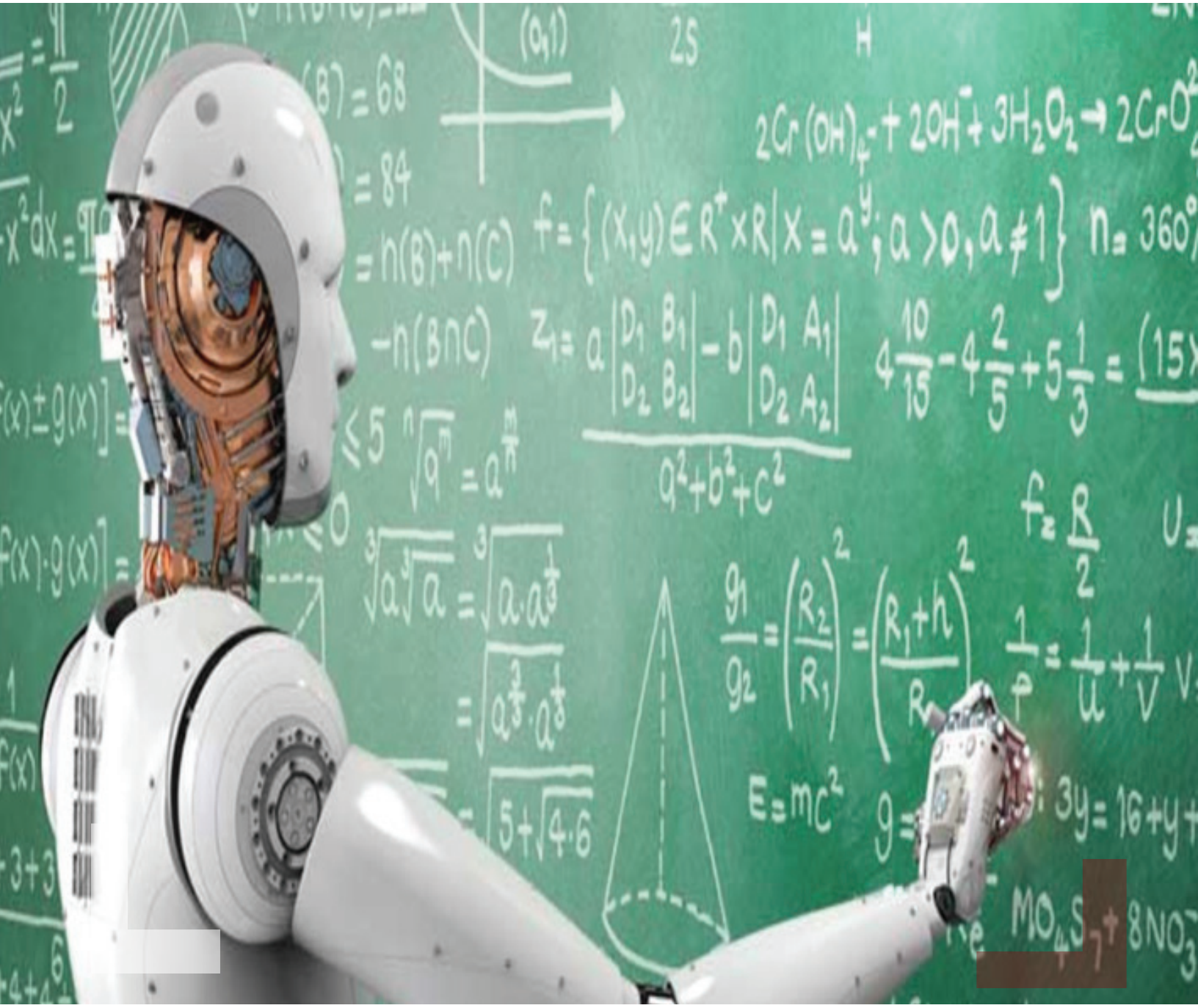


# ARTIFICIAL INTELLIGENCE IN PROFESSIONAL ACCOUNTING EXAMINATION ASSESSMENT: A COMPARATIVE ANALYSIS OF TRADITIONAL AND AI-BASED GRADING METHODS

## ICAG AND WACAR RESEARCH REPORT

# ABOUT **ICAG**

ICAG (Institute of Chartered Accountants, Ghana), established in 1963, is the premier national organization dedicated to advancing the accountancy profession and serving the public interest in Ghana. With over 10,000 members and 16,000 aspiring professionals, ICAG represents a vibrant community of accounting and finance experts committed to the highest standards of integrity, professionalism, and excellence.

ICAG equips professionals across Ghana and the Sub-Region for rewarding careers in accountancy, finance, and management. Through our top-tier educational programs and professional development initiatives, we cultivate our members' financial expertise, business acumen, and digital skills, preparing them to thrive in a dynamic global environment.

Our members, employed across diverse industries, drive economic growth and social progress. ICAG firmly believes that the accountancy profession is a pillar of society, fostering the growth and prosperity of Ghana's economy, businesses, and citizens. By upholding robust financial management practices, combating fraud, promoting ethical leadership, and championing sustainable development, our members lead positive transformation.

ICAG drives accountancy innovation through rigorous research and thought leadership. Our studies address current challenges and anticipate trends, maintaining our position at the forefront of the field. This research-driven, non-profit approach allows us to focus on long-term sector needs, making ICAG a key catalyst for evidence-based progress in Ghana's financial landscape and beyond.

Find out more at: https://www.icagh.org/

# ABOUT THE **WACAR**

The West African Centre for Accountancy Research (WACAR), established in Ghana in April 2023 by the Institute of Chartered Accountants, Ghana (ICAG), stands as the premier hub for financial research in West Africa. WACAR's mission is to revolutionize public financial management, governance, and accountability across the region through rigorous, data-driven inquiry.

WACAR's multidisciplinary team employs advanced methodologies to produce authoritative, evidence-based recommendations in financial reporting, auditing, governance, and tax policy. These high-caliber insights directly inform policy decisions, enhance standards, and foster sustainable economic growth, tailored to West Africa's unique socioeconomic landscape.

Guided by principles of integrity, collaboration, innovation, and measurable impact, WACAR stands at the forefront of accounting research. The Centre's commitment to academic excellence and practical application positions it as an emerging global thought leader, driving transformative change in financial governance.

WACAR's outputs are poised to make significant, quantifiable contributions to West African financial ecosystems. By addressing critical challenges, WACAR's work promises to strengthen institutional frameworks, enhance transparency, and ultimately improve economic outcomes for millions across the region.

# ARTIFICIAL INTELLIGENCE IN PROFESSIONAL ACCOUNTING EXAMINATION ASSESSMENT: A COMPARATIVE ANALYSIS OF TRADITIONAL AND AI-BASED GRADING METHODS

## ICAG AND WACAR RESEARCH REPORT

Osei Adjaye-Gyamfi, FCA (Director, Technical and Research, ICAG and WACAR)

Samuel Koranteng Fianko, P.h.D (Research Manager, ICAG and WACAR)

Frederick Agropah (Research officer, ICAG and WACAR)

## ACKNOWLEDGEMENT

The authors of this report would like to express our profound gratitude to the Members of the Examination Department for their support in this study.

# EXECUTIVE **SUMMARY**

### Study Overview

This comprehensive study investigated the effectiveness of four leading AI models (GPT-4, Claude 3.5, Perplexity, and DeepSeek) in assessing professional accounting examination scripts compared to human markers, both with and without standardized marking schemes across nine accounting subjects.

### Research Questions

1. What is the effectiveness of AI models In grading Ghana's professional accounting examination responses?
2. How do AI models" grading compare with human examiners" assessments with and without marking schemes?
3. What are the implications of Implementing AI-assisted grading for professional accounting examinations In Ghana?

### AI Models Evaluated

- C Claude 3.5
- D DeepSeek
- G GPT-4
- P Perplexity

### Methodology

1. Sample Selection 27 scripts from 9 subjects
2. AI Assessment Without Marking Schemes
3. AI Assessment With Marking Schemes
4. Comparison with Human Markers (Benchmark)
5. Statistical Analysis (p=0.014)

## Key Findings

### AI vs Human Performance Without Marking Schemes

- Most AI models scored more generously than humans
- Claude showed strongest consistency with humans
- GPT-4 and Perplexity typically overscored
- DeepSeek showed most inconsistent patterns, with dramatic over/underscoring in some areas.

### Impact of Marking Schemes on AI-Human Alignment

- Claude demonstrated significant improvement in alignment with human benchmarks (p=0.014)
- GPT-4 generally maintained or increased its deviation from human standards
- Perplexity and DeepSeek showed inconsistent alignment patterns with marking schemes

## AI Model Performance Comparison

| AI Model | Alignment with Human Markers | Overall Rating |
|---|---|---|
| Claude 3.5 | Excellent alignment, especially with marking schemes (p=0.014) | Excellent (90%) |
| Other AI Models | Variable alignment, often overscoring and inconsistent across subjects | Fair (45%) |

## Subject-Specific AI-Human Comparison

| Subject | Key Findings |
|---|---|
| **Audit & Assurance** | Claude maintained exceptional consistency with human standards (57, 65, 62) across all three categories, even without marking schemes. GPT-4 showed reasonable alignment in some cases but significantly overscored in others (79). DeepSeek showed considerable deviation from human standards, particularly underscoring In Dummy 1 (45) while dramatically overscoring in Dummy 2 (89). |
| **Financial Accounting** | Claude achieved Impressive alignment with human benchmarks, coming within a single point in dummies 1 and 2 when using marking schemes. DeepSeek shifted from overscoring to underscoring when given marking schemes. GPT-4 Increased overscoring with marking schemes across all dummies. Perplexity also moved toward better human alignment but underscored in dummy 3. |
| **Business & Corporate Law** | Claude demonstrated remarkable convergence with human assessment patterns when given marking schemes, achieving near-perfect alignment in dummy 3. All AI models gave higher scores than human graders without marking schemes. GPT-4 showed Improved adjustment but continued to award higher marks. DeepSeek and Perplexity maintained considerable overscoring with marking schemes. |
| **Business Management & Info Systems** | Claude showed significant improvement with marking schemes, addressing its previous underscoring tendency in dummy 3 and achieving near-perfect alignment with human benchmarks. In dummy 2. Perplexity consistently demonstrated the highest overall scoring across all categories, substandally exceeding human assessment benchmarks. GPT-4 became more generous with structured criteria. |
| **Corporate Reporting** | Claude's scoring pattern underwent a remarkable reversal with marking schemes, shifting from consistently higher scores to slightly more conservative assessment than human benchmarks. DeepSeek similarly adopted a more conservative approach with marking schemes. GPT-4 became even more generous, while Perplexity reduced its previously excessive scoring with better human alignment. |
| **Financial Reporting** | Claude moved to better alignment with human benchmarks with marking schemes across all dummies, only slightly overscoring in dummy1 and maintaining close alignment in dummies 2 and 3. GPT-4 Increased overscoring with schemes in dummies 1 and 3. DeepSeek maintained its underscoring in dummy1 but reduced overscoring in dummy. |

# Conclusions, Implications & Recommendations

## Conclusions

### 1. AI Models Show Strong Assessment Potential

**AI models possess substantial potential for supporting assessment in professional accounting education, with Claude consistently exhibiting the strongest alignment with human assessment standards across most subject areas.**

### 2. Marking Schemes Significantly Impact AI Performance

The Introduction of marking schemes had divergent effects across different AI models. Claude typically showed Improved alignment with human benchmarks when given marking schemes, while other models like GPT-4 often increased their deviation.

