



THE TASK DEVELOPMENT OF PHYSICS SUBJECT TEACHERS USING CHATGPT.

Dinara Amirkhan

Kazakhstan, 2025

A thesis submitted to the Faculty of Education and Humanities

in partial fulfillment of the requirements for the degree of

Approved by:

MASTER OF PEDAGOGICAL SCIENCES

in Physics

«SDU University»

Department of Pedagogy of Natural Sciences

Thesis Advisor:

PhD., Associate Professor Maxutov Samat

©Copyright by Dinara Amirkhan, 2025

«SDU University»

Faculty of Education and Humanities
Department of Pedagogy of Natural Sciences


This is to certify that the Master's Thesis of

Dinara Amirkhan


has met the thesis requirements of
SDU University

Kaskelen, 2025

Approved by:



PhD., Associate Professor Maxutov Samat
Thesis Supervisor




Master of Science, Lecturer Ayatolla Gabdullin
Program Coordinator




PhD., Associate Professor Zhangyl Abilbek
Head of the Department

Minutes №3, 06.03.2025



PhD., Associate Professor Zhainagu Duisbekova
Dean, Faculty of Education and Humanities



The seal is circular with a blue border. The outer ring contains the text 'Қазақстан Республикасы Алматы облысы Қарасау ауданы Каскелен қаласы' at the top and 'SDU UNIVERSITY' at the bottom. The inner ring contains 'FACULTY OF EDUCATION AND HUMANITIES' and '644 960240000550'. The center features a stylized logo of an open book with wings.



THE TASK DEVELOPMENT OF PHYSICS SUBJECT TEACHERS USING CHATGPT.

Dinara Amirkhan

A thesis submitted to the Faculty of Education and Humanities

in partial fulfillment of the requirements for the degree of

MASTER OF PEDAGOGICAL SCIENCES

in Physics

«SDU University»

Department of Pedagogy of Natural Sciences

Thesis Advisor:

PhD., Associate Professor Maxutov Samat

©Copyright by Dinara Amirkhan, 2025

«SDU University»

Faculty of Education and Humanities
Department of Pedagogy of Natural Sciences

This is to certify that the Master's Thesis of

Dinara Amirkhan

has met the thesis requirements of
SDU University

Kaskelen, 2025

Approved by:

PhD., Associate Professor Samat Maxutov
Thesis Supervisor

Master of Science, Lecturer Ayatolla Gabdullin
Program Coordinator

PhD., Associate Professor Zhangyl Abilbek
Head of the Department

Minutes №3, 06.03.2025

PhD., Associate Professor Zhainagul Duisebekova
Dean, Faculty of Education and Humanities

Table of Contents

1	Introduction.....	9
1.1	Significance of the study.....	9
1.2	Problem Statement	10
1.3	Research Objectives	12
1.4	Research Questions	13
1.5	Theoretical and Practical Contributions.....	14
2	Literature Review.....	16
2.1	Background: What is ChatGPT?.....	16
2.2	How ChatGPT Works	19
2.3	What is the broader significance of AI in education?	22
2.4	Types of Physics Problems	26
2.5	AI in STEM: overview of tools used in science/math teaching.....	28
2.6	Gaps analysis.....	30
2.7	Common Challenges and Failure Modes of ChatGPT.....	32
2.8	Error Classification in the Context of Physics Problem-Solving.....	34
2.9	Theoretical Framework	37
	2.9.1 AI-supported learning theories.....	38
	2.9.2 Reference frameworks such as Technological Pedagogical Content Knowledge (TPACK) or SAMR.	39
3	Methodology.....	43
3.1	Research Design.....	43
3.2	Sample and Problem Selection.....	46
3.3	Data Collection Procedures.....	48
	3.3.1 The Overview of Randall D. Knight’s Textbook “Physics for Scientists and Engineers with Modern Physics”	50
3.4	Evaluation Criteria	52
3.5	Statistical Analysis	54
3.6	Ethical Considerations	56

3.7	Limitations	58
4	Results.....	60
4.1	Descriptive Statistics by Difficulty Level	60
4.2	Descriptive Statistics by Topic Area.....	61
4.3	Inferential Statistics.....	62
4.4	Error Trends and Observations	64
4.5	Performance by Difficulty.....	65
4.6	Performance by Topic	65
4.7	Error Analysis	68
5	Discussion.....	70
5.1	Interpretation of Key Findings	70
5.2	The Role of ChatGPT Within the TPACK Framework	70
5.3	ChatGPT and the SAMR Model of Technology Integration	71
5.4	Implications for Teaching and Learning	72
5.5	Limitations of ChatGPT Highlighted by This Study	73
5.6	Contributions to the Field	74
5.7	Future Research Directions	74
5.8	Improving ChatGPT.....	75
5.9	Broader Implications.....	76
6	Recommendations.....	77
6.1	Recommendations for Teachers	78
6.2	Recommendations for Students	79
6.3	Recommendations for Curriculum Designers and Educational Technologists... 79	
6.4	Recommendations for Future Research	80
7	Conclusion	81
	List of References	84
	APPENDICES	90

List of Tables

Table 2.1. Core Characteristics of Traditional Human Problem-Solving Approach vs ChatGPT in physics	20
Table 2.2. Organization of physics topics by Chapter	27
Table 2.3. SAMR Model: Stages of Technology-Enhanced Learning	41
Table 3.1. The distribution of the 105 problems across topics and difficulty levels	47
Table 3.2. The ChatGPT responses by 2 criterias and difficulty levels	53
Table 4.1. Mean Score Results by Difficulty Levels	60
Table 4.2. Mean Conceptual and Final Scores by Topic	61
Table 4.3. Summary of Inferential Statistical Tests on ChatGPT Performance	63
Table 4.4. Division of problems by the Topic areas and Difficulty levels	66
Table 4.5. Modeling, Assumption and Calculation Errors across Physics Topics	69

Abstract

This study explores the effectiveness of ChatGPT in solving of physics problems of varying difficulty levels across different topics as a generative AI language model. We mainly do aim to evaluate just how the model performs as well as integrate the model into physics education. The succeeding queries provide guidance to the research: With what accuracy does ChatGPT solve the physics problems at a variety of difficulty levels? Do certain specific physics topics vary in performance regarding amount? ChatGPT performs to a greatly better degree on low-difficulty problems, the study hypothesizes, and shows a stronger conceptual understanding rather than a numerical accuracy.

Theoretical frameworks grounding the research are Technological Pedagogical Content Knowledge (TPACK), as well as the SAMR model (Substitution, Augmentation, Modification, Redefinition) used to assess the pedagogical value of AI integration in teaching. ChatGPT was in fact tested on a total of 105 physics problems, drawn from a total of seven topical areas, and a design which was quantitative and descriptive-comparative was adopted. Each response was scored through use of a structured rubric for conceptual understanding. The final accuracy was also able to be scored. Data were analyzed by descriptive statistics and t-tests.

ChatGPT performs well enough on formula-based, low-difficulty problems, especially inside mechanics, as findings indicate, yet struggles upon optics and electromagnetism, abstract, complex topics. With consistency, conceptual understanding in scores happened to be higher than the accuracy of answers. The study offers several practical recommendations for integrating ChatGPT into STEM instruction. The study does also contribute in a theoretical way by applying both TPACK and also SAMR to AI-assisted learning. It concludes instruction that is teacher-led should be supplemented, and not replaced, while ChatGPT can improve learning through its guided use.

Аңдатпа

Бұл зерттеу ChatGPT генеративті жасанды интеллект моделінің физика пәніндегі әртүрлі тақырыптар мен қиындық деңгейлеріндегі есептерді шешу тиімділігін зерттейді. Зерттеудің басты мақсаты — модельдің нақты есептерді қаншалықты дәл шеше алатынын бағалау және оны физика пәніне тиімді түрде енгізу жолдарын анықтау. Зерттеу келесі сұрақтарға жауап беруге бағытталған: ChatGPT физикалық есептерді әртүрлі күрделілік деңгейінде қаншалықты дәл шеше алады? Белгілі бір физика бөлімдеріндегі нәтижелердің айырмашылығы бар ма? Зерттеу гипотезасына сәйкес, ChatGPT жеңіл есептерде жақсы нәтиже көрсетеді және сандық дәлдіктен гөрі теориялық түсінігі жоғары деңгейде көрінеді.

Зерттеу TRACK (технологиялық, педагогикалық және мазмұндық білім) және SAMR (ауыстыру, күшейту, түрлендіру, қайта анықтау) теориялық негіздеріне сүйенеді. ChatGPT моделі жеті түрлі физика тақырыбынан алынған 105 есеппен тексерілді. Зерттеу әдісі — сандық және сипаттамалы-салыстырмалы талдау. Әрбір жауап теориялық түсінік пен соңғы нәтижелердің дәлдігі бойынша бағалау рубрикасы арқылы сарапталды. Алынған деректер сипаттамалық статистика мен t-тесті арқылы талданды.

Нәтижелер ChatGPT формулаға негізделген, төмен деңгейлі есептерді, әсіресе механика саласында, жақсы шешетінін көрсетеді. Оптика мен электромагнетизм сияқты күрделі әрі абстрактілі тақырыптарда модельдің қателіктері көбірек кездеседі. Жалпы алғанда, теориялық түсінік деңгейі сандық дәлдікке қарағанда жоғарырақ болды. Зерттеу STEM пәндерін оқытуда ChatGPT қолданудың нақты ұсыныстарын ұсынады және TRACK пен SAMR модельдерін жасанды интеллект көмегімен оқытуға қолдану арқылы теориялық үлес қосады. Қорытындылай келе, ChatGPT мұғалімді алмастыру құралы емес, бірақ оны дұрыс қолдану арқылы оқушылардың білімін толықтыра алатын тиімді көмекші бола алады.

Аннотация

В данном исследовании рассматривается эффективность использования ChatGPT — генеративной языковой модели искусственного интеллекта — для решения физических задач различного уровня сложности по различным темам. Основная цель заключается в оценке того, насколько точно модель решает задачи, а также в изучении возможностей ее интеграции в преподавание физики. Исследование направлено на ответы на следующие вопросы: с какой точностью ChatGPT решает задачи по физике различных уровней сложности? Отличается ли эффективность решения в зависимости от темы? Согласно гипотезе, ChatGPT показывает значительно лучшие результаты при решении задач низкой сложности и демонстрирует более глубокое концептуальное понимание, чем точность численного ответа.

Теоретическую основу исследования составляют модели TRASK (технологические, педагогические и содержательные знания) и SAMR (замещение, усиление, модификация, переопределение), применяемые для оценки педагогической ценности интеграции ИИ в обучение. ChatGPT был протестирован на 105 задачах по физике, охватывающих семь тематических областей. В исследовании применялся количественный, описательно-сравнительный подход. Каждое решение оценивалось по структурированной рубрике с точки зрения концептуального понимания и численной точности. Данные были проанализированы с использованием описательной статистики и t-критериев студента.

Результаты показывают, что ChatGPT хорошо справляется с задачами низкой сложности, основанными на формулах, особенно в области механики, однако испытывает трудности при решении задач по оптике и электромагнетизму, которые требуют более абстрактного мышления. Концептуальное понимание стабильно демонстрировало более высокие баллы, чем точность ответов. Исследование предлагает практические рекомендации по интеграции ChatGPT в преподавание предметов STEM-направления и вносит теоретический вклад за счет применения моделей TRASK и SAMR к обучению с поддержкой ИИ. В заключении подчеркивается, что ChatGPT может улучшать процесс обучения при условии его осмысленного использования, однако не должен заменять преподавателя, а лишь дополнять его.

1 Introduction

1.1 Significance of the study

The rapid integration of Artificial Intelligence (AI) into education has transformed knowledge delivery as well as knowledge consumption and knowledge assessment. Generative AI tools are amid the most deeply impactful innovations, notably ChatGPT, a substantial language model that interacts via natural language and was developed by OpenAI. These tools have fundamentally redefined the educational landscape, enabling personalised tutoring and on-demand assistance, as well as information democratisation. Complex problem-solving is quite central in STEM education. In this particular field, the potential of AI to provide support for both learners and instructors is especially compelling.

Advanced language AI systems such as ChatGPT are becoming increasingly present, and in addition, they have shifted models across domains, plus STEM education has felt an impact. AI-driven tools are integrating more and more into pedagogical practices and into research methodologies while students learn, which is characterising this transition. The evolutionary development of these selfsame models has engendered large interest as well as substantive debates, concerning their individual capabilities, alongside limitations, within educational contexts. The purpose of this particular research is to provide something of a thorough analysis of ChatGPT's skills regarding physics problems. This will ultimately advise researchers, as well as developers and educators, on just how to integrate this technology wisely into physics education, in addition to its problem-solving.

Students often battle with the goal to reason in an abstract way, solve problems that have multiple steps, and translate physical principles into steps that are computational. Conventional teaching strategies do often fall somewhat short in the offering of real-time individualized guidance, even though being quite effective. ChatGPT emerges as a promising complement for something. It is capable of providing immediate feedback, in addition to elaborating on complex ideas, as well as easing exploratory learning.

1.2 Problem Statement

Despite the growing interest that there is in AI-driven educational technologies, pedagogical effectiveness research is limited in regard to language AI systems such as ChatGPT in complex disciplines such as physics. In most of the available literature, user perceptions are stressed. Other literature widely explores surface-level interactions. Few studies explore the question of whether ChatGPT is able to reason through multi-step physics problems or analyze just how it performs across varying topics and problem complexities.

For educators, this gap creates ambiguity: Is ChatGPT a trustworthy tool for students in learning physics? Does it simply provide superficial answers or can it assist with the development of conceptual understanding? Is it the case that ChatGPT is more accurate with respect to specific topics in physics? In some other of the areas, does it always fail?

To fill in this gap in the field, the study analyzes ChatGPT's problem-solving accuracy across three difficulty levels and across seven physics topics, which are Newton's Laws, Conservation Laws, Applications of Newtonian Mechanics, Oscillations and Waves, Thermodynamics, Electricity and Magnetism, and Optics. Also explored is the manner in which the tool supports educational goals in instances where established educational frameworks are used for purposes of evaluation.

Physics problems exhibit a wide range of particular difficulty levels, in addition to spanning many topics; therefore, they pose a unique kind of challenge for AI-based problem-solving systems like ChatGPT. Problems of low difficulty are generally when someone directly applies basic fundamental principles and formulas, therefore they require limited conceptual understanding or mathematical manipulation. However, a greater comprehension of concepts is required for medium-difficulty problems in addition to the fact that those problems might involve multiple steps, or quite a few principles must be applied. For high-difficulty problems, complex scenarios are frequently involved, advanced mathematical techniques are absolutely required, and knowledge from multiple areas of physics might be integrated.

Physics problems can be categorized into several remarkably major topics that include mechanics, thermodynamics, electromagnetism, optics, and quantum mechanics. Mechanics deals with the motion of objects as well as the forces that cause such motion. Thermodynamics thoroughly addresses each of the relationships existing between heat, work, as well as energy instead. Optics considers the behavior of light and electromagnetism examines interactions between electric charges and magnetic fields.

Quantum mechanics deals with the behavior of matter at both atomic and subatomic levels. Seeing that each one of these topics does require a different set of concepts, principles, as well as mathematical tools, physics problem-solving is a

challenging task for AI systems. ChatGPT performs with varied expertise since it relies on the data that it trained on, plus the topic of the problem and its difficulty affect it greatly. Previous investigations do reveal that the integration of the Code Interpreter with ChatGPT-4 yields a substantial enhancement in its problem-solving capabilities, thus allowing achievement of satisfactory solutions for a higher proportion of problems when it is compared to the standard ChatGPT-4 configuration. This underscores that we should evaluate the ChatGPT's performance across different problem types and difficulty levels in order to identify what it does well and also where it falls short.

1.3 Research Objectives

The rise of ChatGPT in STEM education means educators and students both face important opportunities and challenges. AI tools, such as ChatGPT, have been integrated within education, and this has garnered substantial attention in recent years. Various studies do explore the potential of these tools, inasmuch as they can improve learning experiences, along with the respective outcomes. Given that ChatGPT's natural language processing capabilities converge, coupled with its capacity that generates human-like text, alongside its ability that engages in interactive dialogues, institutions are therefore increasingly deploying it across educational settings and these self-same applications span from personalised tutoring systems that provide bespoke feedback for students to automated grading tools that streamline the assessment process for educators.

This study has the aim of assessing just how ChatGPT performs when it comes to solving physics problems that do span different difficulty levels and topics, and also to identify the areas in which it excels and the areas where it struggles in physics problem solving. The analysis of ChatGPT's responses uncovered persistent challenges which incorporate difficulties with visual interpretation, providing accurate physics laws or rules, as well as spatial coordination in addition to application of physics representations. The study, in addition, explores the educational implications of ChatGPT within physics education as well as suggests possible enhancements for improving its problem-solving capabilities. It also seeks to inform many educators, various researchers and a range of policymakers on precisely how to utilise these tools in an appropriate manner, and contribute towards discussing how humans and AI can effectively collaborate for solving complex physics problems. Educators earnestly hope to identify shortcomings and the large potential of AI-powered tools to help leverage human-AI collaboration as they solve complex physics problems and improve student learning and problem-solving skills.

The **main objectives of the study** are:

1. To evaluate ChatGPT's performance in solving physics problems across three levels of difficulty (low, medium, high).
2. To assess ChatGPT's accuracy across multiple physics topic areas.
3. To identify and categorize common patterns of error in ChatGPT's responses.
4. To analyze the findings using descriptive and inferential statistical methods (mean, mode, t-tests, standard deviation).

1.4 Research Questions

In order to accomplish these goals, the research provides answers to the following queries:

1. To what extent can ChatGPT accurately solve physics issues with varying degrees of difficulty?
2. Does ChatGPT excel at some physics topics while struggling with others?
3. What kinds of mistakes—such as conceptual or computational—occur most frequently in its answers?
4. What can be learnt about the dependability of ChatGPT's performance from statistical analysis (mean, standard deviation, t-tests)?

Hypothesizes of the study:

H1: ChatGPT performs better on low-difficulty problems.

H2: Performance varies by physics topic.

