

Comparative Performance Analysis of Generative AI Applications in PLC: An Industrial Electrical Engineering Subject

Amnaj Prajong

Department of Industrial Electrical Engineering, Faculty of Agricultural Technology and Industrial Technology, Nakhon Sawan Rajabhat University, Nakhon Sawan, Thailand

Email: amnaj.p@nsru.ac.th

ORCID iD: <https://orcid.org/0009-0008-7952-2608>

Therdpong Daengsi*

Department of Sustainable Industrial Management Engineering, Faculty of Engineering, Rajamangala University of Technology Phra Nakhon, Bangkok, Thailand

Email: therdpong.d@rmutp.ac.th

ORCID iD: <https://orcid.org/0000-0002-7569-8197>

*Corresponding author

Received: September 14, 2025; Revised: December 15, 2025; Accepted: February 16, 2026; Published: August 8, 2026

Abstract: This study evaluates the performance of six generative artificial intelligence (AI) systems in solving Thai-language multiple-choice examinations in the subject of Programmable Logic Controllers (PLC), a core component of Industrial Electrical Engineering education. Six large language models (LLMs), including ChatGPT, Claude, DeepSeek, Gemini, Copilot, and Grok, were tested using fifteen sets of PLC examination questions. Statistical analysis was conducted using one-way ANOVA and two-sample t-Tests with Bonferroni correction to examine performance differences. In addition, effect size measures, including Eta-squared (η^2) and Cohen's d, were calculated to assess the magnitude of the observed differences. The results show that ChatGPT achieved the highest mean score (77.27%), while DeepSeek followed closely (76.73%) and demonstrated the lowest standard deviation ($\pm 1.83\%$), indicating the most consistent performance across test sets. Claude also performed strongly (74.80%), whereas Gemini, Copilot, and Grok achieved similar mid-tier scores ranging from 72.40% to 72.73%. Although all LLMs achieved scores within the passing grade range, ANOVA confirmed statistically significant differences among systems ($p = 0.0002$). However, after applying the Bonferroni correction, only a subset of pairwise differences remained statistically significant, particularly between DeepSeek and several mid-tier LLMs, while the differences among ChatGPT, Claude, and DeepSeek were not statistically significant under the adjusted threshold. Effect size analysis further indicates that some of these differences represent meaningful practical variation in LLM performance. These findings indicate that contemporary LLMs demonstrate baseline comprehension of PLC concepts and can achieve passing-level performance in technical examinations conducted in a non-English language. The study contributes empirical evidence on AI performance in Thai-language technical assessments and highlights the potential role of generative AI as a complementary learning support tool in vocational and engineering education.

Index Terms: Engineering Education, Programmable Logic Controller, Large Language Model, Statistical Analysis.

1. Introduction

The advancement of generative artificial intelligence (AI), particularly large language models (LLMs), has reshaped how knowledge can be accessed, processed, and applied across disciplines. Beyond their initial role in natural language processing, these LLMs are increasingly integrated into educational contexts, where they demonstrate the ability to comprehend exam-style questions, provide structured answers, and simulate test-taking scenarios. Such capabilities suggest a potential paradigm shift in academic assessment and training, offering new opportunities for both learners and educators. At the same time, this shift has raised important questions regarding fairness, localization, and the extent to which such systems can complement rather than replace traditional teaching and assessment practices. Existing evidence highlights how LLMs can achieve strong results in standardized examinations across medicine,

biology, and other academic domains, with performances often approaching or surpassing human baselines in a few studies [1–5]. Nevertheless, limitations remain: LLMs frequently struggle with high-application tasks that require critical reasoning, problem-solving, or domain-specific contextualization. Moreover, most of the available evidence derives from high-resource languages such as English, which restricts the generalizability of these findings across diverse linguistic and cultural contexts. Recent research on LLMs has increasingly explored their performance in standardized examinations, particularly in medical and professional certification contexts. However, most of these evaluations focus on domains dominated by natural language interpretation rather than technically structured problem solving.

However, for Programmable Logic Controllers (PLCs), which represent a cornerstone of industrial electrical engineering education, mastery requires not only factual knowledge but also applied problem-solving and logical reasoning in technical contexts. PLCs are widely used in automation and control systems, serving as the backbone of industrial processes ranging from manufacturing to energy management. Because PLC tasks rely heavily on rule-based logic and structured engineering reasoning, evaluating LLM performance in PLC examinations provides an opportunity to examine how these systems handle industrial technical knowledge rather than primarily language-driven tasks. Despite their critical importance, little is known about how LLMs perform when challenged with PLC-focused assessments, particularly in non-English settings such as Thai-language examinations. Addressing this gap is essential to determine whether generative AI can reliably support industrial education and contribute to preparing students for real-world technical tasks. In this regard, the evaluation of LLMs on PLC examinations is not simply an academic exercise; it offers insights into whether these systems can facilitate skill acquisition in industrial contexts, support localized education, and extend their utility beyond general knowledge domains.

This study therefore conducts a comparative performance analysis of several leading LLMs on multiple-choice examinations in PLC. By systematically evaluating their accuracy, stability, and consistency, the research aims to clarify the potential and limitations of generative AI in technical education. More broadly, it seeks to contribute to the growing discourse on the role of AI as both an assessment tool and a supplemental learning resource. Understanding how LLMs perform in this niche yet practically significant domain will provide valuable implications for educators, policymakers, and technologists seeking to integrate AI responsibly into industrial and engineering education.

1.1. Contributions

The major contribution of this research lies in its domain-specific and language-localized focus, which contrasts with prior English-language and general-domain studies. It provides the first benchmarking of LLMs on Thai-language PLC assessments, covering both well-established LLMs such as ChatGPT and Claude as well as newer systems like DeepSeek and Grok that have rarely been evaluated in education. By addressing a technical subject in a non-English context, the study fills a critical research gap. The findings reveal both the strengths and limitations of generative AI, offering evidence-based insights for educators and policymakers. These results emphasize that LLMs can serve as supportive tools, rather than replacements, in engineering education and digital learning design, provided they are carefully selected and responsibly integrated.