### 3. Subject Complexity Not Necessarily a Limitation

Contrary to initial expectations, in subjects requiring Integrated judgment such as Strategic Case Study, certain AI models demonstrated effective assessment capabilities.

## Implications

### 1. Potential for Enhanced Assessment Efficiency

AI Integration in professional accounting assessment could improve efficiency while maintaining quality, especially when using specific models (Claude) that show consistent alignment with human judgment.

### 2. Importance of Well-Designed Marking Schemes

The significant Improvement in alignment when using marking schemes highlights the need for carefully structured assessment criteria to guide AI evaluation processes toward human assessment standards.

### 3. Need for Subject-Specific Approaches

Variations in AI-human alignment across accounting disciplines suggest the need for tailored approaches that recognize and accommodate subject-specific assessment.

## Recommendations

### 1. Pilot AI-Assisted Assessment

Accounting professional bodies should consider piloting the use of certain AI models (particularly Claude) as supplementary assessment tools alongside human markers, especially in subjects where AI-human alignment was strongest.

### 2. Implement Well-Designed Marking Schemes

Any Implementation of AI in professional accounting assessment should include well-designed marking schemes, as these significantly improve alignment with human standards for most AI models.

### 3. Develop Hybrid Assessment Approach

A hybrid assessment approach combining AI pre-marking with human verification may be optimal, allowing for efficiency Improvements while maintaining assessment quality and professional judgment.

**4. Subject-Specific AI Assessment Protocols**

Professional bodies should develop subject-specific AI assessment protocols based on observed alignment patterns with human examiners, recognizing that AI performance varies considerably across different accounting disciplines.

**5. Ongoing Monitoring & Evaluation**

Establish ongoing monitoring of AI assessment quality relative to human standards to ensure continued alignment with professional standards.

**6. Examiner Training on AI Prompt Engineering**

Accounting educators and examiners should receive specialized training on effective AI prompt engineering and marking scheme design to maximize alignment between AI and human assessment standards. This should include practical workshops on crafting clear, unambiguous Instructions that guide AI models toward accurate assessment outcomes.

**7. Optimize Marking Schemes for AI Interpretation**

Further research should explore how marking schemes can be optimized specifically for AI assessment, potentially through Iterative design and callbration processes. These optimized schemes should maintain educational Integrity while providing sufficient structure for AI systems to accurately interpret assessment criteria

**8. Phased Implementation Strategy**

Professional bodies should adopt a phased Implementation approach, beginning with subjects where AI-human alignment was strongest (such as Audit & Assurance and Financial Accounting). This allows for controlled Integration, ongoing quality assurance, and Iterative improvement before expanding to more complex subject areas.

# STATEMENT FROM
# OUR PRESIDENT

Dear Esteemed Members and Stakeholders,

As President of the Institute of Chartered Accountants Ghana, I am sharing insights from our study on artificial intelligence models in the assessment of professional accounting examination scripts. Our research reveals interesting possibilities that merit careful consideration.

The study indicates the potential for artificial intelligence to support our assessment processes in the future. Claude 3.5 showed promising alignment with human markers across several subject areas, including more complex assessments like Strategic Case Studies.

Key observations include improved artificial intelligence-human alignment when using marking schemes, with notable consistency in subjects like Business and corporate Law and Financial Accounting. We also observed varying performance across different subjects, suggesting any potential implementation would require tailored approaches.

The Institute remains committed to examination integrity as we explore this technology, looking forward to potentially integrating artificial intelligence in our assessment processes through quality-controlled hybrid models that combine technological capabilities with human expertise.

Rest assured, any potential integration of artificial intelligence would be designed to enhance efficiency while maintaining the rigorous standards that define our profession. These early findings simply open the door to thoughtful exploration of innovative possibilities.

**Mr.Augustine Addo**
**President, ICAG**

# STATEMENT FROM OUR CHIEF EXECUTIVE OFFICER

I would like to share the Institute of Chartered Accountants Ghana's exploratory approach to understanding artificial intelligence-assisted assessment based on our recent study comparing artificial intelligence models with human markers. This research provides valuable insights for potential future consideration.

Our findings suggest possibilities for artificial intelligence to complement our assessment processes, particularly when utilizing structured marking schemes. We have also identified areas requiring further research, especially regarding variations in artificial intelligence performance across different subject areas.

Should the Institute of Chartered Accountants Ghana consider this technology in the future, we would develop a comprehensive framework addressing these considerations. This might include gradual exploration beginning with subjects showing the strongest alignment, such as Audit and assurance and Financial Accounting, robust quality verification processes, and appropriate training for examiners.

We recognize that any significant change in assessment methodology requires extensive research and validation. Our commitment remains to both the exploration of innovation and the preservation of examination integrity.

The study suggests the potential for a hybrid approach that could one day leverage complementary strengths of artificial intelligence systems alongside human expertise, possibly enhancing assessment efficiency while maintaining professional judgment and evaluation quality.

We look forward to continuing this exploration with all stakeholders to ensure any potential digital enhancements would strengthen the Institute of Chartered Accountants Ghana's position as a leading professional accounting body.

**P. Kwasi Agyemang, FCA**
**CEO, ICAG**

# CONTENT


EXECUTIVE SUMMARY ..... iv
STATEMENT FROM OUR COUNCIL PRESIDENT ..... ix
STATEMENT FROM OUR CHIEF EXECUTIVE OFFICER ..... x
LIST OF TABLES ..... xiii
LIST OF FIGURES ..... xiii
ABBREVIATIONS AND ACRONYMS ..... xiv

1.0 Introduction ..... 1
1.1 Background of the Study ..... 2
1.2 Research Questions ..... 4
1.3 Significance of the Study ..... 4

2.0 Methodology ..... 5
2.1 Research Design ..... 6
2.2 Research Phases ..... 7
2.3 Sample Selection ..... 7
2.4 Data Collection ..... 7
2.4.1 AI Assessment Process ..... 7
2.4.2 Human Assessment Process ..... 7
2.7 Data Analysis ..... 7

3.0 Findings ..... 8
3.1 Introduction ..... 9
3.2 Comparative Analysis of Human and AI Assessment Models in Auditing and Assurance Scripts ..... 9
3.3 Comparative Analysis of Human and AI Assessment Models in Advance Auditing and Assurance Scripts ..... 10
3.4 Comparative Analysis of Human and AI Assessment Models in Business and Corporate Law Scripts ..... 11
3.5 Comparative Analysis of Human and AI Assessment Models in Business Management & Info Systems Scripts ..... 12
3.6 Comparative Analysis of Human and AI Assessment Models in Corporate Reporting Scripts ..... 13
3.7 Comparative Analysis of Human and AI Assessment Models in Financial Accounting Scripts ..... 14
3.8 Comparative Analysis of Human and AI Assessment Models in Financial Reporting Scripts ..... 15

3.9 Comparative Analysis of Human and AI Assessment Models in Financial Reporting Scripts ..... 15
3.10 Comparative Analysis of Human and AI Assessment Models in Strategic Case Study Scripts ..... 16
3.11 Statistical Analysis of AI Assessment Performance..... 17

4.0 Discussion of Results ..... 18
4.1 AI Adaptability to Professional Standards ..... 19
4.2 Challenging Assumptions About Technical vs. Judgment-Based Assessment .... 19
4.3The Role of Structured Assessment Criteria ..... 19
4.4 Differences in AI Model Behaviour ..... 19
4.5 Implications for Assessment Methodology ..... 20

5.0 Conclusion ..... 21

6.0 Recommendations ..... 23

7.0 References ..... 25

# TABLES

Table1: Selected Subjects at each Level ..... 6

# FIGURES

Figure 1a: Audit & Assurance marking results without marking scheme ..... 10
Figure 1b: Audit & Assurance marking results with marking scheme..... 10
Figure 2a: Advance Audit & Assurance marking results without marking scheme ..... 11
Figure 2b: Advance Audit & Assurance marking results with marking scheme..... 11
Figure 3a: Business & Corporate Law marking results without marking scheme..... 12
Figure 3b: Business & Corporate Law marking results with marking scheme..... 12
Figure 4a: Business Management & Info Systems marking results without marking scheme... 13
Figure 4b: Business Management & Info Systems marking results with marking scheme..... 13
Figure 5a: Corporate Reporting marking results without marking scheme..... 14
Figure 5b: Corporate Reporting marking results with marking scheme..... 14
Figure 6a: Financial Accounting marking results without marking scheme..... 14
Figure 6b: Financial Accounting marking results with marking scheme..... 14
Figure 7a: Financial Reporting marking results without marking scheme..... 15
Figure 7b: Financial Reporting marking results with marking scheme..... 15
Figure 8a: Principles of Taxation marking results without marking scheme..... 16
Figure 8b: Principles of Taxation results with marking scheme..... 16
Figure 8a: Strategic Case Study marking results without marking scheme..... 17
Figure 8b: Strategic Case Study results with marking scheme ..... 17

# ABBREVIATIONS AND ACRONYMS

AI - Artificial Intelligence

ICAG - Institute of Chartered Accountants, Ghana

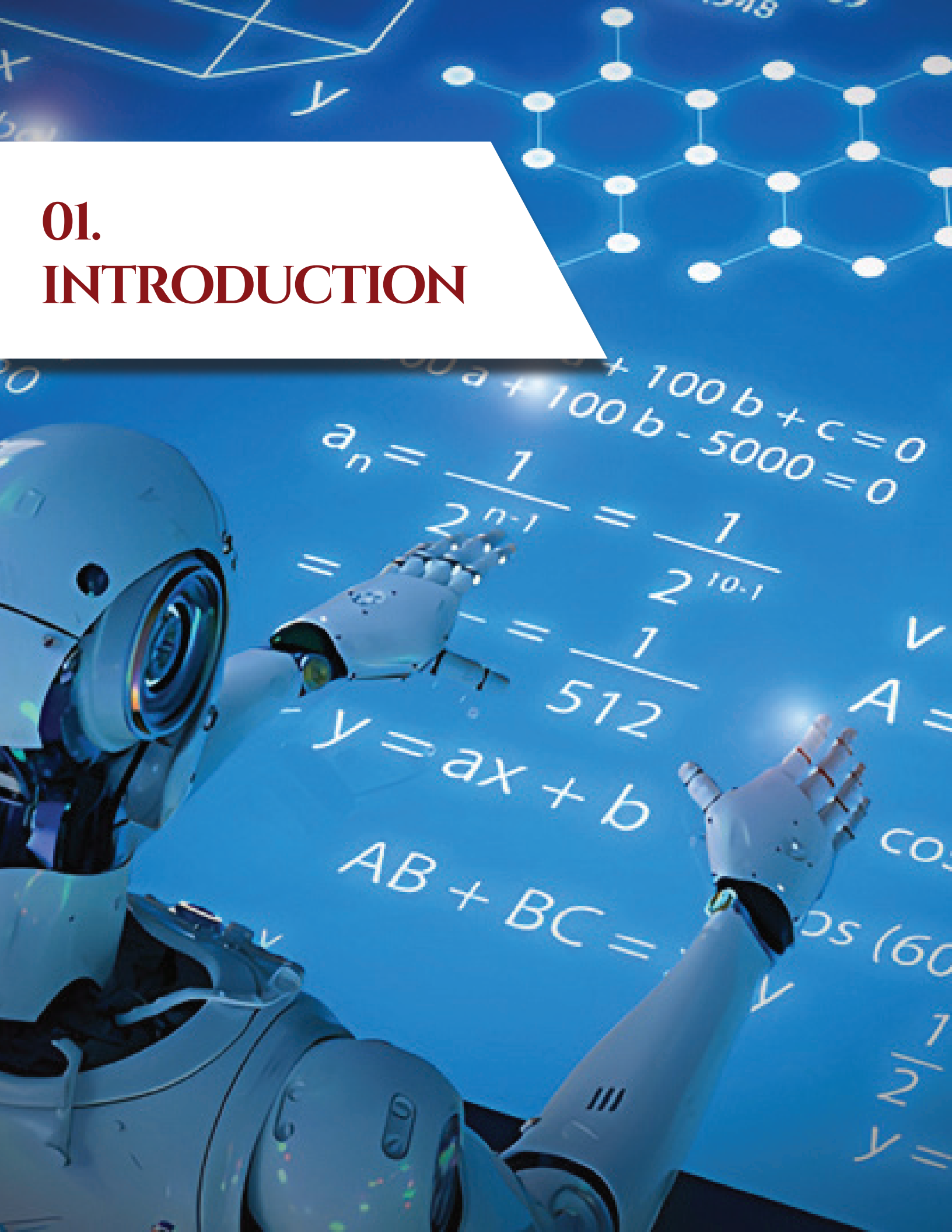
GPT-4 - Generative Pre-trained Transformer 4