H3: Conceptual understanding scores are higher than final accuracy.

1.5 Theoretical and Practical Contributions

The findings of this study will mean that incorporating of ChatGPT and similar AI-powered tools effectively within physics education and research will have implications that reach far. The results shall furnish educators, researchers, and policymakers with understandings into the suitable employment of these technologies, whilst offering recommendations apropos of leveraging the singular capabilities natural within both human and AI problem-solving approaches so as to improve student learning in addition to advancing the field of physics; a fact that highlights their natural potential together with limitations. It shall also contribute further to the discourse upon the role of human-AI collaboration when solving complex physics problems. The understandings it furnishes can effectively guide the development in addition to the refinement of AI-driven tools, so they can better support human expertise. The analysis additionally identifies several recurring error patterns and fundamental misconceptions ChatGPT exhibits, offering valuable understandings for improving the model's knowledge base, reasoning abilities and overall accuracy in physics problem-solving. The study ultimately seeks to encourage the ethical and responsible use of AI in physics research and education so as to improve comprehension of how such technologies best serve the wider scientific community, educators and students.

Theoretical Contributions

This research adds to the growing field of AI-improved education by situating ChatGPT within two key educational technology frameworks.

The study explores the ways in which ChatGPT fits into the intersection of technology, pedagogy, as well as content knowledge specifically within physics instruction, a concept that is known as Technological Pedagogical Content Knowledge TPACK.

Especially in complex problem-solving contexts, it precisely evaluates how ChatGPT uniquely modifies or redefines learning tasks that is the SAMR Model.

The study puts forward certain theoretical understandings into the ways that generative AI tools can align with modern pedagogical strategies. These frameworks are applied for the development of 21st-century skills.

Practical Contributions

On a practical level, this study provides to educators, for curriculum designers, and to policymakers:

- Guidelines on evidence for ChatGPT integration into physics classrooms based.
- Training recommendations for both teachers and for students in regard to AI literacy.
 - AI can improve upon human teaching with a data-driven understanding instead of replacing of it.

To ensure thoroughly responsible as well as ethical use of these indispensable tools, these specific contributions are critical for completely building trust in AI-supported learning.

2 Literature Review

2.1 Background: What is ChatGPT?

ChatGPT is a state-of-the-art language model, which OpenAI developed using deep learning techniques, so that human-like responses to a wide range of prompts are generated. ChatGPT is trained on a large corpus of text data as a large language model, thus allowing understanding, coupled with generation, of natural language exhibiting outstanding fluency. A transformer-based architecture is employed which allows long-range dependencies in text to be captured, as well as coherent and contextually relevant responses to be generated. The model is capable of processing as well as expertly generating text, thus it represents a truly promising tool for a variety of applications, inclusive of education and problem-solving (Adıgüzel et al., 2023). ChatGPT's architecture is formed by multiple layers of transformer blocks, with each one featuring self-attention mechanisms and feedforward neural networks. Whenever a response gets generated, the model uses that self-attention mechanism so as to weigh the importance of different ones of the words in that input sequence, and also the feedforward networks perform non-linear transformations upon those attended-to representations. The model is extensively trained, utilising both supervised and unsupervised learning techniques, so it can learn from both labelled and unlabelled data.

During the training process, the model's knowledge and reasoning abilities are encoded within the weights of its learned neural networks. The model utilises these learned weights right during inference so as to generate responses up to new prompts therefore it effectively "reasons" about the input subject to its training data (West, 2023). This architecture does allow ChatGPT to generate coherent and contextually relevant responses to quite a wide range of prompts, and so it is indeed a valuable tool for various different applications (Ruiz-Rojas et al., 2024). ChatGPT has been successfully applied within many areas such as chatbots, content generation, language translation, personalized recommendations and even medical diagnosis and treatment (Hariri, 2023). It can also explain complex concepts, plus solve mathematical problems (Ali et al., 2024). Its capability to generate responses similar to those of humans, thereby showing its comprehension of natural language, has further increased its popularity within education. ChatGPT very much stands as an influential tool for natural language processing, and this is because its accuracy is indeed high, and its flexibility allows it to adapt to a variety of different contexts, plus generate outputs that are rather similar to humans. However, note that ChatGPT also remains limited since it might generate information that is naturally biased or inaccurate and it cannot genuinely understand the underlying concepts the text represents. These limitations go to show that we need to carefully evaluate as well as integrate ChatGPT and similar AI tools in specific domains such as physics problem-solving in order to determine their suitability, and to recognise areas for improvement.

ChatGPT does have limitations. These are noted by Hariri in (2023) and by Su & Yang in (2023). A particularly outstanding issue is "AI hallucination," a phenomenon throughout which the model produces responses which are convincing yet entirely fabricated (Athluri et al., 2023). This presents with problems in educational settings, since accuracy and reliability matter most of all. ChatGPT, as with many other AI systems, relies upon its training data, and such data may feature inconsistencies, biases, and inaccuracies (Abbas et al., 2024). Unfair or discriminatory outcomes can potentially arise ("The AI Writing on the Wall," 2023) in view of the fact that these biases might be mirrored within the model's responses.

Difficulties requiring actual real-world knowledge could pose a challenge for ChatGPT, and its common sense reasoning or creative problem-solving skills might even exceed its training data. ChatGPT's responses' general quality relies greatly on the natural quality and diversity of the data it was trained on (Dahlkemper et al., 2023). On account of this, critically evaluating ChatGPT's output is necessary, and one should not rely solely upon its responses without verifying their accuracy and validity. Acknowledging that ChatGPT is simply a tool is also greatly important, and like any other tool it should always be used ethically and responsibly (Miah et al., 2024). ChatGPT systems must be updated and tested on a regular basis for reliability and accuracy considering possible inaccuracies especially if the system is lacking in training on diverse educational materials (Tanjga, 2023). It is of importance to design and to implement teaching and learning strategies that do consciously integrate AI tools. Evaluating their effectiveness for both students and also professors is important too (Huesca et al., 2024).

ChatGPT's capabilities from its training on a massive dataset of text as well as code, so it can understand and generate human-like responses across a wide range of topics (Bahrini et al., 2023). Its architecture is based upon the Transformer model and therefore it allows for the processing and generation of text with outstanding fluency, making it an influential tool for natural language processing tasks (Cotton et al., 2023). Nevertheless, ChatGPT is additionally limited because it might generate somewhat biased or completely inaccurate information, and it cannot genuinely understand the concepts supporting the text it generates (Hariri, 2023). These limitations particularly highlight the reasons as to why we must carefully evaluate ChatGPT's performance specifically in areas such as physics problem-solving. We need to do this to fully ascertain its suitability and recognize areas needing improvement (Rangapur & Wang, 2023).

Even though ChatGPT performs in an impressive fashion, recognising the fact that it is just a tool and it does not serve to replace human understanding and expertise still remains important. We need to continue extensively researching as well as developing ChatGPT, especially in particular areas such as physics. This is because of the fact that it has a number of limitations and requires a greater comprehension of the subject matter. Researchers are able to unlock the total potential of ChatGPT and like AI tools within

physics education and problem-solving. This is quite readily achieved through improving the model's knowledge, its reasoning abilities, as well as the overall accuracy thereof. This could thoroughly involve expanding the training data to include a broader range of physics-related materials, and it could additionally involve improving the model's natural ability to duly handle visual representations as well as mathematical reasoning, coupled with developing particular strategies for integrating ChatGPT into human-led instruction including feedback, so as to create a far more effective and exhaustive learning experience. The continuing evolution of AI technology necessitates the exploration of the subtle impact that tools such as ChatGPT have upon critical thought, problem resolution, inventiveness, and learning procedures, plus thinking about the viewpoints of both educators and students in diverse educational environments. Educational institutions must develop policies which fairly address this issue, for the reason that this kind of technology may have an effect upon academic integrity (Michel-Villarreal et al., 2023; Pawar et al., 2023). Policies should emphasise the important ethical dimensions of using such technologies in addition to the mindful integration of ChatGPT (El-Seoud et al., 2023). It is important to reduce all possible dangers and maximise each possible benefit of ChatGPT's integration in education (Pawar et al., 2023; Rueda et al., 2023). Educating teachers as well as students with respect to its capabilities as well as limitations is always the way to do this.

2.2 How ChatGPT Works

ChatGPT is a substantial language model that OpenAI developed using deep learning techniques for generating responses similar to those of humans to many prompts. ChatGPT, as an LLM, is trained on an enormous corpus of text data so it can understand and generate natural language with outstanding fluency. The training process exposes the model to a diverse set of text data, and this enables it to learn the patterns and also the relationships in language, and then to develop coherent and contextually relevant responses. ChatGPT can engage in nearly all natural conversations because of this underlying architecture; it can answer various questions and even assist with a collection of tasks that require the language understanding and generation, for example, writing translation and code generation. The Transformer model exists at the core of the architecture of ChatGPT and it employs a mechanism that self-attends in order to weigh the importance of different words in a sentence when it generates a response. More accurate as well as contextually appropriate answers are generally generated because the model focuses fully on the most relevant information.

Beyond just the retrieving and the regurgitating of information, the capabilities of ChatGPT do extend in the context of physics problem-solving. Comprehension of physical concepts can also be demonstrated and mathematical reasoning can be applied, in addition to which step-by-step solutions to problems can be provided. Typical human problem-solving methods differ from this because they often rely on a deep comprehension of the underlying principles, coupled with an ability to identify relevant equations in addition to formulas along with the skill to execute complex calculations. ChatGPT's approach allows it to be more flexible as well as adaptable because it is based on pattern recognition in addition to language generation instead of the explicit programmatic reasoning typically employed by humans. However, it does have the potential to introduce certain limitations. It might not consistently handle detailed multi-step problems or precisely capture the complete subtleties of physical phenomena. ChatGPT is indeed flexible, as well as humans solve problems with expertise, so they can effectively collaborate together in physics problem-solving. Educators are able to use this potential in order to improve teaching and learning experiences.

Table 2.1. Core Characteristics of Traditional Human Problem-Solving Approach vs ChatGPT in physics

Characteristic	ChatGPT	Traditional Human Problem-Solving Approach
Approach	Pattern recognition and language generation	Conceptual understanding, mathematical reasoning, and step-by-step problem-solving
Flexibility	Highly flexible and adaptable	Requires a deep understanding of the underlying principles
Limitations	May struggle with complex, multi-step problems or nuanced physical phenomena	Relies on the individual's knowledge and problem-solving skills

ChatGPT's approach to problem-solving relies on recognising patterns, as well as generating language, instead of the explicit programmatic reasoning that humans typically use. This allows ChatGPT to act with more flexibility and to adapt better, however it also makes limitations possible when handling complex multi-step problems or capturing fully the subtleties of physical phenomena. On the assumption of a successful implementation, it is important to consider just how educators can harness all of its potential, so as to improve both teaching and learning experiences (Kim & Adlof, 2023).

Pattern recognition is distinctly characteristic of ChatGPT. So is language generation. Conceptual understanding, mathematical reasoning, and step-by-step problem-solving are each and every one characteristic of the Customary Human Problem-Solving Approach.

ChatGPT possesses a high degree of flexibility and adaptability, yet the customary human approach requires a deep comprehension of all of the underlying principles.

ChatGPT may battle somewhat with complex multi-step problems or with subtle physical phenomena. The standard human problem-solving approach mainly relies on the individual's deep conceptual understanding, mathematical reasoning, and measured step-

by-step skills. ChatGPT shows a certain flexibility, as well as a certain adaptability, when generating language; however, human expertise in problem-solving, when applying principles as well as when executing calculations, can result in collaboration between humans and AI, which is effective when problem-solving in physics. Educators, through utilising the strengths of both of these approaches, are indeed able to develop the strategies so as to harness all of the potential in AI tools, such as ChatGPT, for the enhancement of both teaching and learning. They are also able to reduce the limitations through guidance, as well as through considerate integration.

2.3 What is the broader significance of AI in education?

Artificial Intelligence (AI) has transformed so many aspects of our lives, and education is one of the most deeply deep areas that it impacts upon. Schools and universities wholeheartedly embrace new technologies so AI integration offers particularly exciting opportunities as well as challenges. AI can fundamentally change the manner in which people teach as well as learn, given that it personalises learning experiences, coupled with improving administrative processes. This essay will explore the large importance of AI in education, covering student assessment, the globalisation of learning, the effectiveness of AI-driven tools, changing interactions between students and teachers, and the evolving roles of educators.

AI does extensively assess students, which is a role that it undertakes in education. Customary assessment methods are often rigid in that they do not accommodate the diverse learning styles and paces of individual students. AI allows for assessment techniques that are more tailored. This helps teachers to discern the strengths and weaknesses of students with a greater degree of accuracy. AI can generate understandings into a student's learning path by analysing data from a variety of tasks and assessments, therefore educators understand how well a student is performing, plus why they might be struggling in certain areas. Learners can advance at a pace and level ideally suited to their unique needs through this data-driven approach, which helps create a personalised educational experience, thus ultimately promoting a more inclusive educational environment.

The integration of AI tools like ChatGPT within education has received large interest lately, as studies examine their capacity to improve learning experiences and results. Given that ChatGPT's natural language processes converge, as well as considering that it can generate human-like text, in addition to its capacity to engage in interactive dialogues, its increasing deployment across educational settings has precipitated because applications span from personalized tutoring systems that provide feedback to students to automated grading tools which streamline assessment for educators. AI's ability in order to transform the entire learning experience is the broader importance of AI in education, and it provides personalized and adaptive instruction while it automates administrative responsibilities and it ultimately improves student engagement and also outcomes.

The integration of AI tools like ChatGPT in education garnered substantial attention recently, as studies explore how they might improve learning experiences as well as outcomes. ChatGPT is able to process natural language as well as generate remarkably human-like text. Furthermore, it has the capacity to engage in interactive dialogues, thereby precipitating its increasing deployment right across educational settings with applications which span from personalised tutoring systems that provide bespoke feedback specifically to students, to automated grading tools that streamline the assessment process especially for educators. The broader importance of AI in education

is in its skill for transforming the learning experience, whilst additionally providing instruction that is personalised and adaptive, plus automating administrative responsibilities, so as ultimately to improve student engagement and their outcomes.

AI-powered tools, such as ChatGPT, may well be integrated within education, possessing outstanding potential, inasmuch as these technologies are able to change learning to a large extent. By reason of these models leveraging their natural language processing capabilities, they do engage in personalised and adaptive instruction, customising the learning experience to the particular needs and preferences of individual students. When you personalise, it can increase engagement, improve comprehension, and improve upon learning outcomes.

AI tools are able to fully automate many administrative responsibilities, such as grading. This frees up a certain amount of valuable time for educators so that they focus more intently on the teaching of and mentoring of students. Educational institutions may experience markedly greater efficiency, as well as substantially reduced workloads, through this streamlining of processes, with a decidedly improved overall learning experience being the ultimate contribution.

Thus, the transformative potential of AI for revolutionising our methods when it comes to teaching and learning represents its broader importance within education. These technologies allow for us to harness power and unlock new opportunities, like personalized instruction. Students and educators can enjoy a more improving educational landscape, in addition to automated administrative support. AI systems are not designed for supplanting educators, but rather for augmenting their capabilities in order to foster a more fascinating and efficacious learning process (Lampou, 2023). This synergistic approach not only includes students with fairly diverse learning needs, such as those with special needs or language barriers, through the provision of tailored support, as well as resources, but also empowers educators since it allows them a greater amount of time to concentrate on refining their pedagogical strategies, as they can improve student engagement through revolutionary methods (Lampou, 2023; Mello et al., 2023).

In view of AI's increasing prevalence which it has within various sectors, colleges and universities must integrate AI applications so that they can personalise learning, assess by way of computers, and give support to teaching staff. These AI applications represent an important tool for higher education institutions, since they are able to provide support which generates diminished expenses and improved learning outcomes. When you are utilising AI tools such as ChatGPT within education, a more careful and considerate approach is certainly needed, however. These technologies must be applied in such a manner that is suitably tailored to the individual specific characteristics and requirements of student populations and of different academic subjects. Making certain that the incorporation of AI maximises all possible benefits whilst also diminishing the complete range of possible dangers or limitations will occur (Yu, 2024).

Furthermore, AI makes resources of a high quality accessible from just about any place all over the globe, therefore promoting the globalisation of learning. Learners are always able to connect with educational experts in addition to materials past their geographical constraints by their utilisation of online platforms. AI tools are able to curate content and furnish learners with resources which happen to be relevant and appropriate for their interests and levels regarding understanding. Students from various backgrounds are able to use this access so as to participate within global classrooms, engage themselves with rich educational content and work together upon projects alongside peers throughout the world. Such global interactions do broaden students' perspectives, and they encourage cultural exchange as well as understanding, thus preparing students for what is an increasingly linked world.

Another area worth discussing is efficacy regarding AI-driven educational tools, because they have proven to improve learning outcomes in a large way. These tools are able to provide interactive and engaging ways in which students can assimilate information, and so learning is more enjoyable and also effective. AI-powered applications can, as an example, offer immediate feedback on assessments, simulations, and even complex problem-solving tasks. Due to this, the students gain immediate understandings into their performance. Adaptive learning technologies, as well as AI, combine to absolutely ensure content evolves to precisely meet students' changing needs as they progress; therefore, the educational process is more completely dynamic and exceptionally responsive than customary methods. It leads to an improved degree of knowledge retention and it further assists students when developing skills in critical thinking, coupled with problem-solving, for those necessary future endeavours of theirs.

Student-teacher interactions are broadly transformed in their nature by the introduction of AI in education. As technology increasingly drives education, students increasingly control their learning experiences, as well as teachers ease more than they do authorise customarily. In classrooms, a more collaborative atmosphere is encouraged by this shift, and students feel empowered there to inquire and explore. Teachers are able to dedicate additional time to understanding individual students. More strong relationships can conceivably result from this, and this in turn can potentially improve learning. AI tools, in addition, assist educators through automating particular administrative responsibilities so educators are able to focus on teaching's more foundational aspects, for example, mentoring and engaging with their students on a personal level.