1.2. Outline

This article is organized as follows. After the introduction, Section 2 provides the background, covering three aspects: LLMs, their development and applications; PLCs, their role in industrial automation and education; and the literature review, which identifies research gaps in non-English, domain-specific technical education. Section 3 describes the methodology, including the design of Thai-language PLC test sets, the evaluation of six LLMs (ChatGPT, Claude, DeepSeek, Gemini, Copilot, and Grok), and the use of statistical tools such as ANOVA and t-Tests. Section 4 presents the results (e.g., highlighting performance differences, mean scores, and grade classifications) and additional analyses (e.g., Bonferroni correction, and Cohen's approaches). Section 5 discusses these findings in relation to prior research, emphasizing both the potential and the limitations of LLMs in vocational and technical education. Finally, Section 6 offers the conclusion and future work, summarizing contributions and suggesting further research on diverse exam formats, practical applications, and cross-language benchmarking.

2. Background

2.1. Overview on Assessment Theory and Educational Measurement

Before presenting the performance evaluation of LLMs in PLC (an industrial electrical engineering subject), which is linked to scoring, grading, and student assessment, it is necessary to understand the background information on assessment theory and educational measurement. The relevant background information is outlined below.

Assessment theory has evolved significantly from a traditional measurement-centered paradigm toward a broader, learning-oriented framework. Contemporary assessment frameworks no longer depend exclusively on traditional, summative, and memorization-focused methods [6]. In the past, assessment systems were designed primarily for certification, selection, and accountability, relying on standardized procedures to ensure comparability and integrity. In

such contexts, validity and reliability were conceptualized as statistical properties of test instruments [7]. Within traditional large-scale testing paradigms, assessment quality was examined through test-centered indicators external to classroom practice [7]. These approaches emphasized standardized administration, psychometric modeling, and measurement of achievement outcomes. However, contemporary studies recognize that the quality of classroom assessment differs from that required for large-scale evaluation [6]. In the last decade, theoretical developments introduced a critical distinction between summative and formative purposes of assessment. Summative assessment determines the quality or merit of learning outcomes, whereas formative assessment occurs early enough in the learning process to influence instruction and student development [8]. This conceptual shift reframed assessment from a terminal judgment to an instructional resource embedded within ongoing pedagogical activity or during instruction [8]. Formative assessment produces more favorable perceptions of learning and greater student engagement compared with traditional grade-focused approaches [6]. From a measurement perspective, educational assessment operates within the behavioral sciences and therefore differs from measurement in the physical sciences. Constructs (e.g., aptitude, motivation, achievement, and self-regulation) are latent and cannot be directly measured; instead, they are inferred from observable behaviors [7]. Assessment tasks function as operational definitions of these constructs, and the legitimacy of decisions depends on the strength of the inferences drawn from performance evidence. Modern validity theory emphasizes that tests themselves are not validated [7]. Instead, the interpretations and uses of test scores are subject to validation. Despite advances in psychometric modeling, contemporary analysis highlights an increasing tension between statistical measurement and classroom reality. Educational assessment is informed by multiple disciplines that often operate in isolation, resulting in a separation between psychometrics, curriculum, psychology, and classroom instruction [8]. Therefore, classroom assessment prioritizes diagnostic insight, instructional responsiveness, and holistic understanding of student learning rather than classroom test performance indicators [7]. As mentioned in [9], formative assessment fosters ongoing dialogue between teachers and learners and enhances students' self-regulatory capacities. These findings reinforce the view that assessment is not only a measurement event, but it is also a process that supports learning development [6]. Indeed, contemporary perspectives indicate that assessment has become integrated into the learning process [6].

Overall, assessment has become more integrated into the student learning process and is no longer limited to scoring and grading. It now connects curriculum, instruction, feedback, and learner development within a coherent framework that prioritizes continuous improvement over mere score production [6]. This paradigm shift reflects a broader reconceptualization of educational measurement, moving from the static quantification of outcomes to dynamic, context-sensitive designs aimed at enhancing learning quality.

2.2. Large Language Models

Large language models (LLMs), which fall under the umbrella of Generative Artificial Intelligence (Gen AI) (see Fig. 1), are developed using deep learning architectures such as the transformer. The transformer employs self-attention mechanisms to enable contextual understanding and the generation of human-like responses [10]. They are pretrained on massive text corpora and later fine-tuned for domain-specific tasks, allowing them to perform across a wide range of applications, including medical education, professional examinations, and technical assessments [11]. Specifically, LLMs are designed to generate human-like text through extensive exposure to vast datasets [12], with fine-tuning enabling effective performance in domain-specific tasks. Nevertheless, limitations remain, as these LLMs sometimes fail to follow prompts accurately and may hallucinate, producing confident yet incorrect responses [12]. Popular implementations include OpenAI's ChatGPT, which was initially launched with GPT-3.5 containing about 175 billion parameters, followed by GPT-4 with a vastly increased scale estimated between 500 billion and 100 trillion parameters, enabling greater reasoning and accuracy [10]. These LLMs are further enhanced by iterative reinforcement learning and human feedback, which continuously refine their outputs [11].

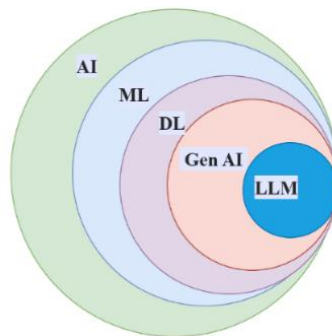


Fig. 1. Overview of AI domain.

2.3. Programmable Logic Controllers (PLCs) and Engineering Education

PLCs are digital control devices designed for industrial environments, enabling logic operations, sequencing, timing, counting, and arithmetic to control machinery and processes through sensors and actuators. They are regarded as the cornerstone of industrial automation and an essential skill for engineering education in the context of Industry 4.0. Teaching PLCs thus requires not only theoretical grounding but also practice-oriented approaches to ensure students gain both technical and problem-solving skills [13]. Innovative educational strategies have emerged to strengthen PLC teaching and training. For example, pinball machines have been leveraged as gamified PLC-controlled platforms, providing learners with practical exposure to ladder logic, input/output handling, and real-time troubleshooting in an engaging environment [14]. Similarly, PLC trainer kits have been developed for electric motor practice, offering students hands-on experience in circuit components, simulation, and real-plant control, thereby bridging theory and applied automation [15]. In advanced teaching practices, PLCs are integrated into comprehensive industrial process control systems, combining modeling, human-machine interface design, logic programming, and experimental testing with real industrial equipment to cultivate engineering-oriented competencies [16]. At the institutional level, universities in Southeast Asia have integrated PLC curricula aligned with the Conceive-Design-Implement-Operate (CDIO) framework, ensuring accreditation standards that support Industry 4.0 readiness and preparing graduates to meet the demands of industrial automation [17].

