Exploring the Potential of LLM Enhanced Engineering Problems Through Student Performance, Preferences, Mental Workload and Emotion

by

Runu Proma Das

(Under the Direction of Beshoy Morkos)

Abstract

Engineering Problems have long been used in the classroom for educational purposes. The process of generating said problems has not changed significantly in decades. However, the advent of large language models presents an opportunity to explore how such AI tools may be used to support problem generation. This thesis proposes a novel approach to redefine the development of engineering problems using Generative Artificial Intelligence (Gen-AI) and acknowledging its unique features. Students' preferences, performance, mental workload, and emotions are evaluated using a mixed-method research approach with different generation sources. Ultimately, the study identifies substantial impacts of the generation of problems on students. The insights of this research contribute to developing the landscape of engineering pedagogy through the potential of Gen-AI to redefine traditional engineering problems and improve students' overall learning experiences.

Index words: Engineering Problem, Problem-solving, Generative AI, Mixed method research, Student Performance, Student Preference, Student Mental Workload, Student Emotion.

Exploring the Potential of LLM Enhanced Engineering Problems Through Student Performance, Preferences, Mental Workload and Emotion

by

Runu Proma Das

B.S., University of Dhaka, 2021

A Submitted to the Graduate Faculty of the

University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Masters of Science

Athens, Georgia

2024

Exploring the Potential of LLM Enhanced Engineering Problems Through Student Performance, Preferences, Mental Workload and Emotion

by

Runu Proma Das

Major Professor: Beshoy Morkos

Committee: Cheryl Gomillion

Andrew Jackson

Electronic Version Approved:

Ron Walcott

Dean of the Graduate School

The University of Georgia

December 2024

Dedication

"You have the right to work, but never to the fruit of work"

Chapter 2, Verse 47: Gita

I would like to dedicate this work to my dearest father, Priyo Ranjan Das, whose unconditional love, support, and faith made my journey so far. I am deeply grateful to the Almighty, whose grace and presence have uplifted me through challenges, providing me with the courage and endurance to persevere even in the most difficult moments. .

Acknowledgments

I am deeply grateful to my supervisor, Dr. Beshoy Morkos, for providing me with an incredible opportunity at his lab that has truly changed my life. His continuous guidance and support have been instrumental in my growth both as a researcher and as an individual. I would also like to express my gratitude to Dr. Andrew Jackson and Dr. Cheryl Gomillion for generously offering their time and providing valuable feedback to enhance my work. I am also thankful to all my labmates, especially Cody Carroll, David Joy, and Logan Smith, for their consistent support in helping me better comprehend the research. Lastly, I consider myself incredibly fortunate to have the continuous love and support of my mother, sister, and uncle.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Engineering problems are a fundamental element of formal education pedagogy. Traditional engineering problems are formed by acquired knowledge and experience. The process of problem generation served as an essential phase in problem-solving that has the ability to directly impact the outcome (Dinar & Shah, 2014; Eierman & Philip, 2003; Paulus, 1966). The importance of this phase is unintentionally overlooked in engineering pedagogy. Although Thomas and Carroll revealed that designers often prefer approaching problems in an ill-defined state rather than redefining them as well-defined, this approach may hinder effective problem-solving (Thomas & Carroll, 1979). However, Valkenburg and Dorst suggested modifying a problem more often results in a more effective design. Also, reflective practice in each design step can be helpful for navigating any problems and enhancing students' creativity and problem-solving skills (Valkenburg & Dorst, 1998). a deficient problem generation approach can lead to solutions that are unclear, overly simplistic, or incomplete (Cowan, 1986). Additionally, engineering courses are also designed to have homework problems for students. These homework problems, which are generally well-defined, are often derived from traditional textbooks which are in contrast ill-structured workplace problems. There are several studies to demonstrate the benefits of having problems correlate with real-world problems (Arnold et al., 2004; Sloboda, 2019). However, these homework problems and their generation in engineering have not received sufficient attention for research purposes. In an effort to pedagogical limitations, generative AI could be used to redefine some elements of problem generation.

Research with Large Language Models (LLM) generally focused on its problem-solving capabilities (Rodríguez-Echeverría et al., 2024; Tsai et al., 2023). Numerous scholars have evaluated the ability of LLMs to pass specific tests, primarily to assess their capacity to imitate human intelligence. For instance, OpenAI GPT-4 has demonstrated proficiency in passing the LSAT, GRE, Bar Exam, US Medical Licence Examination, and other complex assessments (Gilson et al., 2023; Katz et al., 2024). The capabilities of ChatGPT are consistently validated for each examination. ChatGPT is also recognized as an educational tool for students (Bernabei et al., 2023). The ChatGPT can serve as an assistant for students during problem-solving processes (Tsai et al., 2023). A study indicates that individuals across various fields, such as marketing and education, are utilizing ChatGPT with the intent of enhancing their productivity (Adeshola & Adepoju, 2023). Despite extensive research, the application of LLMs in engineering problem generation and formulation remains fully unexplored. With this research, we propose to redefine the fundamental processes of student learning and engineering problem generation using LLMs, specifically ChatGPT, one of the most widely used models. The application of Artificial Intelligence (AI) has gained significant prominence in recent years. The benefits of implementing AI are numerous; the AI chatbot can function as a peer for students, as a guide, as an expert. The traditional textbook in engineering covers the conventional problems which are necessary for foundation. However, problems generated by AI or assisted with AI can be more engaging, personalized, and related to real-life problems. It is hypothesized that AI-generated problems provide an opportunity to understand engineering content in terms of a contextualized way that is provocative to students. This study intends to evaluate the students' performance and emotions throughout the processes, leading to an understanding of the preferences, cognitive workload, and relative emotions of students. To direct the research systematically, there are several research questions implemented:

Table 1.1: Overview of Research Questions

ID	Research Questions	Hypotheses
RQ1	What distinctions in student performance emerge when presented with an engineering problem generated by a large language model (LLM)?	It is primarily envisioned that student performance will exhibit significant trends across different types of problems, with notable variations linked to the source of problem generation
RQ2	What variations in student preference are observed when students are presented with an engineering problem generated by a large language model (LLM)?	It is anticipated that student preference will indicate distinct trends across different types of problems, with notable variations linked to the source of problem generation
RQ3	What differences in cognitive workload emerge when students engage with an engineering problem generated by a large language model (LLM) compared to traditional engineering problems?	With this question, the researchers expect to determine significant variations in cognitive workload among students, depending on whether they are engaging with LLM-generated problems or traditional engineering problems
RQ4	Are students' emotional responses influenced by the generation of problems?	It is anticipated that the application of LLMs in problem generation will lead to differentiating students' emotions, with some emotions experienced more strongly impacted by the topic and source of the problem.

The implementation of a mixed-method study is anticipated to enhance the robustness of the research in terms of its intellectual merits. The combination of quantitative and qualitative data will provide comprehensive implications and generalizations for future research. From the quantitative data, the expectation is to identify the significance of each parameter. For instance, it is anticipated that there will be a significant relationship between the generation of problems and student preferences. The qualitative data will focus on understanding each parameter in-depth, potentially revealing aspects that may be overlooked in a quantitative approach. To achieve this, students' scripts will be evaluated using a structured coding scheme, and semi-structured interviews will be conducted. Providing a novel approach by conducting a mixed method study to reframing engineering problems beyond conventional learning and understanding methodologies, the core of engineering education will benefit.

The purpose of this study is to gain a fundamental understanding of the way students respond and react to engineering problems generated by large language models (LLMs), especially ChatGPT, in comparison with problems from conventional textbooks. Through assessment of student preferences, performance, cognitive load, and emotional reactions, this study intends to understand how LLM-generated problems can offer more interesting, relevant, and defined learning opportunities. With this novel understanding, educators could redefine their methods to incorporate more dynamic and customized problem-solving exercises. This could enhance students' creativity, problem-solving abilities, and competence in addressing challenging engineering problems in the real world.

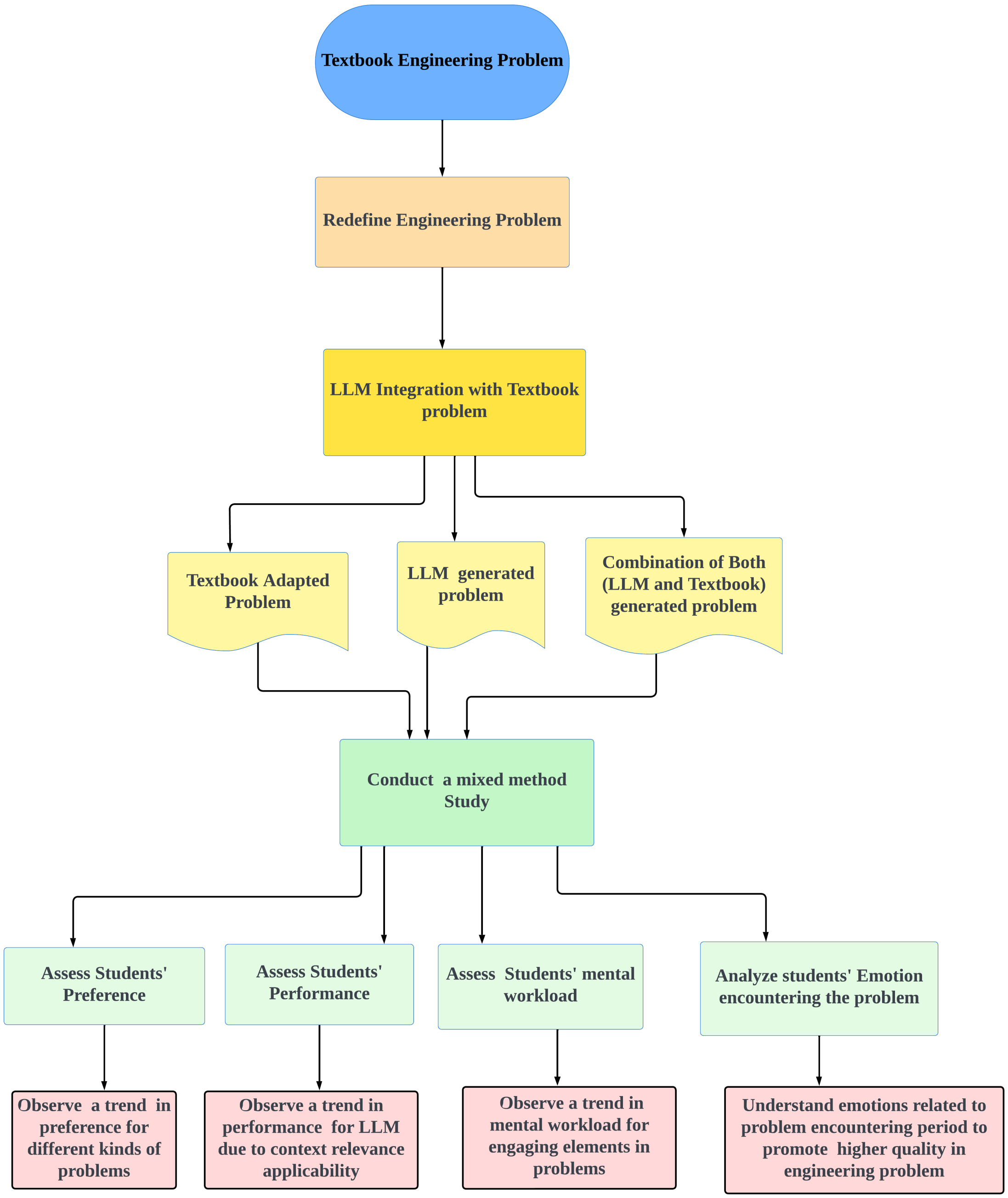


Figure 1.1: Research Exploration Model

Chapter 2

Literature Review

2.1 Overview

As part of that broader effort, this research captures a detailed picture of engineering problems specifically on different ways of generating engineering problems which did not get sufficient attention until now. In this section, the foundation of the research is provided in several sections. Section 2.1 describes an overview of the chapter including the engineering problem formulation and its pedagogy, Section 2.2 delivers the introduction of student output to the Problems, and afterward, Section 2.3 exits with the introduction of Generative AI and its role in engineering education.

2.1.1 Pedagogy of Engineering Problem

A wide range of initiatives were implemented recently to reshape the engineering curriculum and familiarize students with the engineering processes as early as the freshman year. In the initial phase of the engineering design process learning, students must identify the problem as part of the particular requirement statement (Kaushik et al., 2018). Recognizing engineering problems during the design process is a complex yet crucial task. Engineering problem involves effectively communicating a structured proposal to explain an engineering system in uncomplicated terms (Dym, 2013). Traditional engineering problems practiced by students are well-structured, encompassing every detail necessary to effectively solve a problem (Schuelke-Leech, 2021). The process of problem definition and formulation, also named "Problemization" by Harfield, relies heavily on the engineer's mindset, expertise, approach, biases, and personal preferences. (Coyne, 2005; Harfield, 2007).

With the shift towards hands-on activities and outcomes-based education (May et al., 2021; Yang et al., 2023), the types of problems solved in engineering curricula have not received sufficient attention. According to Jonassen (1997), professionals "Recognize different problem states which invoke certain solutions" (p. 71), while non-professionals typically do not adequately represent the problem when searching for a solution, often acting in a "haphazard and incoherent" manner (p. 71) (Jonassen, 1997). By merging engineering principles with practical application, the program intends to maintain the standard of instruction. For instance, Senior Capstone projects provide exposure to real-world engineering practice, where the ability to identify problems plays a crucial role (Akili, 2015; Bessette et al., 2014; Shah, Kames, et al., 2019). Although most engineering programs focus on problem-solving, there is a need to incorporate pedagogical constructivism. Constructivism is vital in professional engineering education because it facilitates students to overcome uncertainty and explore engineering outcomes across disciplinary boundaries (Kames et al., 2018; Rojter, 2009). Research suggests that open-ended engineering problems help students develop critical thinking skills and practical reasoning (de Andrade et al., 2022).

There is a strong argument that effective instruction in engineering problem formulation necessitates flexibility from instructors, as the process of generating engineering problems fosters a creative thinking approach among students rather than just memorization (Schoenfeld, 1980). The instruction for problem formulation or generation strategies is crucial; equally important is knowing when to implement such strategies. However, due to the complexity of problems, providing students with strategies or procedures to help students articulate the problems often leads to inconclusive results (Lecorchick III, 2017). Research indicates that the selection of engineering problems that help students develop both mathematical and engineering skills should be a top priority. The engineering problem must be useful, authentic, and relevant to actual engineering situations (Miyara, 1993).

2.1.2 Problem-based Learning in Pedagogy

In the past 20 years, there is a noticeable shift in higher education from traditional lecture-based learning towards the integration of problem-based learning (PBL) modules into degree programs.PBL was first introduced in medical education at McMaster University in Canada in 1980 and extended to other fields of study. In 1986, Barker defined PBL in civil engineering as a way to help students generate creative ideas and overcome the limitations of general education. According to Chandrasekara et al. (2013), PBL is now widely used in engineering institutions worldwide.(McCrum, 2017). Problem-based learning (PBL) is a practical strategy for instruction that enhances innovation, collaboration, and analytical thinking (Hadibarata et al., 2023). The shift in pedagogy emphasizes student-centered learning and active learning, which is in line with a larger trend in higher education. It is anticipated that PBL will continue to shape the field of engineering education as PBL advances in popularity, producing an entirely new generation of creative and flexible engineers (Jayaram, 2013).

2.2 Student Output to Engineering Problems

To evaluate students with engineering problems, the researchers found the following elements as important to address: Students' performance, Students' mental workload, and the students' preferences and emotions for engineering problems in terms of their generation sources. Addressing elements will guide the pedagogical practitioners to reframe engineering problems effectively.

2.2.1 Student Performance

The performance of students is an essential component of pedagogical research as student performance offers substantial perspectives into the challenges that students encounter during their educational journey. Recognizing such challenges is crucial in establishing productive pedagogical approaches that might amplify educational achievements. Many studies have compared various aspects of student performance, including the effects of individual versus teamwork (Kado & Kames, 2024; Landis & Haley, 2012), the influence of different learning contexts (online versus in-person), different textbook selections, etc. (Easton, 2021; May et al., 2023; Shah et al., 2022; Van Den Ham & Heinze, 2018). Listed studies have made a significant contribution to our understanding of the ways in which various educational settings impact student performance and academic achievement.

There is a significant gap in understanding how different sources of problem generation, such as experts, textbooks, or advanced technologies including large language models (LLMs), impact students' performance in engineering. While previous research has explored various aspects of student performance, the specific influence of problem sources on performance is overlooked. Bridging this gap could lead to improvements in students' engineering performance.

2.2.2 Student Preference

The preference of students has drawn a growing amount of attention, particularly with pedagogical approaches, learning environments, learning styles, and overall achievements. To improve the quality of education, the student's preference for learning is a vital factor (Lee & Sidhu, 2015). Evaluation of students' problem-solving in design contexts can be assisted by assessing the problems according to their preferences. Promoting an understanding of the preferences is key to improving motivation, academic success, and effective engagement with study materials (Benson & Morkos, 2011; Photopoulos et al., 2021).

According to the literature, "learning preference" entails the selection of one learning situation over another (Rezler & Rezmovic, 1981). Students' learning preferences can have a significant impact on their academic performance, as research has demonstrated that they have a variety of learning methods. According to a study on student performance, environments that are incompatible with students' preferred learning styles may result in a decline in students' preferences (Davis & Franklin, 2004). In a similar lens, understanding various learning styles enables educators to modify their pedagogical approaches. Understanding students' preferences can result in more diverse and productive pedagogical tactics, which improve students' academic performance and learning experiences (Madhu & Bhattachryya, n.d.).

Additionally, understanding student preferences is essential to the design process in the field of engineering education. Effective concept generation and screening, and design innovation, are prerequisites for engineering design. A study explored whether student designers' creative preferences could anticipate their capability to generate or select innovative design alternatives during concept development where the preference factors played an important role in predicting the design novelty of generated ideas and selected ideas (C. Toh & Miller, 2019).

Students' preferences for creativity, risk-taking, and ambiguity tolerance significantly form their learning engagement, confidence improvement, and creative thinking. A study reveals that individual perceptions and preferences for innovative ideas for design are influenced by risk attitudes and student education level. The study also suggests that students' engagement in active learning is strongly correlated with their preferences for creative thinking and their level of ambiguity tolerance (C. A. Toh & Miller, 2016a, 2016b).

Evaluation of students' problem-solving in design contexts can be assisted by assessing the problems according to the student's preferences. Promoting an understanding of the preferences is key to improving motivation, academic success, and effective engagement with study materials (Lynch et al., 2017; J. J. Ma et al., 2023; Photopoulos et al., 2021). Limited exploration is conducted in various aspects of problem generation in engineering education, with a primary focus on understanding student preferences. Further investigation into students' preferences and skills is essential in generating problem area (Dinar & Shah, 2014; Miyara, 1993; Shah, Morkos, et al., 2019).

2.2.3 Mental Workload

Measurements of mental workload (MWL) extend beyond World War II, and the development of MWL assessment was influenced by developments in aviation. NASA researchers Cooper and Harper created the Modified Cooper-Harper Scale (MCHS), which paved the way for further breakthroughs including the subjective workload assessment technique (SWAT) and the NASA Task Load Index (NASA-TLX). Workload assessment covers neurophysiological, physiological, and human behavioral assessment. Texts on workload have mentioned historical trends (Hancock et al., 2021). NASA TLX assessment is a typically used subjective method to evaluate mental workload. The NASA TLX consists of six subscales that reflect relatively independent dimensions, including mental, physical, and temporal demands, performance, effort, and frustration. Each dimension carries a task-specific question to figure out the most affected dimension by task (Hart, 2006).

The workload perceived by students is a crucial factor for producing academic stress which affects their performance and physiological health (Gardner & Parkinson, 2011). The complexity of tasks, time management, administrative support, and the relationship with peers and educators are impactors for stress. Expectations and motives also have a role in causing stress. Considering the impactors on the environment, tracking the stress to avoid excessive mental workload in school becomes essential (Rubio-Valdehita et al., 2014).

