ABSTRACT

MCCLURE, JEANNE MARIE. Unveiling the Power of Learning Analytics: Investigating Cognitive Engagement through a Multifaceted Lens. (Under the direction of Shiyan Jiang).

This dissertation investigates the transformative potential of Learning Analytics (LA) and Artificial Intelligence (AI) to enhance cognitive engagement (CE) in high school AI curricula, emphasizing the importance of robust learning theories. The advent of LA represents a significant shift towards a data-informed understanding of learning processes, integrating established educational theories to enhance the applicability and effectiveness of LA in educational settings. This research explores the pivotal role of open-ended questions structured through Bloom's taxonomy in influencing CE and lexical diversity, utilizing Natural Language Processing (NLP) and Large Language Models (LLMs) to analyze and interpret the complexities of student responses within imbalanced educational datasets.

Study 1 (Chapter 2) utilizes an innovative approach to assess how the complexity of questions, informed by Bloom's taxonomy, affects CE and lexical diversity within a Machine Learning Curriculum. This study demonstrates that higher-order cognitive tasks elicit greater CE and lexical diversity, and it reveals how demographic factors influence these dynamics. The integration of LA, AI, and NLP, grounded in educational theories, not only facilitates a deeper understanding of cognitive processes but also enhances the learning experience by fostering greater cognitive and linguistic engagement among diverse student populations. The findings underscore the necessity for curricula that not only engage students cognitively but also cater to diverse linguistic needs, promoting inclusivity within the educational landscape.

Study 2 (Chapter 3) employs an Iterative - In-Context Learning (ICL) Prompt Engineering Design process, significantly outperforming traditional machine learning models. This study focuses on the application of NLP and LLMs, particularly addressing the challenges posed by imbalanced educational datasets through the strategic use of assertion prompts. The results indicate that LLMs with prompt engineering vastly improve the accuracy of classifying nuanced student engagement

levels, particularly in recognizing minority class representations which traditional models often fail to detect.

Together, these studies contribute to the academic discourse on the use of AI in education by demonstrating how advanced AI tools, grounded in robust learning theories and innovative LA techniques, can effectively analyze and enhance learning processes. This dissertation lays a foundation for future research aimed at optimizing educational outcomes through the integration of sophisticated AI technologies, ensuring that educational practices are inclusively and adaptively responsive to the evolving needs of students in a digital age. The integration of LA, AI, and NLP, firmly rooted in solid educational theories, not only facilitates a deeper understanding of cognitive processes but also enhances the learning experience. By systematically incorporating these theories into LA techniques, this research fosters greater cognitive and linguistic engagement among diverse student populations, thereby enriching our approach to developing and applying educational technology in complex learning environments.

Unveiling the Power of Learning Analytics: Investigating Cognitive Engagement through a Multifaceted Lens

by
Jeanne Marie McClure

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Teacher Education and Learning Sciences

Raleigh, North Carolina
2024

APPROVED BY:

Sunghwan Byun	Shaun Kellogg
Kevin Oliver	Shiyan Jiang Chair of Advisory Committee

DEDICATION

To my family—Michael, Elizabeth, Mason, Colin, and Analise—for their unending love and support. Your encouragement and faith in my abilities have been the pillars upon which I built my perseverance. To my parents, who taught me the value of hard work and the joy of learning, thank you for instilling in me the foundations that have carried me through this journey. This achievement is as much yours as it is mine.

BIOGRAPHY

Jeanne McClure, Ph.D., is a driven educator and researcher, who leverages her extensive knowledge in data science and learning analytics to create impactful educational pathways. Residing in Garner, NC with her husband Michael and their four children, Jeanne's journey from North Tonawanda, New York, has been marked by a commitment to hard work and community service—a value instilled by her parents in a modest upbringing.

Jeanne's academic journey is as varied as it is impressive. She holds a Bachelor of Science in Horticulture and Landscape Design and a Master of Education in Learning, Design, and Technology from North Carolina State University, along with several professional certificates in areas ranging from Data Science Methods for Digital Learning Platforms. Her doctoral research focuses on the integration of data science learning pathways through identity-conscious project-based learning, underlining her dedication to inclusivity in educational settings.

Before embarking on her academic career, Jeanne's professional life began in the fields of horticulture and landscape design, where she was involved in projects like the Southern Garden at Johnston Community College and Moore Square in Raleigh, NC. Her transition to education was driven by a passion for social good, leading her to roles with the Girl Scouts and as an educator at a notable elementary school in Clayton, NC. These experiences enriched her understanding of educational needs and the power of community engagement.

Jeanne's research at North Carolina State University has not only been academically rigorous but also practically beneficial. She has developed curricula that integrate machine learning and AI into general high school courses, and has pioneered methods to enhance learning through technology. Her work, funded by significant NSF grants, has led to numerous publications in esteemed educational technology journals and presentations at international conferences.

In addition to her research and teaching, Jeanne is an active member of several professional organizations, including the American Educational Research Association, the Society for Learning Analytics Research, and the International Educational Data Mining Society. She also runs JMcClure

Designs, where she consults on educational development, learning analytics, and AI technologies. Her consulting practice continues her mission to foster educational advancements that are both inclusive and effective, leveraging cutting-edge technology to enhance learning outcomes.

At home, Jeanne is a devoted mother and a proud fan of the Buffalo Bills (Go Bills!). Her journey reflects a seamless blend of personal passion and professional endeavor, making her a respected figure in educational circles and a beloved member of her community. Her story is one of relentless pursuit of knowledge and its application for the betterment of society.

ACKNOWLEDGEMENTS

First and foremost, I wish to express my deep gratitude to my academic advisor, Dr. Shiyan Jiang, for her unwavering support and guidance throughout my PhD journey. Her knowledge and encouragement have been instrumental during both the challenging and triumphant moments of this endeavor.

I also extend my heartfelt thanks to my research supervisor and committee member, Dr. Sunghwan Byun. His mentorship over the past two years has not only fostered my growth as a researcher but also bolstered my confidence in my abilities, even in times of doubt. My appreciation goes to my additional committee members, Drs. Shaun Kellogg and Kevin Oliver, for their guidance and support. Shaun Kellogg, in particular, has played a pivotal role in my development as both a researcher and educator, especially through my involvement with LASER, enhancing my leadership and scholarly skills. Kevin Oliver, whose reasons for my joining the NCSU PhD program were always clear, has always been receptive to my ideas and offered opportunities to test them in meaningful ways.

I am thankful to my co-author, Machi Shimmei, for her support. I have learned a lot from you during our paper-writing process. My gratitude also extends to Noboru Matsuda for his invaluable feedback and suggestions.

Additionally, I extend my thanks to the many professors from whom I've had the privilege of learning throughout my coursework. Drs. Tameica Jones, Megan Manfra, Margareta Thomson, Michael Little, Deleon Gray, Angela White, Samantha Marshal, and others not mentioned here have all provided invaluable feedback that significantly improved my writing and scholarly development. I am deeply appreciative of your contributions to my journey.

I would like to acknowledge an amazing group of friends for their support and feedback throughout the highs and lows of this PhD journey. Shelley Glica, Dr. Cansu Tatar, Amanda Hall, AnneMarie Stephen, Dr. Jennifer Houchins, Doreen Mushi, Yuru Zhang, Franziska Bickle, Laura Fogle, and Hunter Edwards, your encouragement and critical insights have been invaluable. You

have inspired deep thought and significantly contributed to my growth as a researcher and thinker.

Finally, I must thank my family for their unwavering support and love. To my husband, Mike, for his companionship and encouragement; to my wonderful children, Elle, Mason, Colin, and Ana, who inspire me daily and support me in all my endeavors; and to my mother, Cassandra, for her steadfast belief in me. To my siblings, Betsy and Rob, thank you for challenging me to excel. To my in-laws, Joyce and Rich, for your support throughout this process. And to my father, may he rest in peace, I believe I can finally say I am done with college... maybe.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

Introduction and Statement of the Problem

The advent of Learning Analytics (LA) in education signifies a pivotal shift towards a data-informed understanding of learning processes, underscored by the potential to fundamentally alter how learning is facilitated and optimized in digital environments. This shift is not merely technological but represents a holistic transformation, impacting the core of educational methodologies. As Driscoll & Burner (2005) articulates, learning is an evolving process wherein a change in performance or potential is catalyzed by the learner's interaction with their environment. This conceptualization of learning, as both a process and an outcome, underscores the critical importance of embedding LA within a framework of established learning theories to enhance its applicability and effectiveness in educational research (Driscoll & Burner, 2005; Banihashem & Macfadyen, 2021).

Learning theories play a pivotal role in LA by providing a foundational understanding of how individuals acquire, process, and retain knowledge. This understanding is crucial for leveraging LA not just for data analysis but for genuinely enhancing learning experiences through a strategic amalgamation of data science and pedagogical principles (Gašević et al., 2015; Banihashem & Macfadyen, 2021).

Despite its promise, the field of LA faces a significant challenge: a widespread lack of expertise among researchers in applying LA's tools and methodologies effectively, especially within STEM education. This skills gap underscores the need for a multidisciplinary approach that merges aspects of data mining, machine learning, artificial intelligence, and learning sciences with robust pedagogical models to facilitate meaningful educational advancements (Avella et al., 2016; Gašević et al., 2015; Bienkowski et al., 2012; Dawson et al., 2014).

The integration of LA and Artificial Intelligence (AI) in education heralds a transformative potential for enriching student learning processes and cognitive engagement (CE). Yet, current

LA models often remain data-rich but context-poor, gathering extensive metrics without achieving the depth of insight required for effective educational interventions. This limitation arises from a failure to ground the wealth of data within the solid theoretical foundations of education, leading to analyses that, despite their data richness, lack actionable or comprehensive insights into learning dynamics (Fredricks et al., 2004; Blumenfeld et al., 2006).

This issue is acutely evident in the analysis of open-ended responses within high school AI curricula, where traditional LA approaches fall short of capturing the nuanced aspects of CE. Open-ended questions, critical for eliciting higher-order thinking and knowledge construction, offer a unique vantage point into students' cognitive processes. Yet, the challenge of accurately interpreting CE from these responses highlights the inadequacies of existing LA models (Lee & Kinzie, 2012; Smart & Marshall, 2013).

This dissertation seeks to explore the integration of LA and AI-enhanced curricula, focusing on the pivotal role of open-ended questions in high school settings. It aims to employ LA, underpinned by learning theories, to unravel and amplify aspects of cognitive engagement. Recognizing that learning analytics can significantly enhance learning outcomes through predictive models and actionable insights, this research is predicated on a nuanced understanding of learning theories (Avella et al., 2016).

Accordingly, this dissertation endeavors to explore the interplay between learning theories, LA, and Natural Language Processing (NLP) methodologies, including the potential use of Large Language Models (LLMs), to deepen our comprehension of cognitive engagement in digital learning contexts. This investigation strives not only to address the current gaps in assessing cognitive engagement through open-ended responses but also to contribute to a comprehensive, theory-grounded approach to educational research and innovation in the digital age.

Purpose Statement and Research Objectives

The goal of this research is to explore and understand the intricacies of CE in high school students participating in an AI curriculum, with a specific focus on the role of open-ended questions.

This study aims to leverage the potential of LA and NLP, including the application of LLMs where relevant, to offer new insights into the dynamics of student engagement and knowledge construction in digital learning environments.

Key research objectives include:

1. To assess the depth and quality of cognitive engagement elicited by open-ended questions in an AI-enhanced curriculum, thereby identifying patterns, challenges, and opportunities for enhancing student learning experiences.

2. To apply and evaluate learning analytics methodologies, grounded in solid educational theories, for analyzing open-ended responses, aiming to uncover the nuanced aspects of CE that traditional metrics might overlook.

3. To explore the potential of NLP and LLMs in enriching the analysis of student responses, aiming to provide a more nuanced understanding of students' cognitive engagement and its implications for learning and retention.

4. To propose a model for integrating LA tools and methods, informed by learning theories, into the pedagogical design of AI curricula, thereby enhancing the effectiveness of educational interventions and contributing to the broader discourse on the optimization of digital learning environments.

By achieving these objectives, this research intends to bridge the existing gap in assessing and enhancing cognitive engagement through open-ended responses, contribute to the theoretical and practical understanding of LA's role in educational research, and foster a more informed and effective application of AI technologies in K-12 education.

Significance of the Research

This dissertation stands at the confluence of LA, AI, and educational theory, targeting the nuanced domain of CE within AI-enhanced curricula. Its significance stems from several pivotal

areas of contribution:

- This study will provide theoretical integration and empirical insight by developing an integrated approach that marries LA with solid educational theories to bridge the critical gap observed in current educational models—where data richness often lacks contextual and theoretical depth (Fredricks et al., 2004; Blumenfeld et al., 2006). This effort goes beyond mere data collection to offer actionable insights grounded in educational pedagogy, thus enhancing the interpretability and applicability of LA findings in practical educational settings.

- This study will develop advancements in Learning Analytics through NLP by utilizing NLP, especially through the application of LLMs, to forge a pioneering method in analyzing open-ended responses within AI-enhanced curricula. This methodological innovation not only broadens the capabilities of LA in understanding CE but also showcases the transformative potential of blending AI with educational research, potentially establishing new methodological standards for the field.

- This study will promote educational equity and inclusivity by concentrating on the diverse and imbalanced nature of educational datasets to underline the necessity of inclusivity and equity in educational research (Avella et al., 2016; Gašević et al., 2015). The findings from this dissertation are aimed at contributing to more equitable educational practices, ensuring that the advantages of AI and LA are accessible across various student demographics, thereby fostering a more inclusive educational environment.

- This study will offer practical implications for pedagogy and curriculum design through an exploratory analysis of CE among high school students participating in an AI curriculum, providing valuable insights for educators and curriculum developers. This study underscores the significance of question design and pedagogical strategies that encourage deep cognitive engagement, offering practical guidelines for integrating AI into curricula in ways that stimulate critical thinking, problem-solving, and active learning among students.

- This study will lay a foundation for future research by acting as a focal point for interdisciplinary exploration, opening new pathways for further studies into the convergence of LA, AI, and educational theory. It motivates subsequent research into the long-term effects of AI-enhanced learning, the application of NLP methodologies in various educational contexts, and the creation of more detailed models of student engagement.

This research, therefore, extends beyond the academic realm, offering implications for teaching, curriculum design, educational policy, and technology integration in education. By advancing our understanding of CE through the lens of LA and AI, it contributes to developing more effective, engaging, and insightful educational practices equipped to meet the 21st-century challenges (Avella et al., 2016; Gašević et al., 2015).

Theoretical Foundations Supporting the Problem/Issue

Cognitive Engagement

This central theory, as detailed in the literature, offers diverse definitions emphasizing the depth and complexity of learners' psychological and strategic involvement in their educational activities. Cognitive engagement is defined through various lenses, ranging from psychological investment in learning tasks to the employment of cognitive and metacognitive strategies for deeper understanding and mastery of complex skills (Fredricks et al., 2004; Pintrich, 2000; Schunk et al., 2014). It encapsulates students' willingness to go beyond the basic requirements, demonstrating a clear preference for challenging tasks and actively engaging in the learning process. This engagement is characterized by thoughtful and purposeful approaches to school tasks, underpinned by a commitment to exert the necessary effort to comprehend complex ideas and master difficult skills (Fredricks et al., 2004). Moreover, cognitive engagement also touches on the dimensions of motivation, autonomy, and the strategic and self-regulatory processes that facilitate learning (Mayer, 2005, 2008; Helme & Clarke, 2001). Through this multifaceted perspective, cognitive engagement is seen not just as an educational goal but as a critical mechanism through which students engage

deeply with their learning environment, contributing to their overall academic development and success.

Cognitive engagement significantly enhances the learning process by deepening students' understanding of subject matter, promoting higher-order thinking skills, and fostering long-term retention and knowledge transfer. As outlined in the literature, cognitive engagement is instrumental in motivating students to invest greater mental effort, leading to a more profound grasp of complex concepts and the mastery of challenging skills (Fredricks et al., 2004). This type of engagement is closely linked with metacognition and higher-order thinking, where students who actively reflect on and regulate their learning processes are more likely to develop critical thinking, problem-solving, and analytical capabilities (Schunk et al., 2014).

Moreover, the active and meaningful engagement facilitated by cognitive involvement is essential for encoding and storing information, thereby enhancing the likelihood of knowledge retention and its application in real-world contexts (Hattie & Yates, 2014). Through the deployment of deep versus surface-level strategies, students engaged cognitively are more inclined to exert mental effort, create connections among ideas, and achieve a comprehensive understanding of the material, underscoring the pivotal role of cognitive engagement in promoting motivation, enhancing learning outcomes, and nurturing metacognitive abilities (Weinstein & Mayer, 1986, as cited in Fredricks et al., 2004) .

Cognitive engagement has been measured through various methods, reflecting its complex and multifaceted nature. These methods include self-reported scales, observations, interviews, teacher ratings, experience sampling, eyetracking, physiological sensors, trace analysis, and content analysis (Fredricks et al., 2004; Chi & Wylie, 2014; McCoach et al., 2013; Greene, 2015). For example, the use of self-reported scales and surveys, such as the Science Activity Questionnaire (SAQ) by Meece et al. (1988) and the Motivated Strategies for Learning Questionnaire (MSLQ) by Pintrich et al. (1993), evaluates cognitive engagement through students' motivational lens and strategic learning approaches. Observational measures, like those conducted by Lee and Anderson (1993) in science classrooms, seek observable cues of cognitive engagement, including students'

initiation of activities to comprehend concepts, thus offering direct insights into engagement levels.

Open-ended questions are advocated for use in measuring cognitive engagement due to their ability to elicit in-depth and thoughtful responses, revealing students' higher-order cognitive processes such as analysis, synthesis, and evaluation (Fredricks et al., 2004; Chin & Chia, 2004). They allow students to showcase their creativity, originality, and metacognitive awareness, offering educators a richer, more nuanced view of students' cognitive engagement. Open-ended questions also enable the identification of misconceptions and personalized insights, facilitating a deeper engagement and ownership in the learning process. The flexibility and depth afforded by open-ended questions surpass the limitations of closed-ended questions, promoting critical thinking and facilitating the construction of knowledge in a more meaningful and personally relevant manner (Lee & Kinzie, 2012; Çakır & Cengiz, 2016; Svanes & Andersson-Bakken, 2023).

Therefore, employing open-ended questions in educational settings is crucial for enhancing cognitive engagement, fostering critical thinking, and supporting a deeper understanding of complex topics. Open-ended questions can play a significant role in designing personalized and adaptive learning experiences, allowing for a more individualized approach to learning, nurturing lifelong learning skills and promoting student agency (Sharkins et al., 2017), thus contributing significantly to the learning process and the assessment of cognitive engagement.

Blooms' Taxonomy

Bloom's Taxonomy, originally developed by Benjamin Bloom and collaborators in the mid-20th century, is a hierarchical framework for categorizing educational goals, objectives, and outcomes. This taxonomy has been extensively used in education to guide the development of learning objectives, assessments, and instructional strategies. It outlines a progression of cognitive skills that range from basic to more complex and abstract forms of thinking (Anderson et al., 2001).

The original taxonomy delineates six cognitive domains: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. These domains represent a move from simple recall of facts, through understanding and applying those facts, to more complex processes such as

analyzing information, synthesizing ideas to create something new, and evaluating information or ideas critically (Bloom et al., 1956).

In 2001, a revised version of Bloom's Taxonomy was published by Anderson and Krathwohl, who were among Bloom's former students. The revision aimed to update the taxonomy to reflect contemporary understanding of educational psychology and pedagogy. It introduced a two-dimensional framework: the cognitive process dimension, which revises and renames the original six cognitive domains, and the knowledge dimension, which categorizes the type of knowledge to be learned. The cognitive process dimension includes Remembering, Understanding, Applying, Analyzing, Evaluating, and Creating, thereby placing 'Creating'—a process that involves generating new ideas, products, or ways of understanding—at the highest level of cognitive complexity (Anderson et al., 2001).

Bloom's Taxonomy serves multiple roles in education. It acts as a tool for curriculum planning, instructional design, and assessment. Educators use the taxonomy to structure content and activities that promote higher-order thinking skills in learners. By ensuring that learning objectives are aligned with appropriate cognitive processes, educators can better scaffold learning experiences to facilitate students' progression from basic knowledge acquisition to complex, critical, and creative thinking (Krathwohl, 2002).

Furthermore, Bloom's Taxonomy is instrumental in the development of assessment tools that accurately measure learning outcomes across the cognitive spectrum. It helps educators create assessments that not only test for factual knowledge but also evaluate students' ability to understand, apply, analyze, synthesize, and evaluate information (Anderson et al., 2001; Krathwohl, 2002).

Bloom's Taxonomy is a foundational framework in education that categorizes the cognitive objectives of learning into a hierarchy, from basic to complex levels. Its application spans curriculum development, instructional design, and assessment, supporting educators in fostering and evaluating higher-order thinking skills among learners (Bloom et al., 1956).

Definitions

Open-ended questions

Open-ended questions are defined as inquiries that allow respondents to answer in an extended, narrative form, providing depth, nuance, and insight into their thought processes. These questions diverge significantly from closed-ended questions by not limiting responses to predetermined choices, thereby encouraging a broader exploration of topics (Tofade et al., 2013a). They serve as a catalyst for fostering cognitive engagement, as they require students to engage in higher-order thinking skills like critical analysis, synthesis, and evaluation (Lee & Kinzie, 2012; Çakır & Cengiz, 2016; Svanes & Andersson-Bakken, 2023). By prompting learners to connect prior knowledge with new information and to think critically, open-ended questions contribute to a more profound cognitive engagement than their closed-ended counterparts.