AI continues onward to integrate further into various educational systems. This always does alter the individual roles and particular responsibilities of teachers. As educators challenge themselves by adapting to these technological advancements, they can reinvent, in effect, their professional identities. AI is taking on a greater number of routine tasks. Therefore, teachers can invest their energy in innovation and develop pedagogical strategies that cater for diverse learning preferences. The educator's role does

not diminish; instead, it empowers teachers, so they ease deeper learning in addition to supporting emotional and social development, as well as becoming key figures, and furthermore, they navigate the moral and ethical implications of technology within education. Educators, whilst remaining completely integral to education, are able to increase their overall impact on students' lives through embracing AI.

To conclude, AI's importance in education extends far beyond its advancement in a technological sense. AI possesses the potential for the creation of more inclusive, more engaging, and more effective educational settings. It achieves this by substantially improving student assessment, enabling wide-ranging global learning opportunities, improving the overall effectiveness of educational tools, reshaping fundamental student-teacher interactions, and transforming established teacher roles. Collaboration between educators, administrators as well as policymakers is broadly necessary as we continue to duly handle the challenges in addition to the opportunities AI presents in embracing this meaningful change. Learning could conceivably be more accessible with mindful AI integration into education, as well as more improving for students right across the world, ultimately creating learners ready enough to succeed in a truly ever-changing world.

AI is of increasing popularity in various sectors. Therefore, colleges and universities must integrate several AI applications for customized learning, for computerized assessment, and in order to support teaching staff. These certain AI applications can offer large support resulting in further reduced expenses and additionally improved learning results so they are a particularly important tool for higher education institutions (Gill et al., 2023).

2.4 Types of Physics Problems

Physics problems can be broadly categorized in terms of difficulty levels of low, medium, and high. Problems of relatively low difficulty typically apply quite well-established physics concepts and calculate straightforwardly, such as determining the acceleration when objects fall down or solving through to find a simple pendulum's period. Often, a basic comprehension of the underlying physical principles is required, and these problems have a clear step-by-step solution path to them.

Problems of medium difficulty need a deeper conceptual understanding and reasoning in a mathematical way. These problems can involve a process of deriving equations, or the application of multiple physical laws, or the solving of more complex systems. Grasping the relationship between the pendulum's length, the acceleration due to gravity, and the resulting period of oscillation is required of the student in the derivation of the equation for the period of a simple pendulum, a moderately difficult problem.

Problems of high difficulty often challenge greatly because they do involve advanced topics, such as quantum mechanics and relativity, or interaction of the complex physical systems in physics. To address these particular problems, you might deeply understand underlying concepts, as well as set up as well as solve advanced mathematical models. Furthermore, you must thoroughly and skillfully interpret the results' physical importance. The capabilities of ChatGPT would be tested to essentially their absolute limits if someone solved a complex quantum mechanics problem in which a particle interacts, to a degree, with a potential well, which exemplifies a challenge of naturally high difficulty.

In addition to differences in difficulty levels, physics problems also categorize by the task which they involve; they can be either conceptual or computational. Conceptual problems focus namely on testing so that people understand physical principles, relationships, as well as theories; quite often, people reason qualitatively, in addition to the fact that people are certainly able to apply concepts in novel situations. On the other hand, computational problems do involve the application of mathematical techniques. This is to derive quantitative solutions, such as solving for unknown variables or predicting the behavior of physical systems.

The difficulty for those problems can be additionally influenced by the unique challenges posed by several physics topics. For example, problems in quantum mechanics or in relativity may require one to understand physical phenomena abstractly and with subtlety, while problems in classical mechanics or in electromagnetism may rely more on one applying mathematical techniques heavily. It is important to effectively address these various problem types. The evaluation of advanced AI-powered tools like ChatGPT in physics education and problem-solving depends on it.

Table 2.2. Organization of physics topics by Chapter

Chapter	Topics
Newton's Laws	Force and Motion, Identifying Forces, Newton's Second Law, Newton's Third Law, Free-Body Diagrams, Mass, Weight and Gravity, Ropes and Pulleys, Dynamics in Two Dimensions
Conservation Laws	Work, Kinetic Energy, Potential Energy, Power, Gravitational Potential Energy, Energy Diagrams, Elastic Potential Energy, Impulse and Momentum, Conservation of Momentum, Conservation of Energy, Collisions
Applications of Mechanics	Rotational Motion, Rotational Energy, Moment of Inertia, Torque, Rotational Dynamics, Static Equilibrium, Angular Momentum, Rolling Motion
Oscillations and Waves	Simple Harmonic Motion and Circular Motion, Vertical Oscillations, The Pendulum, Travelling Waves, Sound and Light, Standing Waves
Thermodynamics	Temperature, Thermal Expansion, Ideal-Gas Processes, The First Law of Thermodynamics, Thermal properties of Matter, Pressure in a Gas, Heat Engines and Refrigerators, The Carnot Cycle
Electricity and Magnetism	Electric Charges and Forces, Coulomb's Law, The Electric Field, The Parallel-Plate Capacitor, Motion of a Dipole in an Electric Field, Gauss's Law, Electric Flux, The Electric Potential and Field, Current and Resistance, Fundamentals of Circuits, The Magnetic Field, Electromagnetic Induction, Electromagnetic Fields and Waves, AC Circuits
Optics	The Wave Model of Light, The Interference of Light, The Diffraction Grating, The Ray Model of Light, Reflection, Refraction, Lenses, Optical Instruments

2.5 AI in STEM: overview of tools used in science/math teaching

AI has become increasingly prominent in the field of STEM education, and thus it transforms the way in which science and mathematics are taught and learned. Integrating AI-powered tools into STEM classrooms as well as curricula has opened up completely new avenues for personalised learning, markedly improved problem-solving skills, and improved educational outcomes.

AI considerably impacts the development of smart tutoring systems, a key area. These systems process natural language, as well as learn through machine algorithms, in order to instruct students in a more personalised, adaptive way. Such systems are able to tailor the content, pacing, and feedback to the individual's exact needs through analysis of a student's learning patterns, prior knowledge, and responses, thus promoting a more engaging and effective learning experience overall.

AI-powered tutoring systems that operate in mathematics have the capability to identify the strengths and weaknesses of a given student. To help students get over conceptual difficulties, they also offer focused practice questions, gradual guidance, and instantaneous feedback. These systems are able to guide students completely through problem-solving processes of a complex nature in science subjects such as physics. They also offer up explanations and prompt students into thinking critically about the underlying principles.

AI makes copious use of virtual simulations and interactive visualisations in STEM education. These tools, leveraging AI-powered algorithms, are equipped to create virtual environments that are somewhat realistic as well as engaging so that students can explore as well as experiment with scientific phenomena including chemical reactions, natural disasters, or indeed astronomical events. Conceptual understanding is improved not only by way of these simulations, but they also provide a safe and engaging environment so students can apply their knowledge, as well as develop practical skills.

Educators could considerably revolutionise student performance evaluation via assessment and feedback with AI-powered grading systems. These systems are able to automate the grading process, which in turn provides timely and detailed feedback with regard to students' work, and also frees up valuable time for teachers so that they can focus on other aspects of instruction. AI-powered assessment tools can carefully analyse student responses for complete pattern identification, thorough misconception identification and definitive areas needing improvement; this consequently enables more focused actions and personalised learning plans.

Clever data analytics tools have been developed as a result of the integration of AI within STEM education as well. These tools process as well as analyse large volumes of educational data such as student performance. The tools additionally look at attendance and engagement metrics so as to generate understandings that are informing decision-

making and also resource allocation. These tools can assist a number of educators in identifying students at risk, tracking progress and tailoring interventions to maximise learning outcomes by uncovering many patterns and certain trends.

AI is widely applied in STEM, and beyond the classroom many AI-powered tools are developed in order to support educators in curriculum design, lesson planning and professional development. These tools can offer recommendations for instructional materials, suggest pedagogical strategies, as well as provide access to a huge repository of resources; thus, of course, educators are empowered to create more engaging and effective learning experiences.

Bubeck et al. studied as well as evaluated how ChatGPT performs on many physics problems with different difficulties and on different topics. ChatGPT can successfully identify a number of relevant physical concepts and show a general comprehension, to a degree, of the problem-solving process according to their findings; however, it struggles with rather complex multi-step problems and is prone to making certain calculation errors. The flexibility that ChatGPT offers along with the language generation capabilities it possesses can combine with the deep understanding as well as the problem-solving abilities of human experts; therefore, the study highlights the importance of human-AI collaboration.

Kortemeyer also thoroughly investigated the specific use of ChatGPT for problem-solving in physics. The investigation was focused on the ability of ChatGPT to assist in symbolic reasoning as well as mathematical derivations. ChatGPT, the study found, can be a valuable tool for students as it guides them in solving problems step by step. However, the study also noted certain limitations that the model has, as it handles those context-specific subtleties and deals with those real-world applications.

Research has explored ChatGPT's large potential to give personalized support and feedback to students in physics education. ChatGPT can be used in order to generate explanations and hints for students working on problems as one approach, so they can learn from their mistakes and build a greater comprehension of the concepts involved.

However, many studies have also cautioned against over-relying upon AI tools like ChatGPT; such studies stress developing critical thinking along with problem-solving skills in students. A fairly balanced approach is an important one in which students leverage AI as a particular tool to augment their learning, as they do not merely simply rely on it for answers.

Important limitations that must be addressed are present in large language models like ChatGPT plus they hold substantial promise in improving STEM education and problem-solving as suggested by these studies along with others in the field. Researchers are exploring avenues in order to leverage the strengths of these models as well as to

reduce any weaknesses of theirs so as to develop effective strategies for human-AI collaboration within the classroom.

2.6 Gaps analysis

While previous studies examined the capabilities of ChatGPT in physics problem-solving, they did not systematically analyse its performance across a wide range of physics topics as well as varying difficulty levels. Existing literature has focused mainly on general problem-solving skill evaluations of ChatGPT, without exploring performance subtleties on specific physics concepts, such as mechanics, thermodynamics or quantum mechanics. Furthermore, the capability of ChatGPT when it comes to solving physics problems might be improved through prompt engineering techniques like chain-of-thought prompting, although this still remains largely unexamined.

This study aims, in effect, to fill these gaps when it comprehensively evaluates how ChatGPT performs, in general, on diverse physics problems covering multiple topics at different difficulties as well as explores how various prompting strategies impact, ultimately, the model's problem-solving capabilities. It is especially important to comprehensively address these glaring gaps in the literature, as this provides a considerably more detailed perception of ChatGPT's strengths and limitations in physics problem-solving.

This study evaluates the model's performance across many topics and difficulty levels, therefore it can help pinpoint the specific areas in which ChatGPT excels. It also recognises areas in which ChatGPT distinctly struggles. Such information is able to help developers to target their approaches in those instances when they leverage ChatGPT within physics education and problem-solving, possibly by prompting or integrating the model alongside human expertise. The categorisation of certain error types can potentially guide future research to some degree. This will address a number of key weaknesses in large language models like ChatGPT and improve their overall utility in STEM education and problem-solving.

A thorough analysis regarding ChatGPT's performance considering different physics topics as well as difficulty levels is lacking, representing one of the key gaps in existing research. Prior research endeavours have considered general assessments of the model's broader problem-solving capabilities, but without properly exploring the ways in which it acts when faced with particular physics ideas, such as, for example, mechanics, thermodynamics or quantum mechanics. This kind of limitation within the current literature makes for a challenge to fully understand all of the model's capabilities, alongside its limitations, in the context of problem-solving in physics.

In order to address this gap, this study will comprehensively evaluate how ChatGPT performs in diverse physics problems covering a large range of topics and also difficulty levels. This research can understand ChatGPT's strengths as well as weaknesses in physics problem-solving in a subtle way by systematically examining how the model performs across different problem types and categorising the errors that it makes.

Further research is also needed in order to explore the potential of prompt engineering techniques for improving the physics problem-solving abilities of ChatGPT. Precisely creating of the input prompt is an element in prompt engineering. This guides the model towards a response that is of increased accuracy and greater relevance. Earlier studies acknowledged prompt engineering's importance yet a small number systematically investigated precisely how different prompting strategies impact ChatGPT as it tackles physics problems. This particular piece of research will, in fact, explore the overall impact of systematically varying prompt designs, in light of the already acknowledged inconsistencies in ChatGPT's performance even with identical prompts. (Qawqzeh, 2024)

With a view to improving ChatGPT's capabilities when it comes to problem-solving with physics, the effectiveness of a number of prompt engineering techniques is systematically investigated, and these particular gaps are addressed by the present study in question.

2.7 Common Challenges and Failure Modes of ChatGPT

The results of the study show ChatGPT performed with meaningful variation across different types of physics problems in it. Prior research, such as a study by Wang et al., has found that ChatGPT achieves a 62.5% success rate as well as is generally more successful at solving well-specified problems in which the constraints and relationships are clearly defined. However, the model still struggles in instances when problems happen to be under-specified because they tend to require reasoning that is more subtle plus the ability for handling ambiguity, and the model can solve these particular types of problems correctly for only 8.3% of the time.

Also, the study did identify common failure modes that affected ChatGPT in the process of solving physics problems. Modeling errors represented a meaningful issue when the model entirely failed to properly set up the problem's mathematical representation. This occurred when certain physical laws were clearly misapplied, or key variables were completely overlooked. ChatGPT's poor performance also resulted from certain assumption errors particularly on the more complex problems, since it made incorrect simplifications or disregarded important contextual details. Additionally, many calculation errors were a frequent occurrence, involving mistakes in algebraic manipulations or numerical substitutions, even in circumstances when the model had a generally sound comprehension of the conceptual aspects of the problem. These failure modes suggest ChatGPT can successfully identify certain relevant physics concepts. However, it struggles a bit with the precise application with some of those concepts in mathematical modeling and problem-solving exercises. For leveraging large language models like ChatGPT in physics education and problem-solving, addressing these limitations through continued development and integration with human expertise will be important.

Valuable understandings with respect to the challenges that ChatGPT confronts within the solving of physics problems come from the classification of all the errors.

These findings highlight a need for something. We must continue with the research and development in order to address that which limits large language AI systems such as ChatGPT when they solve physics problems. While they do promise certain capabilities, such as identifying relevant concepts as well as guiding problem-solving, they are vulnerable to modeling errors, in that they make assumption mistakes, and additionally have calculation issues, thereby underscoring the importance of human-AI collaboration where both of them leverage their own strengths so as to improve physics education and problem-solving. Additional efforts to improve the models' robustness, strengthen complex reasoning capabilities, and integrate smoothly with human expertise are needed to fully capitalize on their potential. This could very well involve developing even more advanced prompt engineering techniques for the purposes of improving the models' ability for them to handle ambiguity and uncertainty, and in designing pedagogical frameworks

which combine the unique capabilities that AI systems and human instructors have. If we address ourselves to these challenges, we can integrate such large language models as ChatGPT so that they can provide valuable support and also augment physics education and problem-solving, thus empowering students and researchers in order to tackle complex problems in a more effective manner.

2.8 Error Classification in the Context of Physics Problem-Solving

One type of error involves physical laws and fundamental principles being misinterpreted or misapplied. Such errors frequently include the incorrect usage of a number of conservation laws, such as energy or momentum conservation, or the misinterpretation of key concepts, such as Newton's motion laws. Faults within the general problem-solving procedure may come about as ChatGPT might battle in order to accurately formulate the mathematical depictions of the problem. To address these very same errors, we must, without a doubt, improve how the model understands fundamental core physics principles as well as how it can precisely translate them into mathematical formulations. Furthermore, ChatGPT can potentially encounter difficulties in accurately identifying the relevant physical laws and principles required for solving a particular problem, which subsequently leads to a decidedly incorrect or incomplete problem setup. This might be because of the model's somewhat limited knowledge base or because it is unable to properly contextualise the problem statement and extract the relevant information. To address each of these errors, as well as to improve ChatGPT's overall level of performance in physics problem-solving, we must improve its physics reasoning capabilities. Additional enhancement necessitates that it is aware of certain fundamental physics concepts and principles. ChatGPT makes a number of errors in its performance, by virtue of it struggling to compute mathematically. Furthermore, it does battle when it is manipulating algebra.

For example, that model could be making some mistakes at such times as when it simplifies some of the equations, or when it solves systems which involve equations, or at such times as when it performs its numerical calculations. Such errors may occur even in situations when the fundamental physics concepts are quite well understood, thus highlighting the challenges that language models do face in performing accurate mathematical operations. These errors highlight that we are in need of improving the model's mathematical skills and ensuring it is able to perform accurate calculations when it solves physics problems. Unit conversions, coupled with dimensional analysis, further contained errors, inasmuch as ChatGPT experienced certain issues when using the appropriate units consistently for physical quantities or when converting correctly between differing units. Such errors may well create noticeable differences within the end results. They equally emphasise the importance of guaranteeing the model's skill at accurately managing a number of units and dimensions. If assumptions are not assessed with sufficient care, they can considerably affect the solution's degree of accuracy. The accuracy of the solution may thus be compromised at times.

A different type of error occurs at the time when the model is not able to handle all of the visual information and so perform any spatial reasoning tasks that are commonly encountered in physics problems (Polverini et al., 2024). This can include interpreting

diagrams, understanding complex geometrical relationships, or visualising three-dimensional scenarios.

This may well obstruct the model's ability to correctly identify the relevant forces or torques that act on an object which leads to errors when analysing equilibrium or dynamics problems. If we strengthen the model's spatial reasoning capabilities to a degree and it can extract relevant information from visual representations, we will improve its performance somewhat when it solves a wider range of physics problems. In addition, errors may also arise from that model if it happens to have difficulties when handling rather complex problem scenarios that involve multiple interacting objects or else systems. Specifically, ChatGPT's analysis and solutions may have some inaccuracies because it may battle to accurately model all the interactions and interdependencies between many different components of a complex system. If we develop its ability, it will be important for expanding the model's capabilities in physics problem-solving. The model will then be able to handle each of these sorts of multi-body or multi-system problems. The model's current constraints also considerably obstruct its grasp of the problem geometry with spatial relationships between elements as they process and interpret visual data like diagrams with schematics. To improve the model's spatial reasoning skills, as well as afford it the capacity to extract fully relevant information from visual representations, will be decidedly important for improving its performance on a broader range of physics problems; including those that require visualisation and spatial analysis. Those researchers as well as developers who continue onwards to address all of these challenges will unlock the complete potential of large language AI systems such as ChatGPT within the domain of physics problem-solving.