The performance of students is closely linked to their mental workload. Research indicates that both excessively high and low levels of mental workload can have a detrimental effect on performance (Marinescu et al., 2018). Scholars have also found that engineering students' problem-solving strategies correlate with their mental workload (Grigg & Benson, 2012; Grigg et al., 2013). To redefine the engineering problem formation, mental workload is an essential factor to be addressed. By acknowledging the mental workload's dimension particularly, the educators can plan to strategize the problems in enhancing students' mental ease and engagement towards problem-solving experience (Prastawa et al., 2018; Taraban et al., 2019)

2.2.4 Emotion Detection

According to the Dictionary of Cognitive Psychology, there is no formal definition of emotion: emotion is a mental state (Eysenck et al., 1994). Darwin initially made a statement about facial expressions as the facial expressions are universal among human populations (Darwin, 1872). According to research by Ekman and Friesen (1967, 1969a), there is a connection between particular facial muscle movements and particular emotions (Ekman & Keltner, 1970). In 1986, the basic emotions were identified as fear, disgust, anger, sadness, surprise, and happiness which were recognized by human facial behaviors (Ekman & Friesen, 1986). Using the Facial Action Coding System, researchers can utilize this method to figure out the emotions of learners from their facial expressions (Ekman & Friesen, 1978). It is worth mentioning that cultural variability can be a factor in interpreting emotion by facial expression, as emotions can be perceived in different ways in every culture systematically (Matsumoto, 1992).

Students' emotional recognition has achieved adequate attention in terms of their learning states and effectiveness. Generally, if a student emotionally interacts with study materials, there is a higher chance of success. Specifically, students' motivation, attention, problem-solving skills, and decision-making skills are highly dependent on students' emotional states (Kirn et al., 2012; Ochs & Frasson, 2004). Research indicates that emotions play a significant role in influencing students' cognitive functions, such as memory retrieval, information-processing strategies, and attentional resources. (Pekrun, 1992).

Several studies received determined recognition for research on emotions related to problem-solving outcomes. However, there is no specific emotion identified for the problem-solving process, but a suggestion to break the problem-solving process into several steps to indicate a responsible emotion (Pesonen & Hannula, 2014). Although students' emotional recognition of learning outcomes gained significant research attention, the impact of various generation sources in engineering problems on students' emotions is still an under-investigated area.

2.3 Role of Generative AI in Engineering Education

Artificial Intelligence (AI) has become widespread over the last few years because of conversational models such as chatbot models. This language model is designed to provide a real-time conversation with the user. This technology has also served as an effective education tool for the past few years because of its ease of use.

Artificial intelligence (AI) and natural language processing (NLP) have advanced significantly in the twenty-first century, enabling the development of generative artificial intelligent tools that can comprehend user inquiries, offer insightful answers, and automate jobs (Wollny et al., 2021). Generative Artificial Intelligence (Gen-AI) is a form of AI that creates novel material by establishing trends from prior data (Mittal et al., 2024). The foundation of Gen-AI is a combination of autoregressive models, Variational Autoencoders (VAEs), and Generative Adversarial Networks (GANs) (Gatla et al., 2024). Platforms including IBM Watson and Google Dialog Flow have played a critical role in the growth of conversational AI, which is driven by machine learning, deep learning, and natural language processing (Abedi et al., 2023). The creation of robust language models including OpenAI's GPT, which enhances text-based comprehension and generation, is a noteworthy accomplishment in conversational AI. For the purpose of creating responses that are logical, contextual, and human-including, the models are trained on enormous amounts of text data using proficient architectures. More sophisticated models are introduced into the related chatbot with every update, introducing new features such as conditional text generation (GPT 2), translation and summarization (GPT 3), faster outputs and text completion (GPT 3.5), multi-language functions, logical reasoning, and reliable API plugins (including Wolfram Alpha and ScholarAI) (GPT 4) (Liu et al., 2023).

In many different domains, AI has provided a substantial contribution to the creation, prediction, and analysis of synthetic data. AI has the potential to alter several engineering fields as the area of AI has a promising development. Notably, engineering education is redefined by the impact of generative AI tools, such as ChatGPT (Generative Pre-trained Transformer), on research, communication, teaching, and learning practices in engineering education (Johri, 2020; Pursnani et al., 2023).

In order to deliver individualized and successful learning experiences, one of Gen-AI tools such as ChatGPT specializes in developing real virtual simulations for experiential learning and offers individualized answers to student questions (Qadir, 2023). Education becomes essential in an AI-driven society to enable productive conversations with AI (Adeshola & Adepoju, 2023). According to research, AI-driven instructional technologies improved the analytical thinking, problem-solving, and academic achievement of students (SB et al., 2024). As an example of ChatGPT's usefulness in educational contexts, this AI tool is used in creating mathematical problems with various levels of difficulty (Hwang & Utami, 2024). Both students and educators can take advantage of AI-assisted learning, exhibiting its advantages in education (Pham et al., 2023; Smith et al., 2023). Furthermore, by identifying and addressing students' emotions, real-time machine-learning AI chatbots can improve emotional support and understanding in educational settings (Senapati, 2023). Incorporating AI into human-technology roles directs to lower operational expenses and enhances efficiency. A study found Gen-AI users are more productive, less dissatisfied, and have less cognitive demand than software-only users (Schmidhuber et al., 2021).

Educational environments are transformed by Gen-AI by improving intelligent support systems, plagiarism detection, assessment, and student performance prediction (Chaudhry et al., 2023; Khalil & Er, 2023; Pinto et al., 2023; Stojanov, 2023). The use of GenAI programming assistants in computing education is studied in research, providing issues related to advantages, effects on teaching strategies, and possible biases and plagiarism. The Gen AI tool as ChatGPT is a beneficial instrument for learning programming language as the tool offers explanations of code, alternative approaches, and real-time instruction. ChatGPT serves as an online tutor for both developers and novices, especially for people who are unfamiliar with coding languages. Although ChatGPT is useful, its effectiveness varies depending on the intricacy of the work. Because user instructions are crucial to the quality of AI-generated outputs (Amoozadeh et al., 2024; Azaria et al., 2024; Israelsen, 2023). By producing code based on students' textual descriptions, Codex, an LLM capable of annotating existing code in real-time assists students in becoming proficient code writers (Finnie-Ansley et al., 2022). Additionally, the Gen-AI tool enhances essay correction by providing focused guidance on vocabulary selection and logical organization. Despite the usefulness, Gen-AI can use enormous volumes of data to find new patterns in education, offer thorough insights, produce excellent outputs, and offer tailored replies (Wang et al., 2024). Integrating conversational Gen AI in education holds promise for enhancing efficiency and engagement in students' learning experiences. The use of Gen-AI tools in graduate engineering education highlights the potential for problem-solving, self-paced learning, instant feedback, and reducing instructor workload (Abedi et al., 2023). Gen AI is used to prepare intricate math questions due to a time-saving option for experts (Wang et al., 2024).

Though Generative AI (Gen AI) offers numerous benefits, it also presents significant limitations. As highlighted in the literature, the main challenges of using ChatGPT in education are related to the quality, accuracy, and timeliness of the information it provides (Patrício & Gonçalves, 2024). Furthermore, ChatGPT's potential to facilitate plagiarism, fraud, and academic dishonesty, particularly in assessment practices, has raised concerns among educators. Issues surrounding data protection, privacy and security, and even discrimination further hinder its integration into educational contexts (Al Ghatrifi et al., 2023; Farrokhnia et al., 2024; Hsu & Ching, 2023).

A significant concern is the potential misuse of ChatGPT for malicious purposes, such as generating fake news and biased propaganda. This issue highlights the need to revisit traditional assessment methods. For instance, provided with advancements in AI technologies, there may be a case for reintroducing proctored, in-person assessments to reduce the likelihood of misuse of tools like ChatGPT. Instructors might need to implement stricter proctoring measures and prioritize paper-based assessments overseen by humans, ensuring that AI tools are not used improperly during testing. While ChatGPT and similar programs may have a role in general teaching and learning, the classroom environment may need to emphasize academic integrity more strongly during evaluations (Memarian & Doleck, 2023).

While recognizing the challenges and potential solutions, generative AI could significantly enhance the education of the next generation of engineering students by providing them with up-to-date technological knowledge (C. Chen et al., 2020). The integration of generative AI is essential for effectively tackling engineering problem generation and ensuring that students are well-prepared for real-world engineering challenges.

Chapter 3

Methodology

The objective of the investigation is to assess the impact of utilizing large language model (LLM)-generated engineering problems on students in comparison to traditional engineering problems. This study addresses to examine how LLM-generated problems influence students' performance, preferences, mental workload, and emotions. Furthermore, the research's insights will evaluate to whether these LLM-based problems offer any pedagogical advantages or challenges in comparison to traditional problem sets employed in engineering education. Through a comparative analysis of the three-generation types of problems, the study pursues to provide insights into the potential role of Gen-AI in modifying engineering education.

3.1 Overview

This chapter describes each step used to conduct the mixed method study in Figure 1.1. The chapter begins by describing the execution of the study including topic selection, question generation with ChatGPT, and experiment design in Section 3.2. Afterward, Section 3.3 addresses all the dependent variables analysis tools and methods to align with the literature and explains the importance of qualitative data and the analysis method to capture in-depth knowledge as well.

3.2 Research Execution

In this section, each step involved in the experiment is described. The section begins by describing the topic selection of problems performed by participants in Section 3.2.1, the following section outlines the selection criteria of participants in Section 3.2.2. As the section continues, it offers a detailed overview of the experiment design, including the Institutional Review Board (IRB) approval, experiment setting, and experiment steps in Section 3.2.3.

3.2.1 Problem Topic Selection

The study focuses on basic manufacturing topics that significantly impact mechanical engineering students. Its goal is to redefine how engineering problems are generated. To ensure clarity, a standard textbook for mechanical engineering is referenced to derive uniform questions. The topics selected for this research are based on the textbook and lessons from advanced manufacturing classes. The four topics chosen are bending, extrusion, forging, and machining, as they provide fundamental knowledge essential to manufacturing in mechanical engineering. Before carrying out the actual experiment, a pilot study was conducted to effectively define the research plan. Any concerns identified during this pilot study are addressed in the final experiments. In this research, GPT-3.5 (February 2024) is used to effectively generate questions that reference the textbook in a conversational format. Multiple attempts are made to create quality questions while avoiding unnecessary information. Keywords such as "interesting," "real-life engineering," and "engaging" were employed to create these questions with ChatGPT. The Figure3.1 and 3.2 illustrate the effectiveness of the prompts in generating questions. Each question starts with the "New chat" format to eliminate any biases from previous data. The questions were effectively generated and revised using ChatGPT each time followed by a thorough review by an expert. After several attempts, researchers and experts reached a consensus on the quality of the problem.

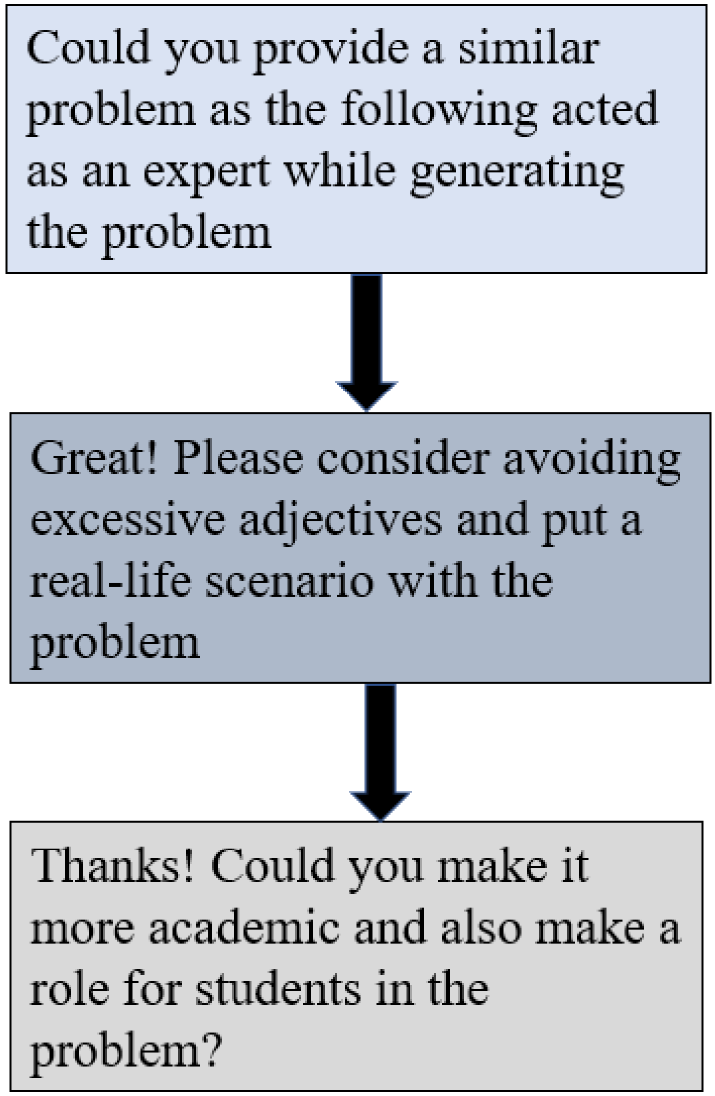


Figure 3.1: Effective Prompt for generating AI problem

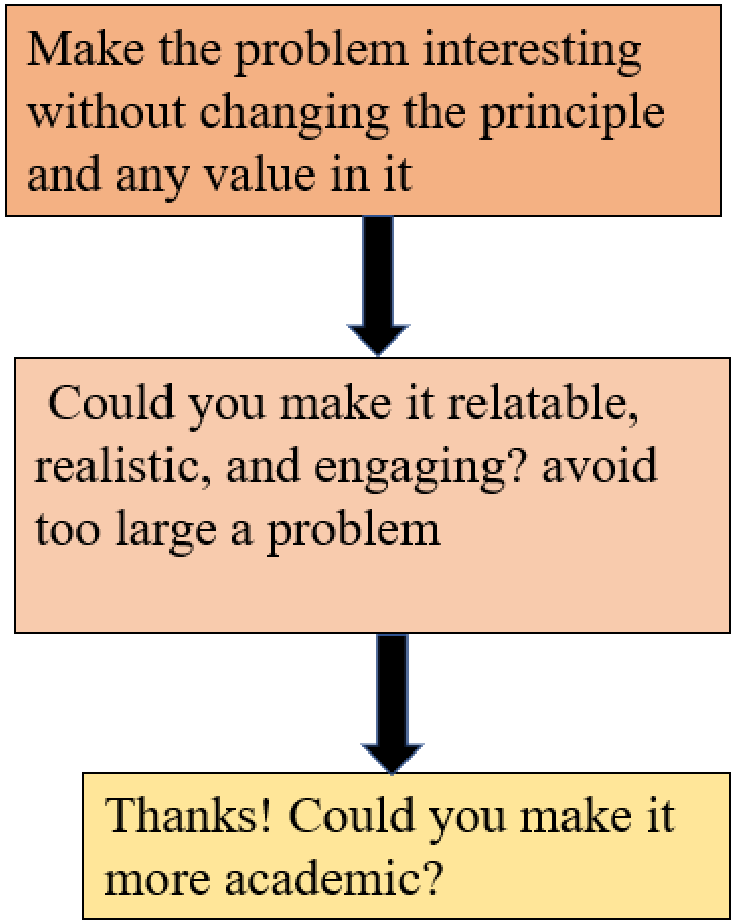


Figure 3.2: Effective Prompt for generating Textbook+AI problem

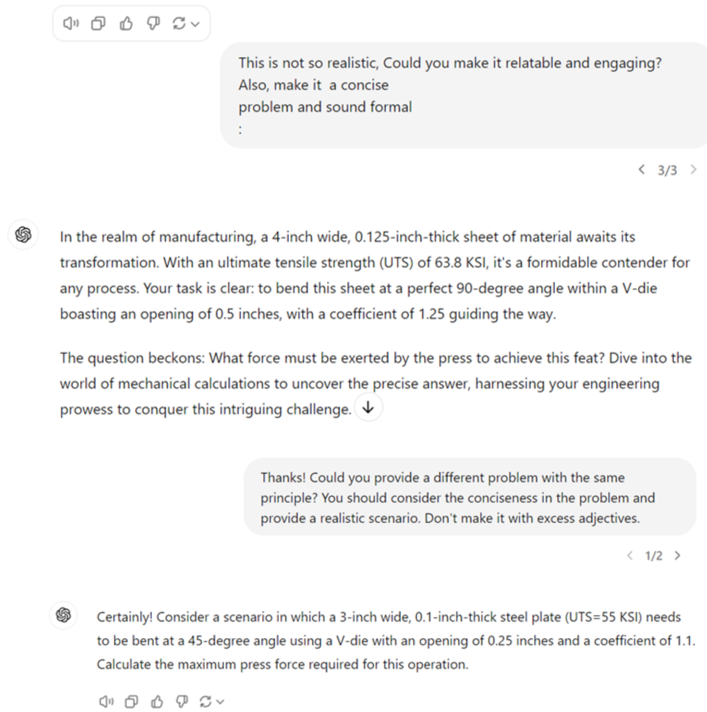


Figure 3.3: Problem Generation with AI

3.2.2 Subject and Recruitment

The experiment is designed based on manufacturing problems. The students from the mechanical engineering department are eligible for the study. At least one manufacturing course was taken by the participants so that they could understand the presented problems. There are no restrictions on the participants' education level. Both undergrad and graduate students were welcomed in this study. Ten male and four female students, a total of 14 participants from the University of Georgia (UGA) participated in the experiment.

3.2.3 Experiment Design

IRB Approval

As the participants are experimental subjects, the permission from UGA Institutional Review Board (IRB) is required and achieved. A consent form approved by the UGA IRB department is delivered to the participants before starting the experiment. Each consent form was read and signed by the participant and the researcher. The participants have the ability to access the forms later on as well.

Experiment Setting

The experiment is carried out in a quiet environment to minimize the distractions. The purpose of ensuring that was to make students completely concentrate on the experiment. Additionally, all studies are conducted one-on-one with identical supplies and a documented instruction is provided to each participant to maintain consistency among participants. Each participant took around 45 minutes to complete the pre-task instructions, problem-solving exercises, and post-experiment interviews.

Experiment Steps

The experiment consists of three portions. All participants are required to complete all three portions of the study as an integral component of the study. In the first portion of the experiment, the participants are offered a demographic survey using Qualtrics on the computer screen which reflects their education level, understanding of manufacturing class, and knowledge of generative AI tools in education. After completing this survey, the participants face a toss for selecting the problem topics. Each participant is tasked with selecting and solving three out of four problems. In addition, the participants encounter a computer screen with a webcam to record their facial expressions correctly. The participants are offered pen and paper to solve the problems and also provide their preference and mental workload after each problem with pen and paper. The participants are advised to provide both their problem-skimming time and problem-solving time. This information is crucial in terms of analyzing the relevant portion of the study to address the research question so that the researchers can provide the correct portion of the analyses. Because only a video is used to capture the complete event per participant. Participants are not informed about the generation of problems during the problem-solving event; this information is disclosed during the interview session.

Almost 35 minutes were taken to complete the problem-solving and provide other experiment parameters. After that, each participant has around 10 minutes of post-experiment interviews to share their thoughts regarding the impact of the generation source of problems. The participants are awarded 50 dollars for their participation.