Furthermore, open-ended questions are characterized by their flexibility, allowing for diverse perspectives and the exploration of subjects without a predetermined answer (Tofade et al., 2013a). This quality promotes critical thinking, creativity, and a deeper understanding among students, as they are free to explore various possibilities and develop complex and comprehensive answers, thus fostering a higher level of cognitive engagement (Adom et al., 2016; Svanes & Andersson-Bakken, 2023).

Moreover, such questions encourage students to articulate and expand upon their thoughts, providing rationales for their ideas and using metacognitive processes to monitor and evaluate their thought processes. This engagement deepens their cognitive involvement compared to the straightforward nature of closed-ended questions (Lee & Kinzie, 2012; Wasik & Hindman, 2013).

In educational settings, embracing open-ended questions can significantly facilitate deeper cognitive engagement and foster critical thinking skills in students, surpassing the limited cognitive demands of closed-ended questions and contributing to knowledge construction and the development of language skills as students engage with newly learned words and concepts (Çakır & Cengiz, 2016; Brock, 1986; Hargreaves, 1984; Bloom et al., 1956; Anderson et al., 2001).

Large Language Models

LLMs, such as, Chat GPT series, represent a significant leap forward in the field of NLP, enabling machines to produce text that closely mimics human writing, answer queries, and perform various language-based tasks with remarkable accuracy. These advancements are largely attributable to the adoption of transformer architectures and attention mechanisms, which enhance the models' capacity to understand the contextual relationships between words in text, regardless of their position in sentences (Vaswani et al., 2018; Devlin et al., 2018; Tay et al., 2022; Wei et al., 2022b). A pivotal strategy in the development of these models is pre-training on extensive text datasets, followed by fine-tuning for specific tasks, a process that significantly boosts their performance across a broad spectrum of NLP applications (Min et al., 2021). For instance, BERT, a model built on the transformer framework, demonstrates how pre-training can be effectively leveraged for tasks such as sentence classification and named entity recognition (Devlin et al., 2018). Furthermore, the emergent capability of LLMs for few-shot learning, allowing them to adapt to various tasks with minimal additional training, marks a notable advancement in the versatility of language models (Brown et al., 2020; Borisov et al., 2022).

Language modeling, a cornerstone of machine language intelligence, aims to predict word sequence likelihoods, thereby estimating future or missing tokens' probabilities. The evolution of language models has seen the progression from Statistical Language Models (SLMs) reliant on statistical learning to Neural Language Models (NLMs) utilizing neural networks for prediction, leading to Pre-trained Language Models (PLMs) that capture context-aware representations. The latest in this evolutionary line, LLMs, are distinguished by their vast scale and emergent abilities in complex task performance, driven by advancements in model and data scaling. This shift from simple text generation to complex task-solving represents a significant leap in AI's capabilities, highlighting the role of LLMs as general-purpose task solvers (Zhou et al., 2023).

The rapid development and application of LLMs, such as ChatGPT, are reshaping the landscape of AI research and its practical applications, challenging traditional models of information retrieval and opening new avenues for AI-enhanced dialogues and multimodal interactions. Despite

their transformative potential, the principles underpinning the unique abilities of LLMs remain an area of active research, with ongoing efforts to understand the mechanisms behind their emergent properties and to align their outputs with human values and ethical standards (Zhou et al., 2023).

Lexical Diversity

Lexical Diversity (LD) is a measure of the range and variety of vocabulary used in speech or writing, offering insights into linguistic complexity, language proficiency, and the sophistication of vocabulary choices. It is an important predictor of language development, literacy achievements, and even socio-economic status, serving as an indicator of a speaker's or writer's proficiency and the complexity of their language use (Chase et al., 2019; Jarvis, 2013; Wood et al., 2019b). LD is closely associated with measures of linguistic complexity, including accuracy and fluency, and is considered alongside cognitive engagement metrics to provide a comprehensive view of student engagement and response nature (Mackey & Gass, 2011).

The assessment of LD involves various metrics such as the Number of Different Words (NDW) and the Type-Token Ratio (TTR), which help in demonstrating vocabulary variation across language samples and ensuring fair comparisons by normalizing vocabulary diversity to the text's length (Johnson, 1944; McCarthy & Jarvis, 2007). These measures are crucial for indicating language development and literacy achievements, particularly in children (Miller et al., 2016; Thordardottir et al., 2001; Price & Jackson, 2015; Farquharson et al., 2014). LD not only highlights linguistic capabilities but also sheds light on the interaction of language with societal, cultural, and historical contexts, thus reflecting and shaping these dimensions (Araujo et al., 2018; Johansson, 2008).

In educational contexts, LD acts as a tool for evaluating how variations in language proficiency among students impact the clarity and coherence of their responses, providing a quantitative, numerical evaluation of a text's linguistic richness. This richness, which includes a weighting for the rarity of words, serves as a lens for exploring the diversity of word choice in speech or writing, illuminating language through variations in vocabulary, grammar, syntax, and dialects

among speakers within a language or across different languages (Daller et al., 2003; Malvern et al., 2004; Hampson & McKinley, 2023).

Learning Analytics

Siemens (2013) asserts that “learning analytics is the measurement, collection, analysis, and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environments in which it occurs” (p. 2).

This multifaceted process encompasses the development of actionable insights through the application of statistical models and analysis against both existing and potentially simulated future data (Cooper et al., 2012). LA represents a bridge between educational data mining (EDM) and broader educational challenges, extending EDM methodologies to tackle questions beyond the reductionist analysis and aiming to influence teaching and learning practices directly.

As a discipline, LA is underpinned by a diverse array of theoretical foundations and practical applications, rooted in the interplay between epistemology, pedagogy, and assessment. This triadic relationship emphasizes the importance of aligning educational theories with analytic techniques to address the inherent tensions between assessment practices and pedagogical goals (Knight & Shum, 2017). LA challenges traditional educational practices by introducing new forms of assessment that can both support and transform education, potentially reshaping educational systems towards or away from technocratic models, and redefining what constitutes "learning" based on what can be measured.

The field of LA is characterized by its application across various organizational levels—from individual classrooms to international education systems—allowing for a nuanced understanding of learning processes through different analytic lenses Buckingham Shum (2012). This comprehensive approach to analytics is informed by historical contributions from fields such as artificial intelligence, statistical analysis, cognitive modeling, and e-learning, among others, which have collectively contributed to the evolution of LA tools, techniques, and applications. Learning analytics tools range from commercial products designed for widespread educational use to research tools focused on

specific analytic tasks. These tools are instrumental in applying LA techniques to improve learning outcomes through applications such as modeling user knowledge, creating learner profiles, and personalizing educational content Baker et al. (2009); Bienkowski et al. (2012).

This definition encapsulates the essence of LA as a field focused on leveraging data to enhance educational outcomes and learning experiences. It underscores the goal of making informed decisions based on insights derived from educational data, thus bridging theoretical knowledge with practical applications to improve learning processes and educational environments (Knight & Shum, 2017).

Chapter 2: Analyzing the Interplay of Question Design and Student Engagement: A Linguistic and Cognitive Perspective

Abstract

Background: The integration of Artificial Intelligence (AI) in K-12 education necessitates an understanding of its impact on cognitive engagement (CE) and lexical diversity. Current research lacks comprehensive insights into how AI-enhanced curricula influence these aspects, particularly across different cognitive tasks and demographic groups.

Objective: This study investigates the relationship between Bloom's taxonomy categories, cognitive engagement levels, and lexical diversity within a Machine Learning Curriculum, examining how these factors interplay with demographic influences to affect student learning outcomes in an AI-enhanced educational setting.

Method: Analyzing data from 28 students using the StoryQ AI curriculum, this research adopted a mixed-methods framework. It applied the adapted ICAP model for assessing CE, McCarthy's MTLD for lexical diversity, descriptive statistics to summarize data, and mixed linear models alongside chi-square tests to investigate the relationship between engagement levels, cognitive task complexity, and student demographics.

Findings: Findings indicate significant variance in cognitive engagement across Bloom's taxonomy categories, with higher complexity tasks correlating with active engagement and increased lexical diversity. Demographic factors showed a nuanced relationship with engagement and linguistic outcomes, suggesting variability in AI-enhanced learning experiences among different student groups.

Implications: The study underscores the need for educational strategies that consider task design and demographic inclusivity in AI-enhanced learning environments. It contributes to a deeper understanding of optimizing AI in education to support diverse cognitive and linguistic development, paving the way for further research in this emergent field.

Introduction

With the exponential growth of AI-enabled systems, tasks traditionally seen as exclusively human (e.g., driving, curriculum development, medical diagnosing, transcribing) are now within the realm of machine capabilities (Dwivedi et al., 2021). This development underscores the necessity for educators to not only introduce AI into their teaching but also to make sure students deeply engage with these sophisticated concepts. In the rapidly evolving landscape of K-12 education,, the deliberate incorporation of Artificial Intelligence (AI) is revolutionizing sectors such as business, manufacturing, and notably, the domain of education, specifically in teaching and learning processes. This revolution, as highlighted by Dumont & Zurn (2007) and Dwivedi et al. (2021), is set to enhance the learning experience significantly and equip students with vital skills for future job markets. Given AI's expanding role in various industries, it's critical for educators to weave AI into their curricula, ensuring that students, particularly those outside computer science disciplines, actively engage with these cutting-edge concepts to develop crucial abilities like critical thinking and problem-solving (Tatar et al., 2023; Zimmermann-Niefield et al., 2019).

To measure how deeply students engage with and learn from these concepts, educators often turn to open-ended questions, a method proven to encourage cognitive engagement (CE) and deeper information processing (Costa & Kallick, 2015; Anderson & Krathwohl, 2001; Bloom et al., 1956). CE, which involves attention, interest, investment, and effort in learning Marks (2000), is greatly influenced by question structure, turning questions into tools for eliciting varied levels of cognitive engagement (Tofade et al., 2013b; Piaget, 2003).

When Lexical Diversity (LD) is considered alongside CE metrics, it provides a detailed view of a student's engagement with material and the nature of their responses (Chase et al., 2019). LD serves as a predictor for several important factors, including language proficiency, complexity, vocabulary knowledge, and lexical proficiency (Jarvis, 2013; Wood et al., 2019a). Recognized as an aspect of linguistic complexity, LD is assessed together with other complexity measures in studies on language proficiency which include complexity, accuracy, and fluency (Mackey & Gass, 2011).

Dillard & Pfau (2002) points out that the complexity brought by LD, when it matches listeners' comprehension abilities, makes messages more appealing due to its engaging nature, indicating sophisticated vocabulary choices. Moreover, it's important to consider that students with varying degrees of LD might exhibit different forms of engagement. Those with lower LD may still be deeply involved in their learning, perhaps focusing on cognitive strategies or content areas that do not necessitate a wide range of vocabulary. This implies that LD should be interpreted in the context of other engagement indicators to fully appreciate the diverse ways students interact with educational content.

Exploring Bloom's taxonomy in relation to these findings reveals complex patterns linking question types to LD and the depth of student CE. By analyzing the interplay between question structure, LD, and CE, educators can gain valuable insights to refine AI curricula, fostering deeper student engagement and understanding.

This paper posits that the nature and structure of questions, classified using a revised version of Bloom's taxonomy, significantly impact both CE and LD in student responses. While prior research connects CE to learning outcomes (Chi & Wylie, 2014), the specific relationship with LD is less examined. Our study aims to bridge this gap by analyzing self-reported student responses within a Machine Learning Curriculum, focusing on how question design influences LD and overall cognitive engagement. Specifically, the primary research questions are:

1. How do Bloom's taxonomy categories, cognitive engagement levels, and demographic factors influence lexical diversity in student responses?

2. How does module/question complexity and sequence influence student engagement levels and LD?

3. How do demographic factors intersect with Bloom's taxonomy categories and cognitive engagement levels to shape lexical diversity in student responses?

The remainder of the paper is as follows: Section 2 looks at the Literature review, while Section 3 describes the theoretical frameworks used to guide the study: Revised Bloom's, Interactive

- Constructive - Active, and Passive (ICAP) framework and Linguistic Diversity. In Section 4, we describe the data used and the construction of the linguistic diversity measure. Results on the analyses are discussed in Section 5. Section 6 discusses the results and the implications and Section 7 provides concluding remarks.

Background

This integration of AI in educational settings presents a unique opportunity to enhance CE and LD within learning environments. CE, extensively explored through various perspectives in literature, encompasses psychological investment in learning and cognition (Fredricks et al., 2004; Mayer, 2008). Defined by Fredricks et al. (2004) as the depth of active mental effort students invest in learning, it includes being thoughtful and purposeful in the approach to school tasks and being willing to exert the effort necessary to comprehend complex ideas or master difficult skills. This is crucial for effective learning, intertwined with the range and complexity of vocabulary used by students (Ransdell & Wengelin, 2003).

CE draws from the dual literatures of psychological investment, characterized by a willingness to embrace challenging tasks (Pintrich, 2000; Schunk et al., 2014), and the cognitive functions involved in learning, such as perception, attention, memory, and problem-solving (Mayer, 2008). Blumenfeld et al. (2006) and Appleton et al. (2006) further elaborate that cognitive engagement intertwines with motivation, incorporating students' willingness to invest effort in learning while employing cognitive, metacognitive, and volitional strategies to promote understanding.

CE in educational settings has been extensively studied, emphasizing its importance in active learning and retention. Marks (2000) defines CE as the intensity of focus and intrinsic interest students demonstrate in learning activities, crucial for deep and enduring learning. Moreover, the definitions of CE in the literature highlight its variability and the broad spectrum of its implications, ranging from task-specific strategic thinking to broader motivational and emotional aspects (Helme & Clarke, 2001; Zimmermann-Niefield et al., 2019). The definitions are rich and underscores the complexity of CE, which is amplified in AI-driven learning environments where strategic and

self-regulating behaviors are crucial Li et al. (2021). Additionally, expanding on the dimensions of CE, researchers like Chi et al. (2018) and Wu (2021) highlight the depth of processing and problem-solving as critical aspects of engagement that involve critical thinking, metacognition, and the application of knowledge to complex scenarios, particularly within AI-facilitated environments.

Chi & Wylie (2014) ICAP framework builds upon the foundational concepts in Bloom's Taxonomy by offering a more granular approach to categorizing and enhancing student engagement and interaction within learning environments. Bloom's Taxonomy categorizes educational goals into cognitive levels that increase in complexity from remembering to creating (Anderson & Krathwohl, 2001). The ICAP framework aligns with and extends these categories by focusing on the nature of student engagement that is necessary to achieve these cognitive levels.

The ICAP framework suggests that questions designed to move students from passive reception to interactive engagement can significantly deepen learning. For instance, transitioning from simple recall questions (passive) to those that require analysis or creation (constructive and interactive), such as designing an experiment or debating a topic's pros and cons based on evidence, can promote higher-order thinking and deeper understanding. Building on this theory are the findings by Garcia-Ponce & Tavakoli (2022), it is critical to integrate insights from the ICAP framework by Chi & Wylie (2014), which provides a structured approach to categorizing and enhancing cognitive engagement through question design.

Furthermore, the ICAP framework serves as a practical tool for educators to design instructional strategies that specifically target and promote higher levels of cognitive engagement. By applying ICAP alongside Bloom's Taxonomy, educators can more effectively scaffold learning activities to not only cover a range of cognitive skills but also engage students at levels that are conducive to deeper learning and retention.

When LD is considered alongside CE metrics, it provides a detailed view of a student's engagement with material and the nature of their responses(Chase et al., 2019). LD serves as a predictor for several important factors, including language proficiency, complexity, vocabulary knowledge, and lexical proficiency (Jarvis, 2013; Wood et al., 2019a). Recognized as an aspect

of linguistic complexity, LD is assessed together with other complexity measures in studies on language proficiency which include complexity, accuracy, and fluency (Mackey & Gass, 2011)(.

Recent advancements have introduced more sophisticated methods to operationalize LD, aiming to minimize construct-irrelevant variations such as sample length. For instance, the Measure of Textual Lexical Diversity (MTLD) (McCarthy & Jarvis, 2007) and the Moving Average Type Token Ratio (MATTR) (Covington & McFall, 2010) have been identified as robust measures, providing reliable LD assessments by maintaining consistency across different sample sizes (Fergadiotis et al., 2015). MTLD calculates the average number of consecutive words that maintain a predefined TTR, showing high validity in reflecting the lexical richness of language samples (Fergadiotis et al., 2015). Similarly, MATTR employs a moving window to estimate TTR, thus adjusting for sample size variability and enhancing measurement reliability (Lissón & Ballier, 2018). These methods contrast with older techniques like the type token ratio (TTR) (Johnson, 1944), which can be disproportionately influenced by sample length, leading to less reliable assessments of lexical diversity (Fergadiotis et al., 2015; Jarvis, 2013).

The study by Lissón & Ballier (2018) found that TTR is highly dependent on text length, which skewed its effectiveness as a measure of lexical complexity. To mitigate this, transformations and more sophisticated metrics like MTLD (Measure of Textual Lexical Diversity) and MATTR (Moving-Average TTR) are proposed. Their findings suggest that while TTR is a simple and historically popular measure, it is less reliable for shorter texts compared to these more advanced metrics. This insight helps in guiding educators and researchers towards using more robust measures of lexical diversity that provide more stable and accurate assessments across texts of varying lengths. Additionally, Koizumi (2012)research examined the impact of text length on lexical diversity measures, including MTLD, in the context of second language (L2) learners' spoken texts. The study highlighted MTLD's robustness in maintaining consistent lexical diversity values across short texts ranging from 50 to 200 tokens, demonstrating that MTLD was least affected by text length compared to other measures such as TTR and Guiraud index. Although most of the literature on Lexical diversity is situated in evaluating L2 and EFL learners the measurement would be consistent

to look at learners' language from a culture perspective (Fergadiotis et al., 2015).

Other studies on LD have demonstrated its value in assessing language development and educational outcomes. For instance, Jarvis (2013) and Yang et al. (2022) highlight how lexical diversity metrics can serve as reliable indicators of language proficiency and cognitive engagement. These metrics provide a quantifiable measure of vocabulary usage that correlates with academic success, particularly in AI-enhanced learning environments where language plays a central role in interactive and adaptive learning systems. For instance, Gómez Vera et al. (2016) conducted a significant study examining the role of lexical quality in 4th-grade students' writing across different text genres—narrative, persuasive, and informative. They found that lexical diversity significantly affects narrative and persuasive texts, evidenced by a robust effect size ($F(1.97, 1338.70) = 360.19$, $p < 0.001$). Additionally, lexical density was shown to have a substantial impact across all genres, with an F-statistic of $F(1.94, 1318.61) = 715.02$, $p < 0.001$. These findings highlight the universal importance of lexical characteristics in influencing writing quality in educational settings.

Building upon the existing literature, a notable gap emerges in the intersectional exploration of Bloom's taxonomy, cognitive engagement levels, and demographic factors in shaping LD within AI-enhanced educational environments. While studies like Lissón & Ballier (2018) have refined our understanding of LD measures, less attention has been given to how these metrics interact with cognitive engagement strategies and demographic variables to influence student responses. This research aims to bridge this gap by integrating ICAP's active learning models with LD assessments, thereby providing a more holistic view of language use and cognitive engagement across diverse student demographics. This approach promises to enhance personalized learning experiences and offer deeper insights into the mechanisms driving effective language learning.

Building on the foundational insights from the literature review, the discussion now transitions to a focused examination of the theoretical frameworks that further examines the mechanisms behind cognitive engagement (Bloom's taxonomy and ICAP) and lexical diversity within AI-enhanced educational settings. We draw on three pivotal models to deepen our understanding of how AI tools can be leveraged to foster enriched learning experiences.