These findings underline the capacity of LLMs to manage structured, knowledge-intensive tasks with considerable consistency. Studies further emphasize their ability to enhance learning environments by tailoring explanations, fostering creativity, and supporting student-teacher interactions [3].

2.4. Literature Review Based on PLC Pedagogy and Training

Prior research in PLC pedagogy consistently emphasizes experiential learning, modular training systems, and empirical evaluation of instructional effectiveness. Early work on embedded PLC development demonstrated the feasibility of low-cost, flexible laboratory platforms to address rapid technological change and equipment constraints [18]. Capacity-building initiatives further highlighted structured training management cycles and measurable pre-post assessment frameworks for enhancing instructor competency in PLC curricula [19]. Portable and compact PLC training kits were shown to improve student achievement when combining simulation with real sensor-actuator wiring [20], while custom-designed instructional kits significantly increased learning outcomes and satisfaction levels through statistically validated testing [21]. Similarly, PLC-based trainer systems for electric motor control enhanced laboratory engagement and conceptual understanding [22]. Beyond hardware innovation, comparative studies between procedural IEC 61131-3 programming and UML-based model-driven approaches revealed differences in cognitive demand and abstraction skills, contributing to structured automation software education [23], with further usability evaluations confirming comparable effectiveness between model-driven and traditional paradigms [24]. Curriculum-oriented modules integrating Ladder Logic and PLC fundamentals addressed industry demand for job-ready competencies [25], and ADDIE-based learning kit development demonstrated systematic instructional design validation [26]. More recently, experimental-based packages integrating PLC with Cloud SCADA and IIoT technologies achieved high instructional efficiency and learner satisfaction, reflecting alignment with Industry 4.0 requirements [27]. Collectively, these studies establish that PLC education research has predominantly focused on laboratory enhancement, pedagogical design, programming methodology comparison, and measurable human learning outcomes. A comparison of these PLC pedagogy and training approaches is summarized in Table 1 [18-30].

2.5. Literature Review Based on LLMs in Academic Assessment

Recent research highlights the growing role of large language models (LLMs) in academic assessment, particularly in multiple-choice testing across medicine and related disciplines, as shown in Table 1 [1-5, 31-35]. Studies consistently show GPT-4 outperforming GPT-3.5 and, in some cases, even surpassing human averages. For example, GPT-4 scored 86.8% on Spain's MIR exam, nearly matching top human scores [2], while Bing AI (GPT-4 class) achieved 78.9-83.7% on the U.S. Orthopedic In-Training Exam, exceeding the resident mean [4]. Similarly, in Canada's CFPC family medicine exams GPT-4 achieved 81-93%, and in Turkey's national TUS exam it reached 88.75%, outperforming both humans and rival models like Gemini 1.5 [1]. These results demonstrate the potential of GPT-4 to rival or exceed human performance in demanding, multilingual assessments.

Despite these advances, prior studies have focused primarily on English language content and conventional academic subjects, leaving gaps in localized, non-English technical fields such as Thai-language assessments in Industrial Electrical Engineering. One exception is Stankovski et al. [5], who explored ChatGPT in PLC logic and programming support, though without rigorous benchmarking across LLMs. Furthermore, the performance of emerging LLMs such as Claude, DeepSeek, Copilot, and Grok remains unreported, with no established metrics in non-English technical domains. Addressing this gap, the present study evaluates six LLMs (ChatGPT, Claude, DeepSeek, Gemini, Copilot, and Grok) on Thai-language PLC multiple-choice tests. It aims to provide empirical evidence on accuracy, fairness, and reliability in vocational and industrial education, offering a novel contribution to localized AI assessment research.

Table 1. Related Research Works Based on PLC Pedagogy and Training.

Authors	Year	Country	Major/Subject/Field	Major Finding
Pratumsuwan & Pongaen [18]	2011	Thailand	Mechatronics education; Low-cost embedded PLC; LabVIEW + ARM	This study proposed a low-cost embedded PLC to address system diversity and laboratory cost issues. By integrating LabVIEW with ARM, the solution was evaluated as simple to use and suitable for foundational training.
Vogel-Heuser et al. [24]	2012	Germany	PLC programming; UML vs IEC 61131-3; Experimental study	They found that UML improves program structure but requires higher abstract reasoning, with usability comparable to IEC 61131-3.
Vogel-Heuser et al. [23]	2013	Germany	PLC programming pedagogy; UML vs IEC 61131-3	This study reported that UML-based instruction produced better program structures, though students required stronger abstraction and problem-solving skills.
Mohd Khairudin et al. [20]	2019	Malaysia	Industrial automation education; Portable PLC training kit; IR4.0	They developed a portable PLC training kit that improved student achievement, particularly when combined with simulation-based learning.
Tran, Yahoui, & Siauve [28]	2019	France	Industry 4.0 automation training; Virtual factory (Factory I/O) + CodeSys	They implemented Factory I/O and CodeSys to enhance engagement and understanding in PLC education through virtual factory simulation.
DeGuglielmo et al. [25]	2020	USA	PLC education; Ladder Logic; Curriculum integration	This study introduced PLC and Ladder Logic modules into engineering curricula to enhance industry-relevant, job-ready skills.
Abdullah et al. [26]	2021	Malaysia	PLC automation; Learning kit; ADDIE model	They developed a PLC learning kit using the ADDIE model. Expert validation confirmed its suitability, and it improved hands-on learning effectiveness.
Meng & Dong [29]	2021	China	PLC teaching reform; OBE; Smart manufacturing	They applied an Outcome-Based Education (OBE) framework with digital platforms (TIA Portal, PLCSIM), improving engineering design and problem-solving skills compared to traditional teaching methods.
Kheowsakul, Senajit, & Maneechay [21]	2022	Thailand	PLC course; Teaching kit; Learning achievement	They developed a PLC teaching kit that significantly improved student achievement and satisfaction.
Khieovongphachanh et al. [30]	2023	Laos	PLC curriculum implementation; Industry 4.0	This study reported successful implementation of a PLC curriculum within an Industry 4.0 context, yielding satisfactory learning outcomes.
Sakanti et al. [22]	2023	Taiwan/Indonesia	PLC trainer; Motor control practice	They developed a PLC trainer kit for motor control practice, improving student understanding and motivation.
Bencheva et al. [19]	2024	EU-SE Asia	Teacher training; PLC curricula; ASEAN Industry 4.0	They conducted intensive teacher training programs that enhanced instructional quality and reflected measurable development through evaluation.
Chumchuen, Pankan, & Akatimagool [27]	2025	Thailand	Mechatronics education; Low-cost embedded PLC; LabVIEW + ARM	They developed an experimental PLC + Cloud SCADA learning package with high efficiency (E1/E2), significantly improving learning achievement and student satisfaction.