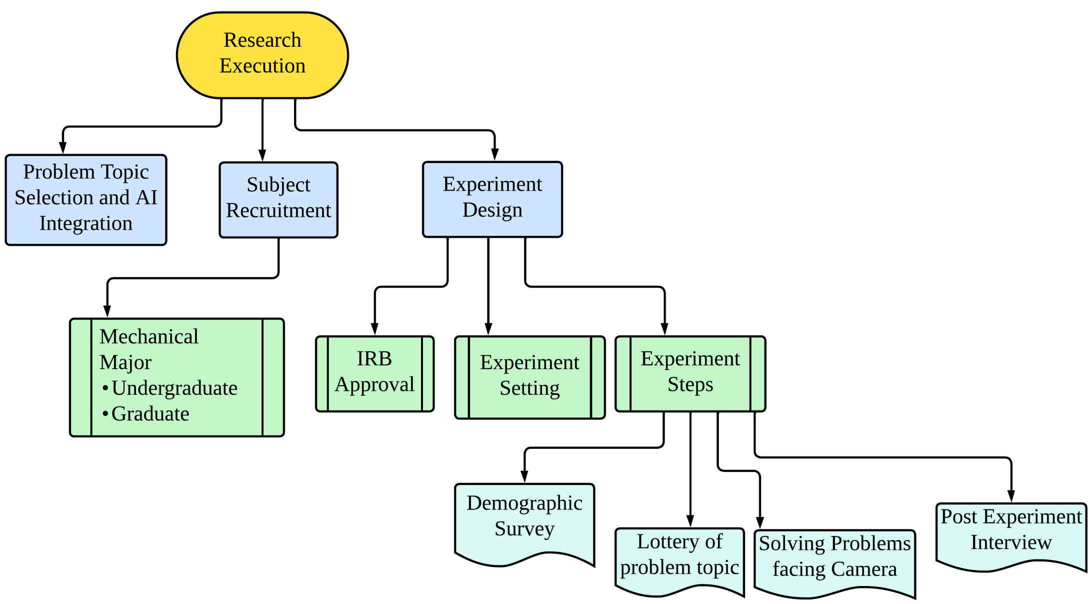


Figure 3.4: Research Execution Model

3.3 Data Collection

The section describes the tools and methods applied to collect the data. It divides into multiple subsections to address each variable with each data type efficiently. This study includes both quantitative and qualitative methods to produce a promising outcome. Section 3.3.1 outlines quantitative data and Section 3.3.2 describes the qualitative data.

3.3.1 Quantitative Data

In this study, most of the data are quantitative. The students' output such as performance, preference, mental workload, and emotion detection are treated as quantitative data. The data collection approaches for each data is different which are listed below:

Student Performance

With the variable, the students' performance towards the different types of problems is evaluated. According to the literature, performance criteria are defined as rules or principles by which students' responses or performance are judged. When the complexity of students' scripts is evaluated, the term 'Rubric' becomes relevant. The literature claims that performance criteria are a set of guidelines or standards that serve as critics for students' performances or responses. When the intricacy of students' scripts is factored into consideration, the term "rubric" becomes relevant (Arter & McTighe, 2001). There are two different kinds of rubrics for performance assessment; holistic rubric and analytical rubric. Holistic rubric typically assigns one grade without breaking the performance down into its components, depending on the overall assessment of the student's performance. The analytical rubric breaks down into multiple particular criteria, and each criterion can be scored separately to provide the student with thorough feedback on various elements of their work (Yulia, n.d.). This study is inspired by analytical rubric but does not provide any feedback to the participants.

The participants' scripts were graded in the following rubrics to reflect the performance evaluation in this research. The rubrics is as below:

Table 3.1: Student Performance Evaluating Rubrics

Content	Score
Assumption	01
Right Equation	01
Right Solution	05
Right Answer	02
Right Unit	01

It should be noted that there was a partial grading only for the right solution content. The highest score anyone can achieve for a problem is 10. The following boxplots reflect the data between the problem generation types and the total score the participants earned, and the problem type and the total score the participants earned.

Student Preference

To record the preference level for each problem from participants, a rating scale is provided at the end of the problem in pen and paper format. The preference rating scale is between the range of -2 to +2, where -2 represents the lowest preferable problem and +2 represents the highest preferable problem. The participants have the flexibility to rate the problem either before solving it or after solving it. In the meantime, they are also advised to be mindful while attempting to rate the scale to ensure that the preference rating is based on their experience of reading the problem rather than the solving process.

Student Mental Workload

Mental workload is a parameter to identify the mental effort to perform a task. In this study, the NASA TLX tool is used to define the mental workload for this experiment. Based on the problems in the experiment, the researchers have decided to adopt an unweighted NASA TLX mental workload with pen and paper. This is a This time-saving approach, supported by the literature provides the same result for the analysis (J. Ma et al., 2021). In the unweighted NASA TLX method, all raw data are used to analyze the mental workload. There are six factors to define the mental workload: mental demand, physical demand, temporal demand, effort, performance, and frustration. Each factor has a directing question and the participant is asked to provide the answer in 0-20 rating points. The directing questions with the rating scale of NASA TLX are illustrated in the following picture:

Mental Demand: The required mental and perceptual activity (e.g., thinking, deciding, calculating, remembering, looking, searching, etc) for completing the task was recorded.

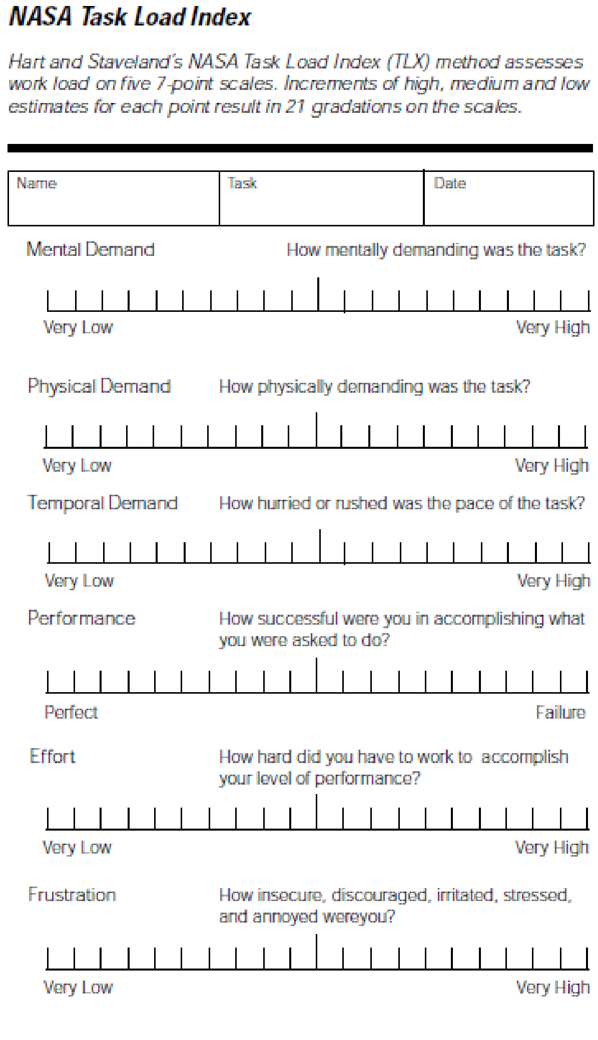
Physical Demand: The required physical activity (e.g., pushing, pulling, turning, controlling, activating, etc.) for performing the task was recorded.

Temporal Demand: The objective is to capture the time pressure experienced by the participants during the completion of the task.

Performance: The participants' level of proficiency in working through the task.

Effort: Participant perspectives on how tough the task was, either intellectually or physically.

Frustration: The amount of the participants experienced tension, discouragement, annoyance, and irritation as opposed to satisfaction, contentment, and relaxation during the task.



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

Figure 3.5: NASA Task Load Index Survey

Emotion Analysis

Based on the literature, human facial expressions are widely used as non-verbal communication which has the potential to identify human emotion (Jain et al., 2018). In this study, facial expressions are recorded on a computer with a Nexigo 60 webcam which has a high resolution. The Open Broadcaster Software

(OBS) is used to record their facial expressions to maintain the video quality. Participants are advised to avoid covering their faces with their hands and ensure their faces are visible to the camera. The video format of the camera is Motion JPEG which is further converted into MP4 for analysis purposes. In order to efficiently manage the huge volume of video data and achieve a balance between the requirement for thorough examination and time limits, a pace of 30 frames per second is implemented. Furthermore, Docker is used to create a more scalable and repeatable workflow that accelerates video processing and enabling expected visualization.

3.3.2 Qualitative Data

To address the research questions and understand the research in-depth, a qualitative approach is considered. A coding scheme is used to analyze students' responses, alongside semi-structured interviews to collect qualitative data in this study. Detailed information is provided in the following subsections:

Student Performance

Additionally, a deductive approach was incorporated to understand the performance of students in-depth. A coding scheme adopted by S.J. Grigg and L.C. Benson (2014) is used to capture each step related to problem solving. The scheme introduced 54 unique codes which are classified as knowledge access, self-management, and error types. It was created via a problem analysis methodology based on a mathematical education hierarchical structure. The coding method enables its potential to evaluate students' problem-solving techniques in detail, emphasizing the significance of comprehending mistakes and self-correction in educational research (Grigg & Benson, 2014).

Post-experiment Interview

To understand the research in depth, the researchers took into consideration participants' thoughts and ideas in the form of an interview. Interview is a broadly utilized tool to define the intricacy of a study. The approach taken to address this study was a one-session Semi-Structured interview. It lasts for almost 10-12 minutes per participant. A semi-structured interview merges a pre-planned list of questions (mostly structured with 'How', and 'Why' starting) with the chance for the interviewer to proceed deeper into topics or answers. The purpose of the interview is to capture the insights that may not be addressed in the quantitative data (Adeoye-Olatunde & Olenik, 2021). The interviews occurred as a post-experimental process in this research, and the questions for the interview were formed initially as open-ended to align with the research questions. The participants have the flexibility to answer the questions from their understanding. To lead the interview effectively, the following questions are asked:

ID	Questions
IQ1	What differences did you observe between the three problems you solved?
IQ2	(Explain the format for the three problems)...What are your thoughts on the problems now that you know how they were generated?
IQ3	How did the way the problem was generated impact your interest and engagement with the problem?
IQ4	How did the way the problem was generated impact your performance on the problem?
IQ5	Do you think the integration of LLM like ChatGPT can help to make engineering students? How?
IQ6	Do you have any final thoughts, information, or comments you would like to share with us?

In this research, the audio clip of the interview is recorded on a phone set. With the audio files, the typed transcripts for each participant are generated using the Restream website. However, there was a revision of automated transcripts by the researchers if the transcripts had captured all information correctly from the audio files. Finally, NVivo 12 Pro software is used to store the transcript data and analyze the transcripts.

3.4 Data Analysis

This section conveys the introduction of the data analysis tool and methods used to address the research question. Section 3.4.1 delivers Quantitative data analysis and Section 3.4.2 describes Qualitative data analysis tools used in analysis.

3.4.1 Quantitative Data Analysis

For quantitative data, two-way Analysis of Variance (ANOVA) is employed to determine differences across the generation of problems (AI, Textbooks, Textbook+AI problems) and types of problems (Bending, Extrusion, Forging, Machining) to address the research questions. The threshold of significance is identified as the P-value of the analysis as 0.05. Before doing the statistical analysis with emotion data, the videos are processed and analyzed with Py-Feat software. Py-Feat is an open-source Python toolbox that aids in facial expression feature detection, prepossessing, analysis, and visualization (Cheong et al., 2023). It is based on Ekman's emotion theory (1978) and examines seven fundamental emotions: anger, disgust, fear, sadness, happiness, surprise, and neutral. The toolkit uses the Facial Action Coding System (FACS) to identify facial muscle movements using Action Units (AUs) (Ekman & Rosenberg, 1997). It uses Open Face architecture to extract a Histogram of Oriented Gradient (HOG) characteristics from facial landmarks. Principal Component Analysis (PCA) minimizes HOG features, and machine learning techniques such as linear Support Vector Machines (SVM) and optimal gradient boosting (XGB) estimate 12 distinct AUs (Baltrušaitis et al., 2016; Chang & Lin, 2011; T. Chen & Guestrin, 2016). Py-Feat classifies fresh photos based on the degree to which a face matches a particular emotional facial expression using emotion detectors trained on intentionally posed or naturally evoked emotional facial expressions. It integrates advanced emotion detection models to provide a strong framework for accurately recognizing emotional expressions (Cheong et al., 2023). In this study, Feat Detectors including RetinaFace as face detector, MobileFacenet as Facial landmark model, Facenet as identity model, img2pose as facepose Model, XGB as Action Units Model, and ResMaskNet as emotion detection are used. Furthermore, with extracted data two two-way ANOVA and regression are also implemented to address a predicted relation between the variables. In the cases with significant (at $p \leq 0.05$) value of F-statistic in ANOVA, a post hoc analysis using Duncan's Multiple Range test (DMRT) was carried out to further test the significance of the observed difference between any given pair of mean results in the groups of three generations of problems (i.e., Textbook, AI, and Textbook + AI), four types of problem (i.e., Bending, Machining, Extrusion, and Forging), or for their 12 combinations ($3 \times 4 = 12$) reflecting interaction effects of both. Conveying this detailed analysis plan ensures the primary and interaction impacts of variables are thoroughly studied.

3.4.2 Qualitative Data Analysis

A thematic approach is introduced to categorize codes from the interview transcripts. The thematic approach generally combines multiple examination flexibility with research questions to strategic codes and themes. Codes generated from the data feature a small scale of intriguing information to address research questions. However, these codes work as a structural element of themes which are the identification of overall analytical observation aligned with research questions (Clarke & Braun, 2017). In this study, the coding scheme with themes is created followed by an inductive approach which is defined as a bottom-up process based on the participants' interview transcripts (Adeoye-Olatunde & Olenik, 2021). Each interview transcript is counted to produce the codes. After the completion of the coding scheme, an inter-rater reliability test is performed to provide reliable and consistent coding analysis and to support the credibility of the study. Different researchers can agree on the same code but interpret it from different angles. Inter-rater reliability becomes essential to maintain the consistency of the process when the inductive approach is utilized (Gisev et al., 2013). For the qualitative data, maintaining the credibility of coding schemes in students' problem-solving responses and interview transcripts, inter-rater reliability is utilized where two coders coded the participants' scripts individually. The development and assessment of a research tool emphasize the importance of inter-rater reliability (Armstrong et al., 1997). Inter-rater reliability is the degree of agreement between two or more raters (or observers, coders, or examiners) that examines the difficulty of applying a rating system consistently. Different researchers can agree on the same code but interpret it from different angles. Moreover, Inter-rater reliability becomes essential to maintain the consistency of the process when the inductive approach is utilized (Gisev et al., 2013). The same two coders are involved in coding both data based on the coding schemes. Numerous statistics may be used to assess inter-rater reliability (Lange, 2011). In this study, the Cohen Kappa coefficient is utilized. Cohen's Kappa coefficient written as lowercase Greek letter k captures the reliable statistic for the inter-rater reliability test. The coefficient's accepted value can be from -1 to +1 where +1 indicates complete agreement between raters and 0 indicates the degree of agreement that might be predicted by chance. Calculation of Cohen's kappa is performed by the following formula (McHugh, 2012):

$$k = \frac{P_o - P_e}{1 - P_e} \tag{3.1}$$

Here, Po represents the actual observed agreement and Pe represents the chance of agreement.

Cohen's Kappa Coefficient Interpretation:

Kappa Coefficient Value	Interpretation
Less than 0	No agreement
0.01–0.20	Negligible
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Strong
0.81–1.00	Almost perfect agreement

3.5 Mixed Method Integration

Mixed method research is a way to produce a more comprehensive and inclusive understanding by integrating quantitative and qualitative methodologies into a single study (Almeida, 2018). It improves the comprehension of the research by offering participants a voice and facilitating inquiry which enables evidence and opens new paths for exploration (Shorten & Smith, 2017). In this study, conducting a mixed method study offers to pinpoint the trend across all generations of problems and types of problems. The student's output is mostly quantitative data, however, integrating students' voices through interviews as qualitative data enhances the exploration of the research. Thus the mixed method approach promotes a deeper insight of students' responses where some significant outcomes might be overlooked if only a single method is considered.

Chapter 4

Result

4.1 Overview

This section covers the results from the analysis of the materials presented in Section 3. The section covers the following information: Section 4.2 outlines the data visualization, and Section 4.3 describes the statistical analysis. Finally, Section 4.4 delivers a coding scheme and the outcomes of the coding scheme with interview data.

4.2 Data Visualization

Box plots are applied to provide a clear picture of the data distribution. To interpret the box plot, key elements such as mean, median, outlier, and length of whiskers are considered. Section 4.2.1 illustrates the box plot of the total score earned by the participants, Section 4.2.2 represents the box plot of the participants' preferences, and Section 4.2.3 represents the box plot of each dimension of the mental workload variable. Section 4.2.4 provides a brief of emotion analysis.

4.2.1 Student Performance

The primary indicators of student performance used in this study are the script evaluation with rubrics and the coding scheme to understand the problem-solving techniques step by step. The two box plots reflect the data between the generation of problems and the total score the participants gained, and the types of problems and the total score the participants gained.

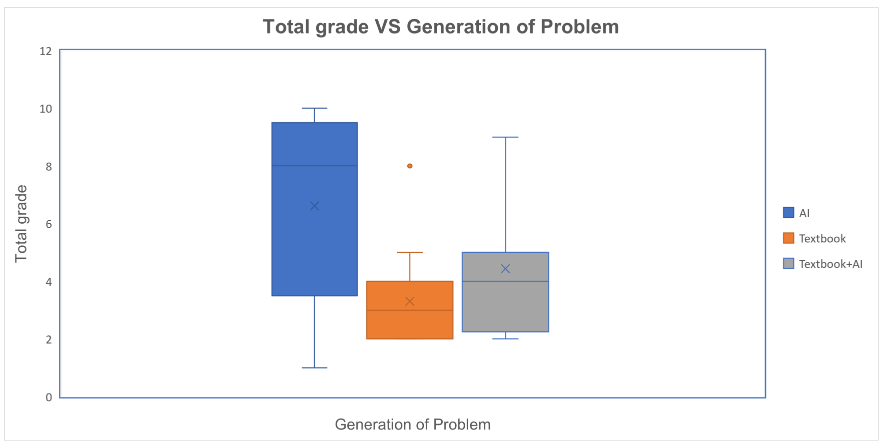


Figure 4.1: Box plot with Generation of Problem and Student Performance

Table 4.1: Generation of Problem and Student Performance Mean Values

Generation of Problems	**Student Performance ($\bar{x} \pm \sigma$)**
AI	6.62 ± 2.68
Textbook	3.31 ± 1.70
Textbook+AI	4.44 ± 2.28

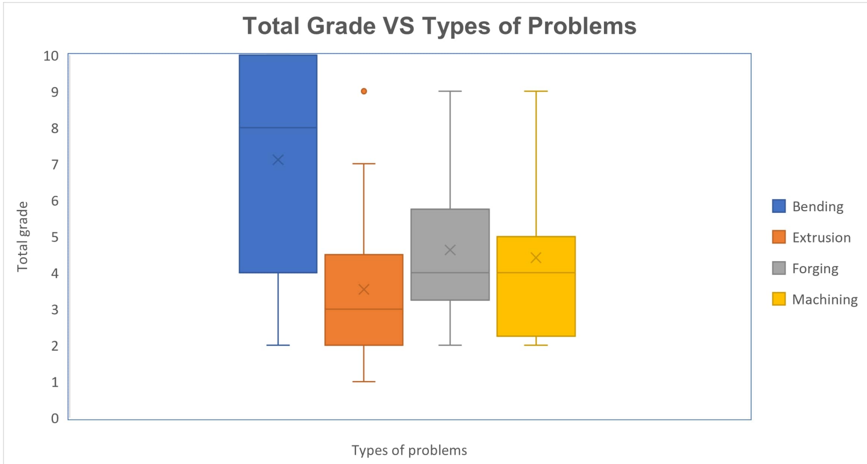


Figure 4.2: Box plot with Types of Problem and Student Performance

Table 4.2: Types of Problem and Student Performance Mean Values

Types of Problems	**Student Performance ($\bar{x}\pm\sigma$)**
Bending	7.11 ± 3.17
Extrusion	3.54 ± 2.25
Forging	4.63 ± 2.13
Machining	4.42 ± 2.43

In Figure- 4.1, the relation between the participants' total grades and generation of problems is observed. Table-4.1 delivers individual mean values of the generation of problems. Figure- 4.1 indicates Textbook+AI is the highest median grade. Textbook+AI and AI problems have higher mean in comparison with Textbook observed in Figure- 4.1 and Table- 4.1. This finding indicates students perform well with the integration of AI into the problem. However, the Textbook problems were helpful for participants to be consistent with their scores as the box plot illustrates the lowest variability. Additionally, in Figure- 4.1, and Table- 4.1, an outlier is observed in the Textbook data and the Extrusion data respectively.