Theoretical Frameworks

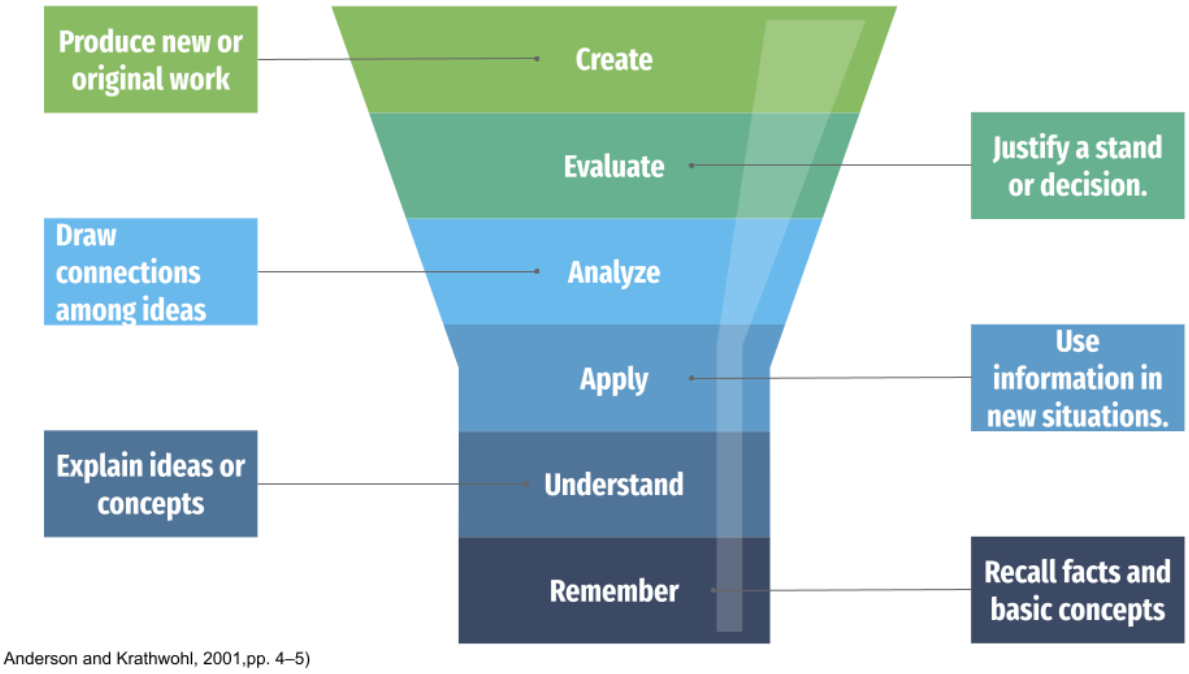
We draw on three theoretical frameworks to examine question-answers in this context: Revised Bloom's taxonomy (Anderson & Krathwohl, 2001), ICAP framework (Chi & Wylie, 2014) informed with Bloom's indicators and lastly, LD framework (Ransdell & Wengelin, 2003). In each of these theoretical lenses, CE plays a pivotal role in learners' learning experiences. This engagement varies, from psychological investment and effort in learning to the depth of cognitive processing learners undertake (Fredricks et al., 2004; Pintrich, 2000; Schunk et al., 2014; Craik & Lockhart, 1972). It can be viewed at both the school level, reflecting overall academic commitment, and the task level, emphasizing engagement in specific activities. Engaged learners not only absorb but actively participate, reflecting, and adapting in their learning journey (Chi & Wylie, 2014).

Bloom's Taxonomy

Bloom's taxonomy offers a structured framework for understanding the different levels of cognitive processes that learners engage in during the learning process. Originated by Bloom et al. (1956) and later revised by Anderson & Krathwohl (2001), this taxonomy outlines a hierarchy of six cognitive tasks ranging from basic recall ("remembering") to higher-order thinking skills such as analysis, evaluation, and creation (see Figure 2.1).

Figure 2.1

Funneling Cognitive Skills: A Visual Representation of the Revised Bloom's Taxonomy



At the Remembering level, Knowledge questions involve basic recall, utilizing verbs like 'know' and 'define,' requiring no creative thought but rather remembering knowledge from long-term memory. Comprehension questions live at the Understanding level which deepen understanding, with verbs such as discuss and compare through one or more forms of explanation. Apply level encompasses the Application questions that use information or a newly learned skill in a new situation, signaled by verbs like apply and demonstrate. Analysis (Analyzing level) involves dissecting information into parts to grasp its structure of how each part relates to one another or to an overall structure or purpose, using verbs like analyze and distinguish. Within the Evaluating level is Synthesis questions which encompasses decision-making and problem-solving to make judgments based on criteria and standards, employing verbs like assess and criticize. Finally, Evaluation demands creativity to generate original ideas, includes action like create and revise, representing the highest cognitive engagement with potentially multiple solutions (Anderson & Krathwohl, 2001; Bloom

et al., 1956).

By designing and categorizing questions based on revised Bloom's taxonomy, educators can purposefully target specific cognitive domains, thereby directing learners towards desired levels of thinking and understanding. Krathwohl (2002) emphasized its utility in educational settings for fostering deeper cognitive engagement. Interestingly, studies in various subjects, including English and Science, have utilized Bloom's taxonomy to categorize and improve the cognitive levels of questions posed in classrooms. However, as Fulford & Ng (2023) identified, there is an existing research gap in applying this taxonomy to foster higher-order cognitive skills, particularly in AI literacy learning, where emphasis is often on lower levels like 'Understanding' and 'Applying'.

ICAP Framework

The ICAP model by Chi & Wylie (2014) provides a detailed framework for understanding cognitive engagement based overt behaviors. It categorizes engagement into four modes based on learner involvement with learning activities from lowest engaged to highest: Passive. Active, Constructive, and Interactive. The Passive mode represents minimal engagement where learners may repeat or recall information. This is the lowest cognitive level. In contrast, Active mode involves integrating new information with prior knowledge, but within similar contexts. Constructive mode goes a step further, involving the generation of new ideas that extend beyond the initial information provided. The framework also has an Interactive mode, which requires mutual constructive communication and adequate back-and-forth interaction where learners will "co-create a tangible product that incorporates each student's ideas" (Wiggins et al., 2017, p.2). Wiggins et al. (2017) hypothesizes that as engagement with the activity increases in each mode, it is predicted that the learning gains will also increase. This model has found applications in various educational settings, guiding instructional strategies and shaping test designs (Chase et al., 2019; Henderson, 2019; Hsiao et al., 2022). For example, research by Morris & Chi (2020) demonstrated that after professional development interventions, teachers shifted their questioning methods from passive to more constructive and interactive formats, significantly enhancing the quality of classroom interactions.

Further studies, like those by Wang et al. (2016); Yogev et al. (2018), have utilized the ICAP model to assess higher-order thinking in online educational settings, such as MOOC discussions and virtual course student comments. In the specific context of AI education, the ICAP framework has been instrumental in guiding learners from passive information absorption to interactive and engaged learning modes, as seen in pharmacology education, where the application of ICAP led to more interactive and constructive pedagogies, significantly enhancing both student engagement and learning outcomes (Quesnelle et al., 2021).

Lexical Diversity

LD sheds light on the range of vocabulary used by speakers or writers, indicating not just linguistic capabilities but also speaker proficiency and potentially even socioeconomic status (Ransdell & Wengelin, 2003). LD is characterized by both the variety and repetitiveness of vocabulary within a text, as Carroll (1938) highlighted, suggesting that the diversity of vocabulary, influenced by the unique words and their frequency, should be normalized to the text's length to allow for fair comparisons.

Particularly, LD is measured by metrics such as the Number of Different Words (NDW), which plays a pivotal role in demonstrating vocabulary variation across language samples. This variation is crucial for indicating language development and literacy achievements in children (T. Thordardottir & Weismer, 2001; Price & Jackson, 2015; Farquharson et al., 2014; Miller et al., 2016). To address the influence of text length on LD measurements, Johnson (1944) proposed using the Type-Token Ratio (TTR), calculated either from a consistent number of words or from the average TTR of equal-sized text segments. McCarthy & Jarvis (2007) method for assessing textual LD, known as the Mean Length of Utterance (MLU), ensures consistency across different sample sizes.

Adjacent to LD is the concept of 'lexical richness' (LR), which, while similar, includes a weighting for the rarity of words (Daller et al., 2003). LD serves as a lens through which the interaction of language with societal, cultural, and historical contexts is explored, highlighting how

it reflects and shapes these dimensions (Araujo et al., 2018). It assesses the breadth of vocabulary usage, thus illuminating language through the variations in vocabulary, grammar, syntax, and dialects among speakers within a language or across different languages (Johansson, 2008). In essence, it evaluates the diversity of word choice in speech or writing (Malvern et al., 2004).

In educational settings, LD acts as a tool for assessing how variations in language proficiency among students affect the clarity and coherence of their responses. Unlike a qualitative, subjective analysis of content, style, or value, LD provides a quantitative, numerical evaluation of a text (Hampson & McKinley, 2023).

Methodology

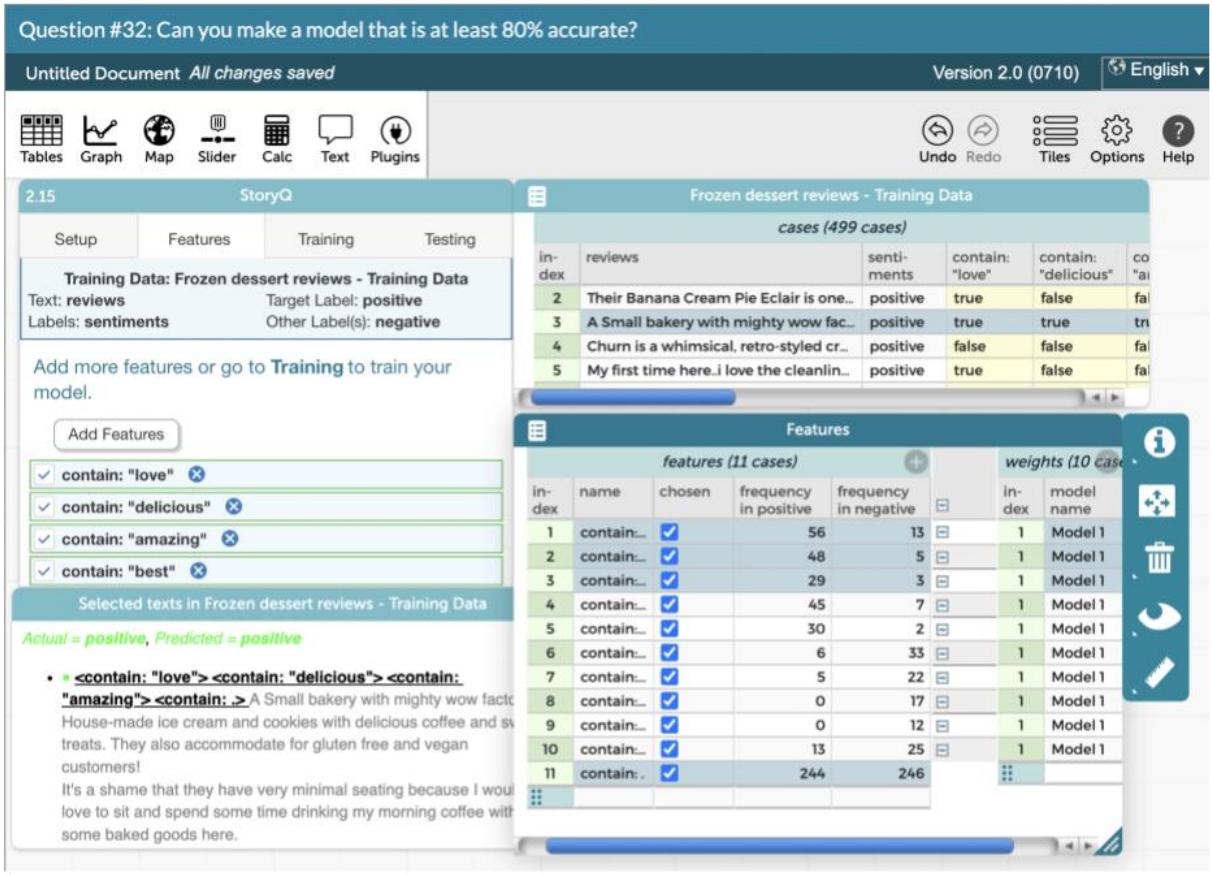
Data Collection

This study was conducted in a high school journalism class in the Northeastern U.S. After undergoing a four-week professional development workshop on AI, the teacher implemented the StoryQ AI curriculum (Chao et al., 2022). StoryQ (see Figure 2.2) is an online tool leveraging AI and machine learning designed for students in grades 6-12. It enables them to explore machine learning concepts and interact with unstructured text data without the need for coding (Chao et al., 2022). This curriculum comprises eight distinct modules that demand no prior coding skills. Opting to integrate seven of these modules, the teacher incorporated them into her three-week journalism course, spanning 45 minutes daily. The elective course attracted twenty-eight learners from various grade levels and ethnic backgrounds, including a majority of 12th graders, followed by 11th and 10th graders. The class demographic was primarily Black/African (54%), with representation from Hispanic (15%), White/Caucasian (20%), and other backgrounds. Learners from diverse grade levels and backgrounds participated in this class, making the dataset rich and varied.

In each session, the teacher introduced AI concepts, followed by hands-on student activities working either individually or in pairs. The hands-on activities included specific questions, both multiple choice, and open-ended. The curriculum also included pre- and post-test and student interviews. For our study we looked at 30 open-ended questions per student, focusing on three

Figure 2.2

StoryQ, a web-based platform for machine learning model exploration and development



modules: Sentiment Analysis, Features and Models, and All Words as Features (n = 840). For instance, in “Sentiment Analysis” learners were asked to evaluate negative or positive reviews with questions like “What features do you think are indicators of negative reviews?” In the "Features and Models" activity learners explored feature engineering and answered questions like how the feature "great" influenced a model. In “All Words as Features” questions probe learners to analyze why the model learned a particular “feature” and its corresponding “weight.”

Data Preprocessing

The data was cleaned by removing rows that did not include a cognitive engagement label. These rows meant that learners did not answer the question and therefore were removed. We also looked for and addressed any empty cell within other variables (gender, race) by including NaN.

To process the textual data in the questions and answers we first removed extra white space, double quotes, back slashes, and characters like emoji or diamonds. We did not remove stop words as we wanted to keep the learners semantic integrity. We then segmented into individual words or tokens. Feature engineering included creating new variables for lexical diversity and answer length analysis. After data processing we ended up with 684 observations.

Analysis

How We Measured Student Engagement

We used the ICAP framework to understand the depth of learners' engagement with their learning materials. However, our study did not encompass Interactive as we did not have collaborative student answers, leading us to exclude the Interactive category from our analysis. The student responses were evaluated based on the adapted ICAP framework, which only included the Constructive, Active and Passive (CAP) levels, omitting Interactive as co-creation and back and forth between learners did not occur. Using a specific scoring rubric where Passive (P) was coded as 0, Active (A) coded as 1 and Constructive (C) coded as 2 (see Table 2.1). It is important to note that if a student did not answer it was initially coded as passive but later removed from the data as it was missing. Initially, two graduate researchers familiarized themselves with the Cognitive Engagement coding system. They then independently assessed 40 observations, after which they calculated the inter-rater reliability using Cohen's Kappa (Cohen, 1968), achieving a score of 65%. To address any discrepancies in their evaluations, they collaboratively assessed another set of 40 observations discussing any disagreements. Subsequently, they individually evaluated 10% of the responses, achieving a satisfactory reliability score (Cohen's kappa .844). Following this, the lead

evaluator took on the task of grading the remaining self-written responses, amounting to 840 in total.

Table 2.1

Cognitive Engagement Label Coding Scheme

Score	ICAP level	Description	Revised Bloom's Indicator	Example
2	Constructive	New information is integrated with activated prior knowledge, and new knowledge is inferred.	Evaluate and Create	"You can look at the data set and find words that really stand out to you or words that have a strong emotional connotation. You can also check the graph and the probability in terms of the features being used or how strongly they correlate with the result."
1	Active	Behaviors that cause-focused attention while manipulating	Apply and Analyze	"You can test multiple reviews with words that you think may be the features to determine if they are actually features."
0	Passive	Overt activities that are carried out mindlessly	Remember and Understand	"Guess words that would most likely fit."

How we measured Linguistic Diversity

We adopted McCarthy's MTLD (2007) as our primary method for assessing linguistic diversity within student responses. The MTLD approach is designed to provide a consistent and

reliable measure of lexical diversity across text samples of varying lengths, thereby addressing the limitations associated with traditional measures such as the Type-Token Ratio (TTR). Specifically, given a text of N words, the MTLD calculation involves dividing the text into sequential tokens and calculating the TTR for these segments. The process continues either forward or backward until the TTR falls below a predetermined threshold, typically set at 0.72. The MTLD value is then computed as the average length of these segments, providing a measure of lexical diversity that is less sensitive to text length. Mathematically, it is represented as:

$$\text{MTLD} = \frac{\text{Total length of the text (in words)}}{\text{Number of segments with TTR} > \text{Threshold}}$$

Statistical Analysis

To analyze the data, we utilized mixed linear models to explore the effects of Bloom's taxonomy categories and cognitive engagement levels on lexical diversity, adjusting for demographic factors such as race and grade. The dependent variable in our model was the logarithm-transformed MTLD score, which normalized the distribution of lexical diversity measures. Independent variables included categorical representations of Bloom's taxonomy levels and engagement scores, along with demographic covariates. Random effects were included to account for intra-student variability, recognizing that multiple responses from the same student are not independent observations.

Additionally, the chi-square tests facilitated an exploration of how demographic factors intersect with cognitive engagement and lexical diversity. By segmenting the data according to demographic variables and applying chi-square tests, we sought to detect significant associations or disparities in engagement and lexical diversity outcomes across diverse student groups.

Results and Discussion

RQ1: How do Bloom's taxonomy categories, cognitive engagement levels, and demographic factors influence lexical diversity in student responses?

Variations in Cognitive Engagement and Bloom's Taxonomy (see Figure 2.3)

We found significant variation in cognitive engagement across different Bloom's taxonomy categories, consistent with the theoretical expectations posited by Anderson et al. (2001). Active engagement predominates in tasks related to Applying, Creating, and Remembering, while Evaluation tasks tend towards passive engagement. This pattern aligns with our theoretical discussions, particularly the implications of task complexity on engagement levels as delineated within the framework of Bloom's taxonomy. Analyzing tasks exhibit a balanced distribution between Passive and Active engagements, with minimal Constructive engagement.

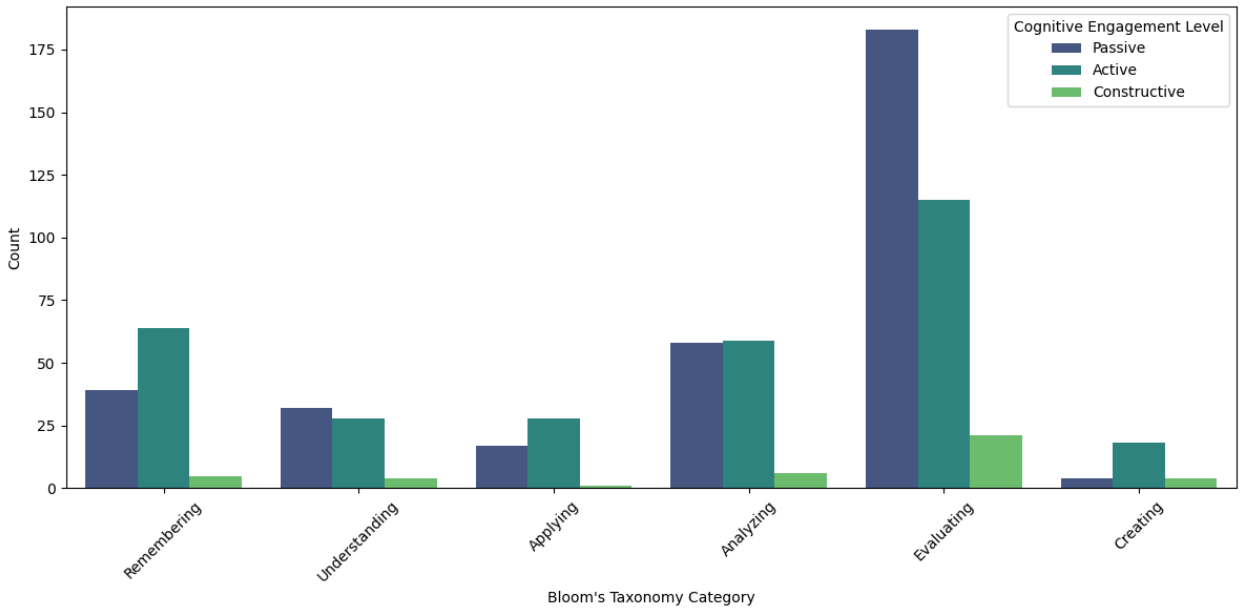
The Remembering category was characterized by a significant portion of Active engagement (59.3%), with Passive engagement at 36.1% and Constructive engagement at 4.6%. This pattern highlights the active engagement that recall tasks tend to elicit from students. Responses in the Understanding category were split between Passive (50.0%) and Active (43.8%) engagements, with a small percentage of Constructive engagement (6.2%). This indicates a tendency towards both passive understanding and active engagement in comprehension tasks.

A majority of responses in the Applying category showed Active engagement (60.9%), followed by Passive engagement (37.0%), and a minimal amount of Constructive engagement (2.2%). This indicates that application tasks predominantly stimulate active involvement. In the Analyzing category engagement was nearly evenly split between Passive (47.2%) and Active (48.0%), with a small portion of Constructive engagement (4.9%). This distribution suggests that tasks requiring analysis prompt a balanced level of engagement, though deeper, constructive engagement is less common.

The Evaluating category had the highest percentage of Passive engagement (57.4%), followed by Active (36.1%) and Constructive (6.6%) engagements. It appears that evaluating tasks

Figure 2.3

Distribution of Cognitive Engagement by Bloom's Taxonomy Category



are more likely to result in passive engagement, with less active and constructive involvement. In the Creating category, Active engagement was most prevalent (69.2%), with equal but lower representations of Passive and Constructive engagements (15.4% each).