INFO SYST. - Information Systems

NMS - No Marking Scheme

WMS - With Marking Scheme

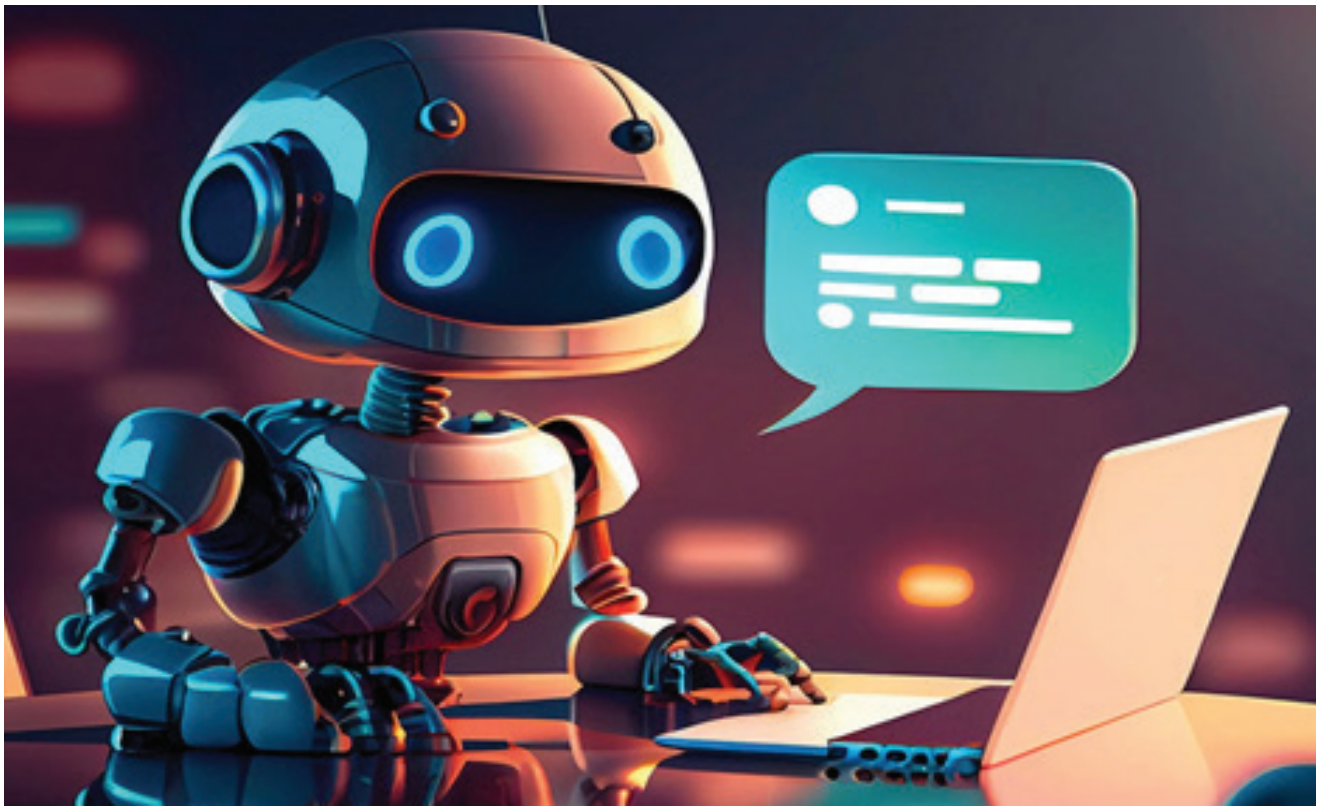
# 01.
# INTRODUCTION



## 1.1 Background of the study

The landscape of education and professional assessment is undergoing a profound transformation, driven by rapid advancements in artificial intelligence (AI) technologies. Over the past few years, AI has made significant inroads into various sectors, with education and assessment emerging as areas of particular interest and potential.[1] The accounting profession stands at a transformative crossroads as sophisticated AI models revolutionize traditional assessment paradigms in professional education.[2] Recent advancements in AI technology, particularly the emergence of sophisticated language models such as AI Chatbots demonstrate unprecedented capabilities in comprehending and analyzing complex accounting scenarios. [3]

Traditional professional accounting examinations, administered by prestigious bodies such as the Institute of Chartered Accountants, Ghana (ICAG), have relied extensively on human markers to evaluate candidates' responses across multiple dimensions: technical proficiency, analytical capabilities, professional judgment, and ethical reasoning competencies. However, this conventional marking process faces several inherent limitations. Specifically, research has identified multiple challenges, including marker fatigue, inconsistency across different markers, scalability constraints, and unconscious biases. [4] Addressing these assessment hurdles, contemporary AI models such as GPT-4o, Claude 3.5, Perplexity, and DeepSeek offer potential solutions through their ability to analyze complex written responses and provide systematic feedback. Beyond this analytical capability, these systems can process extensive examination volumes simultaneously, potentially delivering more consistent and scalable assessment solutions while maintaining rigorous standards.[5]



However, significant empirical gaps persist regarding AI's efficacy in professional accounting examinations. Specifically, in the context of AI's role in professional accounting assessment, three significant research gaps emerge. First, a critical gap exists in understanding the comparative performance of different AI models in assessing complex accounting scenarios, particularly their effectiveness with and without structured marking schemes. Second, another substantial gap involves understanding AI systems' ability to match or exceed human expertise in evaluating nuanced aspects of professional competency. Finally, the third gap concerns our limited understanding of the implications of introducing AI into high-stakes professional certification processes, particularly regarding the reliability and fairness of AI-based assessment systems in accounting education.

These gaps become increasingly critical as professional accounting bodies worldwide contemplate AI integration. While human markers rely on detailed marking schemes, AI models might leverage their pre-trained knowledge of accounting principles alongside these structured assessment frameworks. However, the reliability of AI-based assessments with and without explicit marking schemes requires thorough investigation. The complexity of professional accounting examinations, which test technical knowledge, professional skepticism, and ethical judgment, combined with increasing globalization and demand for qualifications, creates a compelling case for investigating AI-based systems as complementary or alternative evaluation tools.


Therefore, this study aims to investigate the effectiveness and reliability of AI Chatbots (specifically GPT-4, Claude 3.5, Perplexity, and DeepSeek) in grading professional accounting examination scripts in Ghana. Through this research, we seek to understand both the potential and limitations of AI in professional accounting assessment, ultimately contributing to the development of more efficient, equitable, and scalable examination processes while maintaining the rigorous standards essential to the accounting profession.

## 1.2 Research Questions

The research addresses the following question:

RQ1: What is the effectiveness of AI models in grading Ghana's professional accounting examination responses?

RQ2: How do AI models' grading results compare with human examiners' assessments when grading without marking schemes?

RQ3: How do AI models' grading results compare with human examiners' assessments when grading with marking schemes?

RQ4: What are the implications of implementing AI-assisted grading for professional accounting examinations in Ghana and similar developing countries?

## 1.3 Significance of the Study

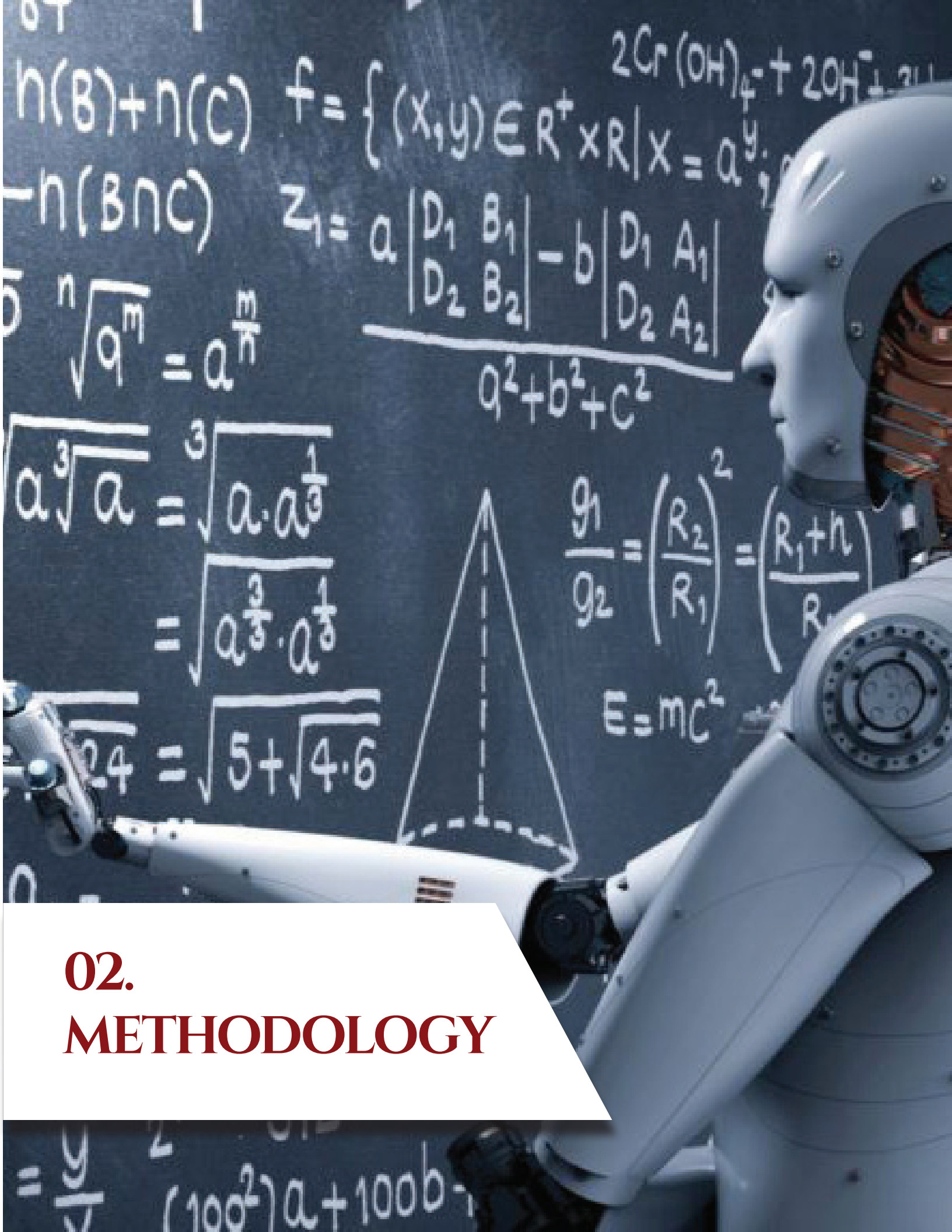
- The significance of this research extends beyond academic interest to offer practical value for the accounting profession in Ghana and similar developing nations. For professional accountancy bodies like ICAG, this study provides critical insights into how AI could potentially transform their examination processes. The findings will help examiners understand AI's capabilities and limitations in assessing complex accounting responses, potentially leading to more efficient and consistent grading practices.