In Figure- 4.2, the relation between the participants' total grades and types of problems is observed. Table- 4.2 delivers individual mean values of types of problems. The lowest variability with the highest median is observed in the Bending problem. Additionally, the participants obtained the highest score with the Bending problem which is reflected in Table- 4.2.

Using Grigg and Benson's (2014) coding scheme, this research identifies 16 codes as significant for students' responses analysis. Problem processing codes, Error codes, and Solution accuracy codes are major themes of this research. Problem processing codes include Knowledge Access, Knowledge Generation, and Self-management elements to solve problems. Error codes include conceptual errors, mechanical errors, and management errors. Solution accuracy includes the correct answer, correct but missing or incorrect units, incorrect answer, and incomplete answer. According to the parent paper, this study fails to assign any codes or sub-codes from strategy themes because of the type of responses received by students.

The frequency of the code is listed in Figure- 4.3. Figure- 4.3 represents the parent codes and code categories of the coding scheme. The most significant code category observed in data is Knowledge generation which consists of drawing a picture, plug in variables into equation, documentation of math, solving an immediate part of a problem, and derivation of units. In error code, conceptual error and management error code share a similar frequency in data. In the assessment code, the answer state shares the second most significance in data which consists of Correct answers, Correct but missing/incorrect units, Incorrect answers, and Incomplete. The findings strengthen the reliability of the coding scheme and their alignment with the parent structure. The findings also indicate a robust foundation for further analysis.

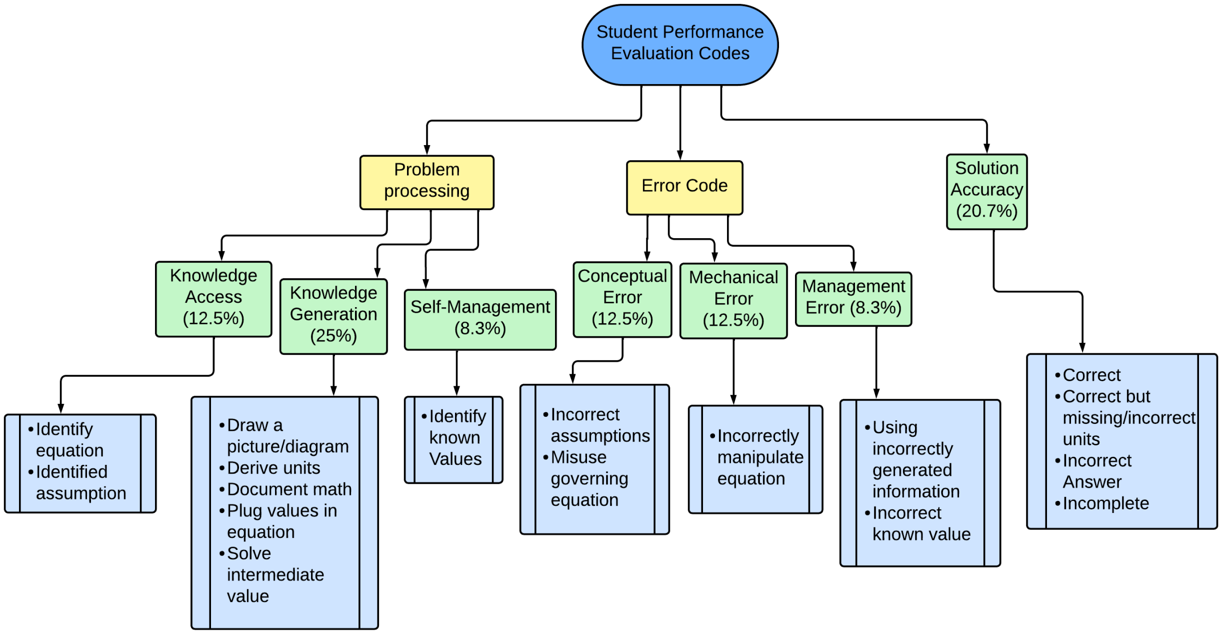


Figure 4.3: Coding Scheme of Student Performance Evaluation

4.2.2 Student Preference

Students' preferences are collected through a rating scale indicating -2 to +2. The lower end is -2 and the upper end is +2 for student preference.

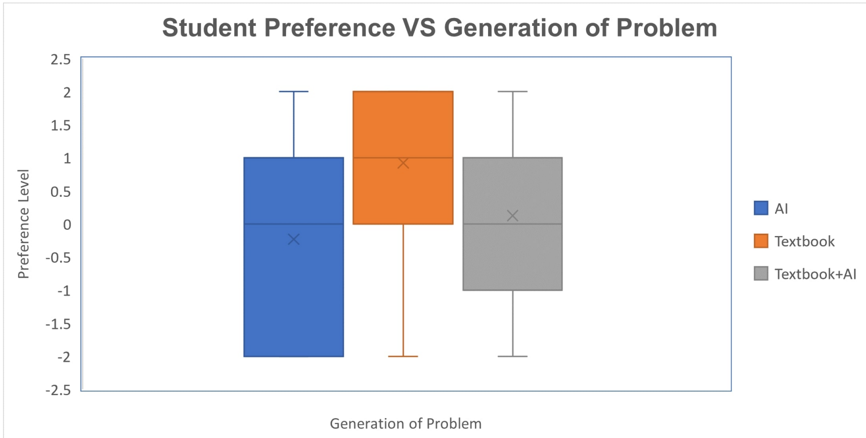


Figure 4.4: Box plot with Generation of Problem and Student Preference

Table 4.3: Generation of Problem and Student Preference Mean Values

Generation of Problems	**Preference ($\bar{x} \pm \sigma$)**
AI	-0.23 ± 1.48
Textbook	0.92 ± 1.20
Textbook+AI	0.13 ± 1.38

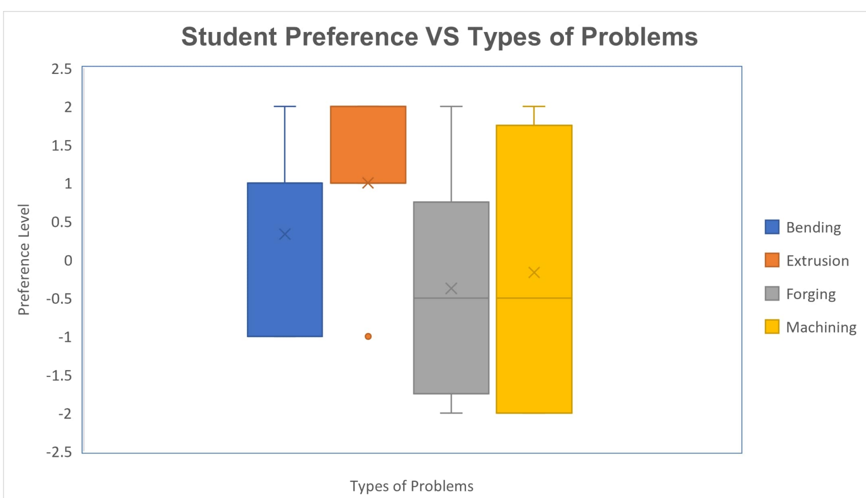


Figure 4.5: Box plot with Types of Problem and Student Preference

Table 4.4: Types of Problem and Student Preference Mean Values

Types of Problems	**Preference ($\bar{x} \pm \sigma$)**
Bending	0.33 ± 1.12
Extrusion	1.0 ± 1.0
Forging	-0.375 ± 1.41
Machining	-0.17 ± 1.78

Figure- 4.4 illustrates the relation between the generation of problems and student preference. Table- 4.3 delivers individual mean values of the generation of problems. The highest preference is observed in Textbook problems indicated in Figure- 4.4 and Table- 4.3. This implies that Textbook problems are more preferred than AI-integrated ones.

Figure- 4.5 illustrates the relationship between the types of problems and the preference level. Table- 4.4 delivers individual mean values of types of problems. The highest preference is observed in Extrusion problems indicated by Figure- 4.5 and Table- 4.4. In Figure- 4.5, an outlier is observed in the Extrusion data.

4.2.3 Students Experienced Mental Workload

Mental workload is collected via the NASA TLX survey. The scale in the survey is 0 (very low) to +20 (very high), and all data are counted from left to right expect performance. To understand the mental workload for individual generations of problems and types of problems, all means of each dimension are considered, and then achieved means are used to observe the overall mental workload. The following table illustrates the overall mental workload:

Table 4.5: Overall Mental Workload with Generation of Problems

AI	**Textbook**	**Textbook+AI**
9.90	8.09	7.66

Table 4.6: Overall Mental Workload with Types of Problems

Bending	**Extrusion**	**Forging**	**Machining**
5.10	7.66	**11.31**	10.16

From Table- 4.5 and Table- 4.6, the generation of problems as AI and types of problems as Forging cause the highest overall mental workload; Some statistical analyses are required for confirmation of statistical significance in the mental workload. To point out the effect of any dimension in mental workload, each dimension is visualized for the generation of problems and the types of problems.

Mental Demand

Mental Demand is the first dimension of mental workload. The data collected with the NASA TLX survey form are provided below:

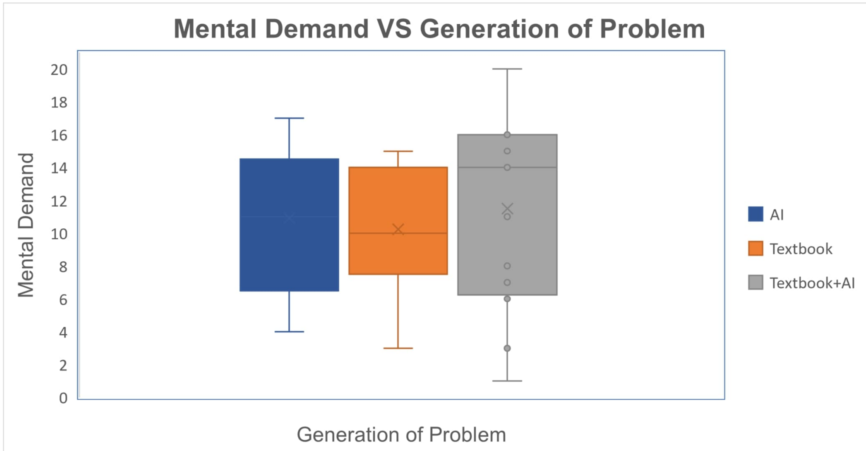


Figure 4.6: Box plot with Generation of Problem and Mental Demand

Table 4.7: Generation of Problem and Mental Demand Mean Values

Generation of Problems	**Mental Demand** ($\bar{x}\pm\sigma$)
AI	10.92 ± 4.48
Textbook	10.23 ± 3.78
Textbook+AI	11.5 ± 5.57

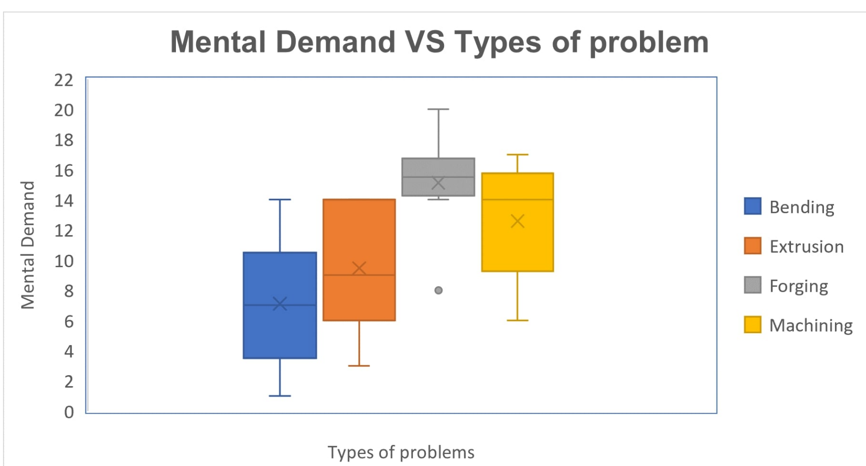


Figure 4.7: Box plot with Types of Problem and Mental Demand

Table 4.8: Types of Problem and Mental Demand Mean Values

Types of Problems	**Mental Demand** ($\bar{x}\pm\sigma$)
Bending	7.11 ± 4.23
Extrusion	9.46 ± 4.01
Forging	15.13 ± 3.39
Machining	12.58 ± 3.57

From Figure- 4.6, the relation between the generation of problems and mental demand is observed. Table- 4.7 delivers individual mean values of the generation of problems. The highest mental demand is observed in Textbook+AI problems in Figure- 4.6 and Table- 4.7. It indicates the integration of AI is structured with cognitive challenges.

On the other, the relation between mental demand and types of problems is observed in Figure- 4.7. Table- 4.8 delivers individual mean values of types of problems. The highest mental demand is observed in Forging problems identified in Figure- 4.7 and Table- 4.8. In Figure 4.6, an outlier is observed in the Forging data.

Physical Demand

Physical Demand is the second dimension of mental workload. The data collected with the NASA TLX survey form are provided below:

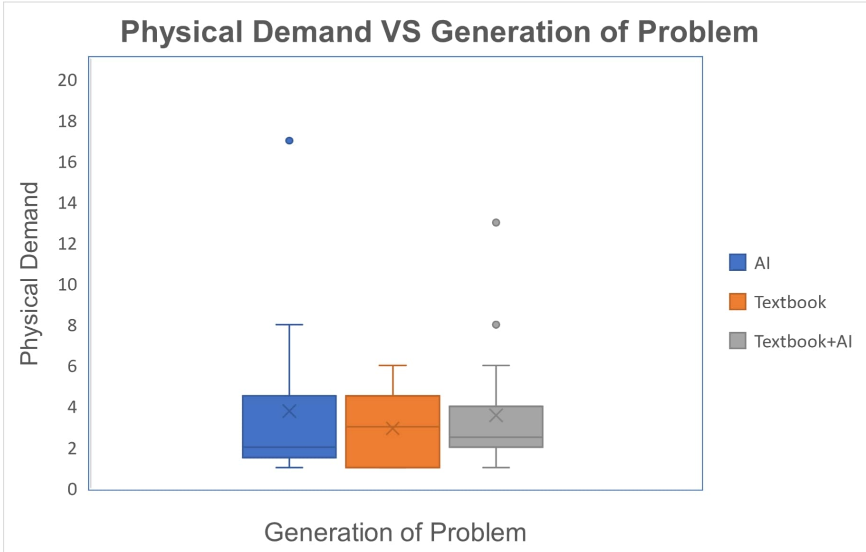


Figure 4.8: Box plot with Generation of Problem and Physical Demand

Table 4.9: Generation of Problem and Physical Demand Mean Values

Generation of Problems	**Physical Demand** ($\bar{x}\pm\sigma$)
AI	3.76 ± 4.48
Textbook	2.92 ± 1.84
Textbook+AI	3.56 ± 3.14

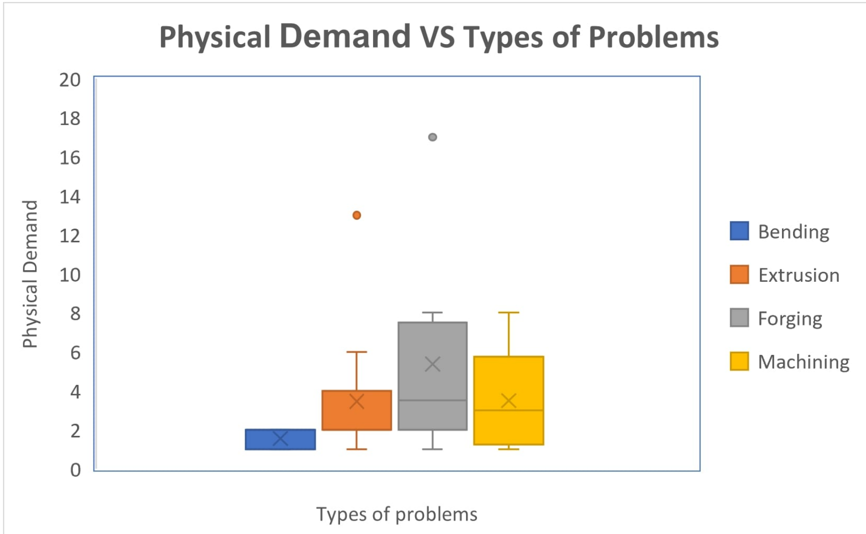


Figure 4.9: Box plot with Types of Problem and Physical Demand

Table 4.10: Types of Problem and Physical Demand Mean Values

Types of Problems	**Physical Demand** ($\bar{x}\pm\sigma$)
Bending	1.55 ± 0.53
Extrusion	3.46 ± 3.17
Forging	5.38 ± 5.24
Machining	3.50 ± 2.27

From Figure- 4.8, the relation between the generation of the problem and physical demand is observed. Table- 4.9 delivers individual mean values of the generation of problems. The highest physical demand is observed in AI problems indicated in Figure- 4.8 and Table- 4.9. Additionally, some outliers are visible in AI and Textbook+AI problems illustrated in Figure- 4.8. From Figure- 4.9, the relation between physical demand and types of problems is observed. Table- 4.10 delivers individual mean values of the types of problems. The highest physical demand is illustrated in Machining problems indicated in Figure- 4.9 and Table- 4.10. Additionally, Extrusion and Forging problems have gained an outlier indicated in Figure- 4.9.

Temporal Demand

Temporal Demand is the third dimension of mental workload. The data collected with the NASA TLX survey form are provided below:

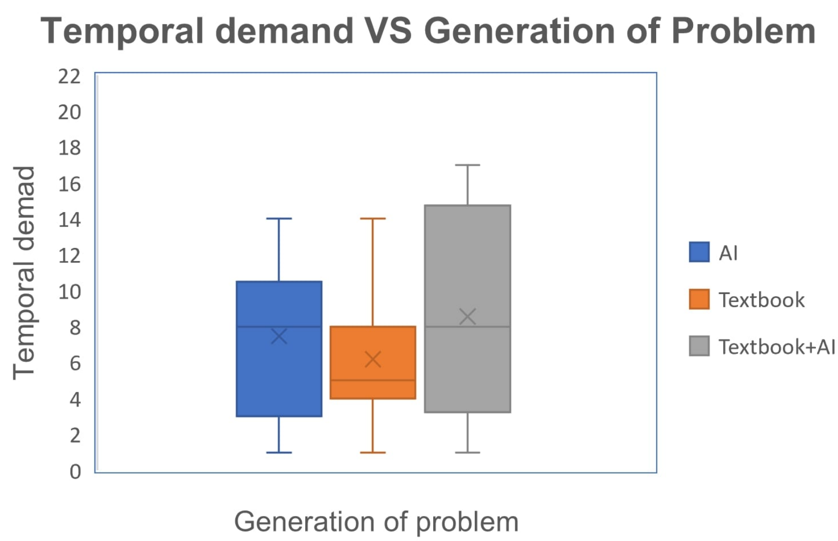


Figure 4.10: Box plot with Generation of Problem and Temporal Demand

Table 4.11: Generation of Problem and Temporal Demand Mean Values

Generation of Problems	**Temporal Demand** ($\bar{x}\pm\sigma$)
AI	7.46 ± 4.29
Textbook	6.18 ± 3.54
Textbook+AI	8.56 ± 5.84

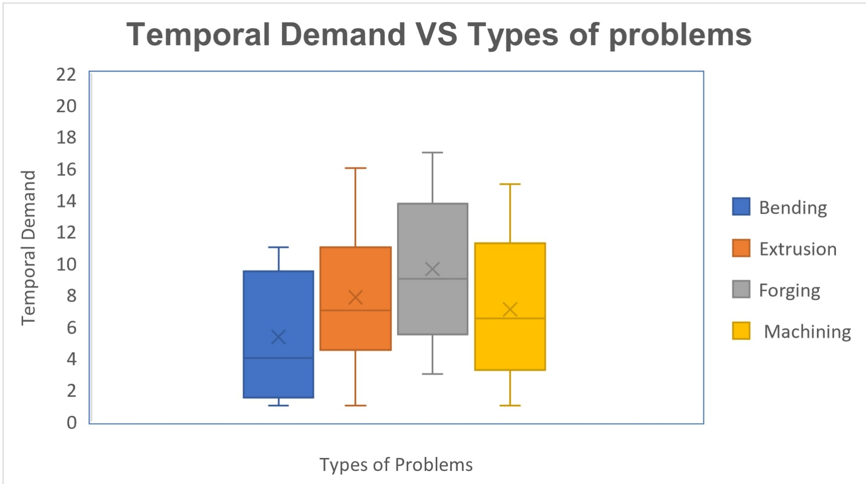


Figure 4.11: Box plot with Types of Problem and Temporal Demand

Table 4.12: Types of Problem and Temporal Demand Mean Values

Types of Problems	**Temporal Demand** ($\bar{x}\pm\sigma$)
Bending	5.33 ± 4.12
Extrusion	7.85 ± 4.77
Forging	9.63 ± 4.77
Machining	7.08 ± 4.83

In Figure- 4.10, the relation between temporal demand and the generation of problems is illustrated. Table- 4.11 delivers individual mean values of the generation of problems. The highest temporal demand is observed in Textbook+AI problems illustrated by Figure- 4.10 and Table- 4.11. It indicates that Textbook+AI problems provide a significant amount of time pressure during the task.