Table 2. Related Research Works Based on LLMs in Academic Assessment.

Authors	Year	Country	Major / Subject / Field	LLM(s) & Performance	Exam Format	Major Finding
Guillén-Grima et al. [2]	2023	Spain	Medicine (MIR residency entrance exam)	GPT-4: 86.81%; GPT-3.5: 63.18%	MCQ (225 Qs)	GPT-4 outperformed GPT-3.5 and approached top human scores.
Massey et al. [3]	2023	USA	Orthopaedics (exams)	GPT-3.5: 29.4%; GPT-4: 47.2%; Residents: 74.2%	MCQ	Residents substantially outperformed both models (GPT-4 > GPT-3.5).
Chen et al. [4]	2024	USA	Orthopaedics (OITE)	ChatGPT-3.5: 50.7–58.9%; Bing (GPT-4): 78.9–83.7%; Residents: 62.1%	MCQ (260 Qs)	Using full OITE item sets, Bing AI (GPT-4 class) exceeded the resident mean.
Stankovski et al. [5]	2024	Serbia	PLC in industrial engineering and in mechatronics	Not applicable	Questions/Tasks (NOT purely short-answer)	ChatGPT-assisted performance significantly exceeded traditional scores; improvement increased with experience.
Mesnard et al. [31]	2024	Europe	Urology (European Board exam)	ChatGPT-3.5: 59%; Bard: 68%; GPT-4: 80%	MCQ (100 Qs)	On a European Board urology sample, GPT-4 approached top-decile trainee performance.
Mousavi et al. [32]	2024	Canada	Family Medicine (CFPC SAMPs)	GPT-3.5: ~66–74%; GPT-4: 81–93% (~100% predicted pass)	SAMP (short answer)	On CFPC SAMPs, GPT-4 reached pass-level ranges across most sections.
Kadoya et al. [33]	2024	Japan	Medical Physics (Board Exam)	GPT-3.5: 42.2%; GPT-4: 72.7%	MCQ	In Japanese medical physics board items, GPT-4 clearly exceeded GPT-3.5.
Sood et al. [34]	2024	UK/US A	Radiology Exams (FRCR, ABR)	GPT-3: 31.7–61%; GPT-4: 70.7% (SBA), 83% (T/F)	MCQ (SBA + T/F)	For FRCR/ABR theory, GPT-4 made the largest gains over GPT-3 on single-best-answer.
Sorenson & Hanson [35]	2024	USA	General Chemistry Exams	ChatGPT-3.5 (~42% overall vs. students ~55%)	MCQ (40–50 Qs per exam)	Rasch modeling revealed a distinctive ChatGPT signature, over-performing on some 'hard' items.
Koçak et al. [1]	2025	Turkey	Medical Specialty Training Exam	GPT-4: 88.75%; Llama 3: 79.2%; Gemini 1.5: 78.1%; Command R+: 50%	MCQ (240 Qs)	On the Turkish Medical Specialty Training Exam, GPT-4 led across basic and clinical domains.

3. Methodology

This study was designed to assess and compare the performance of LLMs in the subject of PLC, a core subject in Industrial Electrical Engineering. To achieve this, the researchers developed a comprehensive evaluation framework using Thai-language multiple-choice questions (MCQs). A total of 15 test sets were constructed, each containing 100 randomized questions (examples of the questions are presented in Appendix A). Importantly, the same pool of 100 questions was used across all 15 test sets. The content, difficulty level, and question composition remained identical in every set. The only variation among the test sets was the ordering of the questions. This approach ensured content consistency while minimizing potential positional or ordering bias. The questions were thoroughly reviewed to ensure they covered essential PLC topics, including fundamental structure and operating principles of PLCs, ladder diagram programming, logical functions, timers and counters, and industrial applications of PLCs. The order of both the questions and answer choices was shuffled across sets to avoid repetition bias.

According to Fig. 2, which is the overview of the processes in this study, six generative AI models were selected for evaluation: ChatGPT, Claude, DeepSeek, Gemini, Copilot, and Grok. Additional information regarding the free version of each model is presented in Table 3. The intention of using the free version of all LLMs was to simulate realistic usage conditions available to students and educators, who typically access the free or publicly available versions of LLMs. However, to minimize variability of the results from LLMs, all experiments were conducted within the same week (15–21 June 2025), while the same prompt structure was used for all LLMs. In addition, each LLM was instructed to complete all 15 test sets under identical conditions. The questions in the PDF file were provided in the Thai-language, mirroring real-world conditions in which Thai students undertake vocational assessments in engineering subjects. No contextual hints or prior knowledge of the test structure were given to the LLMs, ensuring unbiased evaluation. Although the prompt used to instruct the LLM to begin Test 1 was in Thai, it can be translated and presented as follows:

“Assume the role of a student in the Industrial Electrical Engineering program. Complete Test Set 1 for the subject Programmable Logic Controller (PLC), which consists of 100 questions (Questions 1–100). Select the single most accurate answer for each question without providing explanations, based on the attached file.”