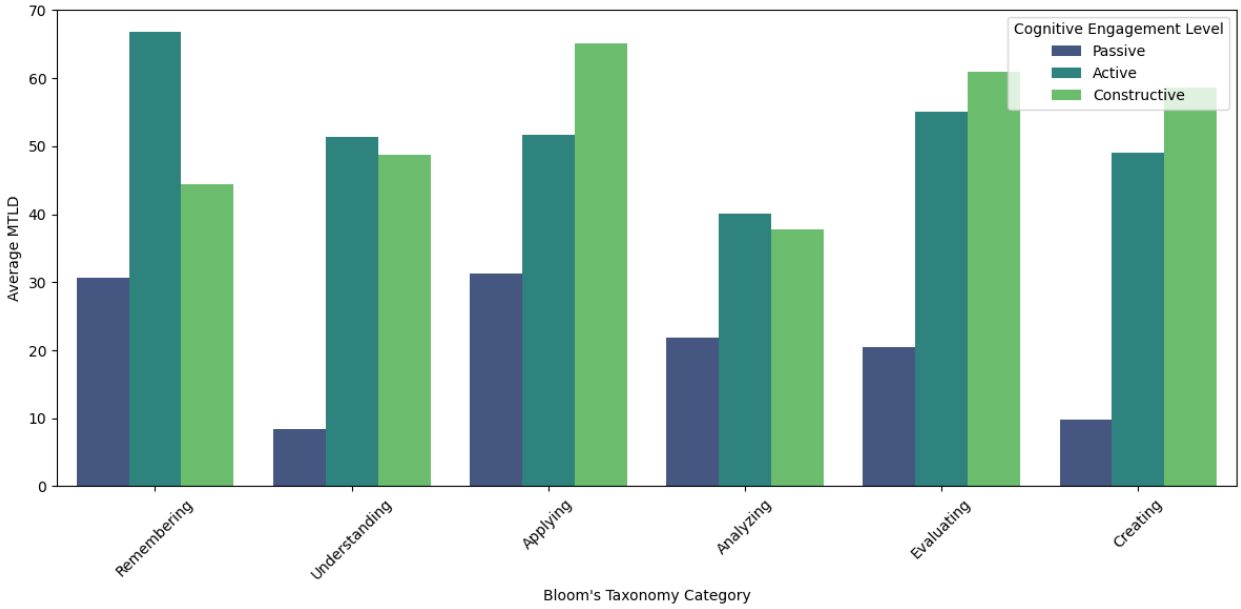
These findings align with the principles outlined by Tatar et al. (2023) and Zimmermann-Niefield et al. (2019), emphasizing the necessity of designing curricula that not only introduce AI concepts but also actively engage students in higher-order thinking processes to foster critical thinking and problem-solving skills. This suggests that creative tasks strongly encourage active engagement while also allowing for both passive reception and constructive contributions.

Relationship between Bloom's Taxonomy Categories, Cognitive Engagement, and Lexical Diversity

Analysis of the relationship between Bloom's taxonomy categories, cognitive engagement levels, and lexical diversity revealed intriguing patterns. These findings are in line with the hypotheses derived from studies such as those by Jarvis (2013) and Mackey & Gass (2011), which suggest

Figure 2.4

Average MTLD by Bloom's Taxonomy Category and Cognitive Engagement Level



a link between the complexity of cognitive tasks and the richness of linguistic expression observed in student responses. Tasks involving Applying, Creating, and Evaluating showed increased lexical diversity with higher levels of cognitive engagement. Remembering tasks exhibited a peak in lexical diversity with Active engagement, while Understanding tasks did not show a clear pattern between Active and Constructive engagement. (See Figure 2.4). This relationship underscores the importance of task design in promoting lexical diversity, a key component of language development, and highlights the potential of AI-enhanced curricula to support this aspect of student learning (Dwivedi et al., 2021).

The Remembering category presents a unique pattern where Active engagement (66.77) significantly surpasses both Passive (30.66) and Constructive (44.41) levels in terms of MTLD, suggesting that Active engagement in Remembering tasks is particularly conducive to lexical diversity. Lexical diversity scores for Understanding tasks are lowest for Passive engagement (8.47), with a notable increase for Active (51.32) and a slight decrease for Constructive engagement (48.76).

Scores in the Applying category increase across engagement levels from Passive (31.35)

to Active (51.76) and peak at Constructive engagement (65.15), indicating that Applying tasks with Constructive engagement yield the highest lexical diversity. The average MTLD scores in the Analyzing category are 21.92 for Passive, 40.07 for Active, and 37.75 for Constructive engagement. This suggests that Active engagement in Analyzing tasks is associated with higher lexical diversity compared to Passive and Constructive engagement.

Similar to Applying and Creating, Evaluating tasks show an increase in MTLD from Passive (20.52) to Active (55.10) and Constructive (60.97) engagement, underscoring the importance of higher cognitive engagement for lexical diversity. Finally, within the Creating category, starting with the lowest average MTLD for Passive engagement (9.75), there's a significant increase for Active (49.07) and Constructive (58.59) engagement levels, highlighting a strong link between higher engagement and lexical diversity in creative tasks.

The clear linkage between higher levels of cognitive engagement and increased lexical diversity supports the assertion that engaging students in complex cognitive tasks not only enhances their understanding of AI concepts but also contributes to their linguistic development (Anderson et al., 2001; Bloom et al., 1956).

Statistical Analysis

We conducted an analysis using a mixed linear model to explore the impact of Bloom's taxonomy categories and cognitive engagement levels on the lexical diversity of student responses, with lexical diversity measured by the logarithm of the Measure of Textual Lexical Diversity (MTLD + 1) to ensure normality of residuals (see Table 2.2). This analysis was carried out using Python's statsmodels library, considering the transformed MTLD (log_mtld) as the dependent variable. Our model included fixed effects for Bloom's taxonomy categories and cognitive engagement levels, distinguishing between "Passive", "Active", and "Constructive" engagement. Additionally, we accounted for individual differences among students by including user_id as a random effect, acknowledging the non-independence of observations from the same individual. The model specification was as follows:

$$\begin{aligned}\text{MTLD}_{ij} = \beta_0 + \beta_1 \text{Bloom's Type}_{ij} + \beta_2 * \text{Active engagement}_{ij} \\ + \beta_3 * \text{Constructive engagement}_{ij} + \beta_4 * \text{Race}_{ij} + \beta_5 * \text{Grade}_{ij} \\ + u_j + \varepsilon_{ij}\end{aligned}$$

Where log_MTLD_{ij} represents the transformed lexical diversity score for the *i*th response from the *j*th student, β_0 is the intercept, β_1 , β_2, and β_3 are the coefficients for the effects of Bloom's taxonomy categories and "Active" and "Constructive" cognitive engagement levels, β_4 and β_5 account for the effects of race and grade respectively, u_j represents the random effect associated with the *j*th student, and εij is the error term. This model formulation allowed us to investigate the association between cognitive task types, levels of engagement, and variations in lexical diversity, alongside demographic factors like race and grade.

Our analysis, underpinned by the ICAP framework (Chi & Wylie, 2014), revealed significant associations across different Bloom's taxonomy categories and cognitive engagement levels with transformed MTLD scores. The statistical evidence, particularly the significant positive effects for 'Active' and 'Constructive' engagement, corroborates our initial hypothesis that higher levels of cognitive engagement, as theorized by Chi & Wylie (2014), are conducive to greater lexical diversity. This finding underscores the potential of AI-enhanced curricula to cultivate a rich linguistic environment by engaging students in active and constructive learning processes. Key findings include significant positive effects for "Active" engagement ($Coef. = 1.123$, $z = 16.110$, $p < 0.001$) and "Constructive" engagement ($Coef. = 1.188$, $z = 8.260$, $p < 0.001$), demonstrating that higher levels of cognitive engagement correlate with increased lexical diversity. This indicates the linguistic richness of student responses is significantly enhanced when they are more actively or constructively engaged with the material.

Moreover, tasks categorized under "Applying" ($Coef. = 0.363$, $z = 2.564$, $p = 0.010$) and "Remembering" ($Coef. = 0.340$, $z = 3.141$, $p = 0.002$) also showed a positive association with lexical diversity. The "Applying" category's significance suggests that tasks requiring students to use

information in new situations not only engage them cognitively but also encourage a more diverse use of language. Similarly, "Remembering" tasks, which involve recalling facts, basic concepts, or answers, contribute positively to LD, potentially because they allow for a range of expression in how information is recalled and presented.

Table 2.2

Mixed Linear Model Results for Predicting Lexical Diversity

Variable	Coefficient (B)	Standard Error (SE)	z-value	p-value	95% CI
FIXED EFFECTS					
Intercept	-8.187	36.862	-0.222	0.824	[-80.435, 64.060]
Bloom's Taxonomy Categories					
Applying	**0.363**	**0.142**	**2.564**	****0.010**	[0.086, 0.641]
Creating	0.13	0.178	0.727	0.467	[-0.220, 0.479]
Evaluating	0.094	0.087	1.075	0.282	[-0.077, 0.265]
Remembering	**0.34**	**0.108**	**3.141**	*****0.002**	[0.128, 0.552]
Understanding	**-0.357**	**0.126**	**-2.819**	*****0.005**	[-0.604, -0.109]
Cognitive Engagement Levels					
Active (vs. Passive)	**1.123**	**0.07**	**16.11**	*****<.001**	[0.987, 1.260]
Constructive (vs. Passive)	**1.188**	**0.144**	**8.26**	*****<.001**	[0.906, 1.470]
Race (vs. Black/African American)					
Hispanic	-0.142	0.144	-0.987	0.324	[-0.425, 0.140]
Latinx	0.33	0.247	1.334	0.182	[-0.155, 0.814]
Other	0.113	0.246	0.461	0.645	[-0.368, 0.595]
Prefer not to answer	-0.279	0.238	-1.17	0.242	[-0.745, 0.188]
Unknown	-0.548	0.292	-1.878	0.06	[-1.119, 0.024]
White/Caucasian	0.196	0.132	1.487	0.137	[-0.062, 0.455]
Grade	0	0	-0.953	0.341	[-0.001, 0.000]
RANDOM EFFECTS					
Group Var (user_id)	0.029	0.024			

Note: Significance levels are indicated as $*p < .05$, $**p < .01$, $***p < .001$.

Conversely, the "Understanding" category exhibited a negative association ($Coef. = -0.357$, $z = -2.819$, $p = 0.005$), indicating a more complex relationship between this type of cognitive

task and lexical diversity. It suggests that tasks focusing on understanding may limit the range of language students use, possibly due to the nature of the tasks which may require more direct answers or explanations. Furthermore, the inclusion of race and grade as covariates provided nuanced insights, though these factors did not show universally significant effects, suggesting that the impact of demographic factors on lexical diversity might be complex and warrants further investigation.

The analysis also accounted for considerable variability between students (Group Var $\bar{0}$.029), highlighting the role of individual differences in shaping lexical diversity. This variability underscores the importance of considering individual student characteristics when analyzing the complexity of language use in educational contexts. The presence of significant variability among students suggests that personal factors, possibly including prior knowledge, language proficiency, and engagement styles, play a critical role in determining the lexical diversity of their responses.

RQ2: How does module/question complexity and sequence influence student engagement levels and LD?

Impact of Module/Question Progression on Cognitive Engagement

In examining how the sequence and complexity of modules or questions affect cognitive engagement levels, a Chi-square test of independence indicated a significant association between engagement level and Bloom's taxonomy types. This statistical finding underscores the curriculum's design's effectiveness, as informed by the principles of the StoryQ curriculum (Chao et al., 2022), reflecting its intent to engage students at varying cognitive levels across the curriculum. The analysis revealed a significant association between engagement level and Bloom's taxonomy type, $\chi_2(10) = 38.11$, $p = .000036$, indicating that the type of cognitive process as defined by Bloom's taxonomy influences the engagement level of students.

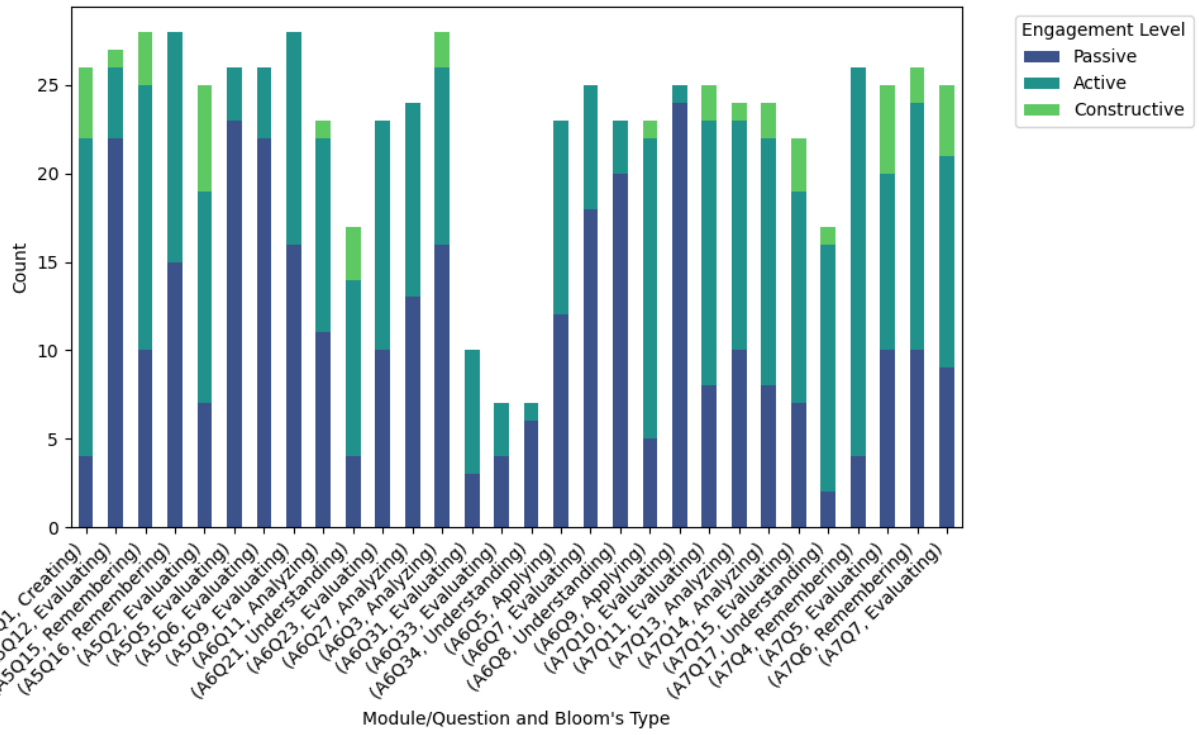
The expected frequencies under the null hypothesis of no association were calculated for each category of engagement level and Bloom's taxonomy type. For questions categorized under Remembering, the expected frequencies of engagement levels were 59.71 for Passive, 55.94 for Active, and 7.35 for Constructive. Similar distributions were observed across the other Bloom's

taxonomy types, suggesting a deviation from these expected frequencies in our observed data, contributing to the significant chi-square statistic.

This significant association underscores the importance of question design in promoting different levels of cognitive engagement among students. Specifically, it suggests that tasks requiring higher-order cognitive processes (e.g., Evaluating, Creating) are more likely to engage students at a Constructive level, while those focused on lower-order processes (e.g., Remembering) may not elicit the same depth of engagement. This highlights the significance of thoughtful question design in fostering varying levels of cognitive engagement, thereby enhancing the educational experience by aligning with students' cognitive processes.

Figure 2.5

Engagement Level Distribution by Module/Question



Identification of Modules/Questions Promoting Higher Engagement and or Higher LD

We identify specific modules or questions that consistently foster higher levels of engagement and explore how these findings can inform instructional design strategies. We employ a pivot table to aggregate MTLD scores by module and question, calculating the mean MTLD for each question within a module. This approach facilitates a module-by-module and question-by-question comparison of lexical diversity across the curriculum. The heatmap visualization (see Figure 2.6) reveals significant variations in LD scores as measured by MTLD across different modules and questions. Modules and questions with notably higher MTLD scores are indicative of tasks that challenge students to utilize a broader vocabulary, potentially signifying more engaging or complex cognitive demands. Conversely, lower MTLD scores may highlight areas where the curriculum could benefit from enhancements to foster greater lexical diversity and engagement.

These findings reveal that modules and questions that generate higher MTLD scores can serve as models for developing tasks that stimulate deeper cognitive processing and language use. On the other hand, areas of the curriculum associated with lower MTLD scores indicate opportunities for improvement, suggesting that these sections may benefit from incorporating more challenging tasks or questions to promote greater lexical diversity and engagement.

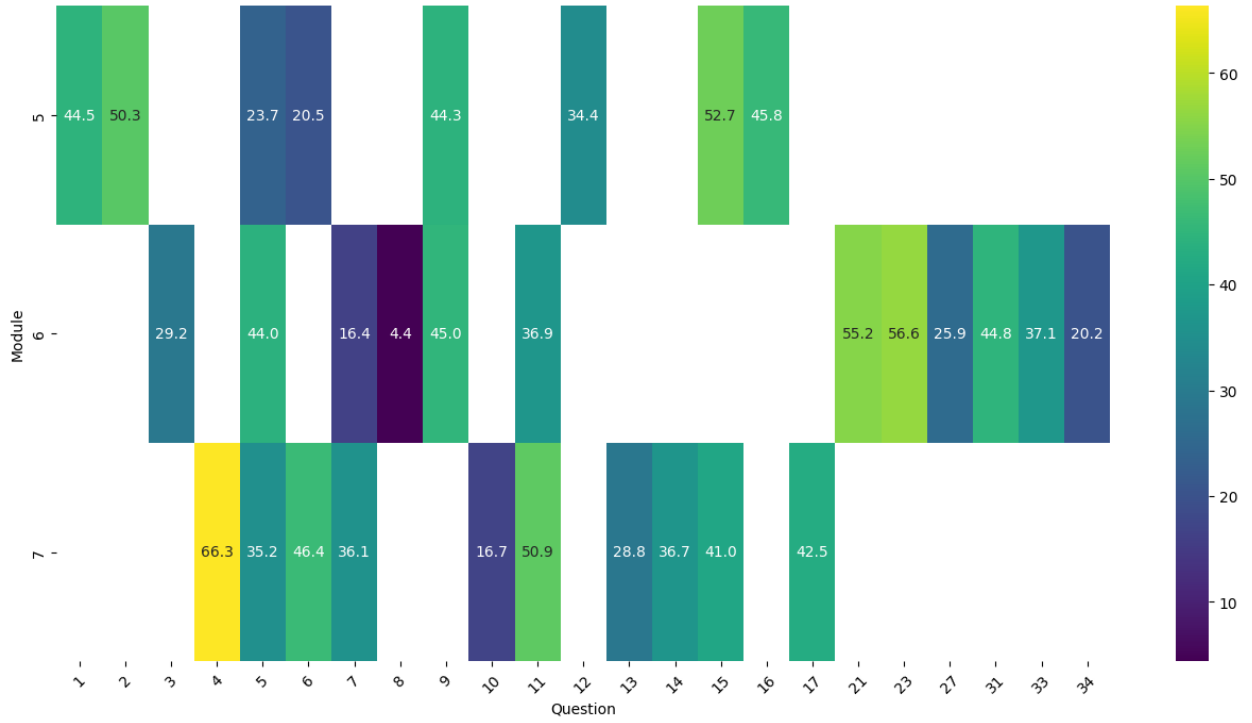
RQ3: How do demographic factors, race/ethnicity, intersect with Bloom's taxonomy categories and cognitive engagement levels to shape lexical diversity in student responses?

Intersection of Demographic Factors with Cognitive Engagement and Lexical Diversity

In our exploration of how demographic factors, such as race/ethnicity (see Figure 2.7), intersect with Bloom's taxonomy categories and cognitive engagement levels to influence lexical diversity, our findings contribute to the growing body of research on this topic. The significant variations in lexical diversity among different racial/ethnic groups highlight the nuanced impact of these demographic factors, suggesting areas for further investigation within the broader context of educational equity and inclusivity. The analysis reveals significant variations in lexical diversity

Figure 2.6

Heatmap of MTLD Scores by Module and Question number



among different racial/ethnic groups.

White/Caucasian participants (n = 129) exhibited the highest average lexical diversity, with a mean MTLD score of 50.99 and a standard deviation (SD) of 48.64, indicating considerable variability. The median MTLD value for this group was 44.80, suggesting a central tendency towards higher lexical diversity. Scores ranged from 1 to 269.08, with a mode of 10, highlighting a broad distribution of lexical richness.

The Black/African American group (n = 304) demonstrated moderate lexical diversity with a mean MTLD of 37.79 and an SD of 35.15, reflecting substantial variability within this group. The median score was 30.81, and the range of MTLD scores spanned from 1 to 252. The mode of 11.0 suggests a common lower bound of lexical diversity among a significant number of responses.

