International Journal of Basic and Applied Sciences, 15 (1) (2026) 44-48

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/vet88k67
Research paper

ChatGPT-Based Preparation Tool for Industrial Engineer Certification Examinations

Cheng-Wen Chang, Chen-Chi Wang *, Yung-Chun Chang, Chun-Ying Lin, Jui-Chan Huang

Department of Industrial Engineering and Management, National Kaohsiung University of Science and Technology, Kaohsiung City 807618, Taiwan
**Corresponding author E-mail: lindawang.921030@gmail.com*

Received: December 18, 2025, Accepted: January 6, 2026, Published: January 9, 2026

Abstract

This study aims to evaluate the performance of artificial intelligence language models in the Industrial Engineer Certification Examination and to analyze their accuracy across different subjects. A total of 750 past exam questions (from 2019 to 2024) across five subjects, namely Quality Management, Production and Operations Management, Operations Research, Engineering Economics, and Human Factors Engineering, were tested using ChatGPT-3.5 and ChatGPT-4o. The models' responses were compared with standard answers to compute accuracy rates and identify performance differences. Additionally, the Chain of Thought (CoT) and Tree of Thought (ToT) reasoning frameworks were applied to incorrect responses from GPT-4o to assess improvements in reasoning. The results show that ChatGPT-4o achieved significantly higher accuracy and reasoning coherence than ChatGPT-3.5, particularly in Quality Management and Production Management. Overall, this study demonstrates the potential of generative AI as an effective tool for professional education and exam preparation, offering insights for future model optimization and educational integration.

Keywords: *AI Language Model; Chinese Institute of Industrial Engineers (CIIE); ChatGPT; Answer Accuracy; Chain of Thought (CoT); Tree of Thought (ToT).*

1. Introduction

This study aims to evaluate the application potential of artificial intelligence language models in the Industrial Engineer Certification Examination, and to compare the answer accuracy and reasoning performance of ChatGPT-3.5 and ChatGPT-4o using the action research method. As professional certifications become a critical means for university students to enhance their competitiveness, more and more candidates utilize ChatGPT for learning assistance. However, the low pass rates across various subjects in the Industrial Engineer Certification Examination over the past five years indicate a high degree of professional difficulty, prompting this study to investigate whether language models possess question-solving capabilities comparable to human experts.

In this study, exam questions from 2019 to 2024 across five subjects, namely Production and Operations Management, Quality Management, Operations Research, Human Factors Engineering, and Engineering Economics, were introduced into both ChatGPT-3.5 and ChatGPT-4o models to analyze their accuracy and reasoning performance across different question types. Additionally, the Chain of Thought (CoT) and Tree of Thought (ToT) strategies were applied to incorrect responses from GPT-4o to assess improvements in reasoning.

Although this study draws upon prior literature on AI-assisted learning, its primary contribution is empirical, based on experimental testing of large language models using real certification examination data

2. Research Method

2.1. Overview of language models and learning applications

ChatGPT, developed by OpenAI, is a large language model with powerful language understanding and generation capabilities, and has been widely applied in educational and self-study contexts. It is noted that artificial intelligence can provide real-time feedback and individualized learning resources to improve learning efficiency (Gocen & Aydemir, 2020). In recent years, there has been an increasing number of studies exploring the application efficacy of ChatGPT in professional examinations. Lewandowski et al. (2023) compared the performance of GPT-3.5 and GPT-4 in dermatology professional examinations, and discovered a significant improvement in accuracy for the newer model, along with a reduction in cross-linguistic discrepancies. Alfertshofer et al. (2023) found that the performance of ChatGPT in medical licensing examinations across different countries was influenced by question type and length, indicating limitations in complex

reasoning tasks. Overall, ChatGPT demonstrates potential for application in education and professional examinations. Particularly, it develops continuously in semantic understanding and knowledge extraction.
Overall, prior studies consistently report performance gains in newer generations of language models; however, domain-specific professional examinations continue to reveal persistent limitations in mathematical reasoning and graphical interpretation. The present study extends these findings by empirically validating such limitations within industrial engineering certification contexts while demonstrating how structured reasoning strategies may partially mitigate them.