After completing Test Set 1, the same prompt with the number of the next test sets (2, 3, ..., 15) was continued. However, this study utilized the default free version of each LLM, which typically imposes limits on file uploads or usage frequency per session. To ensure continuous access, multiple user accounts were utilized to complete the tasks. The models' responses were scored based on the number of correct answers. For each model, the mean score (percentage of correct answers) and standard deviation (S.D.) were calculated to represent its average performance and consistency across the 15 test sets. To further contextualize the results, each model's average score was converted into a letter grade using a Thai university-style grading scale, where: A is excellent ($\geq 80\%$), B+ is very good (75–79%), B is good (70–74%), ..., D is minimum pass (50–54%), and F is fail ($< 50\%$). This grading approach enabled the performance of each LLM to be interpreted in a familiar academic context. The combination of quantitative evaluation, statistical comparison, and grade-based classification provides a comprehensive methodology to benchmark the capabilities of generative AI in domain-specific, Thai-language technical education.

To determine whether the differences in performance among the six LLMs were statistically significant, the researchers conducted a one-way analysis of variance (ANOVA). Additionally, pairwise independent-sample t-tests were used to compare specific model pairs. A significance level of $p < 0.05$ was set for both statistical tests, corresponding to a 95% confidence level.

Table 3. LLMs under this study and their information.

LLM	Version / Release / Model	Remark
ChatGPT	GPT-4o	Accessed 15-21 June 2025
Claude	Claude Sonnet 4	
DeepSeek	DeepSeek-V3	
Gemini	Gemini 2.5 Flash (Released June 17, 2025)	
Copilot	GPT-4o powered Copilot (Microsoft 365 integration)	
Grok	Grok 3	

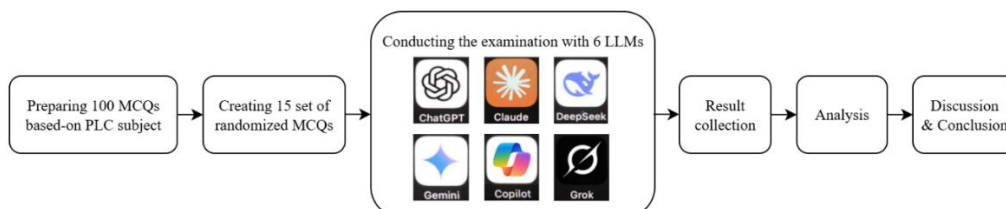


Fig. 2. Overview of the processes in this study.

4. Results and Analyses

4.1. Results and Analysis Using one-way ANOVA and Two-sample t-Test techniques

After obtaining the results from the method mentioned in Section 3, they are presented in Table 4. It provides a detailed comparison of six LLMs—ChatGPT, Claude, DeepSeek, Gemini, Copilot, and Grok—based on their performance in solving Thai-language multiple-choice questions in Programmable Logic Controllers (PLC). Each model was tested on 15 sets of questions, and their performance was evaluated using several statistical measures. ChatGPT achieved the highest average score at 77.27% with a standard deviation (S.D.) of $\pm 5.69\%$, indicating strong performance but with relatively higher variability. DeepSeek followed closely with a mean score of 76.73% and the lowest S.D. of $\pm 1.83\%$, reflecting the most consistent performance across all test sets. Claude earned an average score of 74.80% with an S.D. of $\pm 3.86\%$, showing slightly more variation than DeepSeek but maintaining solid results overall. The remaining three LLMs, Gemini, Copilot, and Grok, achieved similar average scores of 72.40%, 72.73%, and 72.67%, respectively, with standard deviations of $\pm 3.00\%$, $\pm 3.01\%$, and $\pm 2.72\%$. These values suggest that while these LLMs were slightly less accurate than the top performers, they remained relatively stable in their output.

Table 4. Results from six LLMs.

Item	Score (%)					
	ChatGPT	Claude	DeepSeek	Gemini	Copilot	Grok
Min.	68	62	73	68	68	68
Max.	86	78	80	77	77	77
Mean $\pm$ S.D.	77.27 $\pm$ 5.69	74.80 $\pm$ 3.86	76.73 $\pm$ 1.83	72.40 $\pm$ 3.00	72.73 $\pm$ 3.01	72.67 $\pm$ 2.72
95% Confidence Interval	(74.12, 80.42)	(72.66, 76.94)	(75.72, 77.75)	(70.74, 74.06)	(71.07, 74.40)	(71.16, 74.17)
Grade	B+	B+	B+	B	B	B

Table 5. Analyzed results using one-way ANOVA.

Source of Variation	SS	df	MS	F	p-value	F _{critical}
Between Groups	353.9667	5	70.7933	5.5882	0.0002	2.3231
Within Groups	1064.1333	84	12.6683			
Total	1418.1	89				

Table 6. Analyzed results from using one-way ANOVA (p-value only) and Two-sample t-Test.

Hypothesis	p-value ($\alpha=0.05$)	Significant ($\alpha=0.05$)	p-value $\alpha=0.0033$	Significant (adjusted $\alpha=0.0033$)
H1: The scores from all LLMs are the same or different.	0.0002 [*]	Yes	0.0002 ^b	Yes
H2: The scores from ChatGPT and Claude are the same or different.	0.1755	No	0.1755	No
H3: The scores from ChatGPT and DeepSeek are the same or different.	0.7338	No	0.7338	No
H4: The scores from ChatGPT and Gemini are the same or different.	0.0080 [*]	Yes	0.0080	No
H5: The scores from ChatGPT and Copilot are the same or different.	0.0126 [*]	Yes	0.0126	No
H6: The scores from ChatGPT and Grok are the same or different.	0.0104 [*]	Yes	0.0104	No
H7: The scores from Claude and DeepSeek are the same or different.	0.0949	No	0.0949	No
H8: The scores from Claude and Gemini are the same or different.	0.0674	No	0.0574	No
H9: The scores from Claude and Copilot are the same or different.	0.1132	No	0.1132	No
H10: The scores from Claude and Grok are the same or different.	0.0909	No	0.0909	No
H11: The scores from DeepSeek and Gemini are the same or different.	< 0.0001 [*]	Yes	< 0.0001 ^b	Yes
H12: The scores from DeepSeek and Copilot are the same or different.	0.0002 [*]	Yes	0.0002 ^b	Yes
H13: The scores from DeepSeek and Grok are the same or different.	< 0.0001 [*]	Yes	< 0.0001 ^b	Yes
H14: The scores from Gemini and Copilot are the same or different.	0.7634	No	0.7634	No
H15: The scores from Gemini and Grok are the same or different.	0.8003	No	0.8003	No
H16: The scores from Copilot and Grok are the same or different.	0.9497	No	0.9497	No

Note: 1) * means p-value < 0.05 is significant with 95% confidence interval.