On the other, the relation between the temporal demand and the types of problem is observed in Figure- 4.11. Table- 4.12 delivers individual mean values of the types of problems. The highest observed temporal demand is for Forging problems indicated in Figure- 4.11 and Table- 4.12.

Performance

Performance is the fourth dimension of mental workload. The data collected with the NASA TLX survey form are provided below:

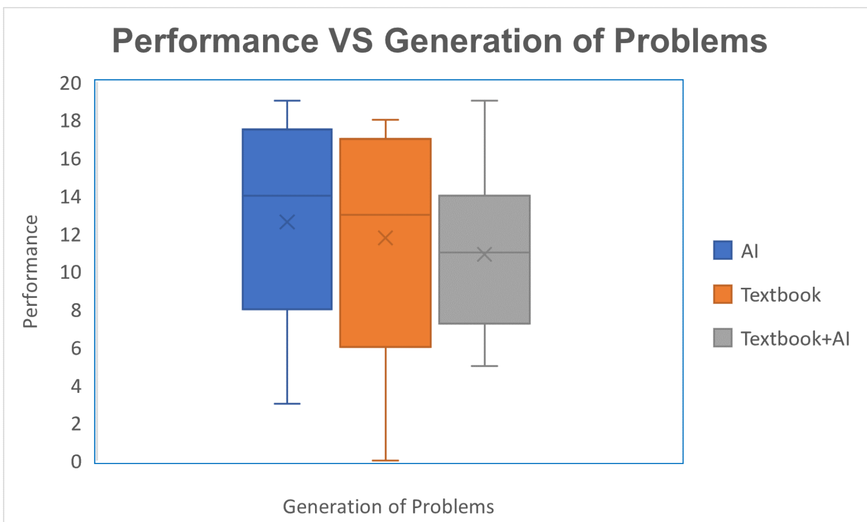


Figure 4.12: Box plot with Generation of Problem and Performance

Table 4.13: Generation of Problem and Performance Mean Values

Generation of Problems	**Performance** ($\bar{x}\pm\sigma$)
AI	12.62 ± 5.41
Textbook	11.77 ± 5.94
Textbook+AI	10.91 ± 4.42

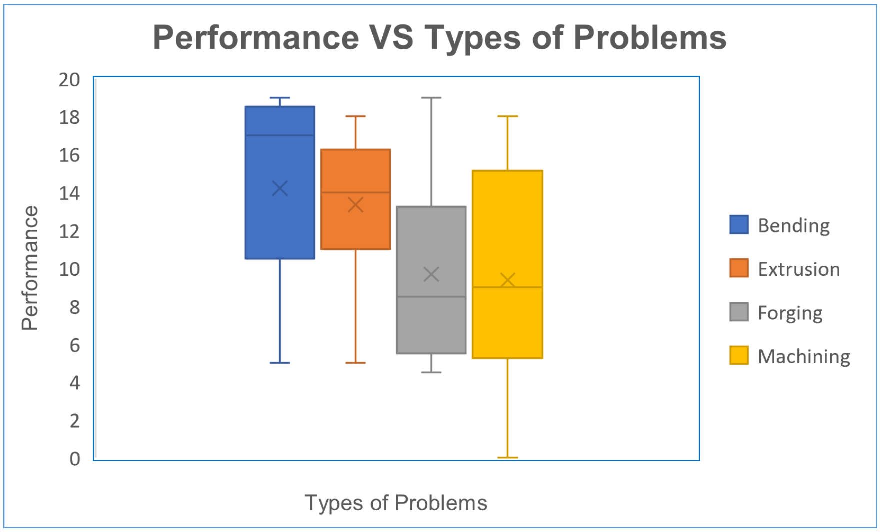


Figure 4.13: Box plot with Types of Problem and Performance

Table 4.14: Types of Problems and Performance Mean Values

Types of Problems	**Performance** ($\bar{x}\pm\sigma$)
Bending	14.22 ± 4.97
Extrusion	13.35 ± 3.76
Forging	9.68 ± 4.93
Machining	9.38 ± 5.73

From Figure- 4.12, the relation between the performance factor and the generation of problems is observed. Table- 4.13 delivers individual mean values of the generation of problems. AI problems deliver the highest performance, as indicated in Figure- 4.12 and Table- 4.13.

From Figure- 4.13, the relation between the performance and types of problems is observed. Table- 4.14 delivers individual mean values of the types of problems. The highest performance observed is in Bending problems illustrated in Figure- 4.13 and Table- 4.14.

Effort

Effort is the fifth dimension of mental workload. The data collected with the NASA TLX survey form are provided below:

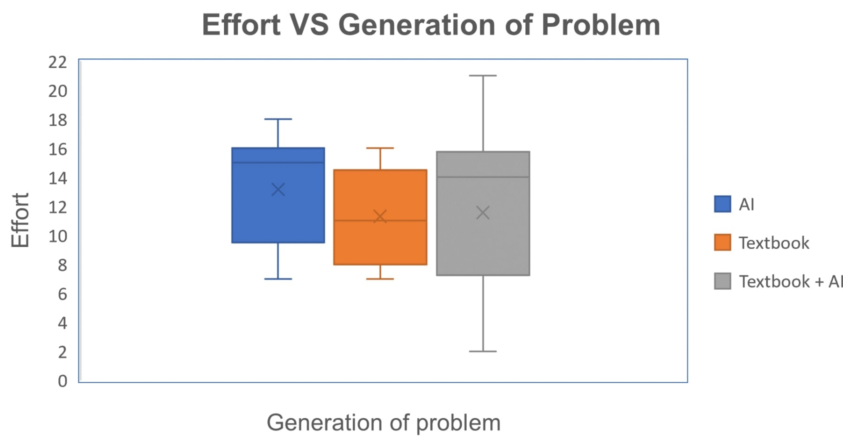


Figure 4.14: Box plot with Generation of Problem and Effort

Table 4.15: Generation of Problem and Effort Mean Values

Generation of Problems	**Effort** ($\bar{x}$±σ)
AI	13.15 ± 3.67
Textbook	11.31 ± 3.17
Textbook+AI	11.56 ± 5.61

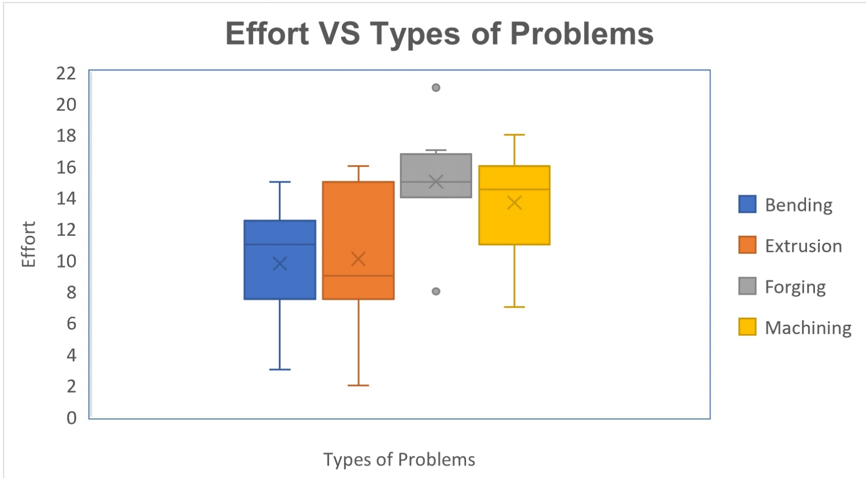


Figure 4.15: Box plot with Types of Problem and Effort

Table 4.16: Types of Problems and Effort Mean Values

Types of Problems	**Effort** ($\bar{x}\pm\sigma$)
Bending	9.77 ± 3.63
Extrusion	10.07 ± 4.46
Forging	15.00 ± 3.63
Machining	13.67 ± 3.65

From Figure- 4.14, the relation between the effort and generation of the problem is observed. Table- 4.15 delivers individual mean values of the generation of problems. The highest effort is observed in AI problems illustrated in Figure- 4.14 and Table- 4.15.

On the other hand, the relation between the effort and the types of problems is observed in Figure- 4.15. Table- 4.16 delivers individual mean values of the types of problems. The highest effort is observed in

Forging problems illustrated in Figure- 4.15 and Table- 4.16. Additionally, two outliers in data are visible in the Forging data.

Frustration

Frustration is the sixth and last dimension of mental workload. The data collected with the NASA TLX survey form are provided below:

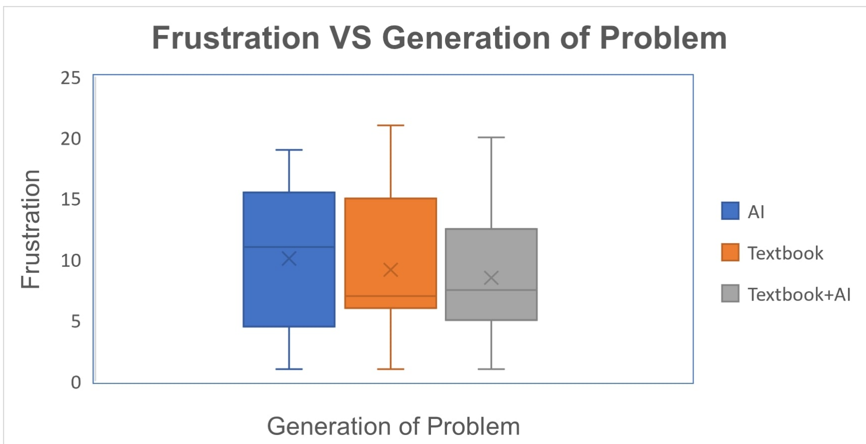


Figure 4.16: Box plot with Generation of Problem and Frustration

Table 4.17: Generation of Problem and Frustration Mean Values

Generation of Problems	**Frustration** ($\bar{x}\pm\sigma$)
AI	10.07 ± 6.24
Textbook	9.15 ± 6.06
Textbook+AI	8.5 ± 5.39

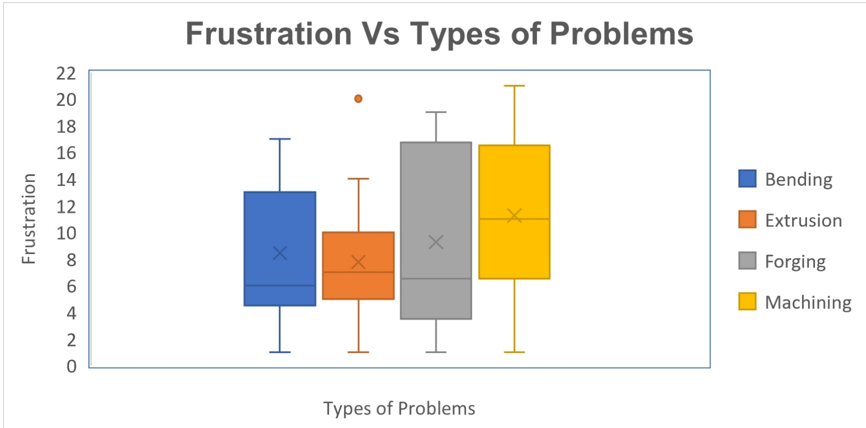


Figure 4.17: Box plot with Types of Problem and Frustration

Table 4.18: Types of Problem and Frustration Mean Values

Types of Problems	**Frustration** ($\bar{x}\pm\sigma$)
Bending	8.44 ± 5.27
Extrusion	7.76 ± 5.02
Forging	9.25 ± 6.98
Machining	11.25 ± 6.14

From Figure- 4.16, the relation between frustration and generation type is observed. Table- 4.17 delivers individual mean values of the generation of problems The highest frustration is observed for AI problems which is indicated by Figure- 4.16and Table- 4.17.

From Figure- 4.17, the relation between frustration and the types of problems is observed. Table- 4.18 delivers individual mean values of the types of problems. The highest frustration is observed in Machining problems indicated in Figure- 4.17 and Table- 4.18. However, there is an outlier with an extrusion problem illustrated in Figure- 4.17.

4.2.4 Emotion Detection

To efficiently represent the emotional response data, the root mean square (RMS) for each problem is considered after processing the data with the help of Py-Feat. The code was set up in a way to achieve the emotion data frame by frame for each problem, the mean of emotion for each problem, the action units of each problem, and their graphical representations. The data does not illustrate any significance in the study. Therefore, no visualization and statistical analysis are included as result.

4.2.5 Interview Data

Using the transcripts, two coders develop a coding scheme. The coders work individually and also as a team during the development period. After developing the scheme, a thematic analysis is employed to extract the promising themes from the data. There are 17 codes in different categories with highlighting themes in Figure- 4.18. The research has produced three themes by conducting the thematic analysis.

Problem Attributes: Codes under this theme explain the nature and structure of the problem. This theme is divided into two categories of codes such as characteristics of problems and challenges faced by students while reading the problems. The characteristics of the problem code describe the potential with three sub-codes: contextualized, concise, and embellished. The challenge code consists of distractions and confounding that are encountered by the students. The significance of this theme is to address the problem perception of the participants which is necessary to address while redefining engineering problems for the future. A note should be mentioned that this theme is generated with the transcripts when the participants are not aware of the generation source of problems.

Impact of generation sources of Problem: This theme delivers a vital role in explaining the influence of problem-generation sources (e.g., conventional or AI-generated) on participants. The insight from this code is valuable in a way that the researchers need to consider using different sources of generation problems. The perspective of the theme focuses on two codes: student performance and learning potential. Performance entitles motivation, time management, and working environment which are the core responsive factors impacting students. Learning potential provides the exposure of different problems and learning enhancement summarizing participants' experiences with the study.

Impressions of AI in Education: This theme summarizes participants' perceptions of AI in education. There are positive and negative sides of AI that are addressed in this theme. On the positive side, the code is entitled as a Guiding tool which reflects three subcategories such as peer, organizer, and facilitator. On the negative side, the code is entitled Threads which inquires with various subcategories namely trust, ethical code, and verification. This theme intends to uncover the overall participants' understanding of AI in education.

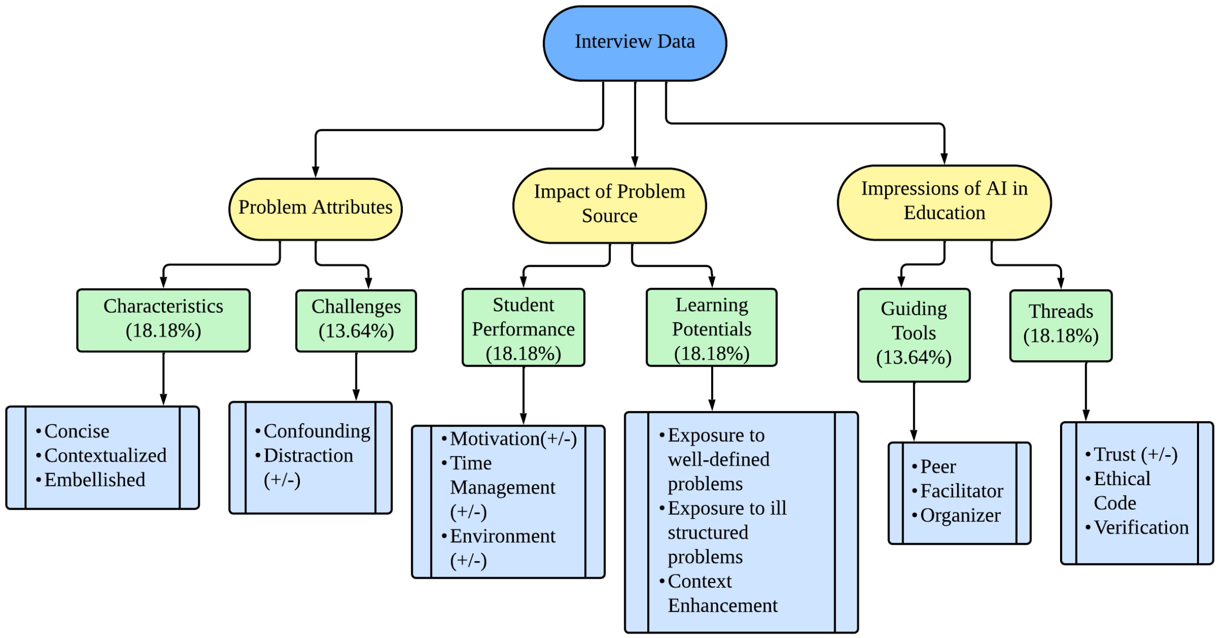


Figure 4.18: Coding Scheme of Interview Data

4.3 Statistical Analysis with Variables

After visualizing the data, statistical analyses are performed to point out the significance. The two-way ANOVA and Duncan's Multiple Range Tests (DMRT) are considered. Only statistically significant variables are mentioned in this section. Section 4.3.1 outlines the result of the ANOVA test, Section 4.3.2 indicates the result of DMRT, and finally, 4.3.3 delivers the result of the inter-rater reliability test.