2.2. Improvement of GPT-4o accuracy

GPT-4o is the latest generation of models from OpenAI, showing significant improvements in language understanding and reasoning capabilities compared to GPT-3.5. In this study, GPT-4o was used as the primary test subject, and the CoT and ToT reasoning strategies were introduced to enhance its performance in professional question types. CoT is conducive to improving logical consistency by guiding the model through a step-by-step reasoning process. Meanwhile, ToT allows the model to explore multiple branches of thought in parallel and select the optimal answer (Yao et al., 2023). Overall, both strategies can enhance the model's reasoning depth and accuracy in complex professional examinations, improving the feasibility of using GPT-4o as an educational aid tool.

2.3. Challenges to language models

Language models may encounter challenges of insufficient context understanding or responses deviating from the topic during long sequence interactions. Although ChatGPT holds significant potential for educational applications, the content it generates may contain errors, which, if unverified, may easily mislead learning (Meyer et al., 2023). Due to the diverse sources of training data, the content generated by the model may also reflect potential biases, involving issues related to academic integrity and educational ethics. Ray (2023) indicated that to make ChatGPT a reliable educational technology, its content accuracy must be enhanced, and data biases must be minimized to ensure fairness and credibility in professional educational settings.

2.4. Research design and procedure

This study targets the Industrial Engineer Certification Examination held by the Chinese Institute of Industrial Engineers (CIIE), covering five core subjects: Production and Operations Management, Quality Management, Operations Research, Engineering Economics, and Human Factors Engineering. The research data consists of examination questions from 2020 to 2024, with a total of 750 single-choice questions with four options.
The research process involves four steps: (1) select five subjects for testing, and ensure the question types cover text comprehension, logical reasoning, and mathematical calculation; (2) input all questions in a consistent format into ChatGPT-3.5 and ChatGPT-4o for answering; (3) compare the model responses with standard answers, calculate the accuracy for each subject with a passing threshold of 60, and compare these pass rates s with historical pass rates; (4) apply CoT and ToT reasoning strategies to the incorrect answers by GPT-4o, and assess their impact on accuracy improvement. All data were compiled and analyzed using Microsoft Excel, with accuracy and pass rates serving as the primary quantitative indicators.
In this study, Chain of Thought (CoT) and Tree of Thought (ToT) reasoning strategies were applied exclusively to incorrect GPT-4o responses. This design choice was made to isolate and measure the incremental impact of structured reasoning frameworks on error correction and reasoning improvement, without inflating baseline model performance.

3. Results and Discussion

3.1. Experimental results and analysis

The average pass rates of CIIE from 2020 to 2024 across five subjects were compared with the exam results from ChatGPT models, as shown in Figure 1. The average pass rate for general candidates was only 38.9%, indicating a high level of difficulty in the exam.
The overall accuracy of GPT-3.5 was 42.3%, slightly above the average level of candidates, and it performed weaker in computation and reasoning questions such as Engineering Economics and Operations Research. In contrast, the overall accuracy of GPT-4o increased to 62.7%, approximately 20% higher than GPT-3.5, demonstrating a significant advantage in subjects such as Production and Operations Management, Quality Management, and Human Factors Engineering. However, GPT-4o had not yet achieved passing scores in Engineering Economics and Operations Research. Particularly, it performed poorly in question types involving formula calculations and tabular interpretations.
Overall, GPT-4o possesses knowledge and reasoning abilities comparable to high-intermediate candidates, but there remains room for improvement in complex mathematical operations and graph comprehension.