- For practicing accountants and firms, this research illuminates how emerging technologies could impact professional certification, continuing education, and competency assessment. Understanding AI's role in professional examination assessment is crucial as firms increasingly integrate AI tools into their practice.

- From an examination quality perspective, this study offers valuable data on whether AI can maintain the high standards required for professional accounting certification. By comparing AI assessment with human examiner grading - both with and without marking schemes - the research provides insights into the capabilities and limitations of AI in professional accounting examination assessment processes.

- For policymakers in professional accounting bodies, this research will inform decisions about the potential integration of AI in examination assessment processes and the development of AI-enhanced grading tools tailored to accounting examinations. The findings will also guide decisions regarding training requirements for examiners working alongside AI systems and the necessary quality control measures needed when implementing AI in assessment. Additionally, the research will help inform resource allocation decisions for examination administration.

- As Ghana's accounting profession continues to grow and align with international standards, these findings will help guide the strategic implementation of AI technologies in professional certification processes while maintaining the rigour and credibility of accounting qualifications. The insights gained will be particularly valuable for developing countries seeking to leverage technology to enhance their professional certification processes while ensuring the maintenance of high professional standards.



# 02.
# METHODOLOGY

## 2.1 Research Design

This study employed a quantitative research approach to investigate the impact of AI Chatbots on the assessment of professional accounting examination scripts in Ghana. The research design focused on statistical analysis of grading outcomes to examine the effectiveness and reliability of AI-based assessment compared to traditional human marking.

## 2.2 Research Phases

The study was structured in three main phases. The first phase involved AI-based assessment of examination scripts without marking schemes. The second phase consisted of AI-based assessment using marking schemes. The final phase encompassed traditional human assessment using marking schemes.

## 2.3 Sample Selection

The study utilized stratified random sampling to select examination scripts from the Institute of Chartered Accountants Ghana (ICAG) professional examinations conducted during the July 2024 diet. Out of the 14 subjects in the ICAG professional accounting examination, 9 were selected for the study. For each selected subject, 3 scripts were randomly selected, resulting in a total of 27 scripts (3 scripts × 9 subjects). These scripts were labeled Dummy1, Dummy2, and Dummy3 for each subject. The stratified random sampling technique ensured representation across different subject areas and performance levels. Given that each script was assessed by four different AI models (GPT-4, Claude 3.5, Perplexity, and DeepSeek), the study generated 108 AI-based assessments (27 scripts × 4 AI models), plus the original human assessments. This expanded assessment pool provided sufficient data points for robust statistical analysis.

| Level 1 (Foundation) | Level 2 (Application) | Level 3 (Professional) |
|---|---|---|
| Business & Corporate Law | Audit & Assurance | Advanced Audit & Assurance |
| Business MGT & Info Syst. | Principles of Taxation | Strategic Case Study |
| Financial Accounting | Financial Reporting | Corporate Reporting |

## 2.4 Data Collection

The data collection process involved multiple rigorous steps to ensure accuracy and ethical compliance. First, examination scripts were obtained from the Institute of Chartered Accountants Ghana (ICAG) after securing all necessary permissions and addressing ethical considerations, including confidentiality protocols and data protection measures.

These handwritten scripts were then carefully transcribed into digital format by trained research assistants to facilitate AI processing. To ensure transcription accuracy, a validation process was implemented following a comprehensive protocol for qualitative data verification. This included independent verification of transcriptions by a second research assistant and resolution of any discrepancies through consensus.

For the second phase of assessment, official marking schemes for each examination were obtained directly from ICAG with appropriate

authorization. These marking schemes provided the structured assessment framework for both the AI-based evaluation and the original human marking processes. All digital materials were stored securely with restricted access to maintain data integrity throughout the research process.

### 2.4.1 AI Assessment Process

Four AI models were employed: GPT-4, Claude 3.5, Perplexity, and DeepSeek, representing state-of-the-art language models. The assessment was conducted in two stages. In the first stage, each transcribed script was fed into each of the four AI models with prompts to assess and grade each question without reference to any marking scheme. Each dummy script was marked three times by each AI model, and all values were recorded. The average score from these three assessments was calculated for each script to provide a more reliable measure of the AI model's performance and to account for any variability in the AI's evaluations. The AIs evaluated the scripts based on their pre-trained knowledge.

In the second stage, the same process was repeated, but this time, the AI models were provided with the ICAG marking schemes and prompted to assess the scripts according to these official grading criteria. Again, each dummy script was marked three times by each AI model, and all values were recorded and averages calculated. This dual-stage process generated two sets of grades from each AI model for each script: one based on autonomous assessment and another based on marking scheme-guided evaluation.

### 2.4.2 Human Assessment Process

The human assessment process involved utilizing the original marks assigned by ICAG examiners for each script, which were obtained directly from the official examination records. Rather than conducting new human assessments, we used the existing results that were provided by the ICAG examiners for each of the scripts selected for the study. These marks, assigned by experienced ICAG examiners during the July 2024 professional examinations using the standard marking schemes, served as the benchmark for comparison with the AI assessments. The examiners followed ICAG's standard procedures and marking criteria in their evaluation, ensuring a consistent baseline against which to measure the AI performance.

### 2.5 Data Analysiss

To analyze the data, we employed both parametric and non-parametric statistical methods. Paired t-tests were used to examine significant differences between AI assessments and human examiner marks across various conditions, while Wilcoxon signed-rank tests provided complementary analysis without assuming a normal distribution. Descriptive statistics, including means, standard deviations, and score distributions, were calculated to establish baseline performance patterns for both AI and human grading across subjects and examination levels.

### 2.7 Study Limitations

Several limitations were acknowledged in this study. The focus on ICAG examinations might limit generalizability to other accounting certification bodies or countries. The transcription process might have introduced minor discrepancies from the original handwritten scripts. The study captured AI capabilities at a specific point in time, recognizing that these technologies were rapidly evolving. These limitations were considered in the interpretation of results and their potential impact on the study's findings.

# 03.
# FINDINGS

### 3.1 Introduction

This section presents findings from our comparative analysis of AI chatbots and human assessors in marking professional accounting examination scripts. The analysis covers nine subject areas from the Institute of Chartered Accountants Ghana (ICAG) professional examinations, examining how four AI models (GPT-4, Claude, DeepSeek, and Perplexity) perform both with and without marking schemes compared to human benchmarks. Each subject area is analyzed through paired figures (labeled 'a and 'b'), showing assessment patterns across three dummy scripts. Results are organized by subject area, with each analysis comprising two distinct assessment conditions: marking without a standardized marking scheme (NMS) and marking with a structured marking scheme (WMS). This approach evaluates how AI systems interpret accounting knowledge both independently and when guided by formal assessment criteria. The analysis examines performance across foundational subjects like Financial Accounting, specialized areas such as Corporate Reporting and Taxation, and complex assessments like Strategic Case Study. This enables the identification of subject-specific patterns and cross-cutting trends in AI alignment with human standards. Particular attention is paid to consistency relative to human benchmarks and how patterns shift when marking schemes are introduced.

### 3.2 Comparative Analysis of Human and AI Assessment Models in Auditing and Assurance Scripts

Figure 1a presents the results of the marking of Auditing and assurance scripts by human markers and various AI models without a marking scheme. The human benchmark scores (57, 65, 62) serve as the standard for comparison. Claude demonstrates the strongest consistency to these human markers, with its assessment pattern maintaining similar proportions across categories despite slight overscoring in Dummy 2. GPT4 shows reasonable alignment in Dummy 1 but significantly overscores in Dummy 3 (79). DeepSeek exhibits considerable deviation from human standards, particularly underscoring in Dummy 1 (45) while dramatically overscoring in Dummy 2 (89). Perplexity consistently scores lower than human benchmarks across all categories. Additionally, figure 1b presents the results when marking schemes are implemented for the same Auditing and assurance scripts. Claude maintains exceptional consistency with human standards (55, 64, 56) across all three categories. GPT4 substantially overscores in all three Dummies. Both DeepSeek and Perplexity show significant deviation, with pronounced overscoring especially in Dummy 3 (91 and 83 respectively).

These findings demonstrate that marking schemes impact AI assessment alignment differently. Claude uniquely maintains consistent harmony with human judgment regardless of marking conditions, while other AI models diverge further when provided with marking schemes—typically through overscoring. The findings highlight Claude's distinctive potential for academic assessment applications, especially in auditing assurance, where reliable alignment with human judgment is essential.

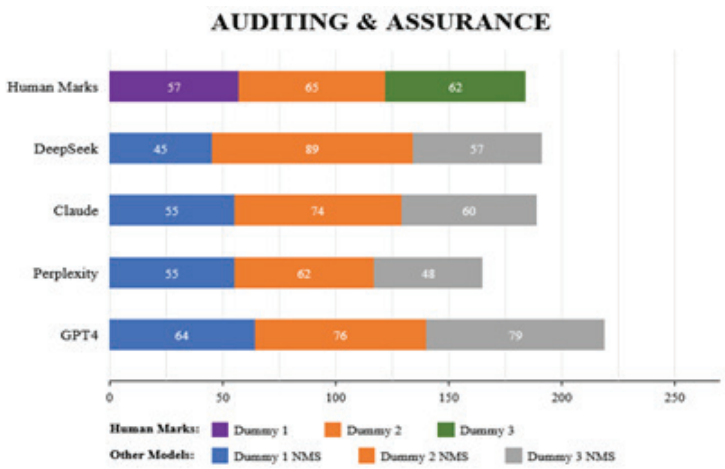



Figure 1a: Audit & Assurance marking results without marking scheme

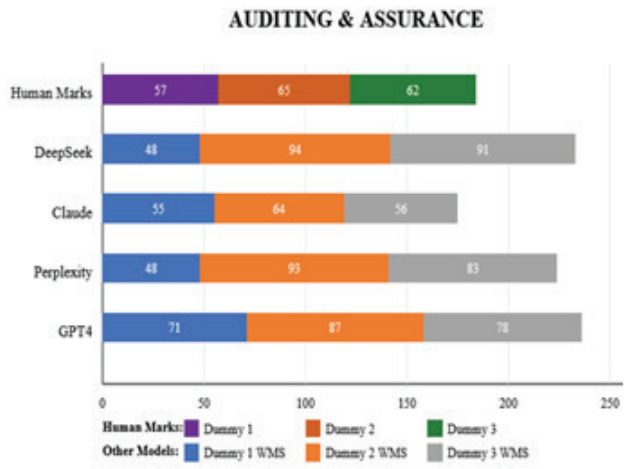



Figure 1b: Audit & Assurance marking results with marking scheme

*Note: NMS = Marking With No Marking Scheme, WMS = Marking With Marking Scheme*

### 3.3 Comparative Analysis of Human and AI Assessment Models in Advance Auditing and Assurance Scripts

Figure 2a displays the assessment results for Advanced Audit & Assurance scripts evaluated by both human markers and a range of AI models without using a marking scheme, establishing human benchmark scores of 55, 86, and 62 as the baseline against which all AI model performances are measured. Most AI models tend to overscore compared to human markers, with Claude and GPT4 showing particularly elevated values in Dummy 1. Perplexity demonstrates significant overscoring across all categories, especially in Dummy 3. DeepSeek shows mixed alignment, slightly underscoring in Dummy 1 but overscoring in Dummy 3. However, the introduction of marking schemes in the marking of Advanced Audit & Assurance scripts reveals shifting assessment patterns amongst the different AI models, as demonstrated in Figure 2b. Claude shows notable improvement in alignment, moving closer to human benchmarks, particularly in Dummy 1 compared to its NMS results. GPT4 shows increased deviation with higher scores across all categories. Perplexity reduces overscoring in Dummies 1 and 2 but increases it in Dummy 3. DeepSeek improves alignment in Dummy 2 but shows increased deviation in Dummy 3.