4.3.1 Two Way ANOVA Test

Student Performance

For performance measurement, the rubric is used and each parameter in the rubric is individually evaluated to examine statistical significance. Three parameters such as right solution, right answer, and total score have statistical significance as their p-values are less than 0.05. As total grade represents overall performance, the ANOVA test with total grade data is listed below. The generation of problems and types of problems are individually statistically significant for performance which indicates they could affect students' performance. Table- 4.19 illustrate the achieved result below:

Table 4.19: ANOVA for Total Grade

Source	**DF**	**Sum of Square**	**Mean Square**	**F Statistic**	**p-value**
Generation of Problem	2	73.84	36.92	6.73	**0.003**
Types of Problem	3	70.71	23.57	4.29	**0.01**
Generation of Problem:Types of Problem	6	0.49	0.08	0.01	1

Student Preference

Considering the importance of student preference based on the literature, this parameter is added to the experiment. After performing the ANOVA analysis, the generation of problems, and the types of problems reveal statistical significance. A point to note is that there is a strong significant interaction of these variables which indicates that both of the independent variables could influence students' preferences. The detailed result is listed in Table- 4.20:

Table 4.20: ANOVA for Student Preference

Source	DF	Sum of Square	Mean Square	F Statistic	p-value
Generation of Problem	2	9.14	4.57	3.94	**0.03**
Types of Problem	3	12.58	4.19	3.62	**0.02**
Generation of Problem:Types of Problem	6	23.62	3.94	3.39	**0.01**

Mental Workload

Similar to student performance, mental workload is treated with individual dimensions to address one or multiple dimensions of significance in the study. The only statistically significant dimensions are mentioned in the following sections:

Mental Demand: To address the significance of mentally demanding work, the ANOVA test is performed. The types of problems illustrate statistical significance. Table- 4.21 provides the details of the test:

Table 4.21: ANOVA for Mental Demand

Source	DF	Sum of Square	Mean Square	F Statistic	p-value
Generation of Problem	2	11.55	5.78	0.41	0.66
Types of Problem	3	332.87	110.96	7.87	**0.0005**
Generation of Problem: Types of Problem	6	123.19	20.53	1.46	0.23

Physical Demand: To address the significance of the physical demands of the study, the ANOVA test is performed. The interaction between the generation of problems and types of problems illustrates statistical significance in data. The detailed result is listed below in Table- 4.22:

Table 4.22: ANOVA for Physical Demand

Source	**DF**	**Sum of Square**	**Mean Square**	**F Statistic**	**p-value**
Generation of Problem	2	5.12	2.56	0.35	0.71
Types of Problem	3	61.96	20.65	2.82	0.06
Generation of Problem: Types of Problem	6	148.18	24.69	3.38	**0.01**

Effort: To address the significance of the Effort, the ANOVA test is performed. The types of problems and interactions between the two independent variables are statistically significant in this study. Table-4.23 illustrates in detail:

Table 4.23: ANOVA for Effort

Source	**DF**	**Sum of Square**	**Mean Square**	**F Statistic**	**p-value**
Generation of Problem	2	26.57	13.28	1.07	0.35
Types of Problem	3	197.83	65.94	5.33	**0.005**
Generation of Problem: Types of Problem	6	185.44	30.91	2.49	**0.04**

4.3.2 Duncan Multiple Range Test

In the ANOVA table if a F-value is significant, we only say that there is significant difference among the means of the variable (e.g., generation of problem, i.e., Textbook, AI, and Textbook + AI) being tested. However, there is a need to make a comparison between specific pairs of means which is not possible from the F-test in the ANOVA but is done by a post-hoc analysis such as Duncan Multiple Range Test (DMRT). It permits to decide which observed differences in various pairs of means are significant and which are not. The DMRT procedure uses significant ranges, each range depends upon several means under comparison. The procedure first calculates the standard error of means as follows:

$$SE = \sqrt{\frac{\text{Mean Square Error (MSE)}}{n}} \tag{4.1}$$

Least Significant Ranges (LSR) are then calculated by multiplying the calculated standard error of means with the significant studentized ranges (SSR) at p = 0.05 and the degree of freedom for the error in the ANOVA table. To determine whether a given pair of means is significantly different, the observed difference between the means is compared with the appropriate LSR. If the observed difference between the means is equal to or lower than the applicable LSR, there is no statistically significant difference between the pair means under comparison, and they are grouped and assigned the same letter ('a'). In contrast, if the observed difference between the pair of means is greater than the applicable LSR, they are significantly differing means from a statistical standpoint and are allocated separate letters in alphabetical order. The lettering system provides a visual representation of the relationships among group means.

Student Performance

Referring to the previous section 4.3.1, after getting the significant result for the total score in student performance, the Duncan Multiple Range Test (DMRT) is calculated. Considering the statistical significance, the research directs to specify the different groups of significance. The mean from each group (e.g., AI, Textbook, Textbook+AI, Bending, Extrusion, Forging, Machining) is compared with the specific LSR and illustrated the most significant one in Table 4.24 and after assigning the letters, AI as generation of problems and Bending as types of problem reflected as the most significant among student performance.

Table 4.24: DMRT for Student Performance

Source	Student Performance ($\bar{x} \pm \sigma$)
Generation of Problem	AI (6.62 ± 2.68)
Types of Problem	Bending (7.11 ± 3.17)

Student Preference

Referring to the previous section 4.3.1, after getting the significant result for student preference, the Duncan Multiple Range Test (DMRT) is evaluated. Considering the statistical significance, the research directs to specify the group of significance. The mean from each group (e.g., AI, Textbook, Textbook+AI, Bending, Extrusion, Forging, Machining) is compared with the specific LSR and illustrated the most significant one in Table 4.25 and after assigning the letters, Textbook with Extrusion reflected as the highest as a preference among student.

Table 4.25: DMRT for Student Preference

Source	Student Preference ($\bar{x}\pm\sigma$)
Generation of Problem	Textbook (0.92 ± 1.2)
Types of Problem	Extrusion (1.0 ± 1.0)

Mental Workload

To get a more definite evaluation, the DMRT analysis is done with the significant mental workload dimension. Based on the p-value gained in section 4.3.2, there is no significance related to the generation of problems in mental workload. In types of problems and the interaction of both independent variables (generation of problems and types of problems), mental demand, physical demand, and effort have statistical significance. Therefore, DMRT analysis is done with them.

Mental Demand

According to section 4.3.1, the ANOVA test demonstrates a significant result with mental demand and types of problems, the DMRT test is performed with this variable. Considering the statistical significance, the mean from each group (e.g., AI, Textbook, Textbook+AI, Bending, Extrusion, Forging, Machining) is compared with the specific LSR. After the comparison between the mean difference and LSR, the Forging problem illustrates higher mental demand among students as Table **??**.

Table 4.26: DMRT for Student Mental Demand

Source	Student Mental Demand ($\bar{x}\pm\sigma$)
Types of Problems	Forging (15.13 ± 3.39)

Physical Demand

According to section 4.3.1, as the ANOVA test demonstrates a significant result with physical demand in terms of the interaction of generation of problems and types of problems, the DMRT test is performed with the dimension of mental workload. Considering the statistical significance, the mean from each group (e.g., AI, Textbook, Textbook+AI, Bending, Extrusion, Forging, Machining) is compared with the specific LSR and illustrates the most significant one in Table 4.27 and after assigning the letters, AI with Forging reflected as the highest as a preference among student.

Table 4.27: DMRT for Student Physical Demand

Source	Student Physical Demand ($\bar{x}\pm\sigma$)
Generation of Problem	AI (3.76 ± 4.48)
Types of Problem	Forging (5.38 ± 5.24)

Effort

According to section 4.3.1, as the ANOVA test demonstrates a significant result with Effort in terms of the types of problems and the interaction of generation of problems and types of problems, therefore, the DMRT test is performed with this dimension of mental workload. Considering the statistical significance, the mean from each group (e.g., AI, Textbook, Textbook+AI, Bending, Extrusion, Forging, Machining) is compared with the specific LSR and illustrates the most significant one in Table 4.28 and after assigning the letters, AI with Machining reflected as the highest as a preference among student.

Table 4.28: DMRT for Student Effort

Source	Student Effort ($\bar{x}\pm\sigma$)
Generation of Problem	AI (13.15 ± 3.67)
Types of Problem	Machining (13.67 ± 3.65)

4.3.3 Inter-rater Reliability Test

Student Performance

To avoid unexpected findings, and promote credibility with the parent coding scheme, the inter-rater reliability test is implemented. The codes from Figure-4.3 are applied. The calculation of the Cohen Kappa coefficient is performed by the equation mentioned in Section 3. The expected agreement (Pe) is 0.95 and the observed agreement (Po) is 0.98. By employing this value in the equation, the Kappa Coefficient is obtained as 0.71 which is a strong agreement between two coders indicated in section 3.4.2 and also resembles the parent coding scheme's kappa coefficient.

Interview Data

After establishing the coding scheme, several revisions between researchers are performed. To define the credibility of the scheme, the same inter-rater reliability test with the Cohen Kappa coefficient is taken into consideration. All transcripts and codes in the scheme are used to calculate the reliability test. The codes from Figure-4.18 are applied. The expected agreement (Pe) is 0.96 and the observed agreement (Po) is 0.97. Using the similar formula mentioned in Section 3.4.2, the Cohen kappa coefficient is achieved at 0.99 reflecting the almost perfect agreement between coders.

Chapter 5

Discussion

5.1 Overview

This section serves as a crucial guide for the discussion chapter, ensuring a comprehensive understanding of the material it contains. Section 5.2 addresses the research questions based on the result of data analysis, 5.3 describes the broader impact of the research and the chapter ends with Section 5.4 pointing to the limitations and recommendations.

5.2 Addressing Research Questions

RQ1: What variations in student performance are observed when students are solved with an engineering problem generated by a large language model (LLM)?

The ANOVA test for performance clearly proves that there are significant differences in student performance based on the generation of the problem. Particularly, significance is observed in the right solution, right answer, and total grade in performance rubrics. Notably, the total score, which combines the right solution and right answer, reveals a statistically significant relationship with both problem generation ($p = 0.003$) and types of problems ($p = 0.012$). In addition to the significance, Duncan's Multiple Range Test (DMRT) reveals that AI constantly exceeds textbook problems, and the combination of Textbook and AI falls in between. Moreover, the type of problem affects considerably, with "Bending" problems producing significant performance. However, as indicated by p-values considerably over 0.05, the interaction between problem generating and problem type was not statistically significant for any of the three factors. This suggests that as each variable has an impact on performance separately, taking the two variables simultaneously has no effect that increases or decreases the performance of students. To observe the data from another angle, students' responses are evaluated with the coding scheme. A strong agreement between coders is verified with the Cohen's Kappa Coefficient of 0.713 which also represents the credibility of the coding scheme. This strong coefficient highlights the consistency of coding between coders and the reliability of the coding scheme for this research. From the analyses, the most beneficial approach for problem-solving is knowledge application strategies. These strategies ignite the understanding of applying earned knowledge to specific problems. The error codes mentioned in this study leverage the aspect of uncovering the identification of particular shortcomings of students during problem-solving. In this study, the conceptual error and mechanical error are the highest observed errors which directs to strengthen the fundamental knowledge for solving problems. The most successful scorers of the study have not encountered these sorts of errors which interprets the necessity of earning fundamental knowledge before starting problem solving. In summary, the analysis justifies the hypothesis, indicating that AI-generated problems significantly improve performance by helping students avoid errors, as proved by the coding scheme.

RQ2: What differences in student preferences emerge when presented with an engineering problem generated by a large language model (LLM)?

The analyses reveal that students' preference has significant differences with generations of problems. The ANOVA test uncovers that the generation of problem variables is statistically significant, which identifies that the generation of problems independently is an influencing factor in forming student preference. Additionally, the types of problems factor is also statistically significant, notably, the interaction between these variables has achieved the most significance, which reveals that these variables are equally influential for student preference. The DMRT analysis suggests that students prefer the textbook-based problem at most. The variation in preference implies that LLM-generated problems are welcomed but less preferable in comparison to textbook problems. The reason behind the student's preference for the problem is to have the structure of a conventional setting in terms of familiarity, clarity, and conciseness. This observation points out that students prefer less complexity in the problem which does not reflect any characteristics of practical engineering problems. In other words, AI-generated problems have the potential to reflect practical problems where many challenges are introduced before starting any task.

RQ3: What differences in cognitive workload emerge when students engage with an engineering problem generated by a large language model (LLM) compared to traditional engineering problems?

The most significant trend emerges in mental demand, physical demand, and effort dimension from the ANOVA test. The generation of problems variable does not demonstrate any significance throughout the dimensions, however, the types of problems and the interaction between the generation of problems and types of problems illustrate significance. For instance, from the ANOVA test, the types of problems are significantly responsible for mental demand rather than the generation of problems. From the DMRT test, the forging problem illustrates different mental demands than other problems with all generation types. The intricacy of the forging problem ignites the student's required high mental demand. For physical demand, the ANOVA test reveals that the interaction effect of these variables is promising in terms of significance. According to the DMRT test, forging with AI problems requires more physical demand on students compared to other problem types and generation methods. This is because of the complexity and the visualization the students need to do with the AI-generated problems. For the effort dimension, the ANOVA test and DMRT test reveal that types of problems are significantly responsible for igniting this dimension, and consequently, forging and machining problems demand higher effort than any other problems. This trend continues across all generations of problems (AI-generated, textbook problems, Textbook+AI problems) and with the interaction between problem generation and problem type producing significant findings. In summary, the mental workload of students reading and solving AI-generated problems is not significantly different from traditional textbook problems, however, the complexity of the types of problems signifies the difference. The generation of problems may influence interactions, and the confounding element of AI-generated problems causes a higher (mean 9.909) overall mental workload. However, further analysis is required to prove that the generation of the problem could be a key component of mental workload differences.

RQ4: Are students' emotional responses influenced by different generations of problems?

The research question investigates whether students' emotional responses to problem-reading and solving activities are influenced by diverse problem-generation sources with a variety of topics. During the ANOVA and regression analyses with facial expression data, each emotion is treated as a dependent variable whereas the generation of problems and types of problems are independent variables. From the analyses, no emotion plays a significant role in students' responses about either the generation of problems or the types of problems. This dearth of significance could result from the sensitivity of emotion detection methods, which may overlook minor facial changes with emotion that occur during the activities. This study also indicates that there may be other factors such as time, students' culture, or influence the emotional responses rather than the generation of problems and types of problems.

5.3 Broader Impact of Research

The purpose of the study is to redefine the engineering problem with the Generative AI tool to promote the students' experience. After analysis, the study is not limited to this confined box. This study offers instructors and curricula designers vital insights by thoroughly examining differences in student performance, preferences, cognitive workload, and emotional responses when faced with generative AI problems.

This study demonstrates that incorporating AI-generated problems could increase student performance and prove the usefulness of Gen AI for educators to create engineering problems. This tool can make students quick problem solvers in the engineering profession as they will be familiar with the complexity ahead of time. Acknowledging the student preference and performance engineering curricula may be modified to bridge the gap between theory and practice. The emotion analysis did not produce statistically significant effects, however, the emotion analysis may provide insights for creating problems that reduce stress while inhibiting students from becoming overwhelmed. The study reveals that the level of difficulty significantly affects mental workload, with complex problems such as forging requiring higher mental, and physical demand; and effort. These findings should be considered when creating any classroom tasks or any tests that successfully challenge students and nurture cognitive growth.

By integrating Generative AI tools into problem generation, students across the field of engineering can be benefited. There are engineering students and educators who are deprived of access to resources from prestigious books. With the integration of Generative AI, all engineers can have access to the same resources without discrimination. The Generative AI tool can provide customized problems to strengthen students' fundamental knowledge and prepare students for practical experience. This tool can also assist educators in promoting experiences such as enhancing context with critical thinking elements, planning engaging class activities, and developing curricula beyond traditional textbooks.

5.4 Limitations and Recommendations

The research presents promising results by identifying the significance of all dependent variables. Some limitations can not be ignored. The findings suggest that though AI-generated problems are acceptable but less preferable, the performance is improved when they solve AI-integrated problems which are contradictory. A point should be noted that the participants were not aware of the generation source of the problem during the problem-solving event. However, the integration of AI in problems provides more context to the problem and helps to visualize the problem which results in less error occurrence in responses. Additionally, the problems solved by the students have only a solution which leads to further exploration of open-ended engineering problems. For mental workload, the interesting finding is observed that the generation of problems did not impact students more than the types of problems did. This finding indicates further exploration of the setting complexity level during problem generation. Based on the literature, the unweighted NASA TLX is implemented, however, there is a fine line of physical demand dimension that may be observed by implementing the weighted NASA TLX. The emotion detection analysis suggests that a particular emotion as a dependent variable is not sufficient for the exploration as the emotion naturally changes over time.

The recording of facial expressions is a difficult job with a camera but the researchers are optimistic that if multiple cameras can be used for capturing a participant that can be beneficial. An increased sample size can influence the outcome of a study significantly which is another observation of this study and recommended for future exploration. In the future, a multi-modal approach (i.e., audio analysis, text analysis, etc.) can be used to enhance the outcome of emotion detection.

Chapter 6

Conclusion

This study advocates the application of Generative AI in engineering problem generation, indicating its potential as an educational support tool. This mixed-method study focuses on observing the impact of different problem generations on students and finds the potential impact of student performance, preference, and mental workload. The findings suggest that relying completely on either traditional textbook problems or Generative AI-generated problems is inadequate. The current method reveals that incorporating the traditional textbook structure and the contextualization of Generative AI can redefine engineering problems to improve engineering problem-solving techniques. Improving the development of AI-generated problems to accurately mimic real-world engineering problems along with multiple solution possibilities is the most promising direction for future research. Additionally, different engineering problem topics with different complexity levels can be another direction for the future.

Appendix A

Problems Solved by Participants

Bending (Text Book)

A 4-inch wide×0.125-inch-thick sheet (UTS=63.8 KSI) is bent at 90 degrees in a V-die (K=1.25) of opening 0.5 inches. Calculate the maximum press force required.

Bending (AI + text book)

Picture a 4-inch wide and 0.125-inch-thick sheet. You aim to bend it at a 90-degree angle using a V-die (K=1.25) with a 0.5-inch opening. The material can handle a maximum force of 63.8 KSI. To figure out the force needed, consider it like the "push" required to make the bend. This force is applied where the sheet meets the V-die, where the magic happens in shaping the material and understanding the strength of the material. Basic geometry helps engineers create all sorts of things efficiently!

Bending (AI)

In the workshop of a metalworker named Sam, a new project awaits. Sam's task is to bend a 4-inch wide and 0.125-inch-thick sheet of metal, with a tensile strength of 63.8 KSI, at a precise 90-degree angle. To achieve this, Sam will utilize a V-die with an opening of 0.5 inches, where the curvature plays a crucial role, characterized by a factor K of 1.25.

Now, imagine yourself as Sam's apprentice, ready to assist in this bending process. Your challenge is to determine the maximum press force required. Dive into the realm of metalworking mechanics, apply your understanding of material properties, and embark on this engaging journey alongside Sam.

Figure A.1: Different generation of Bending problem

Machining (Textbook based)

0.05 inch on radius is removed from a 4-inch diameter bar rotating at 200rpm. The tool is feeding at 10 inches per minute. What is the material removal rate?

Machining (AI+ Textbook)

Imagine a 4-inch diameter bar rotating at 200 revolutions per minute (rpm). A precise tool removes a 0.05-inch radius from the spinning bar, while simultaneously advancing at a rate of 10 inches per minute. Dive into the mechanics and compute the material removal rate, a crucial parameter in manufacturing. The orchestrated interplay between rotational speed and feed rate shapes the evolution of the bar's dimensions in this controlled removal process.

Machining (AI)

In aerospace manufacturing, precision machining plays a critical role in producing components that meet stringent quality standards. Consider a scenario where you're part of a team responsible for manufacturing turbine shafts for jet engines. These shafts undergo machining processes to achieve the necessary dimensional accuracy and surface quality.

You're tasked with determining the material removal rate for a machining operation involving a 4-inch diameter bar, which represents the initial stock material for the turbine shafts. The bar rotates at 200 revolutions per minute (rpm), and a cutting tool removes 0.05 inches on the radius while feeding at a rate of 10 inches per minute.

Figure A.2: Different generation of Machining problem

Extrusion (Textbook)

A round billet made of 70–30 brass is extruded at a temperature of 675°C. The billet diameter is 125 mm, and the diameter of the extrusion is 50 mm. Find: Calculate the extrusion force required.

Extrusion (AI+ textbook)

Consider a round billet crafted from 70–30 brass, warmed up to 675°C. This billet starts with a diameter of 125 mm, and you've got plans to extrude it down to a sleeker 50 mm diameter. Your task: Calculate the extrusion force required. This task will give you an understanding of the mechanical aspects behind reshaping the brass through a controlled process.

Extrusion (AI)

Let's step into the workshop of an adept metalworker named Morgan, who specializes in the art of extrusion. In this scenario, Morgan is working with a round billet made of 70–30 brass. The ambitious project involves extruding the billet at a toasty temperature of 675°C. The initial diameter of the billet measures 125 mm, and the extrusion is set to achieve a diameter of 50 mm.

As a curious engineering apprentice, your task is to calculate the extrusion force required for Morgan's brass masterpiece. Dive into the principles of material science, embrace the challenge, and unveil the force needed to shape Morgan's brass creation into a refined and precisely extruded form.