2) b means p-value < 0.0033 is significant to an adjusted $\alpha = 0.0033$, approximately 99.67% confidence level under Bonferroni correction

In addition, after conducting the analysis using one-way ANOVA, the results are presented in Table 5. It shows a p-value of approximately 0.0002, indicating statistically significant differences in performance among the six LLMs. A two-sample t-Test was then conducted. The results from the one-way ANOVA and the t-Test are shown in Table 6. Focusing on pairwise comparison, this table presents the results of hypothesis testing that compares the performance scores of six LLMs, consisting of ChatGPT, Claude, DeepSeek, Gemini, Copilot, and Grok, on Thai-language multiple-choice questions in the PLC subject. All hypothesis tests were initially evaluated at $\alpha = 0.05$ (95% confidence interval). To control Type I error due to multiple pairwise comparisons, a Bonferroni correction, as used in [36], was applied. Since 15 pairwise comparisons were conducted, the adjusted significance level was calculated as follows:

$$\text{Adjusted } \alpha = \frac{\alpha}{n} = \frac{0.05}{15} \approx 0.0033 \quad (1)$$

The test for H1, which examines whether overall performance differs among the LLMs, is statistically significant ($p = 0.0002$). This result remains significant under both $\alpha = 0.05$ and the adjusted $\alpha = 0.0033$ approximately, confirming that meaningful overall differences exist among the LLMs. For pairwise comparisons, several differences are statistically significant at the conventional $\alpha = 0.05$ level. For example, ChatGPT shows significant differences compared with Gemini (H4, $p = 0.0080$), Copilot (H5, $p = 0.0126$), and Grok (H6, $p = 0.0104$) at $\alpha = 0.05$. However, after applying the Bonferroni correction (adjusted $\alpha = 0.0033$), these differences are no longer statistically significant. Therefore, these results should be interpreted cautiously. In contrast, DeepSeek demonstrates statistically significant differences compared with Gemini (H11, $p < 0.0001$), Copilot (H12, $p = 0.0002$), and Grok (H13, $p < 0.0001$), with all p -values below the adjusted threshold of 0.0033. These differences remain significant even under the more stringent Bonferroni-adjusted criterion, indicating stronger statistical evidence of performance separation. The remaining comparisons (H2, H3, H7–H10, H14–H16) (p -values: 0.1755, 0.7338, 0.0949, 0.0674, 0.01132, 0.0909, 0.7634, 0.8003, and 0.9497, respectively) are not statistically significant under either $\alpha = 0.05$ or the adjusted $\alpha = 0.0033$, suggesting relatively comparable performance levels between those model pairs. Overall, while several pairwise differences appear under the conventional 95% confidence level, only the strongest contrasts, primarily those involving DeepSeek, remain statistically robust after controlling for multiple comparisons. This strengthens the validity of conclusions regarding performance hierarchy among the LLMs.

These findings show that although several LLMs achieve relatively high average performance scores, statistically robust performance separation is observed primarily between DeepSeek and the mid-performing LLMs (Gemini, Copilot, and Grok). While ChatGPT, Claude, and DeepSeek do not differ significantly from one another under the Bonferroni-adjusted criterion, the strongest and most statistically consistent contrasts involve DeepSeek compared with Gemini, Copilot, and Grok. These results support the presence of a measurable performance hierarchy among the LLMs, while also demonstrating that some apparent differences at the conventional 95% confidence level do not remain significant after controlling multiple comparisons. Overall, the statistical analysis confirms that model selection can meaningfully influence outcomes in domain-specific, Thai-language technical assessments, particularly when performance differences are substantial and statistically robust.

4.2. Analysis of Effect Sizes Using Eta-squared and Cohen's d

For more reliability of the analyzed results as presented in Table 6, additional analyses using Eta-squared (η^2) and Cohen's d, as employed in [36], were also conducted. η^2 is used to evaluate the effect size associated with one-way ANOVA and can be calculated using the sum of squares (SS) values presented in Table 5, as follows:

$$\eta^2 = \frac{SS_{\text{Between}}}{SS_{\text{Between}} + SS_{\text{Within}}} = \frac{353.9667}{1418.1} \approx 0.2496 \quad (2)$$

The effect size calculated using Eta-squared was $\eta^2 = 0.2496$, indicating that approximately 25% of the variance in PLC test scores can be explained by differences among the evaluated LLMs. According to Cohen's benchmarks, where η^2 values of 1%, 6%, and 14% represent small, medium, and large effects, respectively, this result indicates a large effect size. In the context of AI benchmarking, this magnitude suggests that the observed differences reflect systematic performance variation among the LLMs rather than random fluctuations across test sets.

Similar to Eta-squared, Cohen's d is used to evaluate the effect size associated with pairwise t-tests. It can be calculated using the raw score data presented in Table 3. The formula for Cohen's d is given as follows:

$$d = \frac{\text{Mean}_1 - \text{Mean}_2}{SD_{\text{pooled}}} \quad (3)$$

After the pairwise effect sizes were evaluated using Cohen's d, the results are presented in Table 7. Focusing on high values of Cohen's d, very large differences were observed between DeepSeek and several LLMs, including Gemini, Copilot, and Grok, with d values of approximately 1.75, 1.61, and 1.76, respectively, indicating very large effect sizes. According to Cohen's benchmarks, d values of 0.20, 0.50, 0.80, and greater than 1.20 represent small, medium, large, and very large effect sizes, respectively. The results also show that comparisons involving ChatGPT and lower-performing LLMs yielded large effect sizes ($d \approx 1.0$). In contrast, comparisons among Gemini, Copilot, and Grok yielded negligible effect sizes, indicating similar performance levels. These results suggest that the observed differences among LLMs are not only statistically significant but also substantial in practical magnitude.

Table 7. Analyzed results of pairwise effect sizes using Cohen's d.