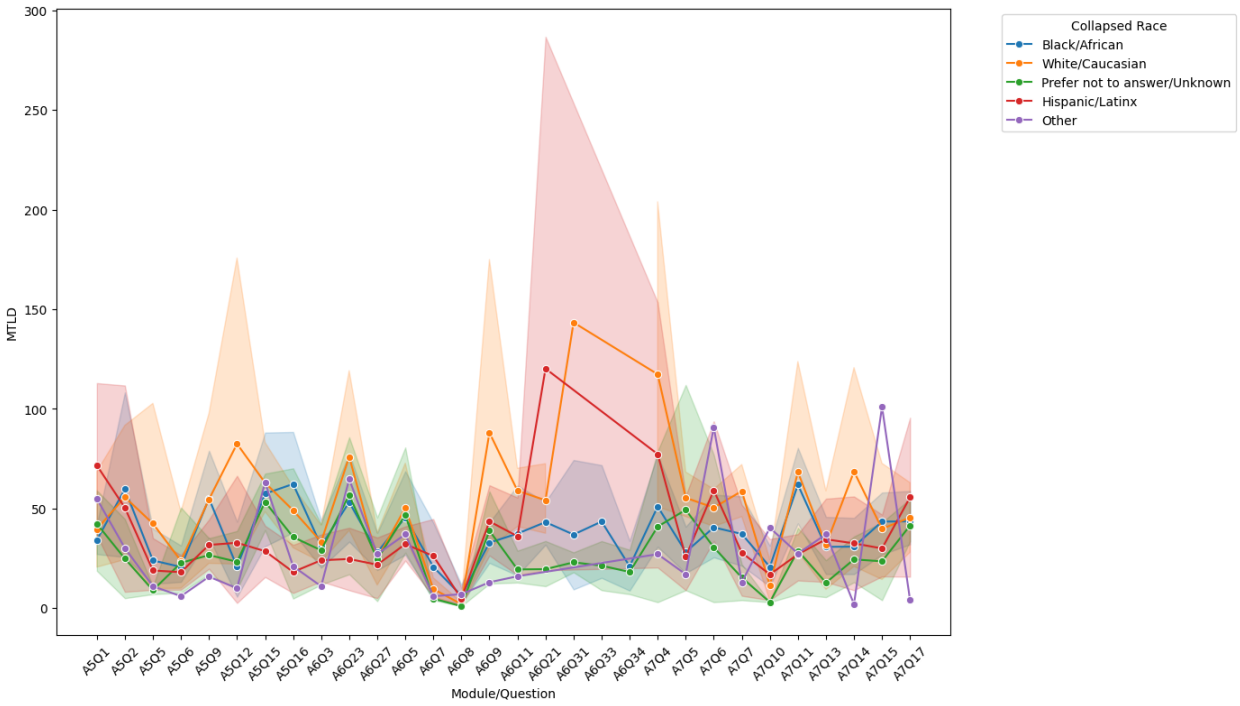
Hispanic/Latinx respondents (n = 119) showed an average MTLD of 35.77 with an SD of 41.34, indicating a high degree of variability. The median MTLD of 23.36 and a range of 1 to

286.72 point to a wide dispersion of scores. The mode of 9.0 implies a frequent lower level of lexical diversity.

For the Other category (n = 26), the mean MTLD was 28.98, with an SD of 26.34, suggesting moderate variability. The median value stood at 19.00, and scores ranged from 2 to 101.08. The mode of 6.0 indicates a concentration of responses towards lower lexical diversity.

Figure 2.7

MTLD Progression by Module/Question and Race



The Prefer not to answer/Unknown group (n = 108) presented the lowest mean MTLD of 27.38, with an SD of 25.67, indicating moderate variability. The median MTLD was 19.08, with scores ranging from 1 to 112. The mode of 1.0 suggests that the lowest lexical diversity was the most common among responses in this category.

Discussion and Implications

This study set out to explore the interplay between CE levels, Bloom's taxonomy categories, and lexical diversity in student responses within a Machine Learning Curriculum, underscored by with a focus on the evolving role of AI in education as delineated by Dumont & Zurn (2007); Dwivedi et al. (2021), and others. Through rigorous analysis, our findings contribute nuanced insights into the complex dynamics at play, affirming and extending upon the theoretical underpinnings posited by Anderson & Krathwohl (2001), Chi & Wylie (2014), and frameworks exploring linguistic diversity (Ransdell & Wengelin, 2003; Jarvis, 2013).

The variance observed in LD across different demographic groups within our study underscores the imperative of inclusivity in curriculum design, highlighting the transformative potential of leveraging Natural Language Processing (NLP) in educational research. This discrepancy in LD necessitates pedagogical strategies that actively bridge linguistic diversity gaps, ensuring equitable learning opportunities for all students. By prioritizing demographic inclusivity at the core of AI-enhanced learning environments, educators and curriculum developers commit to a framework that respects and responds to the diverse backgrounds and cognitive abilities of students, thereby fostering a truly inclusive educational landscape (Fredricks et al., 2004; Pintrich, 2000).

Moreover, our findings on the engagement levels across Bloom's taxonomy categories resonate with the ICAP framework's delineations, revealing significant variation in CE with active engagement being particularly predominant in tasks related to Applying, Creating, and Remembering. This pattern not only supports the notion that different types of cognitive tasks elicit varied levels of engagement but also highlights the potential of sophisticated AI-enhanced curriculum designs to foster deeper cognitive and linguistic engagement among students (Chi & Wylie, 2014; Tatar et al., 2023). The intricate relationship between cognitive engagement and LD, especially the heightened LD associated with active and constructive engagements, underscores the linkage between engagement depth and linguistic richness, echoing the insights provided by Jarvis (2013) and Mackey & Gass (2011) regarding the role of task complexity in fostering rich linguistic outputs.

This comprehensive analysis points to the necessity of inclusive and adaptive curriculum designs that cater to the unique linguistic and cognitive profiles of a diverse student body. The significant associations found between engagement levels, Bloom's taxonomy types, and LD through our statistical analysis further validate the efficacy of AI-curriculum design principles that embrace a wide spectrum of cognitive task complexity. Such an approach not only enriches the educational experience by promoting critical thinking and problem-solving skills but also highlights the crucial role of AI in tailoring educational content to meet individual learning needs, thereby maximizing cognitive engagement and linguistic diversity across diverse student populations (Jarvis, 2013; Mackey & Gass, 2011).

In light of these insights, the critical need for educational interventions that address language development disparities among students from different racial/ethnic backgrounds becomes evident. Implementing diverse and inclusive instructional strategies that cater to the unique needs of each student group could significantly mitigate these disparities. Moreover, our findings emphasize the importance of further research to explore the underlying causes of these differences and to develop pedagogical approaches that promote linguistic richness and cognitive engagement equitably across all demographics. Ultimately, addressing the variations in LD is essential for advancing educational equity, ensuring that every student has the opportunity to deeply engage with the curriculum and realize their full academic potential, thereby contributing to a more equitable and effective educational landscape (Fredricks et al., 2004; Pintrich, 2000).

Thus, these findings can guide educators in creating more effective and engaging learning experiences by aligning instructional strategies with the cognitive demands that drive engagement and language development. The implications of this study for educational practice and policy are multifaceted, touching on curriculum design, inclusivity in pedagogy, professional development, and the need for ongoing research. Firstly, the emphasis on curriculum design suggests a strategic integration of tasks across Bloom's taxonomy, advocating for the use of AI to craft dynamic and engaging learning experiences. Such an approach aims not only to deepen cognitive engagement but also to enhance students' linguistic skills by accommodating a variety of cognitive abilities and

learning styles. This is pivotal in fostering a learning environment where students are challenged both cognitively and linguistically.

Secondly, the study highlights the critical importance of inclusive pedagogy, as evidenced by the observed demographic variations in LD. It calls for instructional strategies that are attuned to the linguistic and cultural diversity of the student body. Educational policies and practices must strive to diminish disparities by embedding diverse linguistic and cognitive tasks within the curriculum, thereby ensuring a fair and equitable learning experience for students from all backgrounds.

Professional development emerges as another crucial area, underscoring the need for programs that prepare educators to effectively integrate AI into their teaching methodologies. These programs should concentrate on fostering pedagogical approaches that promote active and constructive engagement, thereby enhancing both CE and LD. A significant focus should be placed on creating inclusive learning environments that celebrate and utilize the diverse backgrounds of students, preparing educators to meet the needs of all learners.

Lastly, the study opens avenues for further research in the domain of AI-enhanced education. It encourages future investigations into the long-term effects of cognitive engagement and LD within AI-supported curricula, the exploration of personalized learning paths enabled by AI, and the potential of collaborative AI technologies to facilitate interactive student engagement. This direction for future research underlines the evolving nature of educational technology and its impact on teaching and learning processes, urging a continual reassessment of how AI can be harnessed to improve educational outcomes.

In acknowledging the limitations of this work, it's important to recognize that the findings, while insightful, may not be broadly generalizable across different educational contexts or demographic groups. The study's specific setting, within a single high school journalism class employing the StoryQ AI curriculum, offers a focused but narrow lens through which to examine the interplay between CE, LD, and AI in education. This specificity, coupled with the cross-sectional design and primary reliance on quantitative measures, limits our ability to infer causality or fully capture the qualitative dimensions of student engagement and learning experiences. Additionally, the adaptation

of the ICAP framework, excluding its Interactive category, may not encompass the full spectrum of student engagement facilitated by collaborative AI technologies. These constraints underline the need for cautious interpretation of the results and highlight the importance of future research that broadens the scope, employs longitudinal designs, and integrates qualitative insights to more comprehensively understand and leverage AI's potential in diverse educational settings.

While this study provides valuable insights, it also opens avenues for further research. Future studies could explore longitudinal effects of engagement and linguistic diversity within AI-enhanced curricula, potentially incorporating interactive tasks to examine the role of collaborative learning. Additionally, further research might delve into personalized learning pathways that leverage AI to adaptively enhance CE and LD based on individual student profiles.

Conclusion

In conclusion, this study bridges a crucial gap in the literature by elucidating the relationships between CE, Bloom's taxonomy, and LD in the context of AI-enhanced education. Our findings affirm the transformative potential of AI in enriching the educational landscape, not only by equipping students with vital future-ready skills but also by fostering an inclusive environment that celebrates linguistic and cognitive diversity. As we move forward, the insights gleaned from this research will undoubtedly inform the development of more engaging, equitable, and effective educational practices that leverage the power of AI to meet the diverse needs of learners.

Chapter 3: Leveraging Prompts in LLMs to Overcome Imbalances in Complex Educational Text Data

Abstract

Background: The study addresses the challenge of imbalances in educational datasets, which is prominent in the education sector due to the varied cognitive engagement levels among students in their open responses. Traditional machine learning (ML) models often struggle with the complexity and nuanced nature of this data, leading to inadequate analyses, especially for minority data representations (Karimah & Hasegawa, 2022; Radwan & Cataltepe, 2017; Yun et al., 2011). Understanding students' cognitive engagement is vital as it reflects their mental investment in learning activities, which is closely linked to academic success (Fredricks et al., 2004; Blumenfeld et al., 2006; Corno & Mandinach, 1983; Pintrich, 2000; Schunk et al., 2014). Objective: The objective of this paper is to investigate the efficacy of Large Language Models (LLMs) enhanced with assertions in tackling the complexities of imbalanced educational datasets, with a special focus on the precise classification of cognitive engagement levels from student texts. This exploration is underpinned by two critical research questions. The first seeks to evaluate how LLMs equipped with Prompt Engineering fare in comparison to conventional ML algorithms when dealing with the inherent challenges of imbalanced educational data. The second question delves into the specific contributions of integrating assertions into LLMs, examining how such augmentations can improve the models' effectiveness in handling the nuanced difficulties presented by imbalanced textual educational datasets. Through this inquiry, the study aims to shed light on the potential of LLMs and assertions in enhancing the accuracy and reliability of cognitive engagement classification, thereby addressing a significant gap in educational data analysis.

Methods: The study employed an 'Iterative - ICL PE Design Process' to compare traditional ML models against LLMs augmented with assertions (N=135). A sensitivity analysis on a subset (n=27) examined variance in model performance concerning classification metrics and cognitive engagement levels. This process involved the utilization of assertion-based prompt

engineering, comparing the performance of traditional ML models to LLMs with assertions in classifying cognitive engagement from student texts in an educational setting (Shahriar et al., 2023; Brown et al., 2020; Wei et al., 2022a).

Findings: LLMs with assertions significantly outperformed traditional ML models, especially in recognizing cognitive engagement levels with minority representation, showing up to a 32% increase in F1-score. Incorporating targeted assertions into the LLM on the subset enhanced its performance by 11.94%, primarily addressing errors from limitations in understanding context and resolving lexical ambiguities in student responses.

Implications: The study demonstrates the superior capability of LLMs, particularly when augmented with assertions, in addressing the nuanced challenges of imbalanced educational datasets. This advancement not only improves the accuracy of classifying cognitive engagement levels but also opens new avenues for data-driven educational research and practice. The findings suggest a potential paradigm shift towards employing advanced LLM techniques in educational settings to achieve a more nuanced and accurate analysis of student engagement, thereby enhancing learning outcomes. Future research should further explore the capabilities of LLMs across broader educational contexts and investigate additional methods to refine and expand their application in analyzing complex educational data (Shahriar et al., 2023; Zeng et al., 2023).

Introduction

Understanding students' cognitive engagement (CE) at both the school and task levels is crucial, as it offers deep insights into their commitment to learning (Fredricks et al., 2004). This form of engagement, characterized by a student's deliberate and intentional approach to schoolwork and their willingness to invest the necessary effort in comprehending complex concepts and mastering challenging skills, serves as a key indicator of academic success (Fredricks et al., 2004; Blumenfeld et al., 2006). CE encompasses the psychological investment and effort driven by student motivation and strategies, alongside their dedication to learning (Corno & Mandinach, 1983; Fredricks et al., 2004; Pintrich, 2000; Schunk et al., 2014).

While analyzing students' CE is crucial for enhancing learning experiences, a significant challenge arises from imbalanced datasets (Radwan & Cataltepe, 2017). These datasets often feature unevenly distributed categories and are typically small, not fitting the 'big data' criteria usually required for effective Machine Learning (ML) training. This size limitation, along with the disproportionate representation of majority and minority data, further complicates the training process in traditional analyses (Yun et al., 2011). Traditional ML methods, commonly employed to classify CE, often struggle to adequately address these imbalances, raising concerns about the accuracy and reliability of their results. This issue presents a major hurdle in accurately assessing and interpreting CE, as the uneven representation of data can lead to skewed insights and potentially overlook critical aspects of student engagement (Karimah & Hasegawa, 2022). This imbalance in datasets not only complicates the analysis but also raises concerns about the reliability and generalizability of the findings in diverse educational settings (Radwan & Cataltepe, 2017).

The exploration of LLMs provides a promising solution to the limitations of traditional ML approaches. Recent studies, including (Wu, 2021), have highlighted the potential of prompt engineering in reducing the need for extensive training of case labeling which is imperative for imbalance data. LLMs employ techniques like In-context Learning (ICL) (Brown et al., 2020) and Chain-of-Thought (COT) prompting (Wei et al., 2022b), enabling more nuanced and context-aware responses. ICL trains models using examples in specific contexts, improving with scaled model and corpus sizes, as seen in N-shot prompting (Brown et al., 2020). This is illustrated by Brown et al. (2020)'s few-shot learning, where LLMs process input-output pairs in-context, leading to better test-time predictions. Similarly, COT, by Wei et al. (2022b), involves logical, step-by-step natural language reasoning. Furthering this, Shahriar et al. (2023) developed Assertion Enhanced Few-Shot Learning, incorporating domain-specific assertions in prompts to enhance accuracy and reduce errors. These innovations significantly boost LLMs' task-specific efficiency, surpassing traditional methods.

While LLMs have shown potential in educational research, their application has predominantly been refined to solve logical reasoning or arithmetic problems (Lee et al., 2024), with

limited exploration in addressing imbalanced datasets of education. Our study breaks new ground by applying LLMs with Prompt Engineering (PE) to this specific challenge. We hypothesize that LLMs, renowned for their nuanced language understanding, will surpass traditional ML algorithms in classifying cognitive engagement levels from student texts. Our exploration is guided by two research questions: RQ1 addresses the comparative efficacy of LLMs against traditional ML algorithms, and RQ2 investigates the role of assertions in overcoming contextual and lexical challenges within imbalanced datasets. Specifically:

1. How do the results obtained from LLMs with PE compare to traditional Machine Learning algorithms in handling imbalanced educational data?

2. In what ways does the integration of assertions enhance the efficacy of models when addressing the challenges associated with imbalanced textual educational datasets?

This paper examines how AEFL mitigates issues in imbalanced educational data analysis, revealing how these technologies can effectively address the challenges posed by uneven dataset distributions. By applying this cutting-edge technique, we uncover new possibilities for analyzing and interpreting complex educational data. Our findings demonstrate the advantage of AEFL in educational settings, especially where traditional ML methods fall short, opening new avenues for data-driven educational research and practice.

The rest of the paper is set up as follows: Section 2 delves into the background, highlighting the emergence of LLMs as a promising solution in education. Section 3 outlines our methodology, including the Iterative - ICL PE Design Process, and the experimental setup. The results and discussions are presented in Section 4, where we compare the performance of LLMs augmented with assertions against traditional ML models and discuss the impact of assertions on model efficacy and limitations. Finally, Section 5 concludes with our findings and future directions.

Background

The exploration of CE within educational research has significantly evolved, transitioning from a simplistic focus on student participation to a complex understanding of mental investment in learning activities. This shift is paramount for fully capturing the essence of engagement, as initially highlighted by Craik & Lockhart (1972) through their distinction between shallow and deep processing. Subsequent work by Appleton et al. (2006) and Fredricks et al. (2004) expanded the concept to encompass behavioral, emotional, and cognitive dimensions, underscoring engagement's multifaceted nature across various educational contexts. A pivotal insight from this exploration is the strong positive correlation between student learning and cognitive engagement, evidenced by Chi & Wylie (2014), which underscores the significant educational outcomes associated with deep cognitive processes.

CE distinguishes itself within the broader spectrum of educational engagement by focusing on the intensity of students' mental investment in learning. This stands in contrast to behavioral engagement's emphasis on participation and emotional engagement's concern with feelings towards learning Blumenfeld et al. (2006). Such a distinction is crucial for educators and researchers dedicated to enhancing learning outcomes through targeted interventions.

Central to understanding and enhancing CE are theoretical frameworks and models like Bloom's taxonomy, Corno and Mandinach's model, and the ICAP model, as well as Wang et al.'s framework for connectivist learning contexts. These models provide comprehensive insights into the various dimensions and components of cognitive engagement, aiding researchers in designing effective studies, developing targeted interventions, and evaluating educational outcomes (Anderson & Krathwohl, 2001; Bloom et al., 1956; Corno & Mandinach, 1983; Chi & Wylie, 2014; Chase et al., 2019; Hsiao et al., 2022; Wang et al., 2016).

Measuring CE, however, presents inherent challenges due to its complex and internal nature. As a latent construct, CE's assessment relies on inferences from behavioral indicators or through self-report measures (Chi & Wylie, 2014; Fredricks et al., 2004; McCoach et al., 2013).

Traditional methods, including self-report questionnaires, surveys, and observational techniques, often inadequately capture the nuanced cognitive processes involved in learning. A variety of measures have been employed in past studies to gauge CE, such as self-reported scales, classroom observations, interviews, teacher ratings, experience sampling, eyetracking, physiological sensors, trace analysis, and content analysis (Greene et al., 2004; Smiley & Anderson, 2011; Lee & Anderson, 1993; Helme & Clarke, 2001; Wigfield et al., 2008; Xie et al., 2019; D'Mello et al., 2017; Bernacki et al., 2012; Ireland & Henderson, 2014). Nonetheless, the complexity of accurately assessing CE through these measures necessitates innovative approaches that more precisely reflect students' cognitive investment in their educational activities Fredricks et al. (2004).

In educational research, traditional ML methods have extensively analyzed student data patterns but face limitations when addressing nuanced aspects like cognitive engagement. The problem is exacerbated by imbalanced datasets, leading to skewed insights and overlooking crucial engagement aspects, thus affecting the findings' accuracy, reliability, and generalizability across diverse educational contexts (Lee & Kinzie, 2012; Fredricks et al., 2004). This issue with imbalanced datasets, characterized by unevenly distributed categories and small sample sizes, highlights the need for specialized techniques to improve model performance and accuracy, ensuring a comprehensive understanding of CE across educational contexts (Chawla, 2010; Fernández et al., 2018; Kulkarni et al., 2020; Japkowicz & Stephen, 2002; Bruce et al., 2020; LemaÃŽtre et al., 2017).

The advent of LLMs presents a promising solution to the issues posed by imbalanced datasets in educational research. Recent breakthroughs in LLMs, particularly with ICL, COT and AEFL prompting techniques, have demonstrated their potential to generate nuanced, context-aware responses beyond the capabilities of traditional ML methods (Brown et al., 2020; Wei et al., 2022b; Shahriar et al., 2023). For example, Savelka et al. (2023) showcased how GPT-3.5 & 4 could effectively classify student help requests in programming courses, illustrating the superior ability of LLMs to handle nuanced educational data. Zeng et al. (2023) delved into the cognitive and reasoning abilities of LLMs, highlighting the necessity for task-specific tuning to address complex reasoning challenges. Cui et al. (2023) introduced the Divide-Conquer-Reasoning (DCR) framework

to enhance the consistency and reliability of LLM-generated texts, vital for creating educational content. These examples reveal the capacity of LLMs to offer more accurate classification and analysis of CE, surpassing traditional ML methods in dealing with the intricacies of educational datasets. Additionally, Lee et al. (2024) explored LLMs' use with CoT prompting to improve automatic scoring systems in science education, further indicating LLMs' potential to enhance the quality and reliability of educational content analysis.

By harnessing the intrinsic capacity of LLMs to interpret and utilize language within specific contexts, researchers can navigate the challenges posed by imbalanced datasets, facilitating a deeper understanding of student CE.