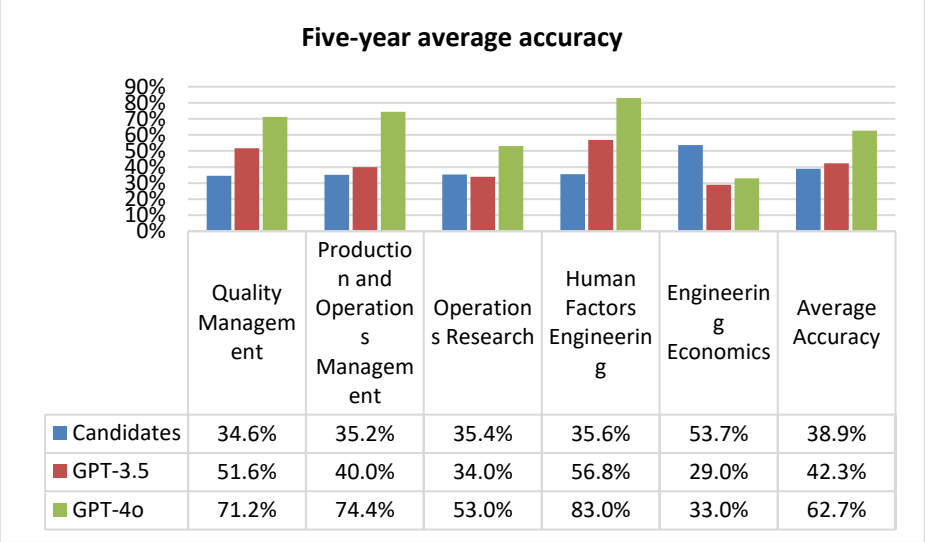


Fig. 1: Comparison of Pass Rates from 2020 to 2024.

Source: Compiled by this study.

3.2. Discussion of results and experimental findings

During the experimental process, three phenomena were observed in this study. These findings can not only facilitate the explanation of the context of the research findings but also provide inspiration and guidance for future research directions.

1) Text Limitations of the Model

The questions in Industrial Engineer Certification Examinations contain substantial graphical information. Since GPT-3.5 lacks image recognition capabilities, its performance on graph-based questions is limited. For instance, in the subject of Operations Research, there were 55 graph-based questions from 2020 to 2024. GPT-4o achieved an accuracy rate of 32.7%, significantly higher than GPT-3.5's 21.8%, as detailed in Table 1.

Table 1: Accuracy Rates of GPT-3.5 and GPT-4o in Graph-Based Questions

	GPT-4o	GPT-3.5
Correct Answers for Graph-based Questions	18	12
Total Graph-based Questions	55	55
Accuracy Rate of Graph-based Questions	32.7%	21.8%

Source: Examination Questions of CIIE Operations Research from 2020 to 2024.

GPT-3.5 could answer some questions correctly. The reason may be that some questions provided textual or tabular information, allowing the model to conduct reasoning based on textual features. The results indicate that multimodal capabilities can significantly impact performance on profession questions, and it would be very difficult for models lacking image processing functions to effectively handle graph-based questions.

2) Input Method of the Model

The input method also significantly affects the accuracy of GPT-4o's responses. For example, in questions from the subject of Production and Operations Management in 2024, GPT-4o achieved the accuracy rate of 80% (the highest) with pure text input, 68% with PDF input, and 64% (the lowest) with image input, as shown in Table 2. The reason is that text input is semantically clear and easily parsed, whereas PDFs and images require OCR or visual recognition, which may result in decoding errors. The results highlight the importance of standardizing input formats when designing AI-assisted learning tools to avoid the interference of non-semantic factors on model performance. AI-assisted learning tools to avoid the interference of non-semantic factors on model performance.

Table 2: Accuracy Rates of GPT-4o by Input Types (PDF, Text, and Image)

	GPT-4o with PDF Input	GPT-4o with Text Input	GPT-4o with Image Input
Correct Answers	17	20	16
Total Questions	25	25	25
Accuracy Rate	68%	80%	64%

Source: Examination Questions of CIIE Production and Operations Research in 2024.

3) Response Style of the Model

Despite identical input method, there remain differences in response styles between models. Some responses are presented solely as options, while others generate complete sentences. Moreover, models may provide lengthy explanations even when explicitly instructed not to do so (Vakilzadeh, 2023). This inconsistency complicates the consolidation and assessment of answers.

3.3 Solutions

In the previous section, examination questions from five subjects over the past five years were analyzed, revealing that GPT-4o's overall performance is clearly superior to that of GPT-3.5. In this section, it would be further explored on how to apply the ToT and CoT reasoning methods to GPT-4o to improve its accuracy in the examination subjects, thereby enhancing the model's reliability in practical applications.

