Please read the plan above and respond to the questions that follow.

1. Will Plan #46A solve the problems in Abeesee? Why or why not?

2. Please describe any unintended consequences that you think might result from this plan

3. What other factors do you think might influence the success of this specific plan?

4. How would you know if this $50,000 was used effectively?

5. One of the steps in Plan #46A is the formation of a review committee. What factors are important to consider in the formation of the committee?

1. Instrument Feedback

1. Please use the space below to tell us anything you would like us to know about the scenario, the questions, and the survey interface. We are particularly interested in knowing about places where question phrasing or terms were not clear.

s