These findings demonstrate that marking schemes have varied effects across different AI models for Advanced Audit and Assurance. Claude shows improved alignment with human benchmarks when provided with structured criteria, while other models show mixed results or increased deviation. This suggests that even in complex subject areas, certain AI models may benefit from structured marking schemes to better align with human assessment standards.

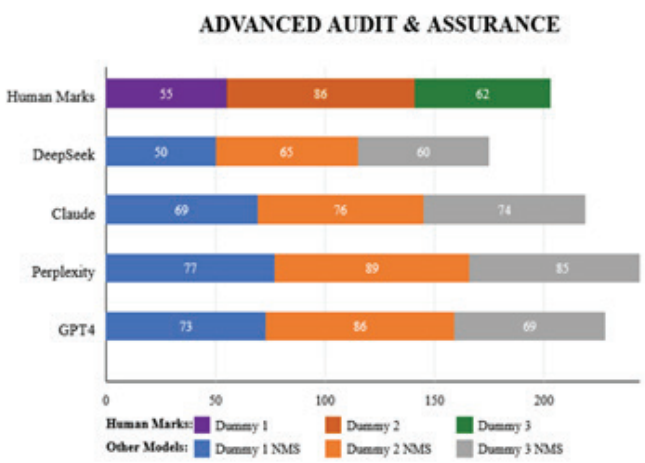
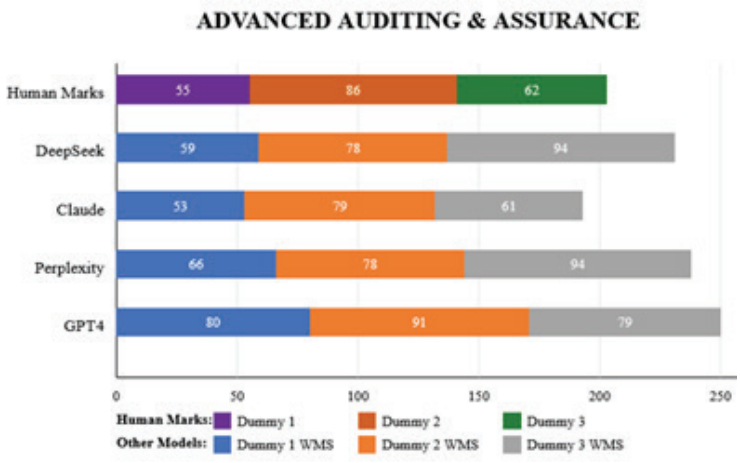



Figure 2a: Advance Audit & Assurance marking results without marking scheme




Advance Figure 2b: Audit & Assurance marking results with marking scheme

### 3.4 Comparative Analysis of Human and AI Assessment Models in Business and Corporate Law Scripts

The results in Figure 3a show the comparative assessment patterns of AI models versus human assessors for Business and corporate Law scripts in the absence of standardized marking criteria. Human assessors established benchmark scores of 37, 50, and 28 across the three dummies. The analysis reveals a consistent pattern of score increase among AI evaluators, with DeepSeek exhibiting the most significant deviation from human standards (, particularly in dummies 1 and 2. All AI models gave higher scores than human graders, showing they tend to be more generous when grading without marking schemes. Figure 3b, on the other hand, illustrates the impact of introducing a standardized marking scheme on assessment outcomes for Business and Corporate Law scripts. Claude demonstrates remarkable convergence with human assessment patterns, achieving near-perfect alignment in dummy 3. GPT4 shows improved adjustment but continues to award higher marks. DeepSeek and Perplexity maintain considerable overscoring in the second and third dummies, though Perplexity shows improvement in the first dummy.

This analysis indicates that in Business and Corporate Law assessment, structured marking schemes substantially enhance alignment for specific AI models—Claude in particular—while yielding less significant improvements for others. Structured criteria appear to help AI systems better interpret the complex requirements of legal assessment, though most models continue to score more generously than human evaluators.

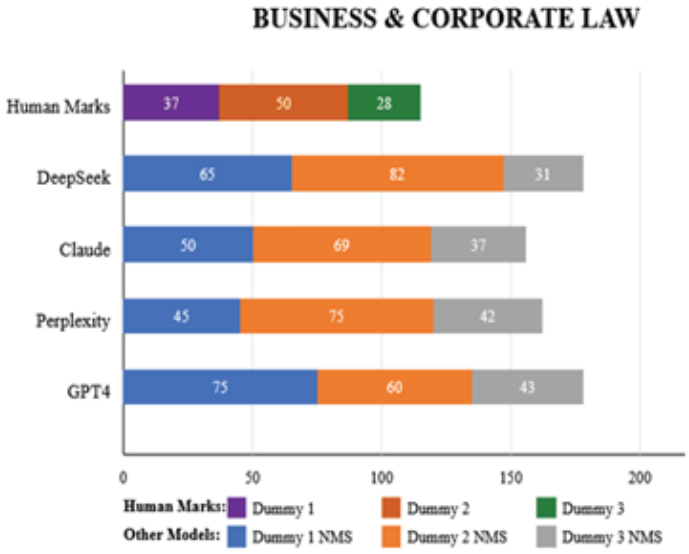
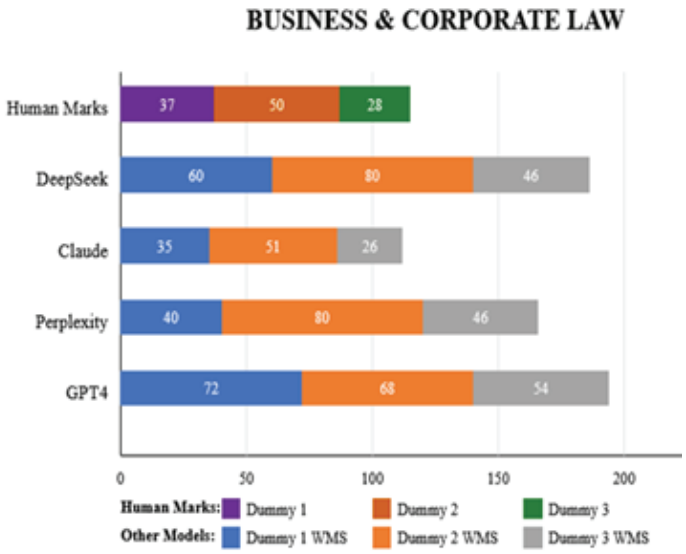



Figure 3a: Business & Corporate Law marking results without marking scheme




Advance Figure 3b: Business & Corporate Law marking results with marking scheme

### 3.5 Comparative Analysis of Human and AI Assessment Models in Business Management & Info Systems Scripts

In evaluating Business Management and Information Systems scripts, Figure 4a reveals diverse assessment patterns among AI models compared to human benchmarks. Perplexity consistently demonstrates the highest overall scoring across all categories, substantially exceeding human assessment benchmarks. DeepSeek and GPT4 scored higher in the first dummy but stayed closer to human scores elsewhere. Claude also scored higher in dummy 1 but significantly lower in dummy 3 compared to humans. However, when marking schemes were introduced (see Figure 4b), Claude's performance showed significant improvement and adaptation. Most impressively, Claude achieved near-perfect alignment with human benchmarks in dummy2 and substantially improved its dummy 3 score, addressing its previous underscoring tendency. GPT4, on the other hand, became more generous with the use of structured criteria.

These results highlight Claude's unique strengths in the assessment of Business Management and Information Systems scripts. Claude shows exceptional responsiveness to the marking scheme guidance and a remarkable ability to align with human judgment standards. While other models increased their deviation from human benchmarks when given marking schemes, Claude demonstrated superior adaptability, suggesting its particular competence for academic assessment applications.

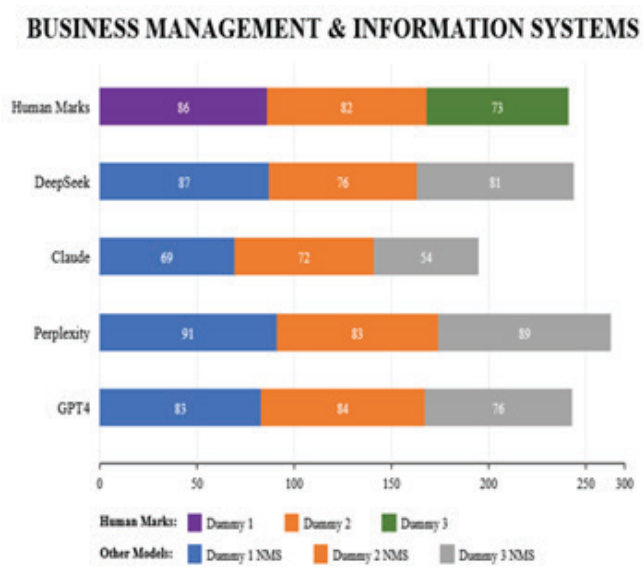



Figure 4a: Business Management & Info Systems marking results without marking scheme

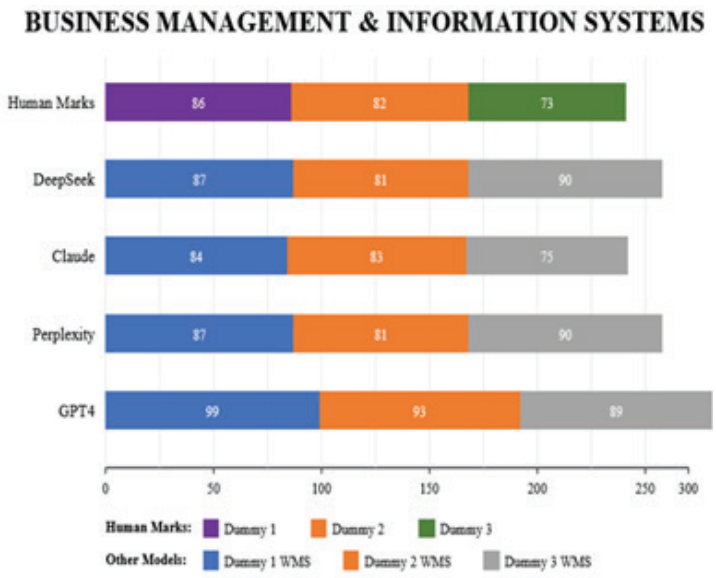



Figure 4b: Business Management & Info Systems marking results with marking scheme

### 3.6 Comparative Analysis of Human and AI Assessment Models in Corporate Reporting Scripts

The Corporate Reporting assessment graphs in Figure 5a reveal striking patterns in AI versus human marking behaviours. All the AI systems exhibit generous scoring tendencies compared to human evaluators. Claude awards higher marks across all categories with particular generosity in the second dummy. Perplexity demonstrates the most liberal scoring approach, especially in the latter two categories. Both GPT4 and DeepSeek allocate substantially higher scores in the third category, with varying patterns elsewhere. However, the use of a marking scheme in marking reveals results in the transformation of the results, as seen in Figure 5b. Claude's scoring pattern undergoes a remarkable reversal, shifting to consistently lower scores than human benchmarks across all dummies. DeepSeek similarly adopts a more conservative approach, which is particularly evident in the second category. Conversely, GPT4 becomes even more generous in the first and second dummies while slightly moderating its dummy 3 scores. Perplexity reduces its previously excessive scoring but maintains better alignment with human standards.