LLM ₁	LLM ₂	SD _{pooled}	Cohen's d	Meaning
ChatGPT	Claude	4.86	0.56	medium
ChatGPT	DeepSeek	4.227	0.15	small
ChatGPT	Gemini	4.545	1.07	large
ChatGPT	Copilot	4.551	1.00	large
ChatGPT	Grok	4.459	1.03	large
Claude	DeepSeek	3.02	-0.69	medium
Claude	Gemini	3.454	0.77	Medium to large
Claude	Copilot	3.461	0.67	medium
Claude	Grok	3.337	0.7	medium
DeepSeek	Gemini	2.482	1.75	very large
DeepSeek	Copilot	2.492	1.61	very large
DeepSeek	Grok	2.317	1.76	very large
Gemini	Copilot	3.003	-0.11	negligible
Gemini	Grok	2.861	-0.09	negligible
Copilot	Grok	2.868	0.02	negligible

5. Discussion

Following the study, data collection, and statistical analysis (including ANOVA, t-tests, and the Bonferroni correction to control for Type I errors in multiple pairwise comparisons), several key findings are discussed below:

1) This study highlights how generative AI models perform when taking Thai-language multiple-choice examinations on Programmable Logic Controllers (PLC), a key subject in Industrial Electrical Engineering. The results show that ChatGPT achieved the highest average score of 77.27%, while DeepSeek followed closely with 76.73% and demonstrated the lowest standard deviation. Claude also performed well at 74.80%, whereas Gemini, Copilot, and Grok obtained slightly lower mean scores (< 73%), though all remained within the passing grade range (B or B+ according to the Thai university grading scale). However, the grading labels (e.g., B or B+) are used only as an illustrative reference within the Thai academic context and should not be interpreted as a theoretical benchmark for AI evaluation.

2) The statistical analysis provides clear insights. The one-way ANOVA confirmed overall significant differences among the six LLMs ($p = 0.0002$). However, after applying the Bonferroni correction (adjusted $\alpha = 0.0033$), only a subset of pairwise differences remained statistically significant. DeepSeek exhibited statistically significant performance differences compared with Gemini, Copilot, and Grok under the adjusted criterion. However, differences between ChatGPT and the mid-tier LLMs, although significant at $\alpha = 0.05$, did not remain significant after the Bonferroni adjustment. Furthermore, ChatGPT, Claude, and DeepSeek did not differ significantly from one another under the adjusted threshold. This indicates that performance among the upper-tier models was comparable. Therefore, the most statistically robust separation observed in this study was between DeepSeek and the mid-tier LLMs rather than within the top three models. This result suggests that although mean scores differ slightly among the top models, their performance levels remain statistically comparable when strict multiple-comparison control is applied.

3) DeepSeek's stability is noteworthy. It exhibited substantially lower variability across test sets, whereas ChatGPT achieved the highest mean score. In technical assessment contexts, consistency may be as important as peak accuracy. This observation is consistent with findings in clinical education research mentioned in [32], where GPT-4 demonstrated stable cross-sectional performance rather than isolated peaks. Therefore, reliability across repeated test administrations should be considered alongside average accuracy when evaluating AI suitability for educational support. However, the present study did not compute formal reliability coefficients or test-retest metrics; therefore, the observed stability should be interpreted as descriptive consistency rather than operational reliability in high-stakes assessment contexts. Such consistency may nevertheless be relevant in educational benchmarking scenarios where stable responses across multiple equivalent test sets are desirable.

4) When compared with previous research on LLMs, Guillén-Grima et al. [2] reported that GPT-4 achieved scores approaching top human performers in medical entrance examinations, while GPT-3.5 performed at a lower level. Similarly, Chen et al. [4] observed that GPT-4-class systems outperformed resident averages in the U.S. OITE orthopedic exam. In the present PLC study, the findings partially mirror this pattern, with higher-tier models demonstrating stronger performance than mid-tier ones. However, unlike studies such as Massey et al. [3], where residents clearly surpassed both GPT-3.5 and GPT-4, the PLC results indicate that leading LLMs already achieve performance levels consistent with a passing student benchmark in vocational engineering. This suggests that structured, rule-based technical examinations such as PLC multiple-choice tests may be more accessible to LLMs than highly interpretative or comprehensive examinations in other fields, such as healthcare or medicine. However, the cognitive

demands of PLC multiple-choice questions may differ from those of clinical case-based examinations, and therefore cross-domain benchmarking results should be interpreted with caution.

5) Compared with the findings of Koçak et al. [1], the results of this study further reinforce the cross-domain adaptability of advanced LLMs. In the Turkish Medical Specialty Training examination, GPT-4 reached approximately 88.75% accuracy, demonstrating strong multilingual capability. The PLC results in this study confirm that high-performing LLMs can maintain competitive accuracy in Thai-language technical contexts. However, as noted in [35], LLMs may exhibit atypical response patterns in certain technical assessments, sometimes succeeding on complex items while missing more straightforward ones. A similar phenomenon may occur in technical engineering questions where pattern recognition and conceptual recall interact differently across problem types. Therefore, careful validation is necessary before integrating such systems into formal educational settings.

6) Although the PLC questions in this study were written in Thai, many technical terms were presented in English. This bilingual structure may influence model comprehension differently, as language plays an essential role in knowledge representation and interpretation. Similar language-related performance variations have been observed in Japanese medical physics examinations [20]. However, this study did not conduct a separate linguistic analysis to isolate the potential effects of language composition on model performance, and future research may further examine how bilingual technical terminology influences AI evaluation in engineering education. In addition, while the models selected correct answers in many cases, they were not required to provide reasoning processes. In logic-based disciplines such as PLC programming, procedural explanation and structured reasoning are essential components of learning. Therefore, the accuracy of answers alone does not necessarily indicate pedagogical suitability.

7) Previous PLC pedagogy studies emphasized experiential learning, laboratory training kits, and curriculum integration to develop job-ready competencies [20,21,25]. In contrast, the present study evaluates LLM performance using structured multiple-choice examinations. Although MCQs enable controlled benchmarking and statistical comparison, they primarily assess recognition-level knowledge rather than practical competencies such as debugging skills, ladder logic construction, or real-time system implementation. In authentic PLC learning environments, students must demonstrate practical competencies such as system debugging, logic development, and real-time control implementation. Consequently, strong AI performance in theoretical PLC assessments does not necessarily indicate practical PLC competence.