Methodology

Context and Participants

This study performs a secondary analysis on a dataset originally gathered to assess CE from student responses in a High School English Language Arts course's AI curriculum. The StoryQ curriculum (Chao et al., 2022), spanned three weeks with daily 45-minute classes, incorporated Machine Learning Practices through open-ended questions in eight modules but our analysis only evaluated three: "Sentiment Analysis," "Features and Models," and "All Words." The initial study's diverse participant group of 28 students included 17 females, 7 males, and 4 non-specified gender individuals, spanning various grades and racial backgrounds. The racial composition was 43% Black/African American, 17% Hispanic/Latinx, 18% White/Caucasian, with others choosing not to disclose. Students' CE was evaluated using a modified Interactive-Constructive-Active- Passive (ICAP) framework by Chi & Wylie (2014), focusing on Constructive, Active, and Passive levels. Their open-ended responses (N = 840) were analyzed using the CE coding scheme (Table 3.1), yielding a Cohen's kappa inter-rater reliability of 0.84.

Table 3.1

ICAP: Why do you think the model learned a large positive weight for this feature?

Score	ICAP Level	Description	Indicator	Example
2	Constructive	New information is integrated with activated prior knowledge, and new knowledge is inferred	Deep reasoning, synthesis of new ideas, or forming hypotheses	"I think that the model learned a large positive weight for the feature because if you came to an establishment then that would indicate that you did like it because you chose to come in the first place."
1	Active	Behaviors that cause-focused attention while manipulating	Apply, Analyze, or Manipulating	"I think this gained a large amount of weight because it is a commonly used word."
0	Passive	Overt activities that are carried out mindlessly	Recalling or Restating	"Amazing, clean, selection, try, regular, seating"

Prompt Engineering Design

Our prompt development process, grounded in the ICL Prompt Engineering Design (see Figure 3.1), begins with drafting an initial few-shot ICL format prompt. This prompt, inputting student responses and outputting CE classifications, undergoes validation testing on a subset (n=27). If benchmarks are met, it progresses to full dataset testing; otherwise, we diagnose misclassifications, realigning LLM outputs with our coding standards through domain-specific CE knowledge integration. Adjustments may involve refining COT processes, FewSHOT learning, or embedding conceptual knowledge assertions. After subset retesting and validation, the optimized prompt is applied to the full dataset (n=135), with iterative refinement ensuring optimal performance. See Appendix B for additional LLM-specific prompt details.

Our engineering approach encompasses three components: General COT, FewShot with Reasoning Sequence, and assertions Prompting. General COT, embeds sequential instructions with "think time" to initiate the model's reasoning on given tasks (Fulford & Ng, 2023). Our General COT prompt follows a seven-step sequence to guide the LLM's task reasoning. Initially, the model attentively reads the provided «Question, Response» (Step 1), laying the foundation for accurate comprehension and subsequent cognitive engagement analysis. Step 2 involves feeding the model CE domain-specific definitions for Passive, Active, and Constructive levels, requiring it to discern the appropriate engagement level based on the initial input. Progressing to Step 3, the model assesses the rationale behind the assigned cognitive engagement label, ensuring it reflects the response's depth and nature. In Step 4, the LLM reevaluates the response to prevent misclassification and assesses if a different CE level is more aligned. Steps 5 and 6 prompt the model to consider ways to enhance the CE level, crucial in the validation and diagnostic phases, particularly when integrating assertions. The final step (Step 7) circles back to the initial input, where the LLM reexamines the cognitive engagement level to verify the accuracy and consistency of its prediction. This structured approach is key in sharpening the model's evaluative and analytical capabilities.

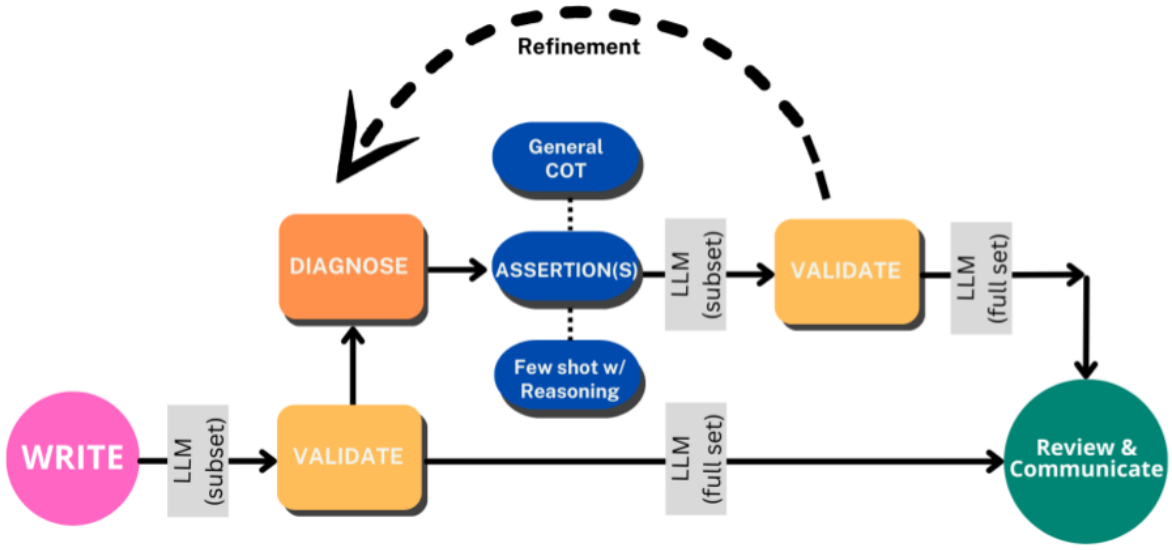
FewShot with Reasoning, guided by gold standard examples (Wang et al., 2023; Shahriar et al., 2023), includes a four-element structure: «Question, Response, Label, and Reasoning». This method enhances LLM's task-specific learning, incorporating reasoning sequences in the examples. Finally adding assertion Prompts, is crucial for knowledge-building explanations, that are domain-specific insights defined from General COT's outputs on misclassified predictions (Shahriar et al., 2023).

Experiment Design

To analyze traditional ML methods (SVM, RF, DT, and ADABoost), we divided our data into training (n=432) and testing sets (n=135), applying default hyperparameters from the Scikit-Learn package (Pedregosa et al., 2011). See Appendix A for hyperparameters. The dataset comprised two majority classes and one minority class (Table 3.2). During data preprocessing, we executed

Figure 3.1

ICL Prompt Engineering Design Process to optimize the accuracy of LLMs in classifying educational data with the use of ICL, COT and AEFL.



text cleaning steps: removing non-alphanumeric/special characters (except periods), new lines, isolated "n" characters, excess spaces, double quotes, and backslashes; converting to lowercase; eliminating stop words; and correcting spelling errors. We transformed the tokenized text using TF-IDF vectorization for ML algorithm suitability. These traditional ML methods served as benchmarks for comparing with LLM prompt results.

Table 3.2

Dataset numbers for Training, Testing and subsets by cognitive level.

ICAP Level	Training	Testing	Subset
C	202	62	10
A	203	66	10
P	27	7	7

In analyzing LLM, we employed GPT-4 through the Colab Python OpenAI API, setting hyperparameters to temperature = 0 and top p= 0.01 for optimal automatic scoring (Wang et al., 2023). The data preprocessing mirrored the traditional ML approach but without tokenization or vectorization. We maintained the integrity of student sentences, ensuring capitalized start and appropriate punctuation, mainly periods. The final prompt (see Appendix B) underwent testing with the same dataset (n=135) used in traditional ML.

In our final experiment, we adopted a subset-based iterative modification approach (n=27) as per the ICL Prompt Design Process (Section 3.2). This involved a sensitivity analysis for precise influence measurement of assertions on LLM performance. Each iteration entailed scrutinizing misclassified data, focusing on informal language nuances in text inputs. This qualitative analysis was pivotal for understanding the impact on model accuracy and response. This systematic approach enriched our comprehension of LLM's interaction with varied prompts and offered insights for enhancing LLM's performance in processing and interpreting informal language, a significant challenge in educational datasets.

Analysis

In our multiclass dataset analysis, we utilize Precision, Recall, and F1 Score to evaluate the performance of LLMs with assertions versus traditional ML models. These metrics are integral for assessing model efficacy in a multiclass environment. Precision gauges the model's accuracy in predicting each class, indicating the reliability of its positive predictions. Recall measures the model's capacity to correctly identify all instances of each class, vital for ensuring comprehensive representation in a multiclass context. The F1 Score, as the harmonic mean of Precision and Recall, offers a balanced evaluation of the model's overall performance, particularly important in our study to address potential class imbalance. Following Pennebaker et al. (2015), we emphasize both precision and recall to minimize false positives and negatives, crucial in multiclass datasets. Additionally, we assess the percentage change in F1 score performance to quantify the impact of assertions, using the following formula:

$$\text{Percent Increase} = \frac{\text{F1 score of LLM} - \text{F1 score of traditional ML}}{\text{F1 score of traditional ML}} x100$$

To further this analysis we examined F1 scores. To differentiate between models, we developed a custom metric, inspired by Cohen's D (Cohen, 2013). However, unlike the traditional Cohen's D, which uses standardized effect sizes (small at 0.2, medium at 0.5, large at 0.8) based on pooled standard deviation, our metric directly compares raw F1 score differences. This modification suits our data, where standard deviation calculations aren't feasible due to single observations per model. We categorized differences in F1 scores as small (up to 10 points), medium (10 to 30 points), and large (over 30 points). We defined a function for calculating pairwise differences in scores mi , mj M represent any two models, and si , sj are their respective scores. The function:

$$f(mi, mj) = si - sj$$

computes the difference in performance scores between each pair of models. For each combination of models (mi , mj), the score of model mj is subtracted from that of model mi. This function calculates the performance difference between each model pair. We then generate a matrix showcasing these differences, allowing for a thorough pairwise comparison of model performances.

To answer RQ 2 and evaluate the ways that the integration of assertions enhance the efficacy of models when addressing the challenges associated with imbalanced textual educational datasets we chose to test on a subset (N=27, P = 10, A = 10, C= 7) as is common in the research to "increase the depth of our analysis, reduce run-time, and decrease cost" (Rodriguez et al., 2023, p. 2). We chose a sensitivity analysis (Akinwande et al., 2023) to critically assess the impact or influence of the assertions. We did this qualitatively by adding two steps (Step 5 & 6 of General COT in Appendix B) into the «General COT» and interpreting for the «model outcome» for recurring themes. Our examination extended to a comparative analysis of the experiments, employing class-wise analysis to measure each experiment against a baseline prompt that did not incorporate assertions.

Results and Discussion

RQ1: How do the results obtained from LLMs with Prompt Engineering compare to traditional Machine Learning algorithms in handling imbalanced educational data?

Performance Metrics

The summary results in Table 3.3 indicated a varied performance across classes. In the Passive class, the LLM significantly outperformed traditional models, showing a 14.9% increase over SVM, 6.25% over RF, 18.0% over DT, and a notable 23.2% increase over AdaBoost. Conversely, in the Active class, traditional models (SVM, RF, and DT) surpassed LLM by 11.1%, while AdaBoost and LLM performances were comparable. The most striking contrast was observed in the Constructive class, where traditional models (SVM, RF, DT, and AdaBoost) failed to effectively identify instances. In contrast, the LLM demonstrated a remarkable improvement with an F1 score of 32, showcasing its superior capability in recognizing elements of the minority class.

Table 3.3

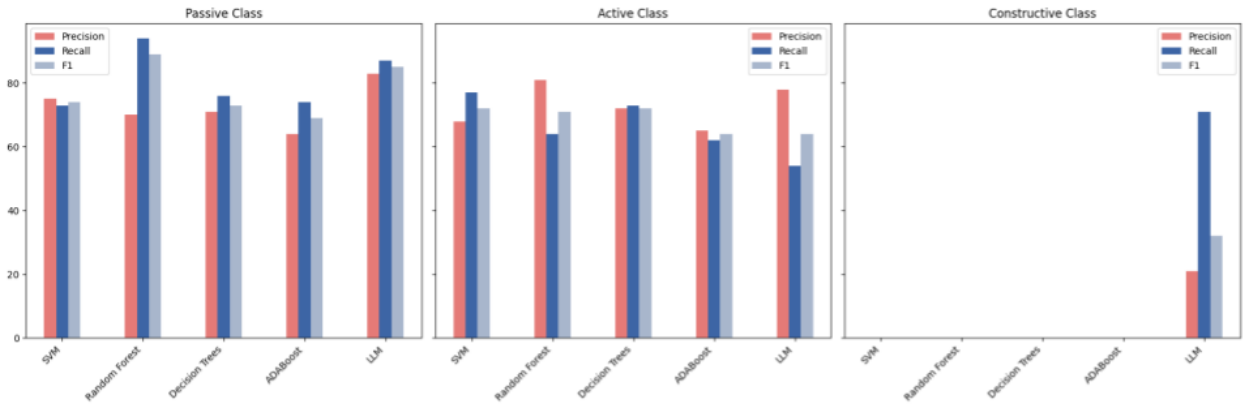
Summary of Performance Metrics by Cognitive Engagement Level

	Passive (62)			Active (66)			Constructive (7)		
Model	P	R	F1	P	R	F1	P	R	F1
SVM	75	73	74	68	77	72	0	0	0
RF	70	92	80	80	65	72	0	0	0
DT	71	74	72	71	73	72	0	0	0
ADABoost	64	74	69	65	62	64	0	0	0
LLM	83	87	85	78	54	64	21	71	32

These results suggest that while traditional machine learning models like SVM, RF, DT, and AdaBoost may perform comparably or better in majority classes, the LLM exhibits superior

Figure 3.2

Performance Metrics Summary by Cognitive Engagement Class showing results for each cognitive class.



capability in dealing with minority class instances, particularly in complex classification tasks like the Constructive class in our dataset (see Figure 3.2). The versatility and adaptability of LLMs in handling imbalanced class distributions highlight their potential in enhancing classification tasks, especially in scenarios where minority classes hold substantial importance. These findings affirm our hypothesis that LLMs, especially when augmented with assertions, offer superior capabilities in classifying cognitive engagement levels from student texts, addressing the core of RQ1.

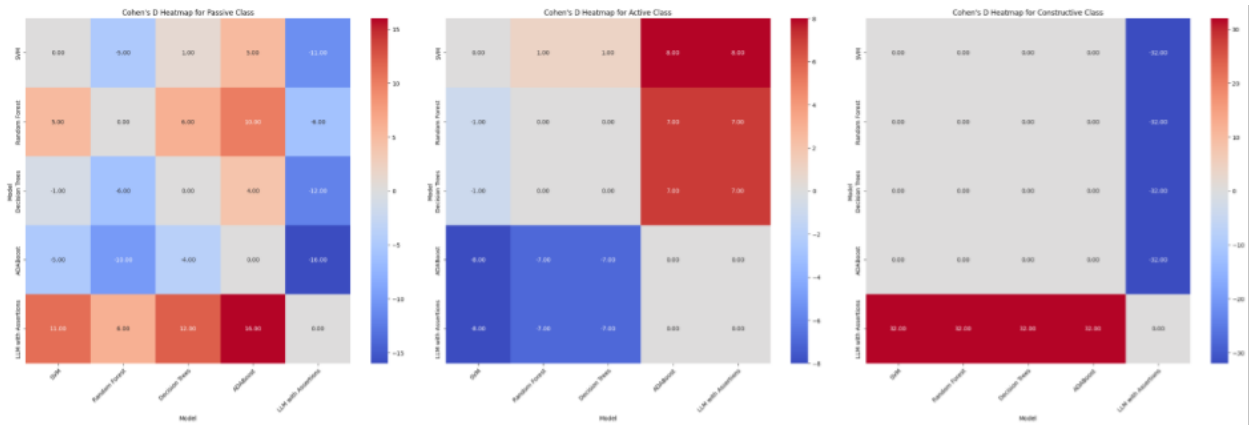
Relative Performance

We see similar results in our custom metric inspired by Cohen's D due to the unique nature of our data, where standard deviation calculations were not applicable, and produced interesting results (see Figure 3.3). The LLM with assertions for the passive class demonstrated noteworthy advantages over traditional models in various comparisons which resonate with the work of researchers (Shahriar et al., 2023), who demonstrated the enhanced effectiveness of LLMs in educational settings.

Against SVM, the LLM had a significant edge, showing an 11-point advantage in the F1 score, categorized as a 'medium' difference according to our threshold range. This indicates a considerably better performance of the LLM over SVM. When compared to DT, the LLM with

Figure 3.3

Relative Performance Heatmap by Cognitive Engagement Class



assertions again showed a 'medium' difference, outperforming DT by 12 points, underscoring its effectiveness in handling complex classification tasks. In a more striking contrast, the LLM outperformed ADA Boost by 16 points, falling into the 'medium' range and highlighting a substantial performance gap where the LLM was far superior.

In the Active class, the LLM with assertions exhibited a mixed performance. It showed a close competition with SVM, trailing by just 2 points, which falls into the 'small' difference category, implying a nearly equivalent performance between the two models. However, the LLM outperformed ADABoost by a margin of 8 points, a 'small' difference that nonetheless underscores its relative effectiveness. This suggests that while LLMs offer substantial advantages in many areas, their performance can vary depending on the specific classification task, echoing the findings of Lee et al. (2024), who explored the use of LLMs in automatic scoring systems. Against RF and DT, the LLM had a slight disadvantage, trailing by 7 and 3 points respectively, suggesting that in certain scenarios, traditional models may have a slight edge over the LLM.

The Constructive class results were particularly striking. The LLM with assertions demonstrated a pronounced superiority in this category. It dramatically outperformed all traditional models (SVM, RF, DT, and ADABoost), each of which failed to identify instances within the Constructive class effectively, as indicated by their zero scores. The LLM achieved an F1 score of 32, which

not only establishes a ’large’ difference according to our threshold but also highlights the LLMs exceptional capability in handling minority classes or complex classification tasks where traditional models fall short. It points to the LLMs’ superior ability to handle imbalanced datasets, a common challenge in educational data analysis, as illustrated by the work of researchers like Zeng et al. (2023), who evaluated the cognitive and reasoning abilities of LLMs.

RQ2: In what ways does the integration of assertions enhance the efficacy of models when addressing the challenges associated with imbalanced textual educational datasets?

Our analysis aimed to augment Active class metrics and foster a more equitable model across cognitive classes. Throughout the course of ten experiments, including the baseline, the implementation of assertions, particularly those delineated in «General COT» (Steps 5 & 6, see Appendix B), was pivotal in surfacing two primary themes post the initial experiment: textual ambiguity and contextual comprehension challenges.

For text ambiguity, the baseline experiment revealed the model’s propensity to misconstrue the depth of student engagement. Instances where contributions appeared analytical but merely constituted a superficial application of known concepts underscored this issue. By systematically applying the assertions detailed in the Methodology, we observed significant improvements in model performance, particularly within the Active and Constructive classes.

With regard to Unusual language, the model’s interpretation of speculative language (e.g., "I think," "possibly," "I believe") as indicative of reflective or analytical thought. Such expressions, particularly when conveying opinions that superficially suggested deeper analysis, were erroneously classified as constructive engagement.

Initially, our approach to integrating assertions was exploratory but became more systematic by the third experiment. For example, between experiments two through four, certain responses intended as "Constructive" were incorrectly classified as "Active":

Misclassified Example 1:

Question: Why do you think the model learned a large negative weight for this feature?

> Student response: “I think the model learned a negative weight for this feature because the model categorized the reviews as negative and categorized the surprisingly negative features as negative too since that was the whole sentiment of the review.”

Misclassified Example 2:

> Question: Why do you think the model learned a large positive weight for this feature?
> Student response: “I feel like it had to do with the words and how much they were used whenever there was a positive review it would contain more than one good word to go along with it”

By incorporating the assertion «Do label the statement as Constructive when they form a hypothesis about why the model learned a weight for a certain feature», these responses were accurately predicted as constructive, enhancing the Constructive class with precision and recall metrics—specifically, a recall increase of 6.33% and an F1-score improvement of 4.30%.

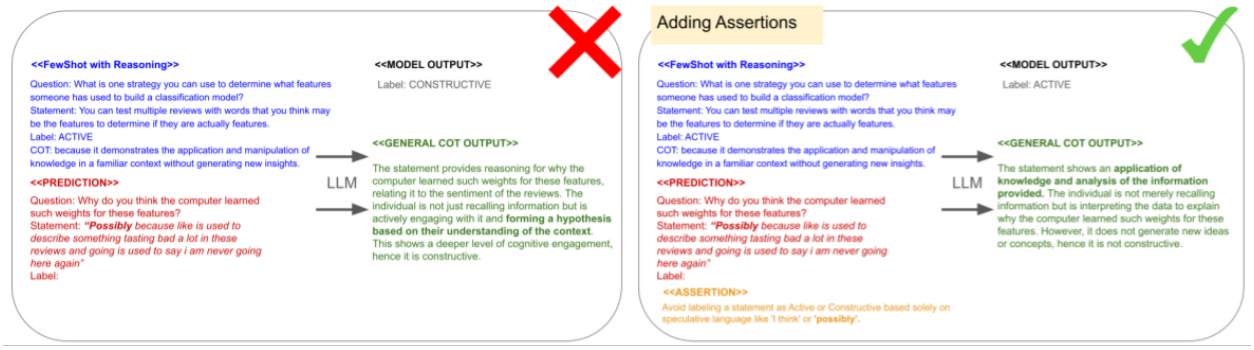
Moreover, addressing the misuse of speculative language through the assertion «Avoid labeling a statement as Active or Constructive based solely on speculative language like ’I think’ or ’possibly’» (see Figure 3.4) led to an increase in precision for the Active class by 15.96% and an F1-score increase by 6.08%. This adjustment resulted in the most balanced model performance observed, despite a slight decrease in recall for the Active class by 2.34%. Further attempts to amplify Active class metrics by refining definitions in «General COT» and enhancing reasoning in «FewShot with reasoning» revealed that, while assertions impacted model performance, their effect varied across classes and metrics.