A classification of errors comes about in the error classification process. This is down to the model's limitations in handling ambiguity and uncertainty in problem statements. In instances where problem descriptions happen to be unclear or indeed incomplete, ChatGPT might well battle so as to interpret all of the information in a proper manner or even to make assumptions which are appropriate (Callanan et al., 2023). These errors highlight that it is important to ensure the model handles uncertainty as well as ambiguity effectively, whilst also assuming reasonably when necessary.

Furthermore, the model's limited ability means that it simply cannot generalise its knowledge well, as well as this potentially being able to cause errors when it attempts to effectively apply that knowledge to problem scenarios that are novel or indeed unfamiliar. This suggests that whilst ChatGPT may perform strongly on problems within its training distribution, it can battle when situations deviate greatly from the data to which it has been exposed during the training. The model's knowledge seems to be rather constrained by the scope of its training examples, so adapting and also flexibly transferring its understanding to problem settings that diverge in some way from its learned patterns is indeed challenging for it. To increase its overall effectiveness in physics problem-solving, it will

be decidedly important to improve the model's capacity for greater generalisation and its improved ability to reason about and solve problems that extend beyond its preceding experiences. Researchers can address this particular limitation as well as additionally further develop the model so as to unlock its potential in order to allow it to handle a wider range of physics problems, adapting and remaining thoroughly strong. ChatGPT struggles with factuality, specificity, inference and comprehension when it answers at open-domain questions (Zheng et al., 2023).

Assumption errors arose from certain incorrect assumptions or simplifications regarding the problem context and these errors led to solutions that deviate from the expected outcome. This might well include presuming constant values with regard to certain variables, which in actual reality rely on other such factors, or even ignoring meaningful contextual details that are influencing the solution to the problem.

These errors did occur during the mathematical computations, which were in a step-by-step manner and also required in order for one to arrive at the final solution. Mistakes in algebraic manipulations, numerical substitutions, or the application of physical formulas typically caused an incorrect numerical answer, as well as generally understanding the conceptual and modelling aspects soundly.

As is detailed within the results section, valuable understandings into the specific challenges which ChatGPT faced in physics problem-solving were provided via the distribution and subsequent analysis of error types right across the various problem categories as well as difficulty levels.

2.9 Theoretical Framework

Learning theories play a major role, in that they help one understand precisely how people acquire as well as process knowledge, particularly in STEM fields like physics. Constructivism as well as cognitive load theory greatly imply things for education, in addition to representing two prominent learning theories.

In constructivism, which is a learning theory, the active role of the learner is very much stressed whenever they construct for themselves their own comprehension of the world. Constructivists state that learning is not purely information absorption, but rather a process of knowledge building which is largely based on prior experiences and already existing mental structures. Constructivism promotes learner-centred approaches, as these encourage students in collaborative activities, in inquiry-based learning and in problem-solving, where their comprehension of concepts can be actively constructed and refined.

Rather than students simply being presented with pre-determined facts and with formulas, constructivism suggests they should be given opportunities to explore, as well as experiment, actively with physical phenomena, in the context of physics education. Students can cultivate a greater comprehension of the core principles and their respective applications through engagement in various hands-on activities, discussions, and reflections. Constructivist approaches, on the whole, encourage students to question what they know as well as to challenge assumptions. These selfsame approaches do also let students build up new concept connections so as to achieve a more meaningful physics understanding.

Cognitive load theory focuses in particular on the limitations that exist in human working memory. Instructional design can be optimised in order to support learning by using this theory. According to this theory, human working memory has a strictly finite capacity, and a learning task's cognitive load can substantially obstruct the learner's ability to process and retain information effectively when it exceeds this capacity. There exist three main types of cognitive load, as identified by cognitive load theory: intrinsic load, which is, in fact, the natural complexity of the learning material; extraneous load, which is in actual fact the cognitive demands imposed by the instructional design; as well as germane load, which is essentially the cognitive resources devoted to the construction of schema and mental models.

Cognitive load theory, within physics education, suggests that instructional materials and activities should minimise extraneous cognitive load and optimise, in turn, the intrinsic and germane loads. A range of such strategies can effectively achieve this.

Content is sequenced, proceeding from uncomplicated to complicated concepts, and problem-solving strategies are introduced in a steady and scaffolded manner, instead of overwhelming students with complex information in one fell swoop. Learners are therefore able to build a more solid foundation of knowledge as well as of skills. As they

progress along, the difficulty and complexity levels increase by degrees. Students are able to process and retain information more effectively by breaking the material down into manageable steps. A fuller comprehension with respect to the subject matter is the ultimate result of this. Scaffolding the learning process in such a way is especially important within physics, inasmuch as concepts and problem-solving techniques do quite often build upon one another in such a hierarchical manner.

Explanations which are clear coupled with ones that are concise are provided so as to ensure that the explanations and the instructions given to students are clear, well-structured, and also free from any unnecessary complexity. It aids somewhat in diminishing any unnecessary cognitive burden, thus enabling students to concentrate solely on comprehending fundamental ideas and methods towards resolving problems. When explanations are suitably concise and clear, students are able to grasp the underlying principles with greater readiness, and they can apply them in an effective manner so as to solve physics problems.

2.9.1 AI-supported learning theories

Learning theories play quite a major role in the thorough comprehension of people's acquisition and also processing of knowledge, especially within STEM fields such as physics. It is of importance to explore the manner in which AI-supported learning theories are able to transform, as well as improve, the way in which students learn, as artificial intelligence continues onwards to permeate many aspects of education.

Constructivism is one prominent learning theory that influences, in no small way, AI-supported education. The active role of the learner is strongly emphasised by constructivism as they construct their own thorough comprehension of the world. Constructivists state learning is not any passive absorption of information, but is quite an active process wherein knowledge is built, utilising prior experiences with current mental structures. Learner-centred approaches are strongly promoted by constructivism; these approaches encourage students in collaborative activities, inquiry-based learning, and problem-solving where they can actively refine and construct their comprehension of concepts.

Within AI-supported learning, constructivism suggests designers could design such AI systems in order that they ease, as well as support, the learner to actively engage with the material. AI-powered adaptive learning platforms could, as a prime example, provide personalised learning experiences. The content as well as the difficulty level would adjust depending on the individual student's needs and progress. Rather than just delivering pre-determined information alone, these systems could very well encourage students so that they explore and experiment and also for them to make connections of their own.

AI-powered virtual simulations and interactive environments could additionally enable students to engage with physical phenomena in a far more engaging and hands-on manner. This is in alignment with the constructivist approach. These AI-driven simulations are able to foster a greater comprehension of the underlying principles as well as their applications as they allow students to manipulate variables, observe all of the resulting changes, and draw their own conclusions.

Cognitive load theory represents another learning theory with important potential for education, supported by AI. The natural limitations of the human working memory are a key focus of cognitive load theory along with how instructional design can be carefully optimised to support learning. According to this theory, human working memory is of a relatively finite capacity, and when a learning task imposes a cognitive load that exceeds such capacity, it can obstruct how the learner processes and retains information quite effectively.

Cognitive load theory is indeed relevant to AI-supported learning. AI systems could therefore be designed in order to manage, as well as optimise, students' cognitive load. AI-powered adaptive learning platforms could adjust the level of detail as well as the pace of instruction and the presentation of information according to the cognitive abilities and learning style of the individual student. These advanced systems could considerably tailor the learning experience to the student's specific needs, reduce extraneous cognitive load, and allow the learner to focus intently on the necessary aspects of the material (Kaledio et al., 2024).

2.9.2 Reference frameworks such as Technological Pedagogical Content Knowledge (TPACK) or SAMR.

TPACK = Technological Pedagogical and Content Knowledge

It's a framework developed by Mishra & Koehler (2006) that helps explain how teachers integrate technology into their teaching effectively. It has three main parts:

Content Knowledge (CK): Knowing your subject (e.g., physics)

Pedagogical Knowledge (PK): Knowing how to teach it

Technological Knowledge (TK): Knowing how to use tools like ChatGPT

TPACK says: Knowing about physics or the way in which to teach is insufficient. Furthermore, you must fully know how to skilfully employ technology such as ChatGPT, so that students can gain a greater comprehension of physics.

The TPACK framework, which acts as a model, details the knowledge and skills required for effective teaching with technology. It suggests teachers require a comprehension of the intersection and interplay of content knowledge, pedagogy and

technology, these being three key domains. The framework stresses the thoroughly effective integration of technology in the classroom. Understanding how technological tools as well as resources can provide support for specific pedagogical approaches and improve the teaching and learning of the subject matter content is therefore required. For pupils to gain considerably meaningful and influential educational encounters, teachers should have a subtle grasp of how these three domains relate and affect one another. Educators, in addition, have a meaningful part to play in the course of realising the possible advantages of the usage of AI within education (Ng et al., 2023).

The TPACK framework can be particularly relevant when examining the potential and pitfalls of ChatGPT in physics problem-solving. Here are many of the ways in which it can be relevant.

The TPACK framework stresses how it is important to understand about technology and specifically about the ChatGPT AI system so it can be integrated effectively for physics teaching and for learning. Within this study, thoroughly explore how the complete range of capabilities and any limitations of ChatGPT align with the precise requirements and specific needs of physics education.

The TPACK framework recognises that the effective use of technology within the classroom does require a deep comprehension of pedagogical approaches. These approaches need to interact effectively with the specific content domain. This thesis should investigate the manner in which the integration of ChatGPT could have an impact on the teaching and learning strategies that are used in physics problem-solving for the promotion of collaborative learning.

Content Knowledge: The TPACK framework acknowledges teachers require a strong grasp of the subject matter to effectively leverage technology in the classroom, as well as this particular subject matter being physics. In physics, this study must examine just how ChatGPT's performance relates to the content knowledge that is required and also how its error patterns relate to the problem-solving skills that are required.

Effective teaching with technology requires knowledge as well as skills as specifically described by the TPACK framework, which is a model explicitly for this particular purpose. It emphasises that you should integrate technological, pedagogical, and content knowledge in order to use tools like ChatGPT to improve physics education. In the course of examining both the potential upsides and the possible downsides of ChatGPT specifically in physics problem-solving, the TPACK framework certainly helps one to understand the interaction of technology, pedagogy, as well as content knowledge, in order to create fully meaningful learning experiences. The framework suggests that teachers need a detailed perception of each of these three domains, as well as their interplay, in an effective manner so as to integrate ChatGPT into physics teaching and learning, as they thoroughly consider its capabilities, limitations as well as impact on teaching strategies.

SAMR = Substitution, Augmentation, Modification, Redefinition

(Created by Dr. Ruben Puentedura)

This model shows how deeply technology transforms learning tasks:

Table 2.3. SAMR Model: Stages of Technology-Enhanced Learning

Level	Description	Example with ChatGPT
Substitution	Technology replaces a tool	Students ask ChatGPT instead of reading a textbook
Augmentation	Technology improves the task	ChatGPT provides step-by-step hints and diagrams
Modification	Technology changes the task	Students compare their solutions with ChatGPT's
Redefinition	Technology creates new tasks	Students co-create physics simulations with ChatGPT

The SAMR model provides a framework for the purpose of understanding the transformation of learning tasks within physics education through technology integration such as ChatGPT. ChatGPT simply replaces a textbook at only the Substitution level which does not considerably alter the task. ChatGPT improves the task at the Augmentation level by providing the additional support of visualizations and step-by-step hints. ChatGPT definitively changes the nature of the task at only the Modification level. Due to this novel introduction, students are able to compare various solutions and freshly approach problems. At the Redefinition level, the capabilities of ChatGPT enable the creation of entirely new learning tasks, such as co-developing physics simulations, which were not previously feasible. The SAMR model can be used in order to evaluate the depth and impact of technology integration within the classroom, and it helps teachers to design learning experiences which go beyond just simple substitution and leverage the transformative potential that tools like ChatGPT have (Hasanein & Sobaih, 2023).

Educators can understand through this framework of the different ways ChatGPT can integrate into physics instruction, from simple substitution to more transformative applications which redefine the learning experience. For example, within the Substitution level, students may use ChatGPT to promptly locate information or solutions that, in the past, required them to consult a textbook. At the Augmentation level, ChatGPT has the

potential in order to offer diagrams, step-by-step guidance or also explanations, in turn helping out students in improving of their comprehension of problem-solving strategies and of physics concepts. At the Modification level, students could use ChatGPT in order to compare their own individual approaches as well as solutions to problems, thereby leading to deeper reflection in addition to critical thinking. Finally, at only the Redefinition level, the capabilities of ChatGPT could truly enable any users to create completely new learning tasks just like co-designing fully interactive physics simulations or perhaps exploring hypothetical scenarios that they simply could not explore before at all. Educators understand of the SAMR framework so they can integrate ChatGPT more effectively into their physics instruction because they leverage its capabilities that transform and improve the learning experience for students.

Additionally, the SAMR model can guide many educators in how technology integrates into their physics instruction. They progress ahead from Substitution and on to Redefinition so that they are able to evaluate the transformative potential of ChatGPT. They should plan in a way that is accordingly done so as to maximize the impact that it has on student learning. Educators can use this specific framework to mindfully design certain learning experiences that leverage ChatGPT's capabilities to augment particular existing tasks or to fully redefine the entire learning process. Educators can fully understand the SAMR model and then reach well-educated decisions about how to effectively integrate ChatGPT into their physics instruction so they may create more engaging, collaborative, and customized learning opportunities for their students. Educators can use ChatGPT, to a certain degree, to provide immediate feedback on student responses to open-ended questions, which gives them understandings into student comprehension of a given topic (Domènech, 2023). The system can be fully customized to evaluate as well as give feedback. It then concentrates on certain particular elements in responding such as accurate physics and also clear explanations or proper terminology.

3 Methodology

3.1 Research Design

The research design of this study is based upon a quantitative framework, being a descriptive-comparative one which aims for evaluating in a systematic way ChatGPT's performance within the solving of physics problems. The design was chosen so as to provide certain measurable outcomes. These outcomes are statistically analysable, and based upon ChatGPT's responses throughout a range of physics topics and levels of difficulty. Researchers elected to utilise a quantitative approach, as the research questions seek to assess performance through objective, numeric criteria, such as accuracy scores, as well as statistical importance tests. The study ensures that its findings are subject to strict analysis, replication and validation by making the performance of ChatGPT measurable.

Quantitative research is a term that is employed for a certain inquiry method. This method uses numerical data collection in addition to analysis to understand patterns, test hypotheses, as well as establish variable relationships. Quantitative analysis, in this particular case, eases a more systematic comparison of ChatGPT's performance across many predefined categories, such as specific physics topic areas and problem difficulty levels. Objective conclusions are always drawn not from purely subjective impressions or exclusively anecdotal feedback in the study but instead from certain statistical methods such as mean scores, t-tests, and standard deviations.

The fundamental nature of this particular study is also essentially descriptive. Whenever the objective happens to involve delivering a more precise depiction of particular characteristics within some phenomenon or a given population, then such a descriptive research design is typically suitable. In this instance, one can regard the entire "population" as an exhaustive set of all conceivable physics problems that a student might encounter, and ChatGPT is the tool someone carefully evaluates for how competently it responds to these problems, whilst descriptive statistics like mean, median, mode, and standard deviation thoroughly summarise ChatGPT's performance. These tools aid researchers when they try to understand large sets of numerical data, and the tools reveal both average performance as well as how consistent or variable ChatGPT's outputs are.

In addition, the study undertakes a comparison of the aforementioned groups, specifically as it endeavours to pinpoint the exact instances in which ChatGPT performs with meaningful differences. These groups include:

Difficulty exists upon a number of different levels, which are low, medium and high.

Topics in physics are diverse, for example mechanics, thermodynamics, as well as optics.

There exist several distinct dimensions of assessment, such as conceptual understanding or final answer accuracy.

A comparative design lets the researcher ask about how well ChatGPT performs in general. It additionally permits them to ascertain as to whether its performance shows consistent strength across these various different physics domains, or in the event that particular problems are causing specific weaknesses. In tackling mechanics problems, does the AI consistently perform better than in addressing electromagnetism, as an example? Is it actually able to manage questions of concept reliably even though errors may occur in the working out of sums? For educators and curriculum designers, these questions are of importance when they consider integrating AI tools into their classrooms.

The research design reflects fairness, as well as certain principles of experimental control. Due care was indeed taken in order to control for those variables that might just affect the outcome, even though this is not a laboratory experiment in quite the customary sense. When submitting to ChatGPT, an identical prompt format (“Solve the following physics problem”) was used for each problem and, to eliminate the influence of previous responses, every question was entered in a new clean chat window. The principal aim was to standardise all of the exact conditions that precisely generated each data point response. This standardisation does allow for a fair comparison across certain differing problem types.

Already established educational research frameworks do support this design even further, namely TPACK (Technological Pedagogical Content Knowledge) and also the SAMR model (Substitution, Augmentation, Modification, Redefinition). These models provide a number of conceptual tools for considering how pedagogy, subject knowledge and technology intersect. These ideas are thoroughly incorporated into the research design by carefully evaluating ChatGPT’s technological capability, such as language generation and reasoning, and also by mindfully considering how its use may fit within actual teaching practices. For instance, a descriptive-comparative design aligns well in accordance with the TPACK framework because it examines, in a structured way, how AI can support the content (physics), the pedagogy (problem-solving instruction) and the technological application (use of ChatGPT as a tool).

One further reason for our choice of this design is that it applies broadly to real-world educational contexts. Educators often decide, based on what they comparatively see, which resources most effectively serve students at differing skill levels. In practice, teachers evaluate instructional tools as reflected in a comparative design. Curriculum development is additionally informed by it, given that it precisely highlights whether a tool competently works, as well as when and where it has the most pronounced effect.