These contrasting responses highlight how differently AI models interpret structured assessment criteria or marking schemes in Corporate Reporting. Claude and DeepSeek appear to apply marking schemes more rigorously than human markers, while GPT4 maintains its generous approach, and Perplexity demonstrates improved alignment with human benchmarks.

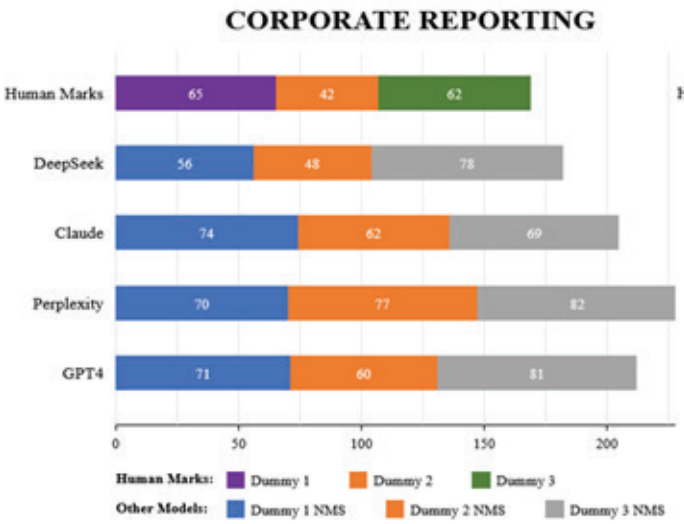



Figure 5a: Corporate Reporting marking results without marking scheme

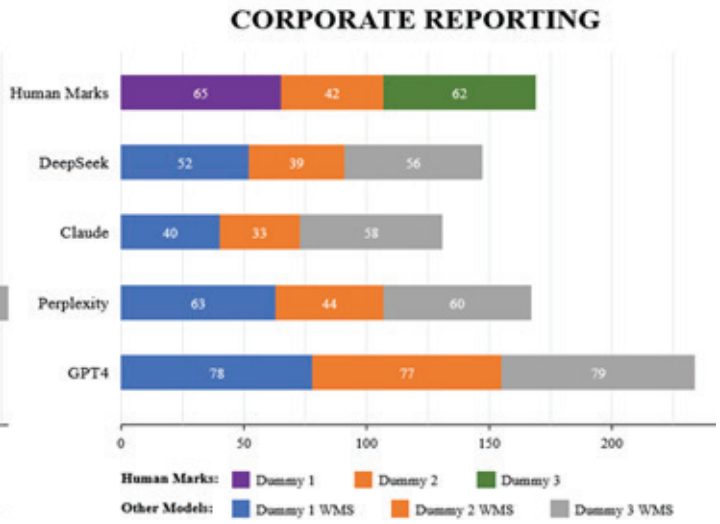



Figure 5b: Corporate Reporting marking results with marking scheme

### 3.7 Comparative Analysis of Human and AI Assessment Models in Financial Accounting Scripts

Comparing the results of the human markers and that of the four different AI marking in which no making scheme were used, figure 6a reveals that most of the AI models demonstrate generous scoring tendencies, with DeepSeek showing dramatic overscoring across all dummies. Claude and GPT4 both award higher marks than humans, particularly in the second dummy. Perplexity strongly overscores in the first two categories but goes below human marks in dummy 3. With the use of the marking scheme, figure 6b demonstrates notable changes. For instance, DeepSeek shifts to giving very low. Claude, on the other hand, achieves impressive alignment with human benchmarks, coming within a single point in dummies 1 and 2. Perplexity also moves toward better human alignment but underscores in the dummy 3. Meanwhile, GPT4 increases its overscoring substantially across all dummies.

Evidence from this results show that the use of marking schemes affects AI models very differently in Financial Accounting. While Claude shows an excellent ability to match human judgment when given a marking scheme, DeepSeek struggles with structured criteria, while GPT4 moves further away from human standards.

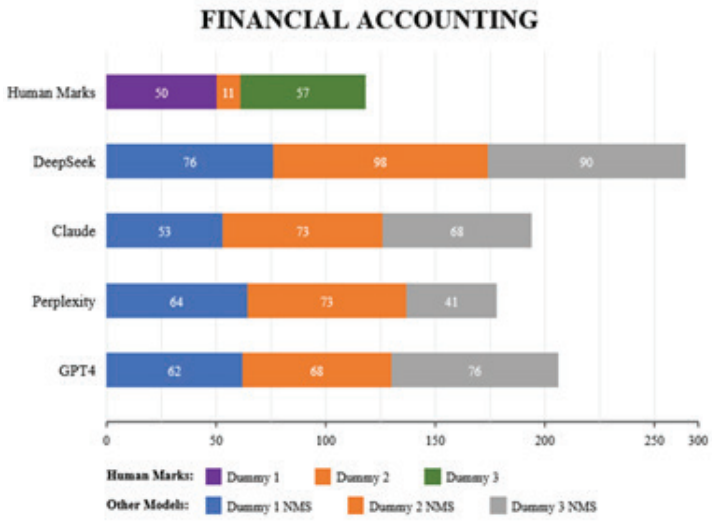



Figure 6a: Financial Accounting marking results without marking scheme

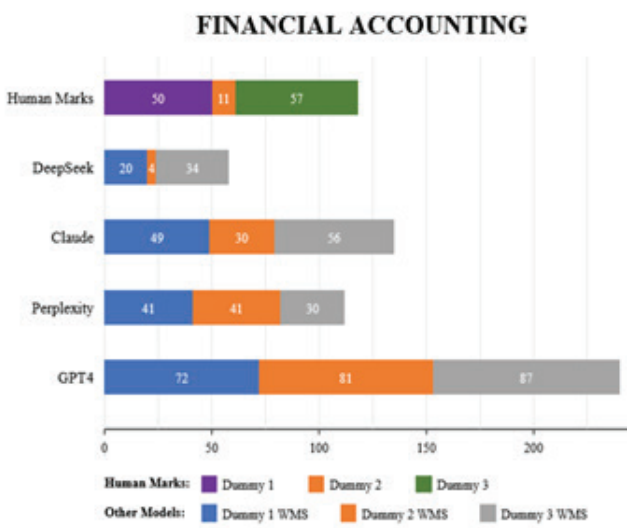



Figure 6b: Financial Accounting marking results with marking scheme

### 3.8 Comparative Analysis of Human and AI Assessment Models in Financial Reporting Scripts

Figure 7a shows that when comparing the marking of financial reporting scripts by AI models without using a marking scheme against human assessors, there are significant variations in assessment patterns. Claude consistently overscores across all categories, particularly in dummy 3. GPT4 shows mixed patterns, underscoring slightly in dummy 1 but substantially overscoring in dummies 2 and 3. DeepSeek also underscores in dummy 1, significantly overscores in dummy 2, and shows close alignment in dummy 3. Perplexity, on the other hand, demonstrates a unique pattern, underscoring in dummies 1 and 2 but overscoring in dummy 3. When the marking scheme were used by the AI models in the marking the financial reporting, Claude moved to better alignment with human benchmarks across all dummies, only slightly overscoring in dummy 1 and maintaining close alignment in dummies 2 and 3. GPT4, however, increases overscoring in dummies 1 and 3 while maintaining high scores in dummy 2. DeepSeek maintains its underscoring in dummy 1 but reduces its overscoring in dummy 2 and shows better alignment in dummy 3. Finally, Perplexity improves its alignment in dummies 1 and 2 while continuing to overscore in dummy 3.

In effect, the marking schemes significantly impact AI assessment behaviour in Financial Reporting, with Claude and DeepSeek showing improved alignment with human standards, while GPT4 becomes more generous and Perplexity demonstrates better consistency with human scoring patterns.

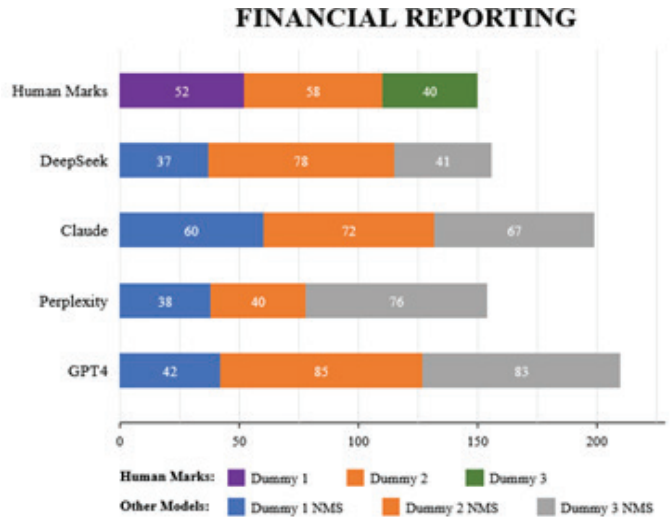



Figure 7a: Financial Reporting marking results without marking scheme

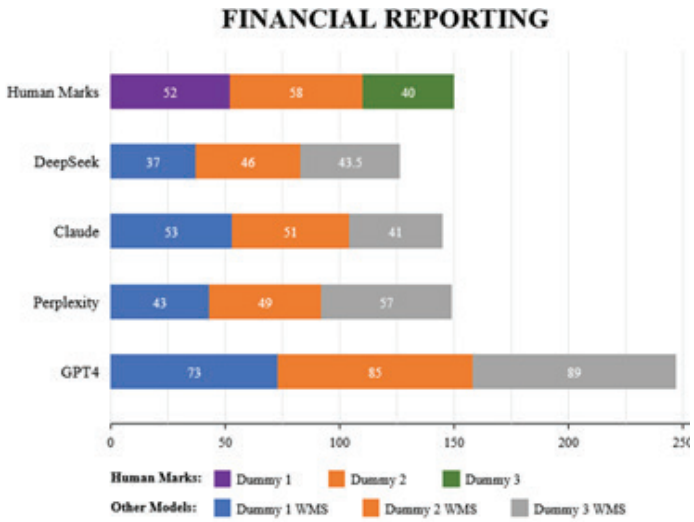



Figure 7b: Financial Reporting marking results with marking scheme

### 3.9 Comparative Analysis of Human and AI Assessment Models in Principels of Taxation Scripts

Results from Figure 8a show that most of the AI models score higher than human benchmarks when they are marked without using marking schemes. GPT4 shows the highest scoring, especially in dummies 1 and 2. Claude scores higher across all dummies. While DeepSeek scores moderately higher, Perplexity shows the closest match with human benchmarks. With the introduction of a marking scheme in the marking, Figure 8b shows that it aligns better with human benchmarks in dummies 1 and 2 but scores lower in dummy 3. Furthermore, while GPT4 reduces its higher scoring but remains above human benchmarks, DeepSeek and Perplexity show mixed results. These results show marking schemes help align AI assessment in taxation topics, with Claude showing the most improvement in matching human standards.