Notably, Experiment 6.1 (see Figure 3.5) emerged as particularly effective, showcasing the significance of tailored assertions in reducing misclassifications linked to textual ambiguity and unusual language use, thereby contributing to a more balanced and accurate model.

These findings highlight the nuanced role of assertions in enhancing model efficacy against the backdrop of imbalanced educational datasets. By meticulously integrating assertions to counter specific challenges—textual ambiguity and unusual language—the experiments demonstrated

Figure 3.4

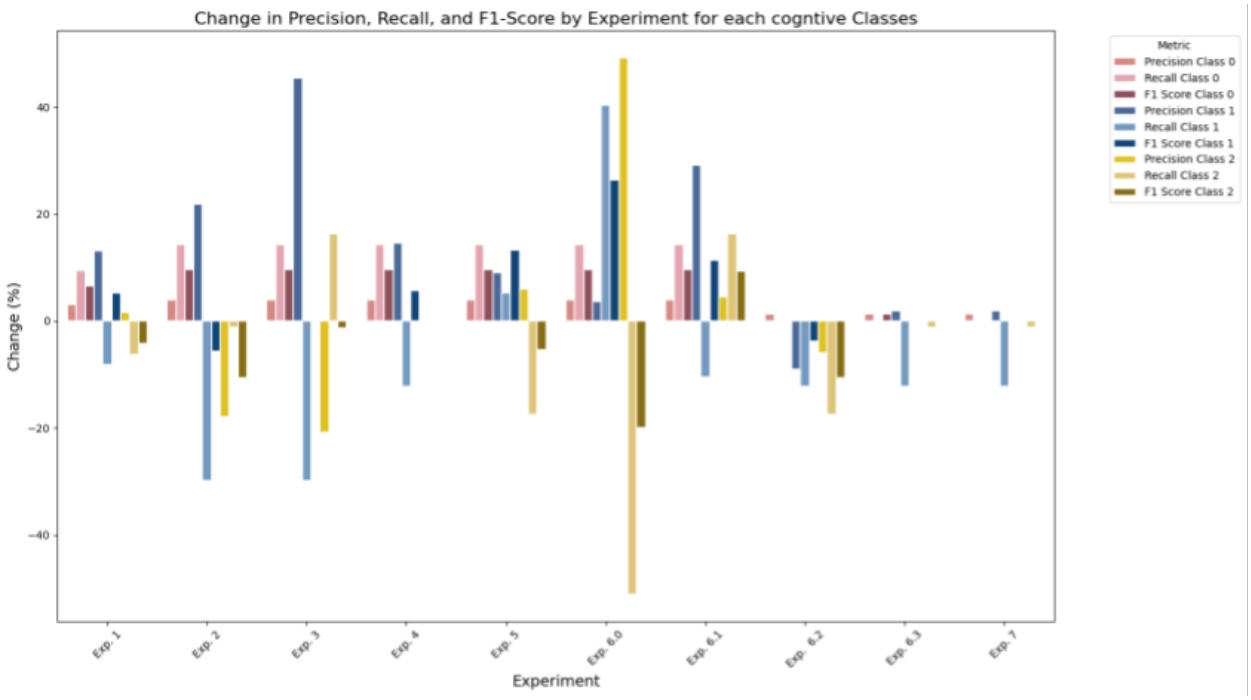
Left image does not include a targeted assertion while the one on the right does and improves the model output to correctly predict the students cognitive level of their text.



a discernible improvement in model precision and balance, particularly within the Active and Constructive classes. This strategic approach underscores the potential of assertions to mitigate inherent dataset imbalances, ultimately contributing to the development of more nuanced and effective educational models.

Figure 3.5

Percentage Change in Metrics for Each Class Across Experiments



To further understand our model we compared accuracy of models to the baseline where Experiment 5 marked an 8.96% improvement but Experiment 6.1 stood out with the highest increase in accuracy, at 11.94% from the baseline. This improvement primarily addresses the challenges identified in RQ2, demonstrating the significant role of assertions in resolving errors related to context understanding and lexical ambiguities. The Active and Constructive classes, associated with focused attention and deeper reasoning, respectively, pose classification challenges due to their subtleties and contextual dependencies (Chi & Wylie, 2014). These classes often require inferring cognitive engagement levels from implicit cues and context, making their distinctions less explicit within student responses.

Limitations

While our study sheds light on the potential of LLMs and AEFL in addressing imbalanced datasets, it also highlights the need for caution in interpreting these findings without consideration of the broader methodological and technological landscape. Firstly, our reliance on specific LLM techniques and AEFL might not capture the full spectrum of potential solutions available within the rapidly evolving field of machine learning. The specific parameters and configurations employed in our LLM applications (Shahriar et al., 2023; Wei et al., 2022b; Zeng et al., 2023), while effective in this context, might not be universally applicable or optimal across different datasets or learning tasks. While our study provides valuable insights, it echoes the concerns raised by Radwan & Cataltepe (2017) and Yun et al. (2011) regarding the challenges of imbalanced datasets in education and the limitations of traditional ML approaches.

Furthermore, our study's focus on a High School ELA course dataset (Zeng et al., 2023), while providing a rich source of cognitive engagement data, also presents a limitation in terms of diversity and representativeness. The linguistic and cognitive patterns inherent in this specific educational setting may not fully encapsulate the variety of cognitive engagement manifestations across different age groups, subjects, or educational methodologies. This limitation underscores the importance of extending research efforts to encompass a wider range of educational contexts, to ensure the findings' applicability and robustness, as indicated by Fredricks et al. (2004) andvBlumenfeld et al. (2006).

Additionally, while LLMs and AEFL present innovative approaches to overcoming the challenges of imbalanced datasets, they also introduce new complexities and considerations (Shahriar et al., 2023; Wei et al., 2022b). The computational demands and resource requirements of these technologies, coupled with the need for specialized expertise to implement and interpret their outputs, may pose barriers to widespread adoption and application in educational research and practice. The dynamic nature of LLM development also means that the models and techniques used today may rapidly evolve, necessitating continuous updates and adaptations to maintain their

effectiveness and relevance.

Lastly, the ethical implications of applying LLMs in educational settings, particularly concerning data privacy, security, and the potential for bias in model training and outcomes, warrant careful consideration (Zeng et al., 2023). As LLMs become more integrated into educational research and practice, it is crucial to develop and adhere to ethical guidelines that prioritize the well-being and rights of students and educators.

These limitations highlight the need for ongoing research and dialogue within the educational and machine learning communities. By addressing these challenges and exploring the vast potential of LLMs and AEFL, we can advance our understanding of cognitive engagement and enhance educational outcomes in diverse and inclusive ways.

Conclusion and Future Studies

Our study makes significant contributions to the evolving landscape of cognitive engagement (CE) research, building upon the foundational work of seminal researchers like Craik & Lockhart (1972), Appleton et al. (2006), and Fredricks et al. (2004). We leveraged the capabilities of Large Language Models (LLMs) and Assertion Enhanced Few-Shot Learning (AEFL), marking a notable advancement in the domain of CE. This approach pays homage to the pioneering efforts that have shaped our understanding of CE while extending these concepts through the integration of cutting-edge LLM technologies.

By adeptly navigating the challenges posed by imbalanced datasets and accurately classifying cognitive engagement levels, this study underscores the potential of LLMs to refine our measurement and analysis of CE, setting a new benchmark for educational research. The integration of AEFL enhances contextual comprehension, improving model accuracy and balance, as highlighted by Shahriar et al. (2023). Experiment 6.1 further illustrates the value of tailored assertions in reducing misclassifications linked to textual ambiguities, offering novel insights into AEFL's effectiveness in managing class-imbalanced data.

The promising outcomes of this research suggest that LLMs hold significant potential for

future educational studies, particularly in complex data analysis tasks. These findings encourage the exploration of LLMs' full capabilities in educational settings, advocating for a paradigm shift towards more sophisticated and nuanced approaches to data analysis. Moreover, the integration of AEFL points to a nuanced method of enhancing model performance, especially in the context of imbalanced textual educational datasets.

Given the multifaceted nature of cognitive engagement and the challenges associated with its measurement, there is a compelling need for further research. Future studies should aim to refine and expand the application of LLMs and AEFL across a broader spectrum of educational contexts. Additionally, exploring additional theoretical frameworks and models could yield deeper insights into cognitive engagement, thereby contributing to the enhancement of educational outcomes. This call for further research not only reflects the complex landscape of CE but also highlights the endless possibilities that LLM technologies and innovative methodologies like AEFL present for advancing our understanding and practices within the educational domain.

Chapter 4: Conclusion

This chapter offers a summary and discussion of each research study, including their scholarly and practical implications. This is followed by a more detailed contribution and discussion of future research in this area.

Summary

Chapter 2

In this study, we investigated the relationship between Bloom's taxonomy categories, cognitive engagement (CE) levels, and lexical diversity (LD) within a Machine Learning Curriculum, examining how these factors interplay with demographic influences to affect student learning outcomes in an AI-enhanced educational setting. This analysis was anchored on the examination of student responses through the lens of Bloom's taxonomy, the augmented Constructive - Active - Passive (CAP) framework, and LD measures.

In doing so, we revealed that the revised Bloom's taxonomy provides a robust framework for differentiating the cognitive demands of questions, which in turn significantly influences both cognitive engagement and lexical diversity in student responses. The study's findings highlighted significant variations in CE levels across different cognitive task complexities, with more complex tasks eliciting higher levels of active engagement and enhanced lexical diversity. Moreover, the nuanced interplay between demographic factors and engagement levels pointed towards a varied AI-enhanced learning experience across different student groups. This study underscores the critical importance of designing curricula that not only engage students cognitively but also cater to the diverse linguistic needs, promoting inclusivity within the educational landscape.

This novel approach enriches our understanding of the dynamic interplay between question design and student engagement, offering valuable insights for optimizing educational strategies in machine learning contexts.

Chapter 3

In this chapter, we investigated the effectiveness of Large language Models (LLMs) enhanced with assertions for analyzing imbalanced educational datasets, particularly for accurately classifying cognitive engagement levels in student texts. In doing so, we developed an "Iterative - ICL PE Design Process," for classifying cognitive engagement from student responses by initially creating and testing a few-shot format prompt. If initial benchmarks are not met, the prompt is refined through adjustments in Chain of Thought (COT) processes, FewSHOT learning, and the integration of domain-specific knowledge. This iterative refinement and validation process, which includes structured reasoning sequences and targeted assertion prompts, aims to optimize the prompt's performance on larger datasets, improving LLMs' accuracy in educational data analysis.

The findings revealed that LLMs with PE significantly outperform traditional machine learning models in classifying cognitive engagement levels, particularly in handling the minority Constructive class. LLMs demonstrated notable improvements in recognition and classification accuracy across various metrics compared to standard models like SVM, RF, DT, and AdaBoost. For example, the use of targeted assertions reduced misclassifications related to speculative language, increasing precision in the Active class by nearly 16% and the F1-score by over 6%. These results underscore LLMs' enhanced capability to effectively manage the complexities associated with imbalanced educational datasets, especially in accurately identifying nuanced and minority class representations which traditional models often fail to detect.

The findings also highlighted two key themes that emerged from integrating assertions into LLMs: textual ambiguity and contextual comprehension challenges. The systematic application of assertions helped address instances where the model misconstrued the depth of student engagement, significantly improving the accuracy of classifications within the Active and Constructive classes. These themes emphasize the potential of assertions to refine the interpretation of nuanced language and context, thereby enhancing the overall model performance in imbalanced educational datasets.

These results underscore LLMs' enhanced capability to effectively manage the complexities associated with imbalanced educational datasets, especially in accurately identifying nuanced and

minority class representations which traditional models often fail to detect.

Contributions

Each of the chapters investigate the understanding of CE in education, particularly within AI-enhanced learning environments. By highlighting the potential of LLMs augmented with assertions, this research paves new avenues for addressing the perennial challenge of imbalanced datasets, enriching the analytical tools available for educational research. Furthermore, the findings underscore the importance of curricular designs that promote deep cognitive engagement and cater to the linguistic diversity of learners, ensuring a more inclusive and engaging educational experience.

Our research has advanced the understanding of CE by employing LLMs and prompt engineering to address the challenge of imbalanced educational datasets. Martin et al. (2023) similarly emphasize the role of innovative methodologies in K-12 education, underscoring the growing integration of AI technologies to enhance educational outcomes and equity. This parallel underscores the value of our approach in addressing imbalanced educational datasets, aligning with broader educational research trends that seek to leverage technology for enhanced learning insights. These insights will guide educators and technology developers in designing more nuanced AI-enhanced learning environments that effectively classify and foster deeper levels of student engagement.

The chapters in this dissertation deepens the academic field's understanding of the intricate relationship between the complexity of questions, structured through Bloom's taxonomy, and student engagement at various cognitive levels. Inspired by Tatar et al. (2021), who observed how teachers integrate unstructured data modeling into their pedagogical practices, our findings provide educators with a robust framework for designing questions that engage students meaningfully while simultaneously challenging them to develop their critical thinking and problem-solving skills. This approach enhances the overall learning experience, mirroring the engaged and analytical discussions observed in Tartar et al.'s study where teachers applied inductive reasoning to model

decision-making (Tatar et al., 2021).

Furthermore, by integrating LD measurements alongside cognitive engagement indicators, our research highlights the significant impact of linguistic complexity on learning outcomes. This integration allows educators to tailor content that effectively meets the diverse linguistic needs of students, fostering inclusivity within the educational landscape, akin to the approaches suggested by Ahmed et al. (2021), where NLP is employed for real-time feedback processing to continually adapt and enhance educational delivery. Additionally, this research pioneers the exploration of cognitive engagement through open-ended questions within AI-enhanced curricula, offering novel insights into how question design can profoundly influence student interactions and learning outcomes. The findings enable educators to strategically employ open-ended questions to elicit deeper cognitive processing, fostering richer student engagement and improving comprehension of complex AI concepts (Ahmed et al., 2021).

By integrating CE measures with LD assessments in responses to open-ended questions, this study reveals complex patterns of language use and cognitive effort. These insights, echoing the findings of Chase et al. (2019), who demonstrated the predictive value of LD in academic settings, pave the way for educators to refine instructional strategies and enhance the effectiveness of AI tools in teaching, ensuring that educational practices are both engaging and linguistically diverse (Chase et al., 2019).

By examining the effects of demographic factors on CE and LD, the studies shed light on educational equity. This aspect of the research draws on the work of Marks (2000), who emphasizes the role of engagement in learning outcomes across diverse student populations, thereby encouraging the development of pedagogical practices that ensure all students, regardless of their background, have equal opportunities to benefit from AI-driven educational innovations.

Finally, these studies not only validated innovative methodologies in educational technology research but also set the stage for future investigations into the longitudinal effects of AI-enhanced education. This forward-looking perspective is inspired by the longitudinal research frameworks suggested by Chi & Wylie (2014), advocating for continuous assessment of learning processes.

Researchers and educators can build upon these findings to explore further the dynamic capabilities of AI in education, aiming for continuous improvement in engagement and educational quality Chi & Wylie (2014).

Future Research

Future research could integrate the exploration of cognitive frameworks like Bloom's taxonomy with advanced measures of LD to assess and enhance learning outcomes across diverse educational settings and student demographics. By examining the combined effects of technology integration, instructional design, and socio-cultural influences on cognitive engagement and lexical diversity, such studies could profoundly inform educational policy and curriculum development, ensuring they are inclusive and effective for a broad range of student populations. This includes specific focus on how different formulations of open-ended questions, real-time cognitive engagement tracking, and adaptations in multilingual contexts can influence learning dynamics.

Incorporating Natural Language Processing (NLP) techniques can further refine the analysis of student interactions and contribute to more nuanced learning analytics. Future studies should continue to explore the application of LLMs with assertions across wider educational contexts, assessing their impact on diverse learning populations. Additionally, employing robust learning analytics can provide deeper insights into the effectiveness of AI-enhanced curricula, enabling educators to tailor educational strategies more precisely. The exploration of other advanced AI and ML techniques in analyzing and enhancing cognitive engagement within imbalanced datasets represents another promising research direction. Longitudinal studies examining the long-term effects of AI-enhanced curricula on cognitive engagement and educational outcomes would provide invaluable insights into the sustained impact of these technologies on student learning. Moreover, investigating the role of NLP in developing robust learning analytics frameworks can lead to improvements in how educational data is processed, interpreted, and utilized, paving the way for more adaptive and responsive educational environments. By continuing to explore these avenues, researchers can harness the full potential of AI in education, ultimately contributing to a richer,

more equitable educational landscape.

By grounding future research in robust learning theories and harnessing advanced learning analytics techniques, this work not only aims to deepen our understanding of educational dynamics but also seeks to revolutionize educational practices, ensuring that they are both scientifically sound and practically effective in enhancing student learning outcomes on a global scale.

REFERENCES

Adom, D., Yeboah, A., & Ankrah, A. K. (2016). Constructivism philosophical paradigm: Implication for research, teaching and learning. *Global journal of arts humanities and social sciences*, 4(10), 1–9.

Ahmed, H., Hina, S., & Asif, R. (2021). Evaluation of descriptive answers of open ended questions using nlp techniques. In *2021 4th International Conference on Computing & Information Sciences (ICCIS)*, 1–7. IEEE.

Akinwande, V., Jiang, Y., Sam, D., & Kolter, J. Z. (2023). Understanding prompt engineering may not require rethinking generalization. *arXiv preprint arXiv:2310.03957*.

Anderson, L. W. & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). A revision of bloom's taxonomy of educational objectives. *A Taxonomy for Learning, Teaching and Assessing. Longman, New York*.

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the student engagement instrument. *Journal of school psychology*, 44(5), 427–445.

Araujo, J. J., Babino, A., Cossa, N., & Johnson, R. D. (2018). Engaging all readers through explorations of literacy, language, and culture. the fortieth yearbook: A double peer-reviewed publication of the association of literacy educators and researchers. *Association of Literacy Educators and Researchers*.

Avella, J. T., Kebritchi, M., Nunn, S. G., & Kanai, T. (2016). Learning analytics methods, benefits,

and challenges in higher education: A systematic literature review. *Online Learning*, 20(2), 13–29.

Baker, R. S., Yacef, K., et al. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of educational data mining*, 1(1), 3–17.

Banihashem, K. & Macfadyen, L. P. (2021). Pedagogical design: Bridging learning theory and learning analytics. *Canadian Journal of Learning and Technology*, 47(1).

Bernacki, M. L., Byrnes, J. P., & Cromley, J. G. (2012). The effects of achievement goals and self-regulated learning behaviors on reading comprehension in technology-enhanced learning environments. *Contemporary Educational Psychology*, 37(2), 148–161.

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *Office of Educational Technology, US Department of Education*.

Bloom, B. S. et al. (1956). Taxonomy of educational objectives: The classification of educational goals, by a committee of college and university examiners. *Handbook 1: Cognitive domain*.

Blumenfeld, P. C., Kempler, T. M., & Krajcik, J. S. (2006). *Motivation and cognitive engagement in learning environments*. na.

Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., & Kasneci, G. (2022). Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*.

Brock, C. A. (1986). The effects of referential questions on esl classroom discourse. *TESOL quarterly*, 20(1), 47–59.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media.

Buckingham Shum, S. (2012). Learning analytics policy brief. *UNESCO Institute for Information Technologies in Education*.

Çakır, H. & Cengiz, Ö. (2016). The use of open ended versus closed ended questions in turkish classrooms. *Open Journal of Modern Linguistics*, 6(2), 60–70.

Carroll, J. B. (1938). Diversity of vocabulary and the harmonic series law of word-frequency distribution. *Psychological Record*, 2, 379–386.

Chao, J., Finzer, B., Rosé, C. P., Jiang, S., Yoder, M., Fiacco, J., Murray, C., Tatar, C., & Wiedemann, K. (2022). Storyq: a web-based machine learning and text mining tool for k-12 students. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 2*, 1178–1178.

Chase, C. C., Marks, J., Malkiewich, L. J., & Connolly, H. (2019). How teacher talk guidance during invention activities shapes students' cognitive engagement and transfer. *International Journal of STEM Education*, 6, 1–22.

Chawla, N. V. (2010). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, 875–886.

Chi, M. T., Adams, J., Bogusch, E. B., Bruchok, C., Kang, S., Lancaster, M., Levy, R., Li, N., McEldoon, K. L., Stump, G. S., et al. (2018). Translating the icap theory of cognitive engagement into practice. *Cognitive science*, 42(6), 1777–1832.

Chi, M. T. & Wylie, R. (2014). The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4), 219–243.

Chin, C. & Chia, L.-G. (2004). Problem-based learning: Using students' questions to drive knowledge construction. *Science education*, 88(5), 707–727.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Cooper, A. et al. (2012). What is analytics? definition and essential characteristics. *CETIS Analytics Series*, 1(5), 1–10.

Corno, L. & Mandinach, E. B. (1983). The role of cognitive engagement in classroom learning and motivation. *Educational psychologist*, 18(2), 88–108.

Costa, A. & Kallick, B. (2015). Five strategies for questioning with intention. *Educational Leadership*, 73(1), 66–69.

Covington, M. A. & McFall, J. D. (2010). Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2), 94–100.

Craik, F. I. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, 11(6), 671–684.