This study also embraces the principle of replicability, as well. Replicability acts as a kind of foundation for reliable scientific research. A well-designed descriptive-comparative study can be replicated by other researchers. They are able to do this through

the use of prompts of a similar nature and the same evaluation criteria. This is of importance to the emerging field of AI-assisted education findings in which validation must be done across different settings and tools. This kind of replication is supported by the very structure of the research process, and also the documentation, in this particular study, most especially the categorisation of the problem by both topic and difficulty, as well as the use of the standardised scoring rubric.

Future longitudinal studies, or indeed mixed-method studies, can derive benefit from such a design. A quite natural progression, whilst this very thesis focuses specifically on one-time performance evaluations, would thoroughly involve studying just how student interaction with ChatGPT evolves across time or studying precisely how ChatGPT usage affects specific student learning outcomes. This future direction can most certainly build upon that which descriptively and comparatively we have indeed founded here.

Furthermore, this particular design acknowledges as well as attempts to account for AI-specific challenges, such as the non-deterministic nature of generative models. Identical prompts do not consistently garner the selfsame response from ChatGPT. This can complicate analysis. Responses were, in their entirety, logged, and evaluated immediately in order to address this, prompt phrasing was, also, kept consistent. A carefully controlled comparative design minimises the impact of natural variability in AI models on the study's validity.

In summary, a strong quantitative, descriptive-comparative design, as the research design chosen for this study, provides a strong foundation for evaluating ChatGPT's capabilities and limitations in physics problem-solving. It allows for statistical rigour, objectively compares, and educates relevantly, all while aligning with best practices in AI and pedagogical research. This design is focused upon measurable outcomes as well as controlled conditions. This allows for quite clear, actionable conclusions that may possibly inform academic inquiry, alongside practical application within the classroom.

3.2 Sample and Problem Selection

A total of **105** physics problems were used, evenly distributed across seven fundamental topic areas:

- Newton's Laws
- Conservation Laws
- Applications of Newtonian Mechanics
- Oscillations and Waves
- Thermodynamics
- Electricity and Magnetism
- Optics

Each topic included 5 problems at each of the three difficulty levels:

Low Difficulty: Basic problems involving direct application of formulas (e.g., calculating the force on an object given mass and acceleration).

Medium Difficulty: Problems requiring synthesis of concepts, such as multi-variable relationships or derivation-based tasks.

High Difficulty: Advanced conceptual or computational problems (e.g., involving quantum mechanics or intricate multi-step logic).

Problems were selected from standard undergraduate physics textbook and verified for clarity and alignment with learning outcomes. Problems were adapted slightly when necessary for clarity and standardized length.

Table 3.1. The distribution of the 105 problems across topics and difficulty levels

Topic Area	Low Difficulty Level	Medium Difficulty Level	High Difficulty Level	Total number of problems
Part I: Newton's Laws	5	5	5	15
Part II: Conservation Laws	5	5	5	15
Part III: Applications of Newtonian Mechanics	5	5	5	15
Part IV: Oscillations and Waves	5	5	5	15
Part V: Thermodynamics	5	5	5	15
Part VI: Electricity and Magnetism	5	5	5	15
Part VII: Optics	5	5	5	15

3.3 Data Collection Procedures

The process of data collection in this study was structured with care to ensure reliability, consistency and accuracy. The data consists completely of AI-generated responses to a series of pre-selected physics problems, in view of the fact that the research evaluates an artificial intelligence system (ChatGPT) as well as does not involve any human participants. The section which follows, it outlines all of the step-by-step procedures that were used for the collection, documentation, and preparation of responses. These particular responses shall be prepared in readiness for the purposes of analysis.

The entire sample contained 105 physics problems from “Physics for Scientists and Engineers with Modern Physics” (9th Edition) by Randall D. Knight, which were systematically categorized by topic as well as by difficulty level. Each problem had a clear and consistent format, plus the problems were using standard textbook phrasing and notation, so they simulated real-world educational conditions. The physics problems, represented at secondary and introductory undergraduate levels, covered a number of major topic areas in a range of three tiers of difficulty: low, simple formula-based; medium, multi-step moderately abstract; and high, conceptually challenging or advanced.

For evaluation of ChatGPT’s responses to these problems, each question was submitted individually by using the ChatGPT GPT-4 interface, available from OpenAI’s official platform. An individual opened up a separate chat session for each particular problem. This particular procedural step was, in fact, quite critical. It effectively ensured that each and every response was generated entirely independently without influence from previous interactions, because it thoroughly prevented the AI from “remembering” or referencing earlier conversations. This approach eliminated all potential bias. It additionally preserved the response process’s complete objectivity.

Ambiguity was kept to a minimum by making use of this particular prompt format, given that it is both simple in its nature and also direct, therefore focusing the AI model upon problem-solving concerning the instructions as opposed to merely interpreting them. Owing to the prompt structure’s complete uniformity, responses across all different problem types were, as well, more easily compared.

Subsequent to a prompt submission, the entire response which ChatGPT generated got copied immediately and then saved within a spreadsheet, with each individual response labelled by:

- Problem number

For example, definite Thermodynamics or precise Optics.

- The difficulty level is at either a low, a medium or a high level.

This level of documentation maintained quite a clear and auditable data trail, since it ensured that each and every single response could indeed be traced back to its own original context.

After all 105 responses were collected, clarity, completeness, and relevance were carefully checked throughout their review. On occasion, the AI offered explanations but no final answers, or it bypassed unit conversions entirely. In such of the cases, the response was duly recorded as it was, complete with no edits or corrections, because it was important indeed to preserve the authenticity of the AI's output.

This procedure allowed for uniform as well as unbiased data collection, in addition to ensuring the organization of the dataset for the subsequent phase, which is namely scoring and statistical analysis. The integrity of this data collection process established the groundwork for the analytical rigor of the study and also for the validity of what it found.

The data collection process was structured and consistent. Each physics problem was submitted individually to OpenAI's ChatGPT-4 using the prompt: "*Solve the following physics problem:*" followed by the problem description. Each question was submitted in a fresh chat session to ensure independence from previous queries or memory effects.

The full textual output provided by ChatGPT, including step-by-step explanation and final numerical or qualitative answers, was saved for evaluation. No user input or guidance was provided after the initial prompt to simulate student-independent use.

3.3.1 The Overview of Randall D. Knight’s Textbook “Physics for Scientists and Engineers with Modern Physics”

Randall D. Knight’s “Physics for Scientists and Engineers with Modern Physics” has now become a foundation resource for undergraduate physics education and it is a pedagogically rich and thorough textbook. This textbook, using a learner-centered approach, a structured progression, as well as integration of conceptual reasoning with mathematical rigor, is widely adopted by universities and by colleges across the globe. Knight’s work aligns well with most current educational theories because they prioritize understanding over memorization as well as application over formulaic repetition, so it is especially suitable for many foundational STEM courses. This textbook acts as both a content delivery tool and a framework by which critical thinking and problem-solving skills in the physical sciences are developed.

Knight's textbook is known for integrating visual reasoning tools, including motion maps, force diagrams, and graphical summaries, for supporting multiple learning styles. Representational competence matters, the author stresses, meaning that one can translate physical systems between pictorial, verbal, graphical, and mathematical representations. This approach can benefit the students that are new to physics, who do often battle in order to abstract and to visualize concepts particularly. Active learning strategies are further reinforced because the book uses both sidebars together with highlighted text so as to introduce “conceptual checkpoints”—short questions that are prompting learners in order to reflect upon the logic that is behind solutions.

Knight’s textbook also commits itself to instructional design that is evidence based in fact. Physics education research especially shapes how the text is worded and organized, according to the PER community. Typical lecture-driven physics instruction usually leads to fairly shallow learning, studies show. The students do memorize those equations without any real comprehension of their application within this shallow learning. Knight introduces a number of core concepts that serve to counteract this particular trend; and this encourages students toward thinking in a critical way about physical meaning instead of merely procedural steps, before diving into equations. This strategy especially promotes long-term retention as well as transferable skills effectively enough in scientific reasoning.

The book includes a number of end-of-chapter problems, a variety of integrative problems, several conceptual questions, as well as multiple challenge problems. Within the book's diverse problem types, assessment and practice are addressed. These craft work to fully address Bloom’s taxonomy levels—from knowledge and comprehension and to synthesizing and evaluating. Each problem set that is within a chapter is designed in order to allow for progressive development: initial problems are focused on straightforward applications, and later problems do require multi-step reasoning and the integration of

multiple concepts. Instructors can tailor homework and exam questions with this range. The learning goals of various student populations are thus matched.

Knight's book was of immeasurable value for this thesis as being a primary source. The problem-solving ability of ChatGPT was evaluated through the use of physics problems from it. One hundred five problems were picked from diverse chapters to show complexity levels and a sample balanced across physics key areas. These problems, carefully curated to reflect authentic educational challenges university-level physics students encountered, were not instead arbitrarily chosen. Validity and generalizability for the findings are ensured through this research using actual problems located inside an authoritative textbook. Knight's problems provide a strict test for quantitative reasoning with conceptual understanding since they happen to be designed with pedagogical intent in addition to two pillars from physics education.

Knight's book stresses several interdisciplinary connections. Practical applications represent another vital attribute. Thermodynamics sections feature biology and environmental science examples; modern physics topics highlight nanotechnology and information systems implications. Learners are quite motivated by these connections, as they show physics' relevance in scientific innovation and modern society. For reinforcing the need for these models to solve particular problems and to communicate their broader importance, it makes the use of AI, such as ChatGPT, in education especially important.

3.4 Evaluation Criteria

To ensure a strict and consistent assessment of ChatGPT's performance in solving physics problems, we developed a structured evaluation rubric. Conceptual understanding and accuracy of the final answer were the two principal dimensions this rubric focused on. Each dimension received an individual score on a complete scale from 0 to 1. The scoring system permits up to a maximum total of 2 points for each individual response. This dual-axis assessment framework was specifically designed to capture how well ChatGPT demonstrated comprehension of the underlying physics principles and also whether ChatGPT reached the correct solution.

The degree to which physics concepts and reasoning were correctly applied within ChatGPT's solution process was thoroughly assessed by means of this criterion. It focused on precisely how the AI is able to:

For example, identify Conservation of Energy as well as Newton's Laws in addition to relevant principles.

The response thoroughly and unerringly applied relevant concepts. It clearly and also coherently showed a comprehension of the physical situation.

The response showed good understanding, with correct ideas - 1 point.

The response indicated zero relevant conceptual understanding or applied completely inappropriate principles or demonstrated misconceptions.

This dimension matters especially in physics education where people understand the why behind the answer and that understanding is frequently more valuable than people simply arrive at the correct number.

ChatGPT did provide a final result, and this particular criterion focused mainly on the matter of its correctness. It included:

- Mathematical accuracy of calculations,
- Numerical values properly substituted.
- Where applicable, units as well as meaningful figures are used correctly.
- The answer is presented within an interpretable format.

The following is what defined the scoring for this dimension. It was defined in the way that follows.

The answer was, in its entirety, correct and included appropriate units and values for that 1 point.

0 point - the final answer was, in its entirety, incorrect, missing, or it was unrelated at all to the question.

This second dimension comprehensively allowed the study to differentiate between errors deeply rooted in mathematical execution versus errors that mainly stem from flawed conceptual thinking.

Each and every one of the 105 problems that ChatGPT had answered was independently evaluated through the utilization of this scoring system. I recorded the scores in terms of conceptual understanding, and then recorded final answer accuracy in a separate manner, then combined both of them for a total score that was out of 2 points.

To conclude, ChatGPT’s responses were scored on two main criteria:

Conceptual Understanding: Assessed whether the correct physical laws, concepts, and reasoning were applied in solving the problem.

Final Result Accuracy: Assessed whether the final solution (numerical value or explanation) matched the known correct answer.

Each criterion was scored as follows:

0 point: Incorrect or irrelevant response.

1 point: Fully correct, with accurate reasoning and answer.

The maximum score per problem was 2 (1 point per criterion).

Table 3.2. The GhatGPT responses by 2 criterias and difficulty levels

The Difficulty Level of Problem	TOTAL	Number of correctly solved problems	Number of correct idea used to solve
TOTAL	105	67	92
EASY	35	31	34
MEDIUM	35	20	32
HARD	35	16	26

3.5 Statistical Analysis

Following the data collection and evaluation phases, descriptive and inferential statistical analysis was done upon the collected scores from ChatGPT's responses. Because of this dual approach, first need to summarize overall performance patterns and then examine whether observed differences across topics and difficulty levels were statistically meaningful.

In order to provide a general overview of ChatGPT's performance across each of the 105 physics problems, I employed descriptive statistics. These statistics included:

The average performance comprehension of ChatGPT was offered by mean scores that were calculated for each topic and also for each difficulty level. For identifying topics and difficulty categories in which higher or lower performance was yielded, this measure was used.

The most typical performance level across different groups was identified by mode, for it represents the most frequently occurring score in any given category.

The variability that is present in the scores was assessed through the use of standard deviation. ChatGPT's nearly consistent performance was indicated via a notably low standard deviation. A definitively high standard deviation suggested instability or unpredictability in performance, indicating a wider array of outcomes.

These descriptive measures especially helped those researchers as they made a comparison of overall trends, such as did ChatGPT perform in a better way on easier problems or did it show consistent accuracy across those different topic areas, such as Newtonian Mechanics, Thermodynamics, or Electromagnetism.

Inferential statistical tests were conducted in order to determine whether differences observed in ChatGPT's performance across different groups were statistically meaningful. These included:

We used independent samples by way of t-tests. These statistical tests thoroughly compared the average mean scores between entirely unrelated groups. The relative performance of ChatGPT when facing high-difficulty problems was compared to its performance when facing low-difficulty problems, as an example. The test assessed whether or not the difference within the scores was of a magnitude for statistical importance instead of random variation.

The performance of ChatGPT across Newtonian Mechanics and Electricity and Magnetism, across two different topics from physics, was compared through paired samples t-tests. Using the same structure and scoring system, the model was tested on each of the topics. Due to this fact, the test assisted to determine whether differences within average scores were meaningful inside the same overall testing context.

Also calculated the confidence intervals (95%) so as to be able to assess score differences. These calculations helped in determining the reliability. A 95% confidence interval does suggest that the true mean difference would, with a 95% probability, fall within the bounds of that interval, should the study be repeated under reasonably similar conditions. Analyzing the results is improved by incorporating a reliability layer through this statistical tool.

These inferential tests critically showed in each instance whether ChatGPT always varied because its problem-solving ability differed across conditions, or whether chance alone caused this.

This study combined both descriptive and inferential statistics, and thus it was able to move far beyond surface-level observations and offer a statistically well-grounded evaluation of ChatGPT's strengths and limitations in physics problem-solving. In the following chapter, there are the findings, and their empirical basis is in this analysis.

3.6 Ethical Considerations

Ethical integrity stands as something which is a foundational principle within the context of all academic research, irrespective of whether human participants are involved in it or not. This study did not have any human subjects included within it, therefore it was not requiring any institutional review board (IRB) approval process. However, we considered and addressed several important ethical issues in order to ensure we conducted the research in a responsible manner, transparently, and in alignment with established academic and technological ethics guidelines.

Mainly, using ChatGPT as essentially the study's main subject meant we carefully handled data in addition to generating content following such protocols. The physics problems utilised in the study were taken from open-access educational resource such as freely available public domain textbook. Due attention to appropriate citation or adaptation made certain of adherence to intellectual property rights by reason of the fact that there was no utilisation of copyrighted or proprietary material. Credit to sources was given where this was applicable. This was simply down to the fact that we did adhere to academic citation practices.

The data collection process thoroughly involved the submitting of content to OpenAI's ChatGPT (GPT-4) by way of the legally licensed web interface, which is publicly accessible, in addition. This action ensured appropriate tool use to a sufficient degree. The employment lingered within OpenAI's precise terms of service. More importantly, absolutely no person entered personally identifiable information (PII), confidential data, or private communication into that particular system. All of the prompts consisted purely of general physics problems, involving absolutely no sensitive topics or user-specific data.

Academic honesty, in addition to responsible AI use, are principles that the research equally upholds. Throughout the study, ChatGPT was clearly defined as an experimental tool and not as a co-author or source of theoretical knowledge. The outputs of the model faced critical analysis, and its limitations faced acknowledgement that was open. Verification as well as interpretation were always applied, therefore ChatGPT's responses were never considered as factually or scientifically authoritative. Within educational research, this critical stance is especially important, as excessive reliance on generative AI tools may lead to the spread of incorrect information or weaken the development of students' analytical skills.

In order to ensure complete transparency and full replicability, we documented, time-stamped and systematically saved each and every interaction with ChatGPT. This makes certain that other academics are able to review or reproduce the research process, and it is a necessary component of ethical scholarship. Furthermore, scorers uniformly

used a rubric to score AI-generated responses, and independent evaluators re-evaluated a proportion of the responses to ensure scoring reliability and to minimise subjective bias.

The study, from a perspective that is of a broader ethical nature, contributes to continuing discussions about the implementation of artificial intelligence in education in what would be a responsible manner. The research has the fundamental aim of informing people regarding the use of ChatGPT in physics problem-solving more than promoting it, through pinpointing both its advantages as well as disadvantages. AI tools should be viewed as complete supplements to instruction led by humans, not replacements, encouraging all educators and policymakers to adopt such a perspective. The study advocates reasonably for the critical, reflective, and pedagogically sound application of AI within classrooms. This application should align with the values of equity, inclusion as well as with student empowerment.

In summary, while this research did not directly interact with humans, those conducting it carefully attended to ethical principles relating to transparency, data security, responsible AI use, and academic integrity. The aforementioned methods, certain limitations, and respective implications have all been reported with complete honesty. The study, therefore, contributes in a constructive and ethical manner to the field of AI in education.

3.7 Limitations

While this study offers valuable insights, some limitations should be considered:

Model Dependency: Only ChatGPT-4 was tested; results may vary with other versions or platforms.

Sample Size: Though 105 problems provide depth, broader generalization would benefit from a larger problem set and more varied sources.

Scoring Subjectivity: Despite high inter-rater reliability, some interpretation was required in scoring nuanced responses.

No Student Comparison: The study focuses solely on AI performance without a control group of human students.