3.3.1. CoT

CoT is a strategy that prompts the model to unfold its reasoning process by guiding it to generate a series of semantically coherent intermediary steps, thus allowing it to systematically construct a logical chain and deduce the final answer (Wei, 2022). Its primary functions

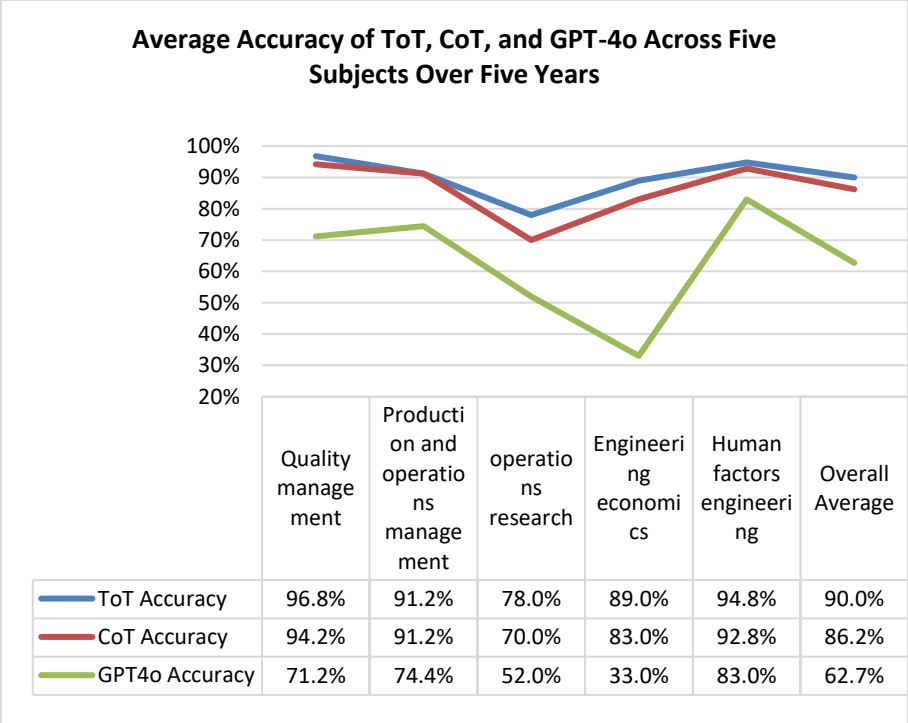
include: enhancing the model's stability in multi-step reasoning and mathematical calculations; reducing the likelihood of skipping steps by the model; providing an interpretable reasoning process to facilitate the educational applications and error diagnosis. In the context of this study, CoT demonstrates particularly significant benefits for Industrial Engineering Certification Examination questions, especially in question types involving formula introduction and tabular data calculations, where it can effectively reduce deviations in intermediate steps.

3.3.2. ToT

The ToT reasoning framework can be regarded as an extension of CoT. Its core concept lies in regarding reasoning as a searching for questions, generating multiple optional thought pathways and incorporating a self-evaluation mechanism for selection (Yao, 2023). ToT is characterized in: providing branched inference instead of linear reasoning, thus allowing for simultaneous examination of multiple potential answers; allowing for forethought and backtracking, thus enhancing reasoning consistency; being suitable for complex question types that require strategic planning or tabular structure analysis. Compared to CoT's single-path reasoning, ToT is more suitable for handling industrial engineering problems that involve high complexity or require cross-step integration of information, such as process scheduling, linear programming, and decision analysis.

3.3.3. Experiments with ToT and CoT

Based on analysis results in the previous section, CoT and ToT methods were further applied to the incorrect responses generated by the model across five subjects over the past five years.
Figure 2 summarizes the overall performance of ToT, CoT, and the original GPT-4o across five subjects from 2019 to 2024. The results indicated that after applying the reasoning-oriented strategies, the model's accuracy was significantly improved: the accuracy of GPT-4o increased from a baseline of 62.7% to 86.2% (CoT) and 90.0% (ToT). Among them, ToT demonstrated the most substantial enhancements, highlighting its clear advantage in multi-step reasoning and strategic choice questions.



	Quality management	Production and operations management	operations research	Engineering economics	Human factors engineering	Overall Average
ToT Accuracy	96.8%	91.2%	78.0%	89.0%	94.8%	90.0%
CoT Accuracy	94.2%	91.2%	70.0%	83.0%	92.8%	86.2%
GPT4o Accuracy	71.2%	74.4%	52.0%	33.0%	83.0%	62.7%

Fig. 2: Average Accuracy of ToT, CoT, and GPT-4o Across Five Subjects Over Five Years.

Source: Compiled by this study.