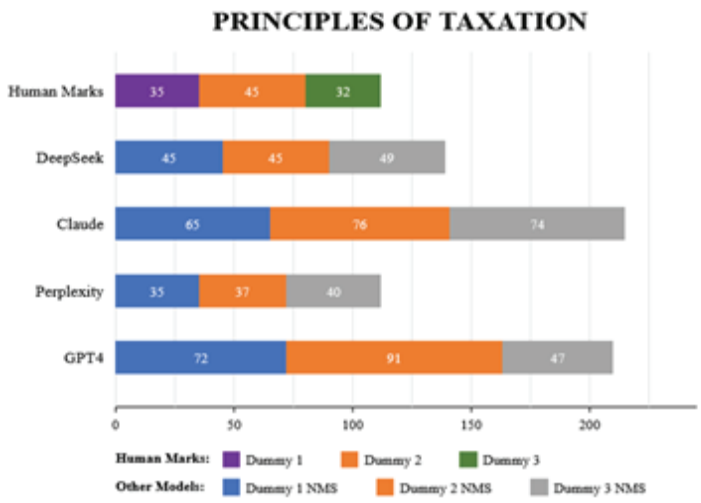



Figure 8a: Principles of Taxation marking results without marking scheme

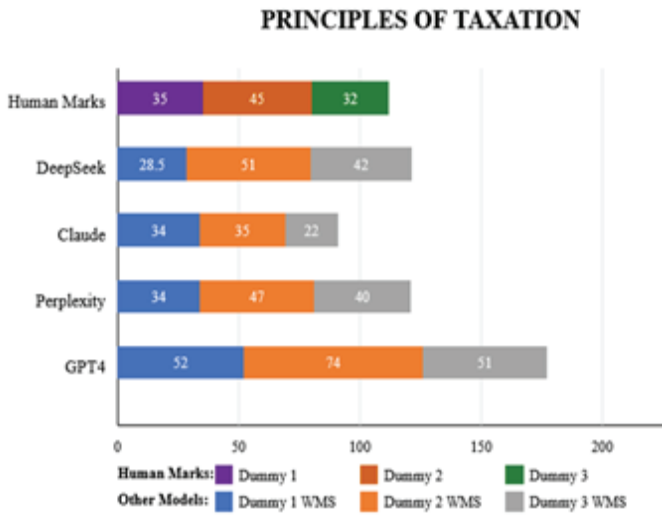



Figure 8b: Principles of Taxation results with marking scheme

### 3.10 Comparative Analysis of Human and AI Assessment Models in Strategic Case Study Scripts

Figure 9a shows the results of the marks of AI models of strategic case studies scripts without the use of a marking scheme and that of the results of human assessors. The results reveal that AI models generally score higher than human benchmarks. While Perplexity shows the most overscoring, especially in dummies, GPT4 consistently overscores across all dummies. Claude shows mixed patterns with both over and underscoring. DeepSeek, however, shows the best alignment, matching humans exactly in dummy 2. With marking schemes by the AI models, Figure 9b shows that Claude demonstrates better alignment with human benchmarks across all dummies, although it still slightly overscores in the dummies 1 and 2 while underscoring in the dummy 3. Furthermore, while GPT4 increases its overscoring tendencies across all dummies, DeepSeek shows increased overscoring in dummies 1 and 2 but improved alignment in dummy 3. Perplexity, however, reduces its overscoring in category 1 but maintains higher scores in categories 2 and 3.

These results indicate that marking schemes have varied effects on AI assessment of Strategic Case Study assignments. While Claude shows improved overall alignment with human standards when provided with marking schemes, the other models show mixed results. This suggests that the complex, integrated nature of case study assessment presents unique challenges for AI evaluation systems, even with structured criteria to guide assessment.

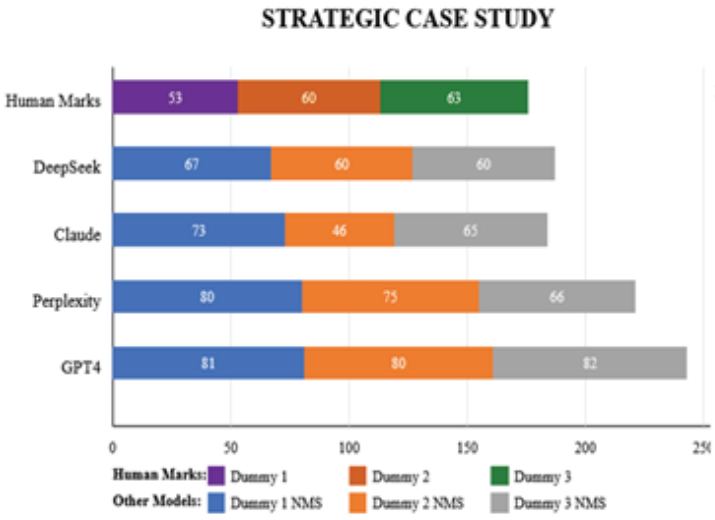



Figure 8a: Strategic Case Study marking results without marking scheme

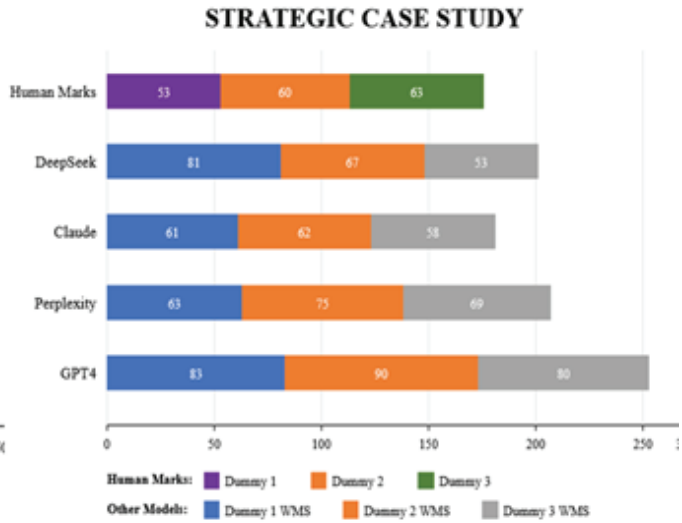



Figure 8b: Strategic Case Study results with marking scheme

### 3.11 Statistical Analysis of AI Assessment Performance

The statistical analysis comparing AI and human assessment with and without marking schemes reveals significant insights into model performance variations. Both parametric (paired t-test) and non-parametric (Wilcoxon signed-rank) tests were employed to examine performance differences across conditions. Claude demonstrates the most statistically significant performance shift between conditions, with a t-test p-value of 0.014512 and a Wilcoxon p-value of 0.01171875. The negative t-statistic (-3.10696) indicates that Claude's performance without marking schemes significantly differs from performance with schemes. This strong statistical evidence confirms the visual analysis that Claude benefits substantially from structured marking criteria, achieving closer alignment with human benchmarks when provided with marking schemes. GPT-4 shows marginally significant differences (t-test p-value = 0.06268, Wilcoxon p-value = 0.12890625) with a positive t-statistic (2.161079), suggesting a tendency toward higher scores when using marking schemes, though not reaching conventional statistical significance thresholds. Both Perplexity and DeepSeek demonstrate no statistically significant differences between conditions (p-values > 0.65), indicating that marking schemes do not consistently alter their assessment patterns across subject areas.

These statistical findings corroborate the subject-specific analyses, highlighting Claude's unique responsiveness to marking schemes and reinforcing its potential value in professional accounting assessment contexts where structured evaluation criteria are employed.

# 04. DISCUSSION OF RESULTS

This study provides novel insights into the application of AI in professional accounting examination assessment, moving beyond the current literature's focus on automated grading in general education contexts. The findings reveal several important patterns with significant implications for assessment in professional accounting education.

### 4.1 AI Adaptability to Professional Standards

The varying degrees of alignment between AI and human markers demonstrate that not all AI systems interpret professional accounting standards similarly. Claude's superior alignment with human judgment across multiple subjects suggests that certain AI architectures may be inherently better suited to professional assessment tasks. This differential performance challenges the notion that AI systems as a category would perform uniformly in specialized professional assessment contexts.

The statistically significant improvement in Claude's alignment when provided with marking schemes ($p=0.014$) indicates that AI systems can effectively adapt to formalized professional standards. This adaptability is particularly noteworthy in a field like accounting where adherence to established standards and frameworks is paramount. The finding extends previous research on AI in education, which has primarily focused on general knowledge domains rather than specialized professional judgment.

### 4.2 Challenging Assumptions About Technical vs. Judgment-Based Assessment

Our findings contradict prevailing assumptions in the literature about AI capabilities in complex assessment tasks. The conventional wisdom that AI would perform better in technical, rule-based subjects than in those requiring professional judgment was not supported by our data. Instead, the provision of clear assessment criteria emerged as the more significant factor in determining AI-human alignment.

The Strategic Case Study results are particularly significant, as they challenge fundamental assumptions about the limitations of AI in evaluating complex, integrated professional scenarios. While prior studies have suggested that AI would struggle with nuanced professional judgment tasks, our findings indicate that certain AI models can effectively assess such material when provided with appropriate guidance. This suggests that the binary distinction often made between "technical" and "judgment-based" assessment may be oversimplified when considering AI applications.

### 4.3 The Role of Structured Assessment Criteria

Perhaps the most significant finding is that the structured nature of assessment criteria appears to be more influential than subject complexity in determining AI-human alignment. This challenges the field's current focus on subject matter as the primary determinant of AI assessment suitability. Instead, our research suggests that the design and implementation of assessment frameworks may be the more critical factor.

This finding has substantial implications for assessment design in professional education. Rather than limiting AI application to supposedly "simpler" technical subjects, efforts might be better directed toward developing more structured assessment frameworks that can guide AI evaluation across all subject areas, including those requiring complex professional judgment.

### 4.4 Differences in AI Model Behaviour

The divergent responses of different AI models to marking schemes suggest fundamental differences in how these systems interpret and apply assessment criteria. While Claude showed improved alignment with human markers when provided with marking schemes, GPT-4's increased deviation indicates that different AI architectures may process structured guidance differently. This finding highlights the importance of model selection in professional assessment applications and suggests that not all state-of-the-art AI systems will be equally suitable for professional accounting assessment.

The inconsistent performance of models like

DeepSeek, which dramatically overscored in some areas while underscoring in others, raises important questions about the stability and reliability of certain AI systems in high-stakes assessment contexts. This variability suggests a need for rigorous validation processes and quality control measures in any implementation of AI-assisted assessment in professional accounting education.

### 4.5 Implications for Assessment Methodology

The study's findings suggest a potential paradigm shift in how assessment methodologies might evolve in professional accounting education. Rather than viewing AI and human assessment as competing approaches, the results point toward complementary strengths that could be integrated into hybrid assessment models. The different patterns of assessment behaviour observed across AI models and human markers suggest opportunities to leverage the consistency and efficiency of AI while maintaining the nuanced professional judgment of human examiners.

The significant variations in AI-human alignment across different subject areas also underscore the need for subject-specific implementation strategies rather than a one-size-fits-all approach to AI integration in professional accounting assessment. This contrasts with some current literature that tends to discuss AI assessment capabilities in more general terms without sufficient attention to disciplinary and subject-specific factors.

Thus this study extends our understanding of AI capabilities in professional assessment contexts beyond the current literature's predominant focus on general education applications. The findings suggest that with appropriate model selection, marking scheme design, and implementation strategies, AI has significant potential to enhance assessment processes in professional accounting education while maintaining the rigorous standards essential to the profession.