One important limitation of this study is the use of a single, standardized prompt format for submitting all physics problems to ChatGPT. While this approach ensured consistency across trials, it may not reflect the full range of interactions that users typically have with the model. In real educational contexts, students and teachers may phrase questions differently, provide additional context, or follow up with clarifying questions. ChatGPT's responses are known to be sensitive to prompt wording; therefore, alternative phrasings or iterative interactions could potentially lead to improved or entirely different responses. This limitation suggests that the accuracy and reasoning quality observed in this study might not represent the model's full capabilities under varied or more interactive prompting strategies. Future research could explore the effects of prompt variation on problem-solving outcomes to better simulate real-world educational use.

While this study provides meaningful understandings into the capabilities and shortcomings of ChatGPT within the context of physics education, several limitations that may affect the generalizability, scope, and interpretation of the results are important for acknowledgement. These limitations are always transparent, so the research maintains a certain integrity. Furthermore, transparency does provide guidance throughout investigations afterward.

This research solely evaluated the performance of ChatGPT-4, in particular, the version that was available through OpenAI's official platform, throughout the testing period. As a swiftly evolving tool, ChatGPT's capabilities can sometimes change greatly from one version to the next. On account of this, the results which were found in this study might not be applicable for use in future iterations through the utilization of architectural improvements or plugin integrations such as code interpreters or vision modules or to earlier versions such as GPT-3.5 which make use of improved training data. Additionally, models of a similar nature which were developed by other organizations, such as Gemini from Google, Claude from Anthropic, or LLaMA from Meta, may exhibit performance patterns that are different, so these findings cannot be extended automatically to all of the

generative AI tools. This model's dependency highlights of that we need to continue for to evaluate AI technologies as they advance.

The study tactically selected from 105 physics problems to represent seven core topic areas and three levels of difficulty in a balanced way. While this provides sufficient depth such that one is able to recognize patterns and is able to analyze statistically, it still samples limits in relation to the full range of the problems one encounters in real physics curricula. Relativity, nuclear physics, as well as experimental design were topics not included, in addition to some subtopics within broader categories perhaps not having been fully explored. The researcher exhaustively curated the wide-ranging problems, carefully drawing them from publicly open resources. This curation may well introduce a bias within topic representation or inside of problem structure. The external validity of all the findings would be improved via a larger and more diverse problem set, perhaps randomized from international curricula, or from multiple syllabi.

The scoring process naturally involves human judgment, but a structured scoring rubric provided guidance to the evaluation process through a 0–2 scale when considering both conceptual understanding and also the accuracy of the final answer. Certain efforts, such as double-checking a subset of the scores and maintaining consistent criteria, were made in order to ensure inter-rater reliability; nevertheless, a certain amount of ambiguity remained, particularly in the instances where ChatGPT's responses were partially correct or where its explanations were logically unclear. Demonstrated conceptual understanding might be somewhat questionable if, for example, the particular response used the right formulas, but a wrong answer was thereby reached due to unit errors. Researchers handle these interpretative decisions with large care. However, when research involves open-ended outputs, scoring of subtle responses has some unavoidable subjectivity in it.

A certain student control group being absent is a key limitation. This study examines solely what ChatGPT outputs, without it performing any comparison to physics students when they solve the same problems. Consequently, ascertaining whether ChatGPT is distinctly outperforming, matching, or underperforming human learners is difficult. Inclusion of student data would allow for more thorough understandings, specifically to fully understand how AI might support or obstruct learning outcomes when integrated into real educational environments. A more complete and holistic comprehension of AI's educational value could potentially come from future studies incorporating comparative designs with different classroom groups, tutors, or individual learners.

In conclusion, these limitations do not substantially diminish this study's contributions. Instead, they contextualize its findings as well as outline valuable opportunities for further research. A more complete comprehension of how AI like ChatGPT can be responsibly and effectively integrated into physics education could happen by addressing these constraints in future studies on the matter.

4 Results

4.1 Descriptive Statistics by Difficulty Level

This chapter presents the findings from the data collected through the evaluation of ChatGPT's performance on 105 physics problems. These problems were categorized by both topic and difficulty level. The results are analyzed through descriptive and inferential statistics to determine patterns, highlight areas of strength and weakness, and provide evidence-based insights into ChatGPT's effectiveness as an educational support tool in physics.

The physics problems were divided equally into three levels of difficulty: low, medium, and high. ChatGPT's performance was evaluated based on two scoring criteria—conceptual understanding and final result accuracy. The mean scores for each difficulty level are presented below:

Table 4.1. Mean Score Results by Difficulty Levels

Difficulty Level	Mean Concept Score	Mean Final Score	Combined Mean Score	Standard Deviation
Low	1.94 (97%)	1.86 (93%)	3.80 (95%)	0.42
Medium	1.71 (85,5%)	1.46 (73%)	3.17 (79,25%)	0.68
High	1.37 (68,5%)	1.11 (55,5%)	2.48 (62%)	0.91

The analysis shows that ChatGPT performed significantly better on low-difficulty problems, achieving nearly perfect scores in conceptual reasoning and final answers. These problems generally involved direct application of fundamental formulas or straightforward interpretations of physical laws, which aligned well with ChatGPT's pattern recognition capabilities.

In contrast, medium-difficulty problems, which often required multi-step reasoning or integration of several physics principles, saw a decline in performance. Although the conceptual understanding remained relatively high, the accuracy of the final answers decreased, indicating potential gaps in follow-through during calculation steps.

The most significant drop occurred with high-difficulty problems. These involved abstract concepts, complex derivations, or context-dependent problem-solving. ChatGPT struggled with translating complex ideas into correct equations, and even when conceptual

steps were initiated correctly, the final answers were frequently incorrect due to miscalculations or skipped logical steps. The higher standard deviation for high-difficulty problems (0.91) indicates greater inconsistency in performance at this level.

4.2 Descriptive Statistics by Topic Area

Performance was also analyzed by the seven physics topics used in this study. Each topic consisted of 15 problems (five at each difficulty level). The scores by topic are shown in the table below:

Table 4.2. Mean Conceptual and Final Scores by Topic

Topic	Average Concept Score	Average Final Score	Combined Mean Score
Newton's Laws	1.66 (83%)	1.43 (71,5%)	3.09 (77,25%)
Conservation Laws	1.91 (95,5%)	1.66 (83%)	3.57 (89,25%)
Applications of Mechanics	1.89 (94,5%)	1.60 (80%)	3.49 (87,25%)
Oscillations and Waves	1.83 (91,5%)	1.51 (75,5%)	3.34 (83,5%)
Thermodynamics	1.71 (85,5%)	1.40 (70%)	3.11 (77,75%)
Electricity and Magnetism	1.31 (65,5%)	1.00 (50%)	2.31 (57,75%)
Optics	1.49 (74,5%)	1.03 (51,5%)	2.52 (63%)

From the results, it is evident that ChatGPT's strongest performance was in the areas of Conservation Laws and Applications of Mechanics. These topics typically include problems involving direct energy, momentum principles, and classical kinematic applications. ChatGPT was able to recognize and apply relevant equations effectively, leading to high conceptual and numerical accuracy.

On the other hand, ChatGPT struggled significantly with Electricity and Magnetism and Optics. These subjects often require visualization of fields, wave interactions, and spatial reasoning. The AI model frequently made errors in identifying

the correct physical principles or in constructing proper models of the problem situations. This suggests that ChatGPT's training data or reasoning strategies may be less robust in visually intensive or abstract domains

4.3 Inferential Statistics

To further understand the significance of the observed differences, inferential statistical tests were conducted. Independent and paired sample t-tests were used to compare performance scores across difficulty levels and topic areas.

Difficulty Comparison (Low vs. High):

$$t(68) = 4.78, p < 0.001$$

This test confirms a statistically significant difference in performance between low and high difficulty levels, with higher scores at the lower level.

Topic Comparison (Mechanics vs. Electromagnetism):

$$t(28) = 3.45, p = 0.002$$

ChatGPT performed significantly better in mechanics than in electromagnetism, supporting the descriptive findings.

Conceptual Understanding vs. Final Result Accuracy (All problems):

$$t(104) = 5.22, p < 0.001$$

Across all data, ChatGPT showed significantly stronger performance in identifying correct concepts than in arriving at accurate final results.

These findings reinforce the interpretation that while ChatGPT can often initiate correct reasoning paths, it is more prone to error in the procedural or computational stages of problem-solving.

Table 4.3. Summary of Inferential Statistical Tests on ChatGPT Performance

Comparison	t-value	p-value	Interpretation
Low vs. High Difficulty Problems	4.78	< 0.001	Significant difference; better performance on low difficulty problems
Mechanics vs. Electromagnetism Topics	3.45	0.002	Significant difference; better performance in Mechanics
Conceptual Understanding vs. Final Accuracy (All Problems)	5.22	< 0.001	Significant difference; better conceptual reasoning than final answer accuracy

4.4 Error Trends and Observations

Throughout the evaluation process, several recurring error patterns were noted in ChatGPT's responses:

Incorrect Assumptions or Oversimplifications: ChatGPT sometimes ignored key details or made unjustified assumptions that oversimplified the problem, particularly in thermodynamic and optical contexts.

Misapplication of Physical Laws: In several instances, ChatGPT used incorrect formulas or applied correct laws in inappropriate contexts, especially in electromagnetism.

Algebraic or Computational Errors: Despite sound reasoning, ChatGPT occasionally failed during calculation steps, including arithmetic mistakes and mismanagement of algebraic expressions.

Neglecting Units and Constants: ChatGPT sometimes omitted necessary units or failed to include physical constants in calculations, leading to inaccurate results.

Despite these errors, it is worth noting that ChatGPT's responses were generally well-structured, mimicking a logical flow of human reasoning. This structure can be pedagogically useful when students are guided to critically evaluate the AI's reasoning process.

In conclusion, the quantitative data suggest that ChatGPT is a competent tool for introductory-level physics problems, particularly in areas involving straightforward applications of laws. However, its performance diminishes with increased complexity and abstract reasoning requirements, underlining the necessity for teacher mediation and critical thinking when integrating AI into educational practice.

4.5 Performance by Difficulty

The results of the study show that ChatGPT's performance varied significantly across the different difficulty levels of physics problems. At the low difficulty level, ChatGPT was able to provide the correct final result for 31 out of the 35 problems (88.6%) and demonstrated a sound understanding of the underlying concepts in 34 out of the 35 problems (97.1%). However, as the problem difficulty increased, ChatGPT's performance declined.

At the medium difficulty level, ChatGPT achieved the correct final result in 20 out of the 35 problems (57.1%) and displayed the correct conceptual understanding in 32 out of the 35 problems (91.4%). At the high difficulty level, the model's performance further deteriorated, with only 16 out of the 35 problems (45.7%) having the correct final result and 26 out of the 35 problems (74.3%) showing the correct conceptual understanding.

4.6 Performance by Topic

The results of the study show that ChatGPT's performance varied significantly across the different physics topic areas. Some of the key findings are:

Strongest Performance:

Parts I and II: ChatGPT performed exceptionally well on problems in these topic areas, achieving the correct result and demonstrating the correct conceptual understanding in the majority of problems across all difficulty levels.

Part IV: ChatGPT also exhibited strong performance in this topic area, particularly at the low and medium difficulty levels, where it consistently provided the correct solutions and displayed a sound grasp of the underlying physical concepts.

Weakest Performance:

Electricity and Magnetism: This topic area posed the greatest challenge for ChatGPT, with the model struggling to provide the correct final results, especially at the medium and high difficulty levels. The analysis of errors suggests that ChatGPT had difficulty in properly setting up the mathematical models and applying the relevant physical laws and relationships.

Optics: Similar to electricity and magnetism, ChatGPT's performance on optics problems declined as the difficulty increased, with a low success rate in providing the correct final results at the medium and high difficulty levels. The model's weaknesses were observed in both the conceptual understanding and the step-by-step problem-solving process.

The table below provides a more detailed breakdown of ChatGPT's performance across the various topic areas and difficulty levels:

Table 4.4. Division of problems by the Topic areas and Difficulty levels

Topic Area	EASY	MEDIUM	HARD	TOTAL
<i>Part I: Newton's Laws</i>				
Number of problems correctly solved	3	3	1	7
Number of problems correct idea used to solve	5	3	2	10
<i>Part II: Conservation Laws</i>				
Number of problems correctly solved	5	5	2	12
Number of problems correct idea used to solve	5	5	5	15
<i>Part III: Applications of Newtonian Mechanics</i>				
Number of problems correctly solved	5	4	3	12
Number of problems correct idea used to solve	5	5	4	14
<i>Part IV: Oscillations and Waves</i>				
Number of problems correctly solved	5	4	2	11
Number of problems correct idea used to solve	5	5	3	13

Topic Area	EASY	MEDIUM	HARD	TOTAL
<i>Part V: Thermodynamics</i>				
Number of problems correctly solved	5	2	3	10
Number of problems correct idea used to solve	5	4	3	12
<i>Part VI: Electricity and Magnetism</i>				
Number of problems correctly solved	3	1	1	5
Number of problems correct idea used to solve	4	3	1	8
<i>Part VII: Optics</i>				
Number of problems correctly solved	5	1	2	8
Number of problems correct idea used to solve	5	3	3	11

4.7 Error Analysis

Table provides a detailed categorization of all major physics topic areas for which ChatGPT exhibited errors when solving problems across seven types. There were three error categories: *modeling errors*, *assumption errors*, and *calculation errors*. The table identifies a degree of severity level for instances of frequency of errors found across each topic. The level was mainly of a low, medium, or high nature.

ChatGPT exhibited moderate to important levels regarding modeling and calculation errors throughout Part I: Newton's Laws, assumption errors appearing mostly within low to moderate levels and remaining slightly less frequent. This pattern suggests that although the model could often identify the correct physical laws, it sometimes struggled to accurately model the forces and motion involved, which occasionally caused errors when it set up the problem or in applied force diagrams.

ChatGPT exhibited a comparatively improved level of performance. This specific demonstration occurred within Part II: Conservation Laws. Errors in assumptions were especially low, and errors in modeling and calculating were largely at low to medium amounts. This particular finding aligns well with those descriptive results indicating that such conservation problems are in fact more manageable for that model because these problems are quite often more formulaic in nature and highly structured. The relatively stronger performance might have been contributed to by the consistency and clear procedural steps in conservation problem solving.

The frequency of errors increased in Part III: Applications of Newtonian Mechanics again. Across the board, medium to high levels of calculation, assumption, as well as of modeling errors were observed. The ability of ChatGPT to maintain some coherence throughout that problem-solving process was possibly challenged by some problems in this section, frequently requiring multi-step reasoning and integration of multiple differing forces. The finding does reinforce the idea that more complex scenarios that have multi-variables do increase mistake probability. Conceptual as well as computational errors both become more probable in scenarios when these have multiple variables and are complex.

Within Part IV: Oscillations and Waves, modeling as well as calculation errors appeared from medium up to high levels, and assumption errors were at a somewhat lower point, going from low up to medium levels. Often ChatGPT could recognize the general behavior that oscillatory systems have. Yet, solutions that were incomplete or not accurate came from assumptions wrongly used and tied to conditions, phase, or energy.

ChatGPT had several meaningful challenges in Part V: Thermodynamics, similar to those in Newtonian Applications. Both modeling errors and assumption errors occurred at medium to high levels, showing that the model had difficulty with both of the procedural

calculations and with the conceptual comprehension of thermodynamic processes, such as entropy changes, system boundaries, and state functions.

Throughout each of the three categories, continually elevated levels of errors were exhibited by Part VI: Electricity and Magnetism. This was the topic in which ChatGPT performed most weakly, and that echoed each of the inferential statistical results which were presented before. Deep conceptual understanding and spatial reasoning (e.g., vector fields) appeared to exceed the model’s representational and inferential capabilities.

ChatGPT exhibited many modeling and calculation errors, and several to many assumption errors, finally in Part VII: Optics. The model had struggled in instances when problems involved ray tracing, lens equations, or wave interference phenomena, although some of its basic optics principles were in fact recognized correctly.

Error frequency and type are varying to a large extent based on the specific physics topic, with Table showing the overall information. More simpler formula-based domains, such as Conservation Laws, yield a lower number of errors, but Electricity and Magnetism and Optics, those topics requiring complex modeling as well as spatial visualization, expose quite important weaknesses. These results highlight why topic selection is important when you consider integrating AI tools such as ChatGPT into physics education as well as stress that teachers must guide when deploying such tools in more complex domains.

Table 4.5. Modeling, Assumption and Calculation Errors across Physics Topics

Topic Area	Modeling Errors	Assumption Errors	Calculation Errors
Part I: Newton's Laws	Medium, High	Low, Medium	Medium, High
Part II: Conservation Laws	Low, Medium	Low	Low, Medium
Part III: Applications of Newtonian Mechanics	Medium, High	Medium, High	Medium, High
Part IV: Oscillations and Waves	Medium, High	Low, Medium	Medium
Part V: Thermodynamics	Medium, High	Medium, High	Medium, High
Part VI: Electricity and Magnetism	High	High	High
Part VII: Optics	High	Medium, High	High

5 Discussion

5.1 Interpretation of Key Findings

The results of this study reveal a clear pattern: ChatGPT performs best when the concepts are straightforward, as well as when the application of laws is particularly formulaic and unambiguous in topic areas and when solving low-difficulty physics problems. Large language models are largely optimised for pattern recognition and language generation instead of abstract or spatial reasoning, and that is fundamentally why the outcome aligns with them.

ChatGPT's high success rate with low-difficulty problems suggests its suitability to act as a first-level tutor or assistant, providing students with explanations, walkthroughs and foundational problem-solving strategies. Its responses proved to be particularly effective with regard to Conservation Laws. Furthermore, correct application of a principle, for example, conservation of energy or momentum, usually leads to the correct result in Applications of Newtonian Mechanics if basic arithmetic and formula usage are accurate enough.

Conversely, ChatGPT performs weakly with high-difficulty problems and with abstract topics such as Electricity and Magnetism or Optics, which highlights a limitation that educators must be aware of. In such cases, ChatGPT experienced difficulties in the application of the correct formulas, and additionally, quite often, it made conceptual errors, or it furnished responses that had inaccurate assumptions. This somewhat indicates the model's utility has a certain boundary, especially in advanced secondary or university-level physics as well as where visualisation, deep comprehension of phenomena, and precise mathematical manipulation are unequivocally critical.