A comprehensive comparison revealed that both ToT and CoT could effectively and consistently enhance the answer accuracy of large language models in industrial engineering examinations. However, ToT and CoT exhibited complementary characteristics across different questions. ToT exhibited excellent performance in question types involving complex reasoning, cross-concept integration, and multi-stage decision-making. In contrast, CoT was more efficient in questions with clear structures, linear steps, or single-path derivations.
These results suggest that a flexible combination of ToT's breadth of search and CoT's depth of reasoning based on the characteristics of the question types and reasoning requirements could further enhance the model's overall performance in professional examinations and knowledge application contexts.

4. Conclusion

Future research may explore hybrid human–AI learning models, adaptive selection of CoT or ToT strategies based on question type, and real-time AI tutoring systems integrated into engineering curricula.
The performance of ChatGPT-3.5 and ChatGPT-4o was compared across the five major subjects of the Industrial Engineer Certification Examination to assess the applicability of large language models in assisting learning during preparation for professional exams. The results revealed that GPT-4o's overall accuracy rate (62.7%) was significantly higher than that of GPT-3.5 and exceeded the historical average pass rate of candidates (38.9%). Notably, the accuracy rate of GPT-4o in several units in the subjects of Quality Management, Production

and Operations Management, and Human Factors Engineering exceeded 80%, demonstrating strong conceptual understanding and question interpretation abilities. In contrast, GPT-3.5 exhibited weaker performance in questions involving multi-step calculations or graph interpretation. After incorporating CoT and ToT reasoning strategies, GPT-4o's average accuracy was improved to 86.2% and 90.0%, respectively, reflecting that structured reasoning frameworks could effectively enhance the model's reasoning capacity and answer efficacy in complex questions.

Despite its effectiveness, excessive reliance on AI-generated solutions without human verification may pose risks to learning integrity and examination ethics. Therefore, generative AI should be positioned as a supplementary educational aid rather than a substitute for independent reasoning and professional judgment.

Overall, GPT-4o demonstrates potential as a professional learning aid tool. However, it remains constrained by unstable performance in complex calculations, potentially resulting in learning risks of generating erroneous content and learners' excessive dependency. AI tools should serve as supplements rather than substitutes and should be supported by human verification and metacognition training.

Acknowledgements

This research is supported by the Research Support Scheme of National Science and Technology Council, Taiwan, R.O.C. under Grant no. 114-2813-C-992-089-E.

References

[1] A. Gocen, F. Aydemir. (2020). Artificial Intelligence in Education and Schools. Research on Education and Media Vol. 12, N. 1. https://doi.org/10.2478/rem-2020-0003.

[2] A. Vakilzadeh, S. Pourahmad Ghalejoogh, M. Hatami. (2023). Evaluating the potential of large language model AI as project management assistants: a comparative simulation to evaluate GPT-3.5, GPT-4, and Google-Bard ability to pass the PMI's PMP test. Evaluating the Potential of Large Language Model AI as Project Management Assistants: A Comparative Simulation to Evaluate GPT-3.5, GPT-4, and Google- Bard Ability to Pass the PMI's PMP Test(August 1, 2023). https://doi.org/10.2139/ssrn.4568800.

[3] Chinese Institute of Industrial Engineers. (2024). Certification Examination Overview. https://www.ciie.org.tw/about1-c10h3.

[4] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, and D. Zhou. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

[5] J. G. Meyer, R. J. Urbanowicz, P.C.N. Martin, K. O'Connor, R. Li, P. C. Peng, T. J. Bright, N. Tatonetti, K. J. Won, G. Gonzalez-Hernandez & J. H. Moore. (2023). ChatGPT and large language models in academia: Opportunities and challenges. BioData Mining, 16(1), 20. https://doi.org/10.1186/s13040-023-00339-9.

[6] M. Alfertshofer, C.C. Hoch, P.F. Funk, K. Hollmann, B. Wollenberg, S. Knoedler, L. Knoedler. (2024). Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Annals of Biomedical Engineering 52 (6), 1542-1545. https://doi.org/10.1007/s10439-023-03338-3.

[7] M. Lewandowski, P. Łukowicz, D. Świetlik, W.B. Rybak. (2023). ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology. Clinical and Experimental Dermatology, Volume 49, Issue 7, July 2024, Pages 686–691. https://doi.org/10.1093/ced/llad255.

[8] P. P. Ray (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems 3, 121-154. https://doi.org/10.1016/j.iotcps.2023.04.003.

[9] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan. (2024). Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems, 36.