# 5.0 CONCLUSION

## 5.1 Conclusion

This study provides comprehensive insights into the capabilities and limitations of AI chatbots in assessing professional accounting examination scripts compared to human markers. Through systematic analysis of nine accounting subject areas across multiple AI models, with and without marking schemes, several significant conclusions emerge.

The findings demonstrate that AI models possess substantial potential for supporting assessment in professional accounting education, though with important qualifications. Claude consistently exhibited the strongest alignment with human assessment standards across most subject areas, particularly when provided with structured marking criteria. This suggests that certain AI models have developed sophisticated capabilities to interpret and apply professional accounting standards in ways that closely mirror human expert judgment.

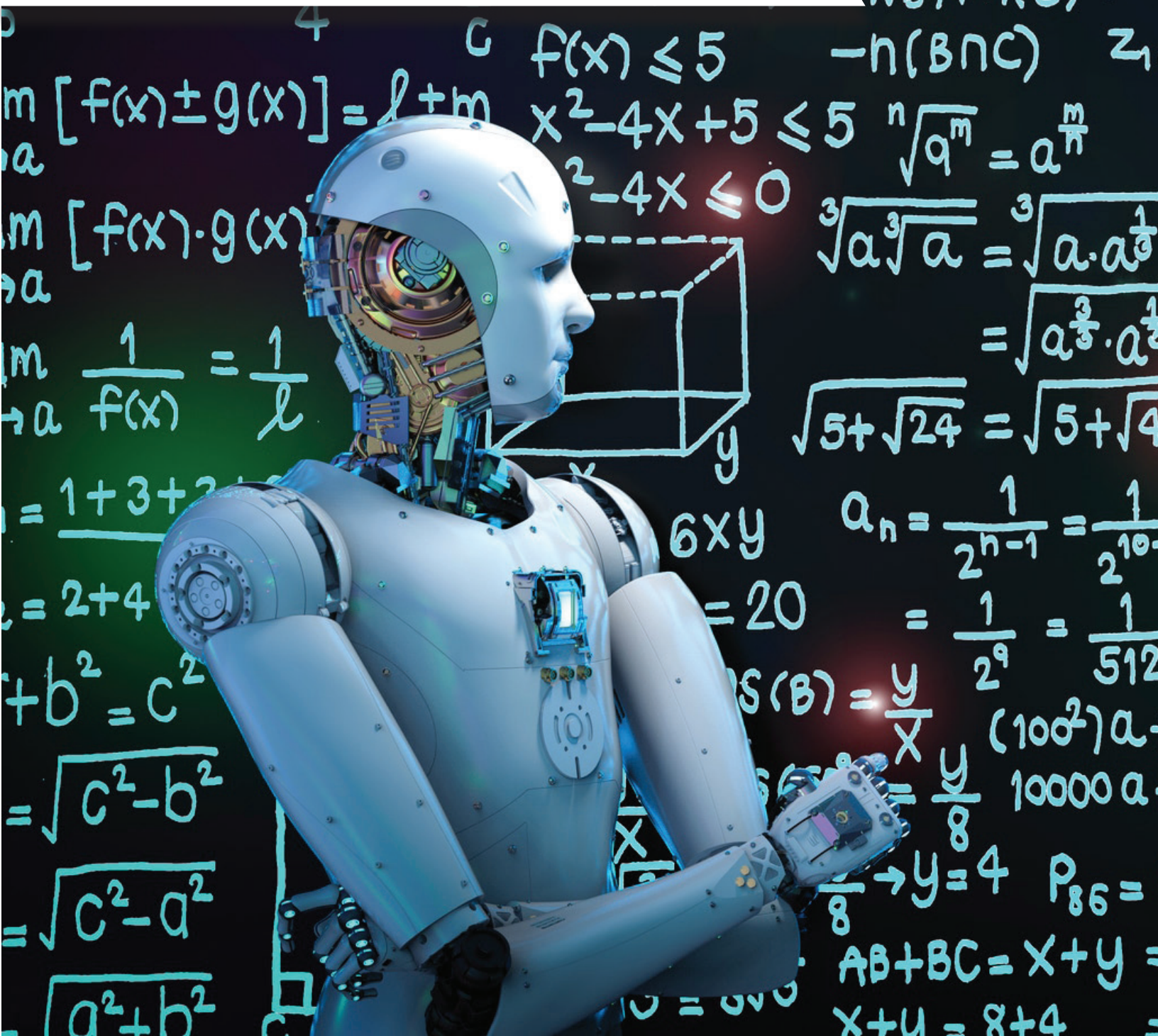
The introduction of marking schemes had divergent effects across different AI models. While Claude typically showed improved alignment with human benchmarks when given marking schemes, other models like GPT-4 often increased their deviation from human standards. This underscores the importance of both AI model selection and well-designed assessment frameworks when considering implementation in professional educational contexts.

Contrary to initial expectations, subject complexity was not necessarily a limiting factor for AI performance. In subjects requiring integrated judgment, such as Strategic Case Study, certain AI models demonstrated effective assessment capabilities when properly guided. This challenges assumptions about the limitations of AI in complex professional assessment tasks.

The consistent patterns observed across multiple subject areas provide robust evidence that AI chatbots can serve as valuable supplementary tools in professional accounting assessment processes. However, the variability in performance across different AI models and subject areas highlights the need for careful implementation, ongoing validation, and subject-specific approaches.

As professional accounting bodies face increasing assessment demands, these findings offer promising avenues for enhancing assessment efficiency while maintaining quality standards. The future of accounting assessment likely involves thoughtful integration of AI capabilities with human expertise, leveraging the strengths of each to create more robust, efficient, and fair assessment systems for the next generation of accounting professionals.

# 6.0
# RECOMMENDATIONS



Based on our findings, we recommend the following:

1. Accounting professional bodies should consider piloting the use of certain AI models (particularly Claude) as supplementary assessment tools, especially in subjects where AI-human alignment was strongest.

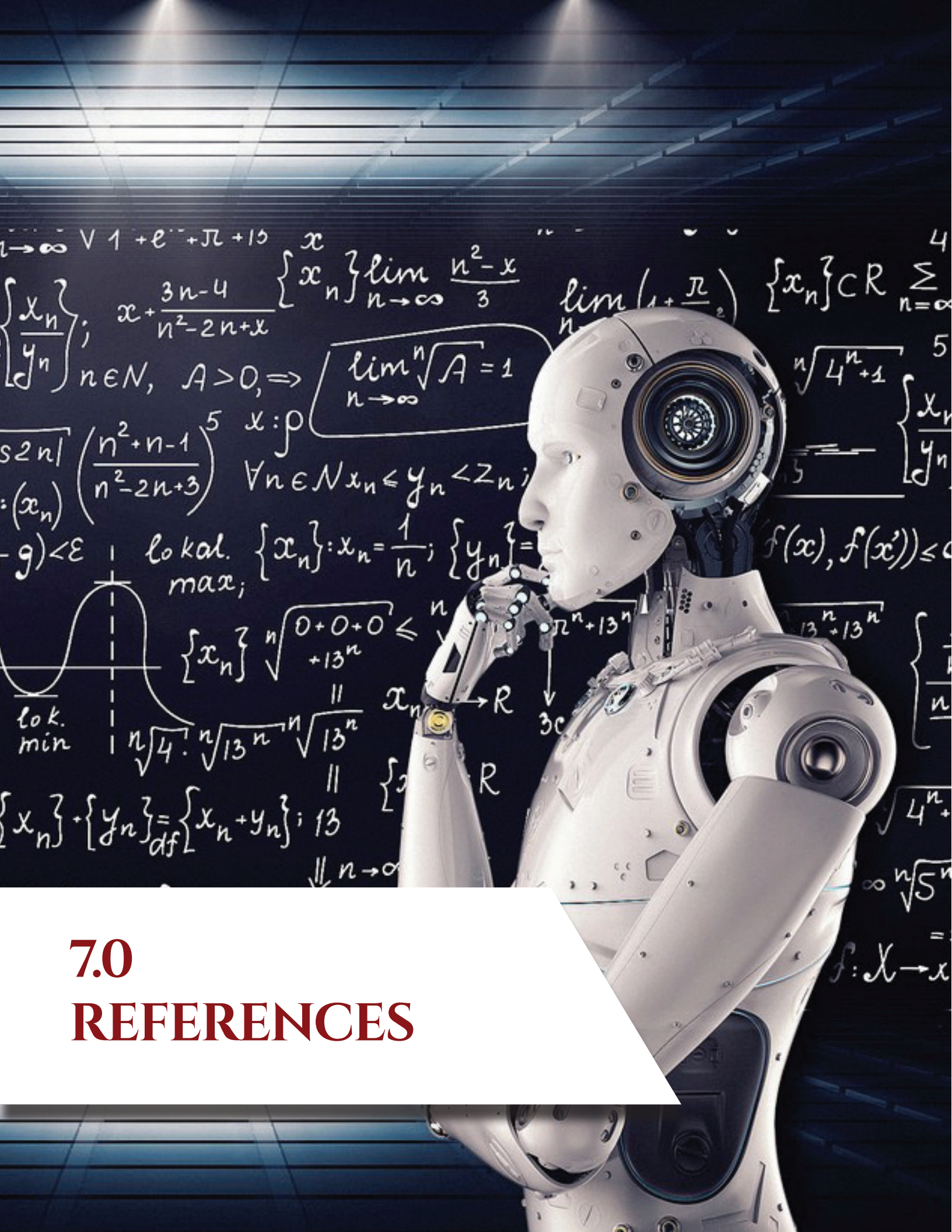
2. Any implementation of AI in professional accounting assessment should include well-designed marking schemes, as these significantly improve alignment with human standards for most AI models.

3. A hybrid assessment approach combining AI pre-marking with human verification may be optimal, allowing for efficiency improvements while maintaining assessment quality and professional judgment.

4. Further research should explore how marking schemes can be optimized to guide AI assessment more effectively, potentially through iterative design and calibration processes.

5. Accounting educators and examiners should receive training on effective AI prompt engineering and marking scheme design to maximize alignment between AI and human assessment standards.

6. Ongoing monitoring and evaluation of AI assessment quality should be implemented to ensure continued alignment with professional standards as both AI capabilities and accounting practices evolve.



# 7.0 REFERENCES

1. Holmes, W., Bialik, M. and Fadel, C. (2019) *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning.* Boston, MA: Center for Curriculum Redesign.

2. Rawashdeh, A., (2023). The consequences of artificial intelligence: an investigation into the impact of AI on job displacement in accounting. *Journal of Science and Technology Policy Management*. 16(3), 506-535. https://doi.org/10.1108/JSTPM-02-2023-0030

3. Wood, D. A., Achhpilia, M. P., Adams, M. T., & Zoet, E. (2023). The ChatGPT artificial intelligence chatbot: How well does it answer accounting assessment questions? Issues in Accounting Education, Forthcoming, 1-28.

4. Mahshanian, A., & Shahnazari, M. (2020). The effect of raters fatigue on scoring EFL writing tasks. Indonesian Journal of Applied Linguistics, 10(1), 1-13.

5. Yesilyurt, Y. E. (2023). AI-enabled assessment and feedback mechanisms for language learning: Transforming pedagogy and learner experience. In *Transforming the language teaching experience in the age of AI* (pp. 25-43). IGI Global.

***All Images Used In This Report Were Sourced From The Internet.***


Accountancy House, off Trinity Avenue, Okponglo East Legon, Accra

054 433 6701 /2 | 027 78014 22/3/4

P.O.Box GP 4268 Accra, Ghana

Digital Address: GA - 416 - 9906

www.icagh.org

@icagh1

samuel.fianko@icagh.com