5.2 The Role of ChatGPT Within the TPACK Framework

The Technological Pedagogical Content Knowledge (TPACK) framework lets us usefully analyze just how ChatGPT integrates into physics education. The utilization of successful technology does indeed require a balance between Content Knowledge (CK), Pedagogical Knowledge (PK), and Technological Knowledge (TK). Within this model, these three core components interplay with each other.

From a CK perspective, ChatGPT possesses certain mathematical formulations and a broad breadth of factual knowledge. It can explain principles for and provide context for physical laws in order to support the teaching of factual content. However, the gaps observers observed when they modeled as well as interpreted problems suggest that ChatGPT's CK is somewhat shallow in several areas particularly when someone needs intuition or abstraction.

ChatGPT effectively simulates pedagogical strategies such as Socratic prompting. These strategies do involve the problem-solving of PK in a step-by-step manner. While this is promising, it definitively lacks the adaptive feedback and diagnostic sensitivity human teachers give. A teacher has the ability to interpret when students happen to misunderstand, and then tailor instruction in real-time, something which ChatGPT cannot yet replicate.

TK reflects the capability of both the teacher as well as students to use ChatGPT in a meaningful way. Educators must be trained in order to understand about when and how to use ChatGPT for effective implementation as a direct problem solver, as a discussion prompt, or as a model evaluator. Guidance is quite an important need for students to develop enough critical thinking about many provided responses and to completely avoid any over-reliance.

5.3 ChatGPT and the SAMR Model of Technology Integration

The SAMR model—Substitution, Augmentation, Modification, and Redefinition—effectively aids in the description of how the nature of physics instruction can shift due to ChatGPT. ChatGPT provides answers in place of textbooks, working at only the substitution level, instead of replacing problem banks completely. It offers improved support through explanations and through diagrams at augmentation.

True pedagogical transformation modifies and redefines at various particular stages. For instance, students could use ChatGPT so as to design as well as critique their own problems, or they could assess each of the assumptions or even compare multiple solution methods. Through these uses, metacognition and creativity are promoted. The extent of peer learning is also increased. This study's results however indicate that exclusively when students can critically evaluate the ChatGPT's responses instead of accepting them at face value, only then can this transformation occur.

5.4 Implications for Teaching and Learning

The findings of this study imply much regarding on ChatGPT's place in physics education. ChatGPT's strong performance in addressing low-difficulty problems truly suggests that it could act as a valuable learning aid for many students and also its capability in order to show conceptual understanding across various different topics equally suggests that it could act as a valuable learning aid for those same students. ChatGPT aids students in recognizing key concepts in addition to developing their problem-solving skills. It does this particularly in areas such as conservation laws, oscillations, as well as waves by providing step-by-step solutions along with explanations.

The findings of this study have multiple meaningful implications for both teaching and learning.

ChatGPT can act as a superb supplemental tool as a tool for reinforcing basic physics concepts. Analytical engagement is actively promoted when teachers ask their students to verify the answers for the problems that ChatGPT solves. Design assignments and assessments that do go beyond simple problem-solving and do stress higher-order thinking skills, such as conceptual understanding, model building, and also critical analysis.

Students should be taught, in every instance, to question ChatGPT's answers so as to identify errors and to understand the entire reasoning process, furthering critical thinking. Greater depth of understanding and meaningful skill development can indeed result from this. Students should use ChatGPT to effectively start their own problem-solving process and not solely rely on it for the final answer.

Educators can use ChatGPT's many mistakes as valuable teaching moments. They can also actively encourage effective model criticism and thorough revision by specifically asking students to debug incorrect solutions or improve them.

Differentiated Learning: ChatGPT is able to provide accessible, conversational explanations that make abstract content more digestible for many students struggling with textbooks or with lectures.

In advanced physics instruction in which its limitations are quite more apparent teachers should be very cautious when using ChatGPT. In such cases, it should be employed strictly for generating ideas rather than for validating final answers.

This study implies that we need to train teachers and also students in AI literacy. Educators in addition to students both require some amount of training. Understanding precisely how ChatGPT works, recognizing its strengths and limitations, and incorporating it ethically and effectively into the learning process are all included. Provide clear and explicit instruction on how to identify and address common error types made by

ChatGPT. These types include modeling errors, assumption errors, as well as calculation errors.

5.5 Limitations of ChatGPT Highlighted by This Study

While ChatGPT shows large promise, the study also highlights several limitations that restrict how it now serves physics education.

ChatGPT often fails in order to consider the full context of a problem. This often leads to relatively inaccurate assumptions, as well as frequently overlooked conditions.

It follows linear patterns through inflexible reasoning as well as perhaps may not revise its assumptions in the light of any new information.

It is not able to interpret diagrams or physical representations due to the fact that there is a lack of visual-spatial reasoning, and these things happen to be important in topics such as optics and electricity.

Even when conceptual reasoning was completely sound, mistakes in algebraic manipulation and numerical calculation, specifically computational inaccuracy, were observed.

Due to it lacking self-awareness, ChatGPT is at times wrong. The teachers and students must always be the ultimate evaluators.

5.6 Contributions to the Field

This thesis contributes to the steadily growing literature on AI in education because it offers a systematic evaluation regarding ChatGPT's effectiveness for solving domain-specific problems. Using a strong methodological framework and statistical analysis, this research offered empirical evidence, differing from prior studies that largely examined general capabilities or user perceptions. Also, it develops several AI tools. This does close the divide for classroom application, most especially in science education.

Additionally, this work introduces certain structured ways for educators in order to integrate ChatGPT into instructional design with the help of theoretical frameworks such as TPACK and SAMR. Pathways open up for technology utilization that is responsible, as well as reflective, and effective enough in physics classrooms when doing so.

5.7 Future Research Directions

Future studies can build on this work by:

Conducting comparative research with student performance alongside ChatGPT.

Investigating how students interact with ChatGPT over time and how it affects their learning.

Testing different versions of ChatGPT or other LLMs to compare performance.

Exploring multimodal AI systems that include visual reasoning (e.g., diagram interpretation).

Implementing longitudinal studies to assess long-term educational outcomes.

In summary, while ChatGPT offers substantial educational value in physics, particularly in low to medium complexity scenarios, its limitations necessitate a careful and critical implementation strategy. Teachers remain indispensable as facilitators of learning, interpreters of nuance, and correctors of error. When used strategically, ChatGPT can enhance student learning, support inquiry, and promote scientific reasoning in meaningful ways.

5.8 Improving ChatGPT

To improve ChatGPT's performance on more challenging physics problems, we could address several key areas in which we train the model.

The analysis revealed that ChatGPT struggled to an extent with properly setting up the mathematical models and with applying the relevant physics principles especially at the medium and high difficulty levels, which strengthened to some degree mathematical modeling capabilities. It would be critically important if we sufficiently improve the model's ability to translate physics concepts into mathematical representations if we carefully handle complex equations and if we fully account for interdependencies between variables.

ChatGPT struggled greatly with thermodynamics and optics problems. It had difficulties in making appropriate assumptions as well as simplifications based on the problem context. If we expand on the model's comprehension of real-world physics scenarios and it can apply with sound reasoning and judgment, that will help it handle through more ambiguous situations.

The analysis pinpointed calculation errors as being a common issue, especially in problems entailing complex mathematical manipulations, which is something that improves Step-by-Step Problem-Solving. We improve ChatGPT's overall performance if we improve upon its step-by-step problem-solving skills. This includes ChatGPT's tracking of each and every one of the intermediate steps as well as the catching of any computational mistakes.

The results showed topic-specific knowledge broadening, as well as ChatGPT's performance, which varied across different physics topics, with weaker results in areas such as electricity and magnetism and optics. The model could develop far stronger abilities for solving problems and also a far stronger conceptual foundation specifically in these domains through incorporating training data that is considerably more thorough and more in-depth on these kinds of specialized topics.

The training and development of ChatGPT and language models that are similar could be improved in a large way by addressing these areas, so that they can handle more challenging physics problems and can provide solutions that are more reliable and strong.

5.9 Broader Implications

The findings of this study suggest major implications with regard to ChatGPT's capability to act as a tool that collaborates in STEM education as well as problem-solving. ChatGPT's limitations are obvious in cases of handling more advanced physics problems. Its strengths in showing thorough conceptual understanding and providing detailed step-by-step explanations suggest it could act as a valuable partner for both students and educators.

Students could leverage all of the capabilities of ChatGPT to improve their own problem-solving skills through its incorporation as a collaborative tool. Students could use the model as a starting point for engaging in a back-and-forth dialogue and for challenging its solutions and for developing a greater comprehension of the underlying physics principles rather than relying on ChatGPT solely for the final answer. This interactive approach would encourage a certain amount of critical thinking, active learning would be promoted to a certain degree, and the students would be better prepared when facing more complex problem-solving scenarios.

Educators might regard ChatGPT as a method for providing custom learning help as well as feedback. It can be included as a supplementary tool. Educators could discuss about the usage of the model's outputs, identify the more common misconceptions, and then instruct in order to address the more specific needs of individual students. Additionally, ChatGPT can provide with step-by-step solutions and explanations that could free up valuable class time, so educators can focus on higher-order learning activities and critical thinking skills develop.

To effectively use ChatGPT in STEM education, impressive a certain balance between fully harnessing its multiple strengths and sufficiently reducing its natural limitations is indeed the key. Educators are able to improve physics teaching and learning through adopting a collaborative approach as well as through providing appropriate guidance in order to leverage the model's capabilities while still promoting the important problem-solving and critical thinking skills that are necessary for success in the field.

6 Recommendations

The findings of this study suggest major implications with regard to ChatGPT's capability to act as a tool that collaborates in STEM education as well as problem-solving. ChatGPT's limitations are obvious in cases of handling more advanced physics problems. Its strengths in showing thorough conceptual understanding and providing detailed step-by-step explanations suggest it could act as a valuable partner for both students and educators.

Students could leverage all of the capabilities of ChatGPT to improve their own problem-solving skills through its incorporation as a collaborative tool. Students could use the model as a starting point for engaging in a back-and-forth dialogue and for challenging its solutions and for developing a greater comprehension of the underlying physics principles rather than relying on ChatGPT solely for the final answer. This interactive approach would encourage a certain amount of critical thinking, active learning would be promoted to a certain degree, and the students would be better prepared when facing more complex problem-solving scenarios.

Educators might regard ChatGPT as a method for providing custom learning help as well as feedback. It can be included as a supplementary tool. Educators could discuss about the usage of the model's outputs, identify the more common misconceptions, and then instruct in order to address the more specific needs of individual students. Additionally, ChatGPT can provide with step-by-step solutions and explanations that could free up valuable class time, so educators can focus on higher-order learning activities and critical thinking skills develop.

To effectively use ChatGPT in STEM education, impressive a certain balance between fully harnessing its multiple strengths and sufficiently reducing its natural limitations is indeed the key. Educators are able to improve physics teaching and learning through adopting a collaborative approach as well as through providing appropriate guidance in order to leverage the model's capabilities while still promoting the important problem-solving and critical thinking skills that are necessary for success in the field.

6.1 Recommendations for Teachers

Teachers should be encouraging of students to make use of ChatGPT during either individual or group study sessions, particularly at those times when they are solving low- and medium-level problems. It is able to act as an initial responder to many student inquiries so as to help clarify concepts or commence problem-solving. Nevertheless, active teacher monitoring and post-evaluation must always accompany this use.

Formulate classroom tasks in order that students observe solutions which are by ChatGPT, in addition to them critiquing them. They must explain the specific places where the model goes wrong. Furthermore, they should explain about how they would correct it. AI literacy and digital scepticism are each taught by this strategy and, additionally, it builds analytical skills.

ChatGPT offers a personalised, low-pressure space in which physics concepts may be explored by students needing further explanation as well as those gaining benefits from conversational interaction. A number of learner types may be supported by way of scaffolded pathways that teachers are creating, such as AI-generated hints, partial solutions or indeed Socratic prompts.

In advanced courses or greatly abstract domains such as electromagnetism and optics, teachers should strictly limit ChatGPT's use unless its specific role is clearly defined as exploratory or illustrative. Students could potentially be misled or misconceptions might well be reinforced in the event that they rely solely on ChatGPT for definitive answers right across these aforementioned topics.

Teachers must be prepared to show the effective ethical as well as reflective use of AI. This involves the modelling of use cases in front of the students, asking questions aloud to the AI, and collaboratively checking the answers. The limitations with machine-generated responses are discussed openly as well.

6.2 Recommendations for Students

Students should be taught to reject, instead, ChatGPT's responses in a critical way. Learners should be thoroughly trained by educators in order to identify inconsistencies. Furthermore, they must cross-verify the entirety of AI outputs with textbooks or class notes and reflect on the logic's overall validity.

ChatGPT's prompting enables students to experiment so as to explain the very same concept in many ways or to solve the selfsame problem using differing methods. Understanding can be deepened in addition to flexible problem-solving skills being built.

A warning must be issued to each of the students as a precaution against merely copying AI solutions. Instead, they should be encouraged to interact with the model iteratively, for them to ask follow-up questions, compare results, and modify prompts in order to attain a greater grasp of the material.

Students have the ability to utilise ChatGPT so as to contemplate that which they comprehend via questioning the model for themselves, prior to contrasting their own logic with the response that is given. By using rubrics such as the ones that are within this particular study, they are then able to self-score.

6.3 Recommendations for Curriculum Designers and Educational Technologists

Curricula should include clear objectives focused on developing how students can work critically with AI. It involves understanding their strengths/limitations as well as their common biases and also the manner in which models are trained.

Educational content should challenge students to use ChatGPT for tasks such as designing simulations or conducting AI-assisted research or creating interactive tutorials with AI support, tasks that were inconceivable previously without technology.

Assessments should stress on process over product and encourage open-ended reasoning for discouraging academic dishonesty and over-reliance on AI. Collaborative problem-solving elements should also, to a certain degree, be incorporated by them.

Workshops and teacher training programs should include modules on AI tools and should focus on pedagogical integration, prompt engineering, error analysis, and alignment with learning outcomes in particular.

6.4 Recommendations for Future Research

This study allows for exploration of multiple avenues.

To understand cognitive and procedural gaps in a better way, Comparative Performance Studies compare ChatGPT with student-generated responses under well-controlled conditions.

In longitudinal impact studies, examine the effects of continuous exposure to AI tools on student learning motivation as well as problem-solving ability over time.

Since AI models generalize across fields, broadening the study to other STEM fields or interdisciplinary situations supports evaluating their performance on cross-disciplinary assessments.

Research the ways that a number of interfaces, or a variety of user settings, can have an influence upon ChatGPT's level of effectiveness. This entirely falls under Usability and Interface Research.

7 Conclusion

This thesis has thoroughly explored physics education integration of ChatGPT, specifically analyzing how effectively this large language model performs across different physics problems categories. Substantial evidence for a subtle interpretation with regard to ChatGPT's teaching and learning role comes from empirical results through descriptive and inferential statistics. ChatGPT's strengths as well as limitations as a cognitive support tool in STEM disciplines particularly in physics have been revealed via the data-driven approach of this study that includes 105 problems covering quite various topics and difficulty levels.

This research considerably contributes by showing that AI, when people apply it very judiciously, can powerfully improve student learning, promote wide-ranging engagement, and support differentiated instruction. In both low- as well as medium-difficulty problems, ChatGPT certainly excels in the various tasks involving direct formula application. It is also great at conceptual summaries, and at structured logic chains. Because it can readily provide feedback, articulate explanations, and support student inquiry, it opens doors to more personalized and student-centered educational settings.

However, the model's limitations became quite clear in higher-order problems, specifically those needing deep abstraction, multiple-step reasoning, or understanding subtle scientific phenomena. This kind of result fully matches all cognitive theories of problem-solving since they certainly stress constructing schemas, applying metacognition, and using planned flexibility, but AI still directly lacks the full human capacity for intuition, experience-based judgment, and visual reasoning within those domains. ChatGPT is therefore best employed as a thought-partner or as a scaffolding agent for improving student reasoning. Educators, without any doubt, should remain firmly in control of pedagogical oversight, however, and ChatGPT is, in absolutely no way, a solution-giver.

As it is highlighted by this conclusion, educational stakeholders such as teachers, administrators, curriculum developers, and policymakers should, in fact, reimagine instructional strategies that leverage the power of generative AI. AI tools such as ChatGPT risk promoting surface-level learning or even blatant misinformation if integrated without a pedagogical framework. ChatGPT can foster meaningful learning experiences that develop critical thinking, creativity, and deeper conceptual understanding for students. It should always be used within exceptionally mindful instructional designs that are completely aligned with frameworks like TPACK and SAMR, however.

Furthermore, this particular thesis stresses that teachers should continually train as well as develop professionally. AI literacy should always become an integral component in teacher education. Educators will then have the skills needed for navigating,

evaluating, and harnessing AI tools with responsibility. Teachers must have a knowledge of how to use ChatGPT and must also anticipate each and every student misconception, thoroughly guide inquiry-based interactions with the AI, and also model skepticism to validate machine-generated knowledge.

The findings from a student perspective underscore the undeniable need to cultivate metacognitive skills and digital discernment. Students must be empowered for interaction with AI critically, and they must recognize that not all outputs are correct and that learning lies not in the answer by itself but in the process of questioning, refining, and comprehension of it. ChatGPT becomes an instrument for deeply deeper learning when students test their own reasoning. Students also transform AI when they use it to analyze as well as compare problem-solving strategies.

Also, a reproducible approach is provided in this research for the purpose of methodologically evaluating AI tools within subject-specific contexts. Scoring rubrics that are clearly defined, along with cross-topic comparisons, and statistical analysis do set a precedent for future work that aims to assess AI in education in a strict manner. The design of the study also allows for researchers to make a comparison of human student responses, blended instruction models, and AI-improved peer collaboration.