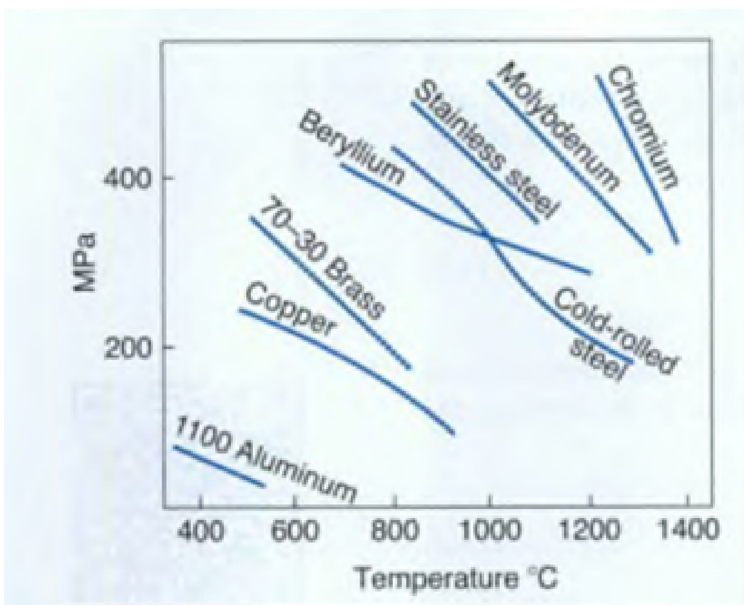


Figure A.3: Different generation of Extrusion problem

Forging (Text Book)

A solid cylindrical workpiece made of 304 stainless steel is 150 mm in diameter and 100 mm high. It is reduced in height by 50%, at room temperature, by open-die forging with flat dies. Assume that the coefficient of friction is 0.2. Find: Calculate the forging force at the end of the stroke.

Forging (AI+ Textbook)

Consider a robust cylindrical workpiece constructed from 304 stainless steel, boasting a diameter of 150 mm and a height of 100 mm. Now, picture this workpiece transforming - a reduction in height by 50%. This operation unfolds at room temperature, facilitated by open-die forging with flat dies. For the sake of analysis, let's assume a coefficient of friction at 0.2. Your task is to investigate the forging force at the end of the stroke in the workpiece's dimensions due to this forging process.

Forging (AI)

Picture a seasoned blacksmith named Riley, working in a traditional forge. Riley's latest project involves a solid cylindrical workpiece made of 304 stainless steels, initially measuring 150 mm in diameter and standing proudly at 100 mm in height. The goal is to reduce its height by 50% through the art of open-die forging using flat dies, all at a comfortable room temperature.

Your role as an apprentice engineer is to step into Riley's workshop and unravel the forces at play. Take into account a friction coefficient of 0.2 as you calculate the forging force necessary for Riley to skillfully reshape the stainless-steel workpiece. Dive into the basics of material science, and join Riley on this hands-on journey in the world of forging.

Typical Values for K and n for Selected Metals

Material	K (MPa)	n
Aluminum		
1100-O	180	0.20
2024-T4	690	0.16
5052-O	202	0.13
6061-O	205	0.20
6061-T6	410	0.05
7075-O	400	0.17
Brass		
70-30, annealed	900	0.49
85-15, cold rolled	580	0.34
Cobalt-based alloy, heat treated	2070	0.50
Copper, annealed	315	0.54
Steel		
Low-C, annealed	530	0.26
1020, annealed	745	0.20
4135, annealed	1015	0.17
4135, cold rolled	1100	0.14
4340, annealed	640	0.15
304 stainless, annealed	1275	0.45
410 stainless, annealed	960	0.10
Titanium		
Ti-6Al-4V, annealed, 20°C	1400	0.015
Ti-6Al-4V, annealed, 200°C	1040	0.026
Ti-6Al-4V, annealed, 600°C	650	0.064
Ti-6Al-4V, annealed, 800°C	350	0.146

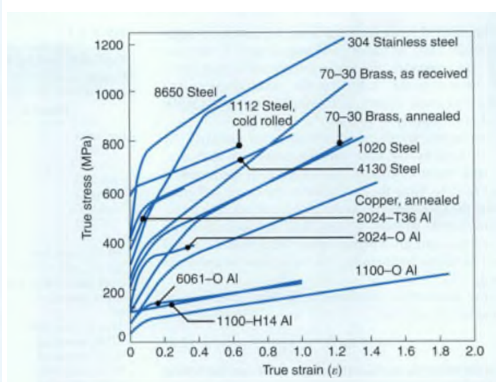


Figure A.4: Different generation of Forging problem

Appendix B

Student Preference Scale

After completion of the above problem, please assess your preference for the problem you addressed in terms of its generation, not the solving preference:

Circle your preference:

-2	-1	0	1	2

Figure B.1: Preference Scale

Appendix C

Codebook Used for Students Performance Analysis

Category	Codes	Definition
Knowledge access	Identify equation	Equation with variables, no values
	Identified assumption	Explicit statement of assumption or self-imposed constraint
Knowledge generation	Draw a picture/diagram	Flow diagram, schematic, sketch, Venn diagram, etc.
	Derive units	Ex: 4 ft12in/1ft = 48 in
	Plug values in equation	Inserting given or derived values
	Document math	Documentation of mathematical calculation
	Solve intermediate value	Getting a sub answer
Self-management	Identify known value	Defining variables by given values from problem statement
Conceptual errors	Misuse governing equation	Error in equation EX: flipped variables or sign
	Incorrect assumptions	Placing or misusing constraints on the system or assumptions not given in problem statement
Mechanical errors	Incorrectly manipulate equation	Algebra problem
Management errors	Incorrect known value	Inserting wrong number for variable
	Using incorrectly generated information	Using incorrect equation or value calculated in previous part of the problem
Solution accuracy	Correct answer	Correctly calculating final answer
	Correct but missing/incorrect units	Correct value with no or incorrect units
	Incorrect answer	Solving for wrong variable, skipped steps
	Incomplete	No final answer produces

Appendix D

Codebook Used for Interview Data Analysis

Category	**Codes**	**Definition**
Problem Characteristics	Concise	The question states as straightforward
	Contextualized	The question adds more than the element requires which helps to understand it well
	Embellished	The question is filled with story-telling elements
Problem Challenges	Distraction	Question is formed to distract such as 'I need to read multiple times to get the point of what I am supposed to do'
	Confounding	The question creates some confusion not telling directly what to do
Student Performance	Motivation	Encourages/ discourages students to be capable of solving the problem
	Time Management	Faces Time-saving or time-consuming problem
	Environment	A relaxed experiment setting, exam settings
Learning Potentials	Exposure to ill-structured problems	Students encounter some practical setting question
	Exposure to well-defined problems	Students encounter some well-structured problem that mirrors a book problem
	Problem Context Enhancement	Adding some contexts to the problem results in learning more knowledge
AI as a Guiding tool	Peer	Works as a fellow with students
	Facilitator	Works in breaking down any complex problem
	Organizer	Plans to complete a task in different small tasks
AI as Threads	Trust	AI lacks some math-solving skills which creates a trust issue with them
	Ethical Code	It interprets the misuse of AI in any settings
	Verification	Having an emphasis on fact-checking before trusting any information from AI

Appendix E

Codes for Facial Expression Extract ion for Emotion Analysis

```python
from feat import Detector
import pandas as pd
import matplotlib.pyplot as plt
import os

# Initialize the Fex detector
detector = Detector()

# Base path for videos
base_path = '/app/videos/'
# Directories to loop through
dir_list = ['AI', 'Textbook', 'Textbook_AI']

# Define the emotion columns
emotion_columns = ['sadness', 'anger', 'happiness', 'disgust', 'fear'
    , 'neutral', 'surprise']
au_columns = ['AU01', 'AU02', 'AU04', 'AU05', 'AU06', 'AU07', 'AU09',
     'AU10',
              'AU12', 'AU14', 'AU15', 'AU17', 'AU20', 'AU23', 'AU25',
                  'AU26','AU28','AU43']

figure_dir = "/app/Figures"
csv_dir = "/app/CSV"
```

```
os.makedirs(figure_dir, exist_ok=True)
os.makedirs(csv_dir, exist_ok=True)

for idx, large_dir in enumerate(dir_list):
    dir_of_interest = os.path.join(base_path, large_dir)
    video_files = [file for file in os.listdir(dir_of_interest) if file.endswith('.mp4')]

    for video_file in video_files:
        video = os.path.join(dir_of_interest, video_file)
        fex = detector.detect_video(video, skip_frames=30)

        au_frame_by_frame_csv = os.path.join(csv_dir, video_file.replace(".mp4", "_au_frame_by_frame.csv"))
        fex[au_columns].to_csv(au_frame_by_frame_csv)

        if set(au_columns).issubset(fex.columns):
            au_means = fex[au_columns].mean()
            au_means_csv = os.path.join(csv_dir, video_file.replace(".mp4", "_au_means.csv"))
            au_means.to_csv(au_means_csv)

            plt.figure(figsize=(10, 6))
```

```
            au_means.plot(kind='bar')
            plt.xlabel('AU')
            plt.ylabel('Mean Intensity')
            plt.title('Mean AU Detections')
            au_means_plot = os.path.join(figure_dir, video_file.
               replace(".mp4", "_au_means.png"))
            plt.savefig(au_means_plot)
            plt.close()

        emotion_frame_by_frame_csv = os.path.join(csv_dir, video_file
           .replace(".mp4", "_emotion_frame_by_frame.csv"))
        fex[emotion_columns].to_csv(emotion_frame_by_frame_csv)

        emotion_means = fex[emotion_columns].mean()
        emotion_means_csv = os.path.join(csv_dir, video_file.replace(
           ".mp4", "_emotion_means.csv"))
        emotion_means.to_csv(emotion_means_csv)

        plt.figure(figsize=(10, 6))
        plt.plot(fex[emotion_columns])
        plt.xlabel('Frame')
        plt.ylabel('Probability')
        plt.title('Emotion Probabilities Over Time')
```

```
        plt.legend(emotion_columns)
        emotion_detection_plot = os.path.join(figure_dir, video_file.
           replace(".mp4", "_emotion_detection.png"))
        plt.savefig(emotion_detection_plot)
        plt.close()

        plt.figure(figsize=(10, 6))
        emotion_means.plot(kind='bar')
        plt.xlabel('Emotion')
        plt.ylabel('Mean Probability')
        plt.title('Mean Emotion Probabilities')
        emotion_means_plot = os.path.join(figure_dir, video_file.
           replace(".mp4", "_emotion_means.png"))
        plt.savefig(emotion_means_plot)
        plt.close()
```

Appendix F

Code for ANOVA test with Emotion data

```python
import pandas as pd
import matplotlib.pyplot as plt

# Load the Excel file
file_path = '/content/Frame by frame emotion categorized (1).xlsx'
xls = pd.ExcelFile(file_path)

# Define parameters
time_frames = 600  # Restrict to 600 frames
generation_column = 'Generation of Problem'

# Iterate through each sheet (emotion) in the Excel file
for sheet_name in xls.sheet_names:
    sheet_df = pd.read_excel(xls, sheet_name=sheet_name)

    # Skip participant name and ID columns
    sheet_df = sheet_df.drop(columns=['Participant Name', 'ID'])

    # Convert all column names to strings (in case some are integers)
    sheet_df.columns = sheet_df.columns.map(str)

    # Identify frame columns by checking if 'Frame' is in the column
       name
```

```python
    frame_columns = [col for col in sheet_df.columns if 'Frame' in
        col]

    # Extract frame numbers as integers (e.g., 0, 30, 60, ...)
    frame_numbers = [int(col.split()[-1]) for col in frame_columns]

    plt.figure(figsize=(10, 6))

    for generation_type in sheet_df[generation_column].unique():
        subset_df = sheet_df[sheet_df[generation_column] ==
            generation_type]
        emotion_intensity = subset_df[frame_columns].mean(axis=0)  #
            Average emotion intensity across participants

        plt.plot(frame_numbers[:time_frames], emotion_intensity.
            values[:time_frames], label=generation_type)

    plt.title(f'Time Series: {sheet_name} vs. Generation of Problem')
    plt.xlabel('Frames')
    plt.ylabel('Emotion Intensity')
    plt.legend(title=generation_column)
    plt.show()
```

Appendix G

Facial Landmark Detection AU and Emotion Intensity Detection Example

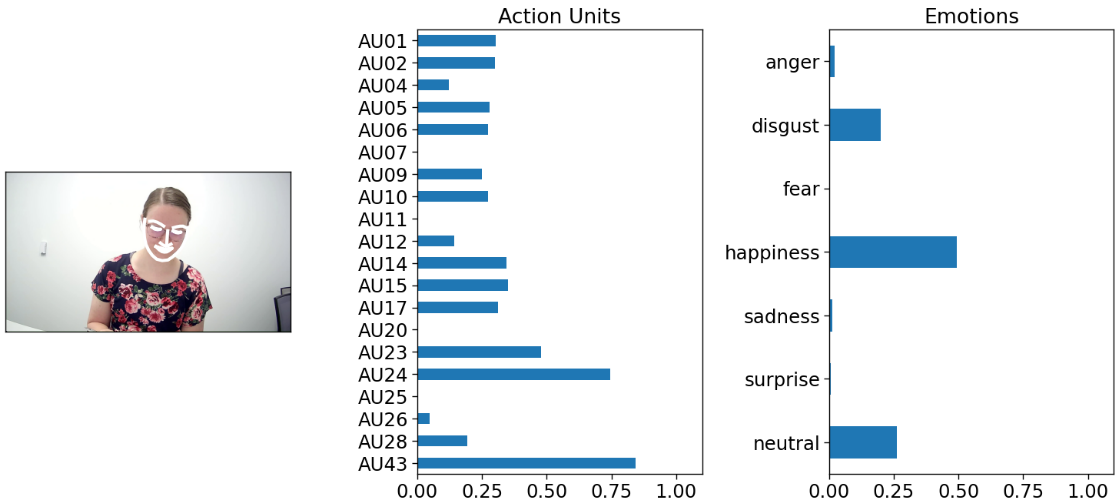


Figure G.1: Landmark detection with AU and Emotion Intensity

Appendix H

ANOVA Result of Emotion Analysis

Emotion	Source	p value
Sadness	Generation of Problem	0.4
Sadness	Type of Problem	0.2
Sadness	Generation of Problem:Type of Problem	0.8
Anger	Generation of Problem	0.5
Anger	Type of Problem	0.6
Anger	Generation of Problem:Type of Problem	0.3
Happiness	Generation of Problem	0.8
Happiness	Type of Problem	0.9
Happiness	Generation of Problem:Type of Problem	0.2
Disgust	Generation of Problem	0.4
Disgust	Type of Problem	0.8
Disgust	Generation of Problem: Type of Problem	0.81
Fear	Generation of Problem	0.48
Fear	Type of Problem	0.61
Fear	Generation of Problem:Type of Problem	0.97
Neutral	Generation of Problem	0.55
Neutral	Type of Problem	0.64
Neutral	Generation of Problem:Type of Problem	0.08
Surprise	Generation of Problem	0.27
Surprise	Type of Problem	0.99
Surprise	Generation of Problem:Type of Problem	0.02

Appendix I

IRB Approval

UNIVERSITY OF GEORGIA

Tucker Hall, Room 212
310 E. Campus Rd.
Athens, Georgia 30602
TEL 706-542-3199 | FAX 706-542-5638
IRB@uga.edu
http://research.uga.edu/hso/irb/

Human Research Protection Program

NOT HUMAN RESEARCH DETERMINATION

January 31, 2024

Dear Beshoy Morkos:

On 1/31/2024, the Human Subjects Office reviewed the following submission:

Title of Study:	Student Evaluation of AI Created Engineering Problems
Investigator:	Beshoy Morkos
Co-Investigator:	Runu Das
IRB ID:	PROJECT00008554
Funding:	None

We have determined that the proposed activity is not designed as research involving human subjects as defined by DHHS and FDA regulations. The activity is designed to determine University of Georgia (UGA) engineering student preferences related to certain design problems and contribute to improvements at UGA. Findings may be shared widely.

UGA IRB review and approval is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these activities are research involving human subjects, please submit a new request to the IRB for a determination.

Sincerely,

Jessica Lasebikan, HRPP Assistant Director
Human Subjects Office, University of Georgia

Appendix J

Approved Consent Form

UNIVERSITY OF GEORGIA
CONSENT FORM
Student Evaluation of textbook and LLM created engineering problems

You are being asked to take part in a research study. The information in this form will help you decide if you want to be in the study. Please ask the researcher(s) below if there is anything that is not clear or if you need more information.

Principal Investigator	**Co-Investigator**
Beshoy Morkos, Ph.D.	Runu Proma Das, MS Student
College of Engineering	College of Engineering
Contact: bmorkos@uga.edu	Contact: RunuProma.Das@uga.edu

The purpose of this study is to determine participants' performance in working on three sources of engineering problems (Textbook adapted, Textbook with LLM, and LLM generated), explore problem preferences, and examine the mental workload and psychological responses related to problem-solving. This will be a one-session study where the researchers will intend to understand the distinguishing factor of Textbook and LLM problems. The scenarios have the goal of mirroring realistic context with engineering problems. You will be asked to solve these scenarios based on your prior engineering knowledge, especially manufacturing. It is necessary for you to offer a solution to each problem; your efforts will be beneficial in the understanding of the significance of LLM in engineering education by researchers. The study will expand scientific knowledge in the field and possibly open up new opportunities for how engineering students are taught to create and solve problems, even though there are no immediate rewards for participants. Your involvement in the study is voluntary, and you may choose not to participate or to stop at any time without penalty or loss of benefits to which you are otherwise entitled. You will be rewarded with a **$50** gift card for participation.

The research team invites you to participate in this study based on the following criteria:

- Currently enrolled at the University of Georgia as an engineering student
- Major in mechanical engineering programs

The study will require a three-step one-time session, including completing an online pre-experiment survey, encountering the engineering problems and solving them with pen and paper, and taking part in a post-experiment interview. The expected duration will vary, but the entire process should not exceed 1 hour. The study will expand scientific knowledge in the field and possibly open up new opportunities for students' engagement, interest, and problem-solving emotions in engineering, even though there are no immediate rewards for participants.
Your involvement in the study is voluntary, and you may choose not to participate or to stop at any time without penalty or loss of benefits to which you are otherwise entitled. If you are interested in participating in the study, please read the additional information on the following, and feel free to ask questions at any point. Your decision to participate will have no impact on your participation in engineering programs.

Study Procedures and Time Commitment

This study will take about 1 hour and consist of three phases: an online pre-experiment demographic survey (7-10 minutes), a problem-solving experiment (30-40 minutes), and a post-experiment interview (15 minutes). The study will be conducted in a lab-based setting at the College of Engineering.

- Pre-experiment survey will ask for your demographic information, engineering challenges, and LLM-using experiences.
- During the experiment, you will be asked to solve the engineering problems and provide your mental workload during the solution process presented on paper and input your solution with pen and paper. You will encounter a camera facing towards you to collect a video of you while solving the problem and it will be used for post-experiment purposes (detecting and analyzing emotion while solving the problem).

1

- After the experiment, you will be asked to participate in a post-experiment interview for 15 minutes. The interview aims to understand your thought process of making a solution and provide your preference for any kind of engineering problems in the future.

Audio/Video Recording/Photographs

Please provide initials below if you agree to have this interview audio-recorded or not. You may still participate in this study even if you are not willing to have the interview recorded.

________I do not want to have this interview recorded.
________I am willing to have this interview recorded.

Voluntary Participation

Participation is voluntary. You can refuse to take part or stop at any time without penalty. Your decision to participate will not impact your grades or class standing.

Risks and discomforts

There are no significant psychological, social, economic, or legal risks involved in this study.
You may skip questions that may make you uncomfortable.

Benefits

There are no probable benefits of participation in the research for the participants. Your responses will be beneficial to us in understanding the significance of LLM in engineering education. The broader impact of the findings concerns the engineering research community by suggesting the potential pedagogical implication to improve course/project offerings for actively student-centered.

Confidentiality of records

We will take steps to protect your privacy, but there is a small risk that your information could be accidentally disclosed to people not connected to the research. To reduce this risk, we will remove all personally identifiable information. The researchers will have access to the identifiable records when recording post-experiment surveys and transcribing them for the research records. Yet, after the transcription, the research team will only use pseudonyms. The research team will only keep information that could identify you long enough to match your responses with your experiment data. The identifiers will be destroyed 5 years after the conclusion of the experiment. The recordings of the interviews will be destroyed 5 years after the study's conclusion. As for the online pre-experiment survey, your confidentiality will be maintained to the degree permitted by the technology used. Specifically, no guarantees can be made regarding the interception of data sent via the Internet by any third parties. We do not plan to share this information with anyone who is not connected to this research study, other than offices identified in this document, without written consent unless required by law. The personally identifiable information attained throughout the research will not be used or distributed for future research.