8) Although this study applied rigorous statistical procedures, including ANOVA and t-tests with Bonferroni correction, assessment theory highlights the tension between psychometric modeling and classroom reality [8]. The analysis provides statistically robust comparisons among LLMs; however, ecological validity in authentic engineering contexts remains limited because the findings were not examined within real instructional environments or compared directly with human student performance. Future studies could strengthen external validity by combining statistical benchmarking with classroom-based experiments or laboratory performance assessments. Furthermore, it is important to acknowledge that LLMs evolve rapidly, and their continuous iteration cycles may limit the strict reproducibility of benchmarking results over time.

9) Ethical considerations must also be addressed. As highlighted in [36], strong exam performance by AI systems introduces potential risks related to academic integrity, fairness, and responsible deployment. Students may rely on AI tools to generate answers that do not fully reflect their own understanding, which may challenge the credibility of formal evaluation processes. If integrated without appropriate safeguards, such tools could undermine the validity of academic assessments and distort learning outcomes. Furthermore, unequal access to AI systems may create additional fairness concerns among students. Therefore, AI adoption in engineering education should be guided by institutional policies, transparency measures, and clearly defined usage boundaries, ensuring that AI serves as a supportive learning aid rather than a substitute for genuine student competency.

10) Effect size measures, particularly Eta-squared (η^2) and Cohen's d, provide additional insights that complement the statistical significance tests applied in this study. While ANOVA and t-Tests indicate whether differences among the evaluated LLMs are significant, effect size measures help quantify the magnitude of those differences. In this study, η^2 provided an overall estimate of how strongly model selection influenced examination outcomes, whereas Cohen's d allowed for a clearer interpretation of the practical separation between specific model pairs. Therefore, the inclusion of Eta-squared (η^2) and Cohen's d strengthens the interpretation of the statistical results by linking statistical significance with practical significance. In AI benchmarking studies, effect size measures help clarify whether observed differences represent meaningful performance variation that may influence educational applications, rather than merely reflecting statistical artifacts caused by repeated testing or sample variability.

11) Overall, the findings from this study suggest that large language models (LLMs) can play a constructive and supportive role in technical and vocational education, even in non-English contexts. The primary contribution of this study lies in extending AI benchmarking to Thai-language PLC examinations and demonstrating statistically controlled performance differences, particularly between DeepSeek and mid-tier LLMs. For educators and policymakers, the results emphasize that model selection, evaluation rigor, and contextual adaptation are important considerations. LLMs should be regarded as complementary learning tools that support preparation and formative assessment rather than replacements for traditional engineering instruction. Careful integration strategies are therefore necessary to preserve rigor, fairness, and the integrity of engineering and vocational education systems.

6. Conclusion and Future Work

This study presented a controlled comparative assessment of six large language models (ChatGPT, Claude, DeepSeek, Gemini, Copilot, and Grok) on Thai-language PLC examinations, which represent a localized and domain-specific engineering context. While all evaluated systems demonstrated competency within a passing academic range, statistical analysis revealed that only selected performance differences remained robust after strict adjustment using the Bonferroni correction for multiple comparisons. Effect size measures (Eta-squared and Cohen's *d*) further examined the magnitude of these differences, highlighting their practical importance. In particular, stability across repeated test administrations emerged as a distinguishing characteristic, emphasizing that consistency may be as critical as peak accuracy in technical assessment environments. The findings indicate that moderate differences in mean performance should be interpreted cautiously unless supported by adjusted statistical validation. The results further confirm that benchmarking outcomes depend not only on raw accuracy but also on variability, effect size interpretation, and statistical rigor. By extending evaluation into a non-English industrial engineering domain, this study contributes empirical evidence to ongoing discussions regarding the reliability of LLMs in specialized educational settings.

Future work should expand beyond structured multiple-choice formats to include open-ended reasoning, ladder logic construction tasks, and applied PLC programming exercises. Cross-language replication and longitudinal evaluation are also recommended to examine performance stability over time. Such investigations will provide deeper insight into the practical role of LLMs as complementary tools in engineering and vocational education. Moreover, the approach used in this study can be applied to and further investigated in other Thai-language subjects within engineering education and related fields. Future studies extending the assessment framework to instructional use should examine its psychometric properties, including expert validation (Index of Item-Objective Congruence, IOC), item analysis, and reliability testing, to strengthen methodological rigor and result reliability.

All the Declarations and Statements

Author Contributions Statement

Amnaj Prajong – Conceptualization and Methodology: Proposed research ideas, Constructed the overall framework; Data Curation and Software Implementation: Handled data acquisition, dataset preprocessing, and implementing the research model; Formal Analysis: Performed analysis of experimental results; Writing – Drafted the initial manuscript, contributed to the literature survey, and documented the technical background of the study; Writing – Review and Editing, and Project Management: Reviewed and edited the manuscript, ensured clarity and coherence, and helped coordinate project milestones and deadlines.

Therdpong Daengsi – Conceptualization, Methodology, and Supervision: Proposed research ideas, Constructed the overall framework, and supervised project execution; Performance Evaluation: benchmarked performance against existing methods; Formal Analysis, Visualization, and Statistical Analysis: Performed in-depth analysis of experimental results, prepared performance charts, and ensured the statistical robustness of the evaluation; Writing – Drafted the initial manuscript, contributed to the literature survey, and documented the technical background of the study; Writing – Review and Editing, and Project Management: Reviewed and edited the manuscript, ensured clarity and coherence, and helped coordinate project milestones and deadlines.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

No funding was received for this research.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author, [TD], upon reasonable request.

Ethical Declarations

No human or animal subjects were involved in this research.

Acknowledgments

We gratefully acknowledge the support of Nakhon Sawan Rajabhat University and Rajamangala University of Technology Phra Nakhon.

Declaration of Generative AI in Scholarly Writing

The authors generated the content, ideas, and findings presented in the manuscript. ChatGPT was then used to assist with proofreading, improving, and refining the English language. Finally, the authors reviewed and validated the final version to ensure its accuracy and integrity prior to submission.

Abbreviations

The following abbreviations are used in this manuscript:

AI - Artificial Intelligence
 ANOVA - Analysis of Variance
 CDIO - Conceive–Design–Implement–Operate
 DL - Deep Learning
 GPT - Generative Pre-trained Transformer
 IEC 61131- International Electrotechnical Commission Standard 61131-3
 LLM - Large Language