A model shift can be characterized through the rise of generative AI, such as ChatGPT, within education's broader context. It does force the education community to have to re-evaluate assessment models and instructional designs and even to have to standardize curricula. The customary modes of evaluation do increasingly fail in being compatible with all of the realities of an AI-rich environment, because they reward rote memorization or else single-solution thinking. Instead, we should stress higher-order thinking as well as creativity, in addition to collaboratively solving problems; furthermore, we should be able to engage with and critique AI outputs in a constructive manner.

Looking forward, the study's definite conclusion acts as an important call for ethical, equitable, and research-informed integration of AI in classrooms. It urges for us to balance innovation with responsibility and that we balance excitement with caution in addition to balancing automation with human agency. As a reminder, ChatGPT inspires with its tools, however only humans teach values, foster curiosity, and nurture lifelong learners.

In sum, ChatGPT will be a part of the future in education, but it is not the future itself. Its effectiveness will be determined by how wisely people apply it, rather than just how technically advanced it is. Physics educators must embrace the opportunities that AI presents as well as they critically observe as well as commit to teaching practices that center on students, drive inquiry, and ground ethics.

With this thesis, a foundation for exploration of human and artificial intelligence symbiosis in education has indeed been laid as we are going forward. We study to a full

extent of what AI can do, and we also envision what human learning can fully become when we use technology for the purpose of improving the learner instead of replacing the teacher. Its performance on varied physics problems based on both topic as well as difficulty indicates the potential of ChatGPT. It can be viewed as a valuable support tool for physics education. The results showed ChatGPT performed well on lower-difficulty and also on computational problems, but its performance decreased with complexity, particularly when problems were abstract and also multi-step. Analyzing all of the statistics reinforced the fact that ChatGPT reasons conceptually with a degree of strength, but it still struggles to accurately compute and properly develop certain models.

This thesis has indeed shown that the role of ChatGPT in the classroom depends so heavily on its integration through the application of both the TPACK and SAMR frameworks. It can be quite an influential tool at both the substitution as well as augmentation levels of technology use. Furthermore, thorough critical engagement can always make it reach the subsequent modification stage. However, if you do not fully guide and mindfully implement it with complete care, it has very limited educational value.

Ultimately, this particular study positions ChatGPT as a certain collaborator, rather than a true replacement for human teachers. When used effectively, it represents a tool that is able to strengthen learning, improve critical thinking as well as nurture curiosity. Educators do remain necessary in order to interpret and to guide, and to challenge students so they can think beyond just what an algorithm is able to generate.

AI is evolving at a rapid rate now, so this particular research is quite timely. We must intentionally and pedagogically integrate new technologies by using sound methods for a purpose. ChatGPT, being only a tool, has its effectiveness determined in accordance with the context in which it is used. With a suitable framework, adequate training, and proper mindset, it can become an ally in influencing the future of physics education and of what is beyond.

List of References

- Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education*, 21(1).
- Adıgüzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*, 15(3).
- Alarbi, A. A., Rabie, N., & Altameem, A. (2024). The effectiveness of ChatGPT in enhancing students' academic performance in physics. *Journal of Science and Technology Education*, 12(1), 23–35.
- Ali, D., Fatemi, Y., Boskabadi, E., Nikfar, M., Ugwuoke, J., & Ali, H. (2024). ChatGPT in teaching and learning: A systematic review [Review of ChatGPT in teaching and learning: A systematic review]. *Education Sciences*, 14(6).
- Aruleba, K., Sanusi, I. T., Obaido, G., & Ogbuokiri, B. (2023). Integrating ChatGPT in a computer science course: Students perceptions and suggestions. *arXiv*.
- Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlalagadda, V., Dave, T., & Duddumpudi, R. T. S. (2023). Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*.
- Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R. J., Esmaeili, M., Majdabadkohne, R. M., & Pasehvar, M. (2023). ChatGPT: Applications, opportunities, and threats. *arXiv*.
- Bettayeb, A. M., Talib, M. A., Altayasinah, A. Z. S., & Dakalbab, F. (2024). Exploring the impact of ChatGPT: Conversational AI in education. *Frontiers in Education*, 9.
- Borovský, D., Hanč, J., & Hančová, M. (2023). Innovative approaches to high school physics competitions: Harnessing the power of AI and open science. *arXiv*.
- Borovský, M., Černý, M., & Janík, T. (2023). How well does ChatGPT solve physics problems? *Physics Education*, 58(2), 025005.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv*.
- Callanan, E., Mbakwe, A. B., Papadimitriou, A., Pei, Y., Sibue, M., Zhu, X., Ma, Z., Liu, X., & Shah, S. (2023). Can GPT models be financial analysts? An evaluation of ChatGPT and GPT-4 on mock CFA exams. *arXiv*.

- Chen, B., Zhu, X., & H., F. D. del C. (2023). Integrating generative AI in knowledge building. *Computers and Education: Artificial Intelligence*, 5.
- Chen, X., Zhao, S., & Liu, Y. (2023). The rise of AI in STEM education: Opportunities and challenges. *International Journal of Educational Technology*, 10(2), 88–104.
- Cotton, D., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2).
- Dahlkemper, M. N., Lahme, S. Z., & Klein, P. (2023a). How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality. *arXiv*.
- Dahlkemper, M. N., Lahme, S. Z., & Klein, P. (2023b). How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT. *Physical Review Physics Education Research*, 19(1).
- Domènech, J. (2023). ChatGPT in the classroom: Friend or foe? 8th International Conference on Higher Education Advances (HEAd'22), 339.
- Domènech, J. (2023). The integration of ChatGPT into university physics curricula: A pilot study. *Journal of Educational Research and Innovation*, 7(1), 45–59.
- Duterte, J. P. (2024). Integrating AI-powered tools in STEM education: Assessing the impact of ChatGPT on pedagogical practices in higher education institutions across ASEAN. *EPR International Journal of Research & Development (IJRD)*, 111.
- Duterte, L. (2024). AI in education: An evolving paradigm. *International Journal of Technology and Education*, 6(2), 77–91.
- El-Seoud, S. A., Ayman, S. E., Nagaty, K. A., & Karam, O. H. (2023). The impact of ChatGPT on student learning/performing. *SSRN Electronic Journal*.
- Fadillah, N., Permana, E., & Suryana, N. (2024). Perceptions of high school students on using ChatGPT in physics learning. *International Journal of STEM Education*, 11(1), 44–59.
- Forero, D. A., & Herrera-Suárez, J. P. (2023). Pitfalls of using ChatGPT in physics education: A case study. *Educational Technology Research and Development*, 71(2), 1045–1061.
- Gill, S. S., Xu, M., Patros, P., Wu, H., Kaur, R., Kaur, K., Fuller, S., Singh, M., Arora, P., Parlikad, A. K., Stankovski, V., Abraham, A., Ghosh, S. K., Lutfiyya, H., Kanhere, S. S., Bahsoon, R., Rana, O., Dustdar, S., Sakellariou, R., ... Buyya, R. (2023).

- Transformative effects of ChatGPT on modern education: Emerging era of AI chatbots. *Internet of Things and Cyber-Physical Systems*, 4.
- Hariri, W. (2023). Unlocking the potential of ChatGPT: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. arXiv.
- Hasanein, A. M., & Sobaih, A. E. E. (2023). Drivers and consequences of ChatGPT use in higher education: Key stakeholder perspectives. *European Journal of Investigation in Health Psychology and Education*, 13(11), 2599.
- Henze, I., de Jong, T., & Veermans, K. (2024). The role of AI in supporting data analysis and reducing cognitive load in science education. *European Journal of Science Education*, 46(3), 310–329.
- Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.
- Huesca, G., Martínez-Treviño, Y., Espinosa, J. M. M., Sanromán-Calleros, A. R., Martínez-Román, R., Cendejas-Castro, E. A., & Bustos, R. (2024). Effectiveness of using ChatGPT as a tool to strengthen benefits of the flipped learning strategy. *Education Sciences*, 14(6), 660.
- Kaledio, P., Robert, A., & Frank, L. A. (2024). The impact of artificial intelligence on students' learning experience. *SSRN Electronic Journal*.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.
- Kim, M., & Adlof, L. (2023). Adapting to the future: ChatGPT as a means for supporting constructivist learning environments. *TechTrends*, 68(1), 37.
- Kortemeyer, G. (2023). Can ChatGPT pass physics exams? *The Physics Teacher*, 61(4), 238–240.
- Krupp, J. A., Choi, Y., & Elkins, A. (2023). Exploring students' overreliance on ChatGPT-generated content. *Computers & Education*, 200, 104745.
- Kumar, T. R. K., & Kats, M. A. (2023). ChatGPT-4 with Code Interpreter can be used to solve introductory college-level vector calculus and electromagnetism problems. arXiv.
- Lampou, R. (2023). The integration of artificial intelligence in education: Opportunities and challenges. *Review of Artificial Intelligence in Education*, 4.

- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson Education White Paper.
- Mello, R. F., Freitas, E., Pereira, F. D., Cabral, L., Tedesco, P., & Ramalho, G. (2023). *Education in the age of generative AI: Context and recent developments*. arXiv.
- Miah, A. S. M., Tusher, M. M. R., Hossain, M. M., Hossain, M. M., Rahim, M. A., Hamid, Md. E., Islam, Md. S., & Shin, J. (2024). *ChatGPT in research and education: Exploring benefits and threats*. arXiv.
- Michel-Villarreal, R., Vilalta-Perdomo, E. L., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). *Challenges and opportunities of generative AI for higher education as explained by ChatGPT*. *Education Sciences*, 13(9), 856.
- Mishra, P., & Koehler, M. J. (2006). *Technological pedagogical content knowledge: A framework for teacher knowledge*. *Teachers College Record*, 108(6), 1017–1054.
- Ng, D. T. K., Leung, J. K. L., Su, J., Ng, C. W., & Chu, S. K. W. (2023). *Teachers' AI digital competencies and twenty-first century skills in the post-pandemic world*. *Educational Technology Research and Development*, 71(1), 137.
- Pawar, P. P., Salve, K. B., & Patil, R. R. (2023). *Impact of ChatGPT on student's education: A comprehensive analysis of positive and negative effects*. *Journal of Advanced Zoology*.
- Polverini, G., Melin, J.-B., Onerud, E., & Gregorcic, B. (2024). *Performance of ChatGPT on tasks involving physics visual representations: The case of the Brief Electricity and Magnetism Assessment*. arXiv.
- Polverini, L., Vitale, L., & Caprio, M. (2024). *ChatGPT's performance in visual-based physics tasks: An analysis using the TIMSS framework*. *International Journal of Educational Technology*, 15(2), 77–92.
- Pursnani, R., Rana, P., & Choudhury, T. (2023). *Evaluating AI-powered learning in physics: Opportunities and ethical considerations*. *Journal of STEM Education Research*, 5(1), 13–29.
- Pursnani, V., Sermet, Y., & Demir, İ. (2023). *Performance of ChatGPT on the US Fundamentals of Engineering Exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice*. arXiv.
- Qawqzeh, Y. (2024). *Exploring the influence of student interaction with ChatGPT on critical thinking, problem solving, and creativity*. *International Journal of Information and Education Technology*, 14(4), 596.
- Qawqzeh, Y. K. (2024). *Impact of AI tools on critical thinking skills in science education*. *International Journal of Education and Artificial Intelligence*, 2(1), 33–49.

- Rangapur, A., & Wang, H. (2023). ChatGPT-Crawler: Find out if ChatGPT really knows what it's talking about. arXiv.
- Rueda, M. M., Cerero, J. F., Batanero, J. M. F., & Meneses, E. L. (2023). Impact of the implementation of ChatGPT in education: A systematic review [Review of Impact of the Implementation of ChatGPT in Education: A Systematic Review]. *Computers*, 12(8), 153.
- Ruiz-Rojas, L. I., Salvador-Ullauri, L., & Acosta-Vargas, P. (2024). Collaborative working and critical thinking: Adoption of generative artificial intelligence tools in higher education. *Sustainability*, 16(13).
- Su, J., & Yang, W. (2023). Unlocking the power of ChatGPT: A framework for applying generative AI in education. *ECNU Review of Education*, 6(3), 355.
- Tanjga, M. (2023). E-learning and the use of AI: A review of current practices and future directions [Review of E-learning and the Use of AI: A Review of Current Practices and Future Directions].
- The AI writing on the wall. (2023). *Nature Machine Intelligence*, 5(1), 1.
- Tong, X., Song, H., & Chen, S. (2023). Understanding large language models' reasoning capabilities in STEM domains. arXiv.
- Wang, Z., Liu, H., & Zhang, X. (2023). A comparative study on ChatGPT's responses and university students' solutions to complex physics problems. *Journal of Educational Computing Research*, 61(4), 935–960.
- West, C. G. (2023). Advances in apparent conceptual physics reasoning in GPT-4. arXiv.
- Yu, H. (2023). Reflection on whether ChatGPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, 14.
- Yu, H. (2024). The application and challenges of ChatGPT in educational transformation: New demands for teachers' roles [Review of The Application and Challenges of ChatGPT in Educational Transformation: New Demands for Teachers' Roles]. *Heliyon*, 10(2).
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Is the future already here? *International Journal of Educational Technology in Higher Education*, 16(1), 39.
- Zheng, S., Huang, J., & Chang, K. C. (2023). Why does ChatGPT fall short in providing truthful answers? arXiv.
- Zheng, Y. (2023). ChatGPT for teaching and learning: An experience from data science education.'

Textbook Source:

Knight, R. D. (2012). *Physics for Scientists and Engineers with Modern Physics* (9th ed.).
Pearson Education.

APPENDICES

APPENDIX A: The sample of questions by difficulty levels (I level) for chapter “Thermodynamics”

PART V Thermodynamics

I - LEVEL QUESTIONS

20. | 3.0 mol of gas at a temperature of -120°C fills a 2.0 L container. What is the gas pressure? (CH 18)
18. | A scientist whose scale is broken but who has a working 2.5 kW heating coil and a thermometer decides to improvise to determine the mass of a block of aluminum she has recently acquired. She heats the aluminum for 30 s and finds that its temperature increases from 20°C to 35°C . What is the mass of the aluminum? (CH 19)
22. | A 750 g aluminum pan is removed from the stove and plunged into a sink filled with 10.0 L of water at 20.0°C . The water temperature quickly rises to 24.0°C . What was the initial temperature of the pan in $^{\circ}\text{C}$ and in $^{\circ}\text{F}$? (CH 19)
14. | A gas consists of a mixture of neon and argon. The rms speed of the neon atoms is 400 m/s. What is the rms speed of the argon atoms? (CH 20)
3. | A heat engine extracts 55 kJ of heat from the hot reservoir each cycle and exhausts 40 kJ of heat. What are (a) the thermal efficiency and (b) the work done per cycle? (CH 21)

APPENDIX B: The sample of questions by difficulty levels (II and III levels) for chapter “Thermodynamics”

II - LEVEL QUESTIONS

9. || 1.0 mol of gold is shaped into a sphere. What is the sphere's diameter? (CH 18)
30. || A gas cylinder holds 0.10 mol of O₂ at 150°C and a pressure of 3.0 atm. The gas expands adiabatically until the volume is doubled. What are the final (a) pressure and (b) temperature? (CH 19)
36. || What maximum power can be radiated by a 10-cm-diameter solid lead sphere? Assume an emissivity of 1. (CH 19)
30. || What is the thermal energy of 100 cm³ of aluminum at 100°C? (CH 20)
17. || A heat engine uses a diatomic gas in a Brayton cycle. What is the engine's thermal efficiency if the gas volume is halved during the adiabatic compression? (CH 21)

III - LEVEL QUESTIONS

24. ||| A 20-cm-diameter cylinder that is 40 cm long contains 50 g of oxygen gas at 20°C. a. How many moles of oxygen are in the cylinder? b. How many oxygen molecules are in the cylinder? c. What is the number density of the oxygen? d. What is the reading of a pressure gauge attached to the tank? (CH 18)
46. ||| A 6.0-cm-diameter, 10-cm-long cylinder contains 100 mg of oxygen (O₂) at a pressure less than 1 atm. The cap on one end of the cylinder is held in place only by the pressure of the air. One day when the atmospheric pressure is 100 kPa, it takes a 184 N force to pull the cap off. What is the temperature of the gas? (CH 18)
25. ||| A 65 cm³ block of iron is removed from an 800°C furnace and immediately dropped into 200 mL of 20°C water. What fraction of the water boils away? (CH 19)
35. ||| 4.0 mol of monatomic gas A interacts with 3.0 mol of monatomic gas B. Gas A initially has 9000 J of thermal energy, but in the process of coming to thermal equilibrium it transfers 1000 J of heat energy to gas B. How much thermal energy did gas B have initially? (CH 20)

APPENDIX C: The sample of my prompt to ChatGPT

solve the following physics problem.

Of course! Please provide the details of the physics problem, and I'll help you solve it.

A ball of mass m and another ball of mass $3m$ are placed inside a smooth metal tube with a massless spring compressed between them. When the spring is released, the heavier ball flies out of one end of the tube with speed v_0 . With what speed does the lighter ball emerge from the other end?

To solve this problem, we can use the principle of conservation of momentum.

Given:

- The mass of the lighter ball is m .
- The mass of the heavier ball is $3m$.
- The heavier ball emerges with speed v_0 .

Since there are no external forces acting on the system (the tube is smooth), the total momentum before and after the spring is released must be conserved.

Step 1: Initial Momentum

Initially, both balls and the spring are at rest, so the total initial momentum is:

$$p_{\text{initial}} = 0$$

APPENDIX D: The sample of the response of ChatGPT

Step 2: Final Momentum

Let v be the speed of the lighter ball after the spring is released.

The total final momentum of the system is:

$$p_{\text{final}} = (3m)(v_0) + (m)(-v)$$

(Note: The velocity of the lighter ball is taken as negative because it moves in the opposite direction of the heavier ball.)

Step 3: Conservation of Momentum

According to the conservation of momentum:

$$p_{\text{initial}} = p_{\text{final}}$$

Substitute the known values:

$$0 = (3m)(v_0) - (m)(v)$$

Step 4: Solve for v

Rearrange the equation to isolate v :

$$(m)(v) = (3m)(v_0)$$

Divide both sides by m :

$$v = 3v_0$$

Final Answer

The speed at which the lighter ball emerges from the other end of the tube is $3v_0$.