Participant rights

Please feel free to ask questions about this research at any time. You can contact the Principal Investigator, Dr. Beshoy Morkos at 706-542-2026, bmorkos@uga.edu. If you have any complaints or questions about your rights as a research volunteer, contact the IRB at 706-542-3199 or by email at IRB@uga.edu.

If you agree to participate in this research study, please sign below:

____________________	____________________	________
Name of Researcher	Signature	Date
____________________	____________________	________
Name of Participant	Signature	Date

2

Please keep one copy and return the signed copy to the researcher.

Bibliography

Abedi, M., Alshybani, I., Shahadat, M. R. B., & Murillo, M. (2023). Beyond traditional teaching: The potential of large language models and chatbots in graduate engineering education. *Qeios*.

Adeoye-Olatunde, O. A., & Olenik, N. L. (2021). Research and scholarly methods: Semi-structured interviews. *Journal of the american college of clinical pharmacy*, *4*(10), 1358–1367.

Adeshola, I., & Adepoju, A. P. (2023). The opportunities and challenges of chatgpt in education. *Interactive Learning Environments*, 1–14.

Akili, W. (2015). Perspectives on engineering design learning: Realities, challenges, and recommendations. *2015 IEEE Frontiers in Education Conference (FIE)*, 1–7.

Al Ghatrifi, M. O. M., Al Amairi, J. S. S., & Thottoli, M. M. (2023). Surfing the technology wave: An international perspective on enhancing teaching and learning in accounting. *Computers and Education: Artificial Intelligence*, *4*, 100144.

Almeida, F. (2018). Strategies to perform a mixed methods study. *European Journal of Education Studies*.

Amoozadeh, M., Daniels, D., Nam, D., Kumar, A., Chen, S., Hilton, M., Srinivasa Ragavan, S., & Alipour, M. A. (2024). Trust in generative ai among students: An exploratory study. *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, 67–73.

Armstrong, D., Gosling, A., Weinman, J., & Marteau, T. (1997). The place of inter-rater reliability in qualitative research: An empirical study. *Sociology*, *31*(3), 597–606.

Arnold, L., Waserman, J., & Boulet, J. (2004). The transition from textbook problems to realistic problems. *2004 Annual Conference*, 9–1297.

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Corwin Press.

Azaria, A., Azoulay, R., & Reches, S. (2024). Chatgpt is a remarkable tool—for experts. *Data Intelligence*, *6*(1), 240–296.

Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). Openface: An open source facial behavior analysis toolkit. *2016 IEEE winter conference on applications of computer vision (WACV)*, 1–10.

Benson, L., & Morkos, B. (2011). Student motivation and learning in engineering. *ASEE 120th Annual Conference*, 13.

Bernabei, M., Colabianchi, S., Falegnami, A., & Costantino, F. (2023). Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances. *Computers and Education: Artificial Intelligence*, *5*, 100172.

Bessette, A., Okafor, V., & Morkos, B. (2014, August). *Correlating Student Motivation to Course Performance in Capstone Design* (Vol. Volume 3: 16th International Conference on Advanced Vehicle Technologies; 11th International Conference on Design Education; 7th Frontiers in Biomedical Devices). https://doi.org/10.1115/DETC2014-35506

Chang, C.-C., & Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, *2*(3), 1–27.

Chaudhry, I. S., Sarwary, S. A. M., El Refae, G. A., & Chabchoub, H. (2023). Time to revisit existing student's performance evaluation approach in higher education sector in a new era of chatgpt—a case study. *Cogent Education*, *10*(1), 2210461.

Chen, C., Olajoyegbe, T. O., & Morkos, B. (2020). The imminent educational paradigm shift: How artificial intelligence will reframe how we educate the next generation of engineering designers. *2020 ASEE Virtual Annual Conference Content Access*.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Cheong, J. H., Jolly, E., Xie, T., Byrne, S., Kenney, M., & Chang, L. J. (2023). Py-feat: Python facial expression analysis toolbox. *Affective Science*, *4*(4), 781–796.

Clarke, V., & Braun, V. (2017). Thematic analysis. *The journal of positive psychology*, *12*(3), 297–298.

Cowan, D. A. (1986). Developing a process model of problem recognition. *Academy of management Review*, *11*(4), 763–776.

Coyne, R. (2005). Wicked problems revisited. *Design studies*, *26*(1), 5–17.

Darwin, C. (1872). The expression of the emotions in man and animals (p. ekman, ed.)

Davis, S. M., & Franklin, S. V. (2004). Assessing the impact of student learning style preferences. *AIP conference proceedings*, *720*(1), 53–56.

de Andrade, M. O., Zurita, M., Burova, I., & Nyamapfene, A. (2022). Assessing higher levels of learning through real-life problems in engineering mathematics. *2022 IEEE Global Engineering Education Conference (EDUCON)*, 2002–2006.

Dinar, M., & Shah, J. J. (2014). Enhancing design problem formulation skills for engineering design students. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, *46346*, V003T04A010.

Dym, C. L. (2013). *Engineering design: A project-based introduction*. John Wiley & Sons.

Easton, T. (2021). Online versus in person student learning outcomes. *2021 ASEE Midwest Section Conference*.

Eierman, M. A., & Philip, G. C. (2003). The task of problem formulation. *International Journal of Information Technology & Decision Making*, *2*(03), 353–372.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Ekman, P., & Friesen, W. V. (1986). A new pan-cultural facial expression of emotion. *Motivation and emotion*, *10*, 159–168.

Ekman, P., & Keltner, D. (1970). Universal facial expressions of emotion. *California mental health research digest*, *8*(4), 151–158.

Ekman, P., & Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (facs)*. Oxford University Press, USA.

Eysenck, M. W., Ellis, A. W., Hunt, E. B., & Johnson-Laird, P. N. E. (1994). *The blackwell dictionary of cognitive psychology*. Basil Blackwell.

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A swot analysis of chatgpt: Implications for educational practice and research. *Innovations in Education and Teaching International*, *61*(3), 460–474.

Finnie-Ansley, J., Denny, P., Becker, B. A., Luxton-Reilly, A., & Prather, J. (2022). The robots are coming: Exploring the implications of openai codex on introductory programming. *Proceedings of the 24th Australasian Computing Education Conference*, 10–19.

Gardner, D. H., & Parkinson, T. J. (2011). Optimism, self-esteem, and social support as mediators of the relationships among workload, stress, and well-being in veterinary students. *Journal of Veterinary Medical Education*, *38*(1), 60–66.

Gatla, R. K., Gatla, A., Sridhar, P., Kumar, D. G., & Rao, D. N. M. (2024). Advancements in generative ai: Exploring fundamentals and evolution. *2024 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC)*, 1–5.

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., Chartash, D., et al. (2023). How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR medical education*, *9*(1), e45312.

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330–338.

Grigg, S. J., & Benson, L. C. (2014). A coding scheme for analysing problem-solving processes of first-year engineering students. *European Journal of Engineering Education*, *39*(6), 617–635.

Grigg, S. J., & Benson, L. (2012). Effects of student strategies on successful problem solving. *2012 ASEE Annual Conference & Exposition*, 25–508.

Grigg, S. J., Van Dyken, J., Benson, L., & Morkos, B. (2013). Process analysis as a feedback tool for development of engineering problem solving skills. *2013 ASEE Annual Conference & Exposition*, 23–987.

Hadibarata, T., Hidayat, T., & Kwabena, J. (2023). Problem based learning in engineering course in malaysia. *Acta Pedagogia Asiana*, *2*(2), 95–105.

Hancock, G., Longo, L., Young, M., & Hancock, P. (2021). Handbook of human factors and ergonomics.

Harfield, S. (2007). On design 'problematization': Theorising differences in designed outcomes. *Design Studies*, *28*(2), 159–173.

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the human factors and ergonomics society annual meeting*, *50*(9), 904–908.

Hsu, Y.-C., & Ching, Y.-H. (2023). Generative artificial intelligence in education, part one: The dynamic frontier. *TechTrends*, *67*(4), 603–607.

Hwang, W.-Y., & Utami, I. Q. (2024). Using gpt and authentic contextual recognition to generate math word problems with difficulty levels. *Education and Information Technologies*, 1–29.

Israelsen, A. (2023). How to use chatgpt to write code. *Learning and Development*.

Jain, C., Sawant, K., Rehman, M., & Kumar, R. (2018). Emotion detection and characterization using facial features. *2018 3rd International Conference and Workshops on Recent Advances and innovations in Engineering (ICRAIE)*, 1–6.

Jayaram, S. (2013). Implementation of active cooperative learning and problem-based learning in an undergraduate control systems course. *2013 ASEE Annual Conference & Exposition*, 23–698.

Johri, A. (2020). Artificial intelligence and engineering education.

Jonassen, D. H. (1997). Instructional design models for well-structured and iii-structured problem-solving learning outcomes. *Educational technology research and development*, *45*(1), 65–94.

Kado, C. J., & Kames, E. (2024). Examining the effect of design stimuli on perception of peer contribution in design teams. *2024 ASEE Annual Conference & Exposition*.

Kames, E., Thiess, D., & Morkos, B. (2018, August). *Learning by doing: A simple educational method for modelling vehicle suspensions* (Vol. Volume 3: 20th International Conference on Advanced Vehicle Technologies; 15th International Conference on Design Education). https://doi.org/10.1115/DETC2018-86327

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, *382*(2270), 20230254.

Kaushik, M., Baligar, P., & Joshi, G. (2018). Formulating an engineering design problem: A structured approach. *Journal of Engineering Education Transformations*, *31*(Special Issue).

Khalil, M., & Er, E. (2023). Will chatgpt g et you caught? rethinking of plagiarism detection. *International Conference on Human-Computer Interaction*, 475–487.

Kirn, A., Morkos, B., & Benson, L. (2012). Work in progress: How differences in student motivation characterize differences between engineering disciplines. *2012 Frontiers in Education Conference Proceedings*, 1–2.

Landis, E. A., & Haley, M. L. (2012). Do students achieve higher grades when working together in learning teams than when working through individualized study efforts? *Journal of Higher Education Theory and Practice*, *12*(5), 74–86.

Lange, R. T. (2011). Inter-rater reliability. Springer New York. https://doi.org/10.1007/978-0-387-79948-3_1203

Lecorchick III, D. G. (2017). *Problem formulation within engineering design: A grounded theory*. North Carolina State University.

Lee, C. K., & Sidhu, M. S. (2015). Engineering students learning preferences in uniten: Comparative study and patterns of learning styles. *Journal of Educational Technology & Society*, *18*(3), 266–281.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., et al. (2023). Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, 100017.

Lynch, P., Sangelkar, S., Demeo, G., & Morkos, B. (2017). The i-c-d-m methodology improving undergraduate engineering student motivation, satisfaction, and performance. *2017 IEEE Frontiers in Education Conference (FIE)*, 1–7. https://doi.org/10.1109/FIE.2017.8190693

Ma, J. J., Bosman, L., Mikhail, M., Tantawi, K. H., & Morkos, B. (2023). How to engage engineering students in teaching linear elasticity through entrepreneurially minded bio-inspired projects. *2023 ASEE Annual Conference & Exposition*.

Ma, J., McCrory, B., & Claudio, D. (2021). Comparison of weighted and unweighted task load indices for single-site surgery tasks. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, *10*(1), 142–146.

Madhu, S., & Bhattachryya, D. (n.d.). Learning styles preferences among the students.

Marinescu, A. C., Sharples, S., Ritchie, A. C., Sanchez Lopez, T., McDowell, M., & Morvan, H. P. (2018). Physiological parameter response to variation of mental workload. *Human factors*, *60*(1), 31–56.

Matsumoto, D. (1992). American-japanese cultural differences in the recognition of universal facial expressions. *Journal of cross-cultural psychology*, *23*(1), 72–84.

May, D., Morkos, B., Jackson, A., Beyette, F. R., Hunsu, N., Walther, J., & Ingalls, A. (2021). Switching from hands-on labs to exclusively online experimentation in electrical and computer engineering courses. *2021 ASEE Virtual Annual Conference Content Access*.

May, D., Morkos, B., Jackson, A., Hunsu, N. J., Ingalls, A., & Beyette, F. (2023). Rapid transition of traditionally hands-on labs to online instruction in engineering courses. *European Journal of Engineering Education*, *48*(5), 842–860.

McCrum, D. P. (2017). Evaluation of creative problem-solving abilities in undergraduate structural engineers through interdisciplinary problem-based learning. *European Journal of Engineering Education*, *42*(6), 684–700.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia medica*, *22*(3), 276–282.

Memarian, B., & Doleck, T. (2023). Chatgpt in education: Methods, potentials and limitations. *Computers in Human Behavior: Artificial Humans*, 100022.

Mittal, U., Sai, S., Chamola, V., et al. (2024). A comprehensive review on generative ai for education. *IEEE Access*.

Miyara, F. (1993). Engineering problem formulation for mathematics courses. *Proceedings of IEEE Frontiers in Education Conference-FIE'93*, 553–558.

Ochs, M., & Frasson, C. (2004). Optimal emotional conditions for learning with an intelligent tutoring system. *International Conference on Intelligent Tutoring Systems*, 845–847.

Patrício, M. R., & Gonçalves, B. F. (2024). Chatgpt: Systematic review of potentials and limitations in education. *International Conference on Information Technology & Systems*, 339–348.

Paulus, W. K. (1966). Methodological aspects of problem formulation. *IEEE Transactions on Systems Science and Cybernetics*, *2*(1), 16–22.

Pekrun, R. (1992). The impact of emotions on learning and achievement: Towards a theory of cognitive/-motivational mediators. *Applied psychology*, *41*(4), 359–376.

Pesonen, J. A., & Hannula, M. S. (2014). Analyzing students' emotional states during problem solving using automatic emotion recognition software and screen recordings. *North American Chapter of the International Group for the Psychology of Mathematics Education*.

Pham, T., Nguyen, T. B., Ha, S., & Ngoc, N. T. N. (2023). Digital transformation in engineering education: Exploring the potential of ai-assisted learning. *Australasian Journal of Educational Technology*, *39*(5), 1–19.

Photopoulos, P., Tsonos, C., Stavrakas, I., & Triantis, D. (2021). Preference for multiple choice and constructed response exams for engineering students with and without learning difficulties. *CSEDU (1)*, 220–231.

Pinto, A. S., Abreu, A., Costa, E., & Paiva, J. (2023). How machine learning (ml) is transforming higher education: A systematic literature review. *Journal of Information Systems Engineering and Management*, *8*(2).

Prastawa, H., Susanto, N., & Nofri, T. (2018). Mental workload of undergraduate student (a case study in industrial engineering department of diponegoro university). *SHS Web of Conferences*, *49*, 02017.

Pursnani, V., Sermet, Y., Kurt, M., & Demir, I. (2023). Performance of chatgpt on the us fundamentals of engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. *Computers and Education: Artificial Intelligence*, *5*, 100183.

Qadir, J. (2023). Engineering education in the era of chatgpt: Promise and pitfalls of generative ai for education. *2023 IEEE Global Engineering Education Conference (EDUCON)*, 1–9.

Rezler, A. G., & Rezmovic, V. (1981). The learning preference inventory. *Journal of Allied Health*, *10*(1), 28–34.

Rodríguez-Echeverría, R., Gutiérrez, J. D., Conejero, J. M., Prieto, Á. E., et al. (2024). Análisis del desempeño de chatgpt en exámenes de ingeniería informática.

Rojter, J. (2009). A review of problem based approaches to engineering education. *2009 Annual Conference & Exposition*, 14–100.

Rubio-Valdehita, S., López-Higes, R., & Díaz-Ramiro, E. (2014). Academic context and perceived mental workload of psychology students. *The Spanish journal of psychology*, *17*, E53.

SB, I. A., et al. (2024). Artificial intelligence effects on student learning outcomes in higher education. *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, 1–5.

Schmidhuber, J., Schlögl, S., & Ploder, C. (2021). Cognitive load and productivity implications in human-chatbot interaction. *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, 1–6.

Schoenfeld, A. H. (1980). Teaching problem-solving skills. *The American Mathematical Monthly*, *87*(10), 794–805.

Schuelke-Leech, B.-A. (2021). A problem taxonomy for engineering. *IEEE Transactions on Technology and Society*, *2*(2), 105–105.

Senapati, A. (2023). Psychological emotion recognition of students based chatbot. *International Journal of Applied and Advanced Multidisciplinary Research*, *1*(3), 267–277. https://doi.org/10.59890/ijaamr.v1i3.565

Shah, D., Kames, E., Clark, M., & Morkos, B. (2019, August). *Development of a Coding Scheme for Qualitative Analysis of Student Motivation in Senior Capstone Design* (Vol. Volume 3: 21st International Conference on Advanced Vehicle Technologies; 16th International Conference on Design Education). https://doi.org/10.1115/DETC2019-98423

Shah, D., Kames, E., Pelham, J., & Morkos, B. (2022). Design Evaluations in Educational Settings: A Neuroscientific Study of Incentivized Test/Retest on Student Performance. *Journal of Computing and Information Science in Engineering*, *22*(3), 031013. https://doi.org/10.1115/1.4053801

Shah, D., Morkos, B., Kames, E., & Clark, M. C. (2019). Examining the differences in student motivation for industry projects and non-industry projects in senior capstone design. *2019 ASEE Annual Conference & Exposition*.

Shorten, A., & Smith, J. (2017). Mixed methods research: Expanding the evidence base.

Sloboda, A. R. (2019). The effect of context on student perceptions of homework-style problems in engineering. *2019 ASEE Annual Conference & Exposition*.

Smith, L., Bowen, M. B., Rezaei, N., Mozaffar, F., Rahmanpour, M., Das, R. P., & Morkos, B. (2023). Industry's take on challenges in the agricultural post-harvesting industry. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, *87295*, V002T02A038.

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

Taraban, R., Anderson, E. E., Schumacher, J. R., Dulli, H., & Lamp, D. (2019). Assessing problem-solving strategy use by engineering undergraduates. *2019 ASEE Annual Conference & Exposition*.

Thomas, J. C., & Carroll, J. M. (1979). The psychological study of design. *Design studies*, *1*(1), 5–11.

Toh, C., & Miller, S. R. (2019). Does the preferences for creativity scale predict engineering students' ability to generate and select creative design alternatives? *Journal of Mechanical Design*, *141*(6), 062001.

Toh, C. A., & Miller, S. R. (2016a). Choosing creativity: The role of individual risk and ambiguity aversion on creative concept selection in engineering design. *Research in Engineering Design*, *27*, 195–219.

Toh, C. A., & Miller, S. R. (2016b). Creativity in design teams: The influence of personality traits and risk attitudes on creative concept selection. *Research in Engineering Design*, *27*, 73–89.

Tsai, M.-L., Ong, C. W., & Chen, C.-L. (2023). Exploring the use of large language models (llms) in chemical engineering education: Building core course problem models with chat-gpt. *Education for Chemical Engineers*, *44*, 71–95.

Valkenburg, R., & Dorst, K. (1998). The reflective practice of design teams. *Design studies*, *19*(3), 249–271.

Van Den Ham, A.-K., & Heinze, A. (2018). Does the textbook matter? longitudinal effects of textbook choice on primary school students' achievement in mathematics. *Studies in Educational Evaluation*, *59*, 133–140.

Wang, L., Song, R., Guo, W., & Yang, H. (2024). Exploring prompt pattern for generative artificial intelligence in automatic question generation. *Interactive Learning Environments*, 1–26.

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet?-a systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, *4*, 654924.

Yang, X., Floyd, A. C., Smith, L. A., & Morkos, B. (2023). Analysis of engineering undergraduates' confidence with hands-on tasks–preparation for collaborative manufacturing environments in the era of industry 4.0. *2023 ASEE Annual Conference & Exposition*.

Yulia, M. F. (n.d.). Assessing performance using scoring rubrics: An overview. *Kajian Linguistik dan Sastra*, *17*(2), 174–182.