Cui, W., Zhang, J., Li, Z., Lopez, D., Das, K., Malin, B., & Kumar, S. (2023). A divide-conquer-reasoning approach to consistency evaluation and improvement in blackbox large language models. In *Socially Responsible Language Modelling Research*.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied linguistics*, 24(2), 197–222.

Dawson, S., Gašević, D., Siemens, G., & Joksimovic, S. (2014). Current state and future trends: A citation network analysis of the learning analytics field. In *Proceedings of the fourth international conference on learning analytics and knowledge*, 231–240.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dillard, J. P. & Pfau, M. (2002). *The persuasion handbook: Developments in theory and practice*. Sage Publications.

D'Mello, S., Dieterle, E., & Duckworth, A. (2017). Advanced, analytic, automated (aaa) measurement of engagement during learning. *Educational psychologist*, 52(2), 104–123.

Driscoll, M. P. & Burner, K. J. (2005). Psychology of learning for instruction.

Dumont, J. & Zurn, P. (2007). Immigrant health workers in oecd countries in the broader context of highly skilled migration. organisation for economic co-operation and development oecd. *International Migration Outlook. Annual Report*.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., et al. (2021). Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57, 101994.

Farquharson, K., Centanni, T. M., Franzluebbers, C. E., & Hogan, T. P. (2014). Phonological and lexical influences on phonological awareness in children with specific language impairment and dyslexia. *Frontiers in psychology*, 5, 98309.

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research*, 58(3), 840–852.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*, volume 10. Springer.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1), 59–109.

Fulford, A. N. I. & Ng, A. (2023). Chatgpt prompt engineering for developers. deeplearning. ai.

Garcia-Ponce, E. E. & Tavakoli, P. (2022). Effects of task type and language proficiency on dialogic performance and task engagement. *System*, 105, 102734.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59, 64–71.

Gómez Vera, G., Sotomayor, C., Bedwell, P., Domínguez, A. M., & Jéldrez, E. (2016). Analysis of lexical quality and its relation to writing quality for 4th grade, primary school students in chile. *Reading and Writing*, 29, 1317–1336.

Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research. *Educational Psychologist*, 50(1), 14–30.

Greene, B. A., Miller, R. B., Crowson, H. M., Duke, B. L., & Akey, K. L. (2004). Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation. *Contemporary educational psychology*, 29(4), 462–482.

Hampson, T. & McKinley, J. (2023). Qualitative and quantitative are data types not paradigms: An mma framework for mixed research in applied linguistics. *LEARN Journal: Language Education and Acquisition Research Network*, 16(2), 1–7.

Hargreaves, D. H. (1984). Teachers' questions: Open, closed and half-open. *Educational Research*, 26(1), 46–51.

Hattie, J. A. & Yates, G. C. (2014). Using feedback to promote learning. *Applying science of learning in education: Infusing psychological science into the curriculum*, 45–58.

Helme, S. & Clarke, D. (2001). Identifying cognitive engagement in the mathematics classroom. *Mathematics Education Research Journal*, 13(2), 133–153.

Henderson, J. B. (2019). Beyond "active learning": How the icap framework permits more acute examination of the popular peer instruction pedagogy. *Harvard Educational Review*, 89(4), 611–634.

Hsiao, J.-C., Chen, S.-K., Chen, W., & Lin, S. S. (2022). Developing a plugged-in class observation

protocol in high-school blended stem classes: Student engagement, teacher behaviors and student-teacher interaction patterns. *Computers & Education*, 178, 104403.

Ireland, M. E. & Henderson, M. D. (2014). Language style matching, engagement, and impasse in negotiations. *Negotiation and conflict management research*, 7(1), 1–16.

Japkowicz, N. & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5), 429–449.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language learning*, 63, 87–106.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*, 53, 61–79.

Johnson, W. (1944). Studies in language behavior: A program of research. *Psychological Monographs*, 56(2), 1–15.

Karimah, S. N. & Hasegawa, S. (2022). Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods. *Smart Learning Environments*, 9(1), 31.

Knight, S. & Shum, S. B. (2017). Theory and learning analytics. *Handbook of learning analytics*, 1, 17–22.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, 1(1), 60–69.

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4), 212–218.

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. In *Data democracy*, 83–106. Elsevier.

Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 100213.

Lee, O. & Anderson, C. W. (1993). Task engagement and conceptual change in middle school science classrooms. *American educational research journal*, 30(3), 585–610.

Lee, Y. & Kinzie, M. B. (2012). Teacher question and student response with regard to cognition and language use. *Instructional science*, 40, 857–874.

LemaÃŽtre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research*, 18(17), 1–5.

Li, S., Lajoie, S. P., Zheng, J., Wu, H., & Cheng, H. (2021). Automated detection of cognitive engagement to inform the art of staying engaged in problem-solving. *Computers & Education*, 163, 104114.

Lissón, P. & Ballier, N. (2018). Investigating lexical progression through lexical diversity metrics in a corpus of french l3. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (23).

Mackey, A. & Gass, S. M. (2011). *Research methods in second language acquisition: A practical guide*, volume 7. John Wiley & Sons.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development*. Springer.

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American educational research journal*, 37(1), 153–184.

Martin, F., Zhuang, M., & Schaefer, D. (2023). Systematic review of research on artificial intelligence in k-12 education (2017–2022). *Computers and Education: Artificial Intelligence*, 100195.

Mayer, R. E. (2005). Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning*, 41, 31–48.

Mayer, R. E. (2008). Applying the science of learning: evidence-based principles for the design of multimedia instruction. *American psychologist*, 63(8), 760.

McCarthy, P. M. & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4), 459–488.

McCoach, D. B., Gable, R. K., Madura, J. P., McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). Defining, measuring, and scaling affective constructs. *Instrument development in the affective domain: School and corporate applications*, 33–90.

Meece, J. L., Blumenfeld, P. C., & Hoyle, R. H. (1988). Students' goal orientations and cognitive engagement in classroom activities. *Journal of educational psychology*, 80(4), 514.

Miller, J. F., Andriacchi, K., & Nockerts, A. (2016). Using language sample analysis to assess spoken language production in adolescents. *Language, speech, and hearing services in schools*, 47(2), 99–112.

Min, S., Lewis, M., Zettlemoyer, L., & Hajishirzi, H. (2021). Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

Morris, J. & Chi, M. T. (2020). Improving teacher questioning in science using icap theory. *The Journal of Educational Research*, 113(1), 1–12.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of liwc2015.

Piaget, J. (2003). Part i: Cognitive development in children–piaget development and learning. *Journal of research in science teaching*, 40.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In *Handbook of self-regulation*, 451–502. Elsevier.

Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (mslq). *Educational and psychological measurement*, 53(3), 801–813.

Price, J. R. & Jackson, S. C. (2015). Procedures for obtaining and analyzing writing samples of school-age children and adolescents. *Language, speech, and hearing services in schools*, 46(4), 277–293.

Quesnelle, K. M., Zaveri, N. T., Schneid, S. D., Blumer, J. B., Szarek, J. L., Kruidering, M., & Lee, M. W. (2021). Design of a foundational sciences curriculum: applying the icap framework to pharmacology education in integrated medical curricula. *Pharmacology Research & Perspectives*, 9(3), e00762.

Radwan, A. M. & Cataltepe, Z. (2017). Improving performance prediction on education data with noise and class imbalance. *Intelligent Automation & Soft Computing*, 1–8.

Ransdell, S. & Wengelin, Å. (2003). Socioeconomic and sociolinguistic predictors of children's l2 and l1 writing quality. *Arobase*, 1, 22–29.

Rodriguez, A. D., Dearstyne, K. R., & Cleland-Huang, J. (2023). Prompts matter: Insights and strategies for prompt engineering in automated software traceability. In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, 455–464. IEEE.

Savelka, J., Denny, P., Liffiton, M., & Sheese, B. (2023). Efficient classification of student help requests in programming courses using large language models. *arXiv preprint arXiv:2310.20105*.

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2014). Motivation in education: Theory, research, and applications. *(No Title)*.

Shahriar, T., Matsuda, N., & Ramos, K. (2023). Assertion enhanced few-shot learning: Instructive technique for large language models to generate educational explanations. *arXiv preprint arXiv:2312.03122*.

Sharkins, K., Newton, A., Causey, C., & Ernest, J. M. (2017). Flipping theory: Ways in which children's experiences in the 21st century classroom can provide insight into the theories of piaget and vygotsky. *Southeast Asia Early Childhood Journal*, 6, 11–18.

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10), 1380–1400.

Smart, J. B. & Marshall, J. C. (2013). Interactions between classroom discourse, teacher questioning, and student cognitive engagement in middle school science. *Journal of Science Teacher Education*, 24(2), 249–267.

Smiley, W. & Anderson, R. (2011). Measuring students' cognitive engagement on assessment tests: A confirmatory factor analysis of the short form of the cognitive engagement scale. *Research & Practice in Assessment*, 6, 17–28.

Svanes, I. K. & Andersson-Bakken, E. (2023). Teachers' use of open questions: investigating the various functions of open questions as a mediating tool in early literacy education. *Education Inquiry*, 14(2), 231–250.

T. Thordardottir, E. & Weismer, S. E. (2001). High-frequency verbs and verb diversity in the spontaneous speech of school-age children with specific language impairment. *International Journal of Language & Communication Disorders*, 36(2), 221–244.

Tatar, C., McClure, J., Bickel, F., Ellis, R., Wiedemann, K., Chao, J., Jiang, S., & Rosé, C. P. (2023). Examining High School Students' Self-Efficacy in Machine Learning Practices.

Tatar, C., Yoder, M. M., Coven, M., Wiedemann, K., Chao, J., Finzer, W., Jiang, S., & Rosé, C. P. (2021). Modeling unstructured data: Teachers as learners and designers of technology-enhanced artificial intelligence curriculum. In *Proceedings of the 15th International Conference of the Learning Sciences-ICLS 2021*. International Society of the Learning Sciences.

Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., Qin, Z., Hui, K., Zhao, Z., Gupta, J., et al. (2022). Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35, 21831–21843.

Thordardottir, T., Susan, E., & Weismer, E. (2001). High-frequency verbs and verb diversity in the spontaneous speech of school-age children with specific language impairment. *International Journal of Language & Communication Disorders*, 36(2), 221–244.

Tofade, T., Elsner, J., & Haines, S. T. (2013a). Best practice strategies for effective use of questions as a teaching tool. *American journal of pharmaceutical education*, 77(7), 155.

Tofade, T., Elsner, J., & Haines, S. T. (2013b). Best Practice Strategies for Effective Use of Questions as a Teaching Tool. *American Journal of Pharmaceutical Education*, 77(7), 155.

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., et al. (2018). Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416*.

Wang, X., Wang, Y., Xu, C., Geng, X., Zhang, B., Tao, C., Rudzicz, F., Mercer, R. E., & Jiang, D. (2023). Investigating the learning behaviour of in-context learning: a comparison with supervised learning. *arXiv preprint arXiv:2307.15411*.

Wang, X., Wen, M., & Rosé, C. P. (2016). Towards triggering higher-order thinking behaviors in moocs. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 398–407.

Wasik, B. A. & Hindman, A. H. (2013). Realizing the Promise of Open-Ended Questions. *The Reading Teacher*, 67(4), 302–311.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022a). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022b). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824–24837.

Wigfield, A., Guthrie, J. T., Perencevich, K. C., Taboada, A., Klauda, S. L., McRae, A., & Barbosa, P. (2008). Role of reading engagement in mediating effects of reading comprehension instruction on reading outcomes. *Psychology in the Schools*, 45(5), 432–445.

Wiggins, B. L., Eddy, S. L., Wener-Fligner, L., Freisem, K., Grunspan, D. Z., Theobald, E. J., Timbrook, J., & Crowe, A. J. (2017). Aspect: A survey to assess student perspective of engagement in an active-learning classroom. *CBE—Life Sciences Education*, 16(2), ar32.

Wood, C. L., Bustamante, K. N., Schatschneider, C., & Hart, S. (2019a). Relationship between children's lexical diversity in written narratives and performance on a standardized reading vocabulary measure. *Assessment for Effective Intervention*, 44(3), 173–183.

Wood, C. L., Bustamante, K. N., Schatschneider, C., & Hart, S. (2019b). Relationship between children's lexical diversity in written narratives and performance on a standardized reading vocabulary measure. *Assessment for Effective Intervention*, 44(3), 173–183.

Wu, J.-Y. (2021). Learning analytics on structured and unstructured heterogeneous data sources: Perspectives from procrastination, help-seeking, and machine-learning defined cognitive engagement. *Computers & Education*, 163, 104066.

Xie, K., Heddy, B. C., & Greene, B. A. (2019). Affordances of using mobile technology to support experience-sampling method in examining college students' engagement. *Computers & Education*, 128, 183–198.

Yang, J. S., Rosvold, C., & Bernstein Ratner, N. (2022). Measurement of lexical diversity in children's spoken language: Computational and conceptual considerations. *Frontiers in psychology*, 13, 905789.

Yogev, E., Gal, K., Karger, D., Facciotti, M. T., & Igo, M. (2018). Classifying and visualizing students' cognitive engagement in course readings. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 1–10.

Yun, Z., Nan, M., Da, R., & Bing, A. (2011). An effective over-sampling method for imbalanced data sets classification. *Chinese Journal of Electronics*, 20(3), 489–494.

Zeng, Z., Chen, P., Jiang, H., & Jia, J. (2023). Challenge llms to reason about reasoning: A benchmark to unveil cognitive depth in llms. *arXiv preprint arXiv:2312.17080*.

Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., & Han, J. (2023). Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

Zimmermann-Niefield, A., Polson, S., Moreno, C., & Shapiro, R. B. (2019). Youth making machine learning models for gesture-controlled interactive media. In *Proceedings of the Interaction Design and Children Conference*, 63–74, New York, NY, USA. Association for Computing Machinery.

APPENDICES

Appendix A: Hyperparameters of Traditional ML

Table A.1

Support Vector Machine (SVM), Random Forest (RF), Decision Trees (DT), ADABoost

ML Classifier	Parameter	Values
SVM	Kernel	'linear', 'sigmoid', 'poly', 'rbf'
	C	0.1, 1, 10, 100, 1000
	Gamma	1, 0.1, 0.01, 0.001, 0.0001
RF	Number of Estimators	100
	Criterion	'Gini'
	Max Depth	none
	Min Samples Leaf	1
	Min Samples Split	2
	Max Features	'Auto'
	Bootstrap	True
DT	Criterion	'Gini'
	Max Depth	none
	Min Samples Leaf	1
	Min Samples Split	2
ADABoost	Number of Estimators	50
	Learning Rate	1.0
	Algorithm	'SAMME.R'
	Base Estimator	DecisionTreeClassifier (max_depth=1)

Appendix B: Final Prompt Full Set

A. Few Shot with reasoning + General COT (step by step) < Black, Green, Red, Blue>

B. Few Shot with reasoning + General COT (step by step) + Assertion (do and don't)<Black, Green, Red, Blue, Orange>

----------Prompt Starts Here----------

Your task is to identify the label of the statement delimited by triple backticks

Read the instructions below:

Step 1: Read the question and statement attentively to understand the context and the nature of the statement provided.

Step 2: Determine the initial cognitive engagement level of the statement using the definitions of the provided cognitive engagement labels - passive, active, and constructive.

1. Passive engagement: a statement is classified as "Passive" when the individual is only receiving information without interacting with it or adding anything to it. Passive engagement typically involves listening, reading, or receiving information without actively processing, manipulating, or reflecting upon it.

2. Active engagement: a statement is classified as "Active" when the response involves applying knowledge, analyzing information, or manipulating information but not generating new ideas or concepts.

3. Constructive engagement: a statement is classified as "Constructive" if it reflects reasoning, justification, or thoughtful consideration based on prior knowledge.

Step 3: Assess why it corresponds to the label you placed it in. Consider the extent to which it demonstrates recall of basic information (passive), application of learned knowledge to slightly different contexts (active), or a deeper level of analysis and synthesis of various concepts (constructive).

Step 4: Critically evaluate whether the statement could potentially belong to other labels. Examine the nuances of the statement to see if there are elements that might indicate a higher or lower level

of cognitive engagement.

Step 5: To upgrade the statement to a higher engagement level, propose alterations that would make it align with the criteria for the "Active" category. This could involve adding details that show the application of learned knowledge to familiar yet slightly different contexts, or demonstrating problem-solving based on previous experiences.

Step 6: Explore how the statement can be restructured to meet the criteria of the "Constructive" engagement category. Consider adding elements that showcase deeper analysis, critical evaluation, or synthesis of multiple concepts to create a more nuanced and thoughtful response. Step 7: Finally, revisit the question and statement to evaluate the original cognitive engagement level making sure the prediction of cognitive engagement is accurate.

Based on your understanding of cognitive engagement and the labeled examples provided, determine the level of engagement for the unlabeled text provided.

‘ ‘ ‘

Question: Why do people write reviews?

Statement: People write reviews to express their feelings on a certain thing to condemn a praise a business, franchise, movie, or book.

Label: <Generate label>

Chain-of-thought: <Generate the chain-of-thought>

‘ ‘ ‘

Use the following examples delimited by triple quotes to understand which label the statement belongs to.

’ ’ ’

Question: What features do you think are indicators of positive reviews?

Statement: Words like love, excellent, greatest, amazing, enjoy, awesome, best.

Label: Passive

Reasoning: because it is a direct response that involves recalling or listing words without further analysis or interaction.

Question: What is one strategy you (as a human) can use to determine if a review is positive or negative?

Statement: I can tell if the person liked something or not. Label: Passive Reasoning: because it does not specify any strategies or reflection to distinguish between positive and negative sentiments.

Question: When you click on the row, the feature in this review will be highlighted in the feature graph (like the one you have seen in the Light On Light Off activity).Which feature do you think is it?

Statement: Because it's associated with positivity.

Label: Passive

Reasoning: because it is simple information without reflection without delving into specific details, analysis, or reflection.

Question: What is one strategy you can use to determine what features someone has used to build a classification model?

Statement: I can use major words that people say in reviews first. Words like 'love,' 'hate,' 'bad,' 'delicious,' and more.

Label: Passive

Reasoning: because it only has recall words and delve into any analysis, reflection, or application.

Question: What is one strategy you can use to determine what features someone has used to build a classification model?

Statement: You can look at the data set and find words that really stand out to you or words that have a strong emotional connotation. You can also check the graphand the probability in terms of the features being used or how strongly they correlate with the result.

Label: Active

Reasoning: because it summarizes and organizes the information in a broad manner

Question: What is one strategy you (as a human) can use to determine if a review is positive or

negative? Statement: One strategy that you can use to determine if a review is positive or negative is looking at diction, which is word choice, and how the words are being used.

Label: Active

Reasoning: because it details a method of analyzing the word choice in reviews, demonstrating the application of acquired knowledge to assess sentiments.

Question: When you click on the row, the feature in this review will be highlighted in the feature graph (like the one you have seen in the Light On Light Off activity). Which feature do you think is it?

Statement: Love is the most defining word in this review, if it were changed to 'hate' it would have a completely different meaning

Label: Active

Reasoning: because it demonstrates the application and analysis of knowledge in a familiar context but does not generate new ideas.

Question: What is one strategy you can use to determine what features someone has used to build a classification model?

Statement: You can test multiple reviews with words that you think may be the features to determine if they are actually features.

Label: Active

Reasoning: because it demonstrates the application and manipulation of knowledge in a familiar context without generating new insights.

Question: Why do people write reviews?

Statement: To share their experience of a certain product or service so that they can either warn or recommend it to people. Sharing experiences is important so that way others who have not experienced it can know what they are getting in to.

Label: Constructive

Reasoning: because it provides an understanding and reasoning of the broader context and implications why sharing experience is important.

Question: If none of the 10 features are present in your review, try again with another review. If some of the 10 features are in your review, examine both your review and the feature graph. What do you think these features are?

Statement: I think these features are key words and numbers. Like the example used the word 'love' which implies a positive reply. The numbers also because if you say 1 out of 10 that's bad but if you say 10 out of 10 that's good.

Label: Constructive Reasoning: because it provides interpretation and application to generate insights about the potential features in reviews.

Question: What is one strategy you (as a human) can use to determine if a review is positive or negative?

Statement: If I am having a conversation with somebody it will be easy to detect if the review is good or bad by word choice and their tone. If they wrote it, I will be able to see key words that point in either a positive or negative direction.

Label: Constructive

Reasoning: because it demonstrates a depth of reasoning and reflection of how to determine if a review is positive or negative.

Question: What kinds of reviews can make our world a better place?

Statement: Some reviews that can make the world a better place is if it's a review about a foreign country then it can give some insight into what is happening within that country. Or even here in the United States, it can share what's happening within their state and let the rest of the world know.

Label: Constructive

Reasoning: because it provides reflection, thoughtful consideration and reasoning about the societal value and potential impact of reviews in fostering global understanding and awareness.

' ' '

A few facts about identifying the cognitive engagement level that you must assert while determining the level of engagement for the unlabeled text provided:

- Do label the statement as Constructive if they are forming an opinion about its usefulness, and providing reasoning for their opinion.
- Do label the statement as Constructive when the statement provides their interpretation and reasoning to the question.
- Do label the statement as Constructive when they form a hypothesis about why the model learned a weight for a certain feature.
- Do label the statement as Constructive when the statement shows active engagement with the information.
- Avoid labeling a statement as Active or Constructive based solely on speculative language like 'I think' or 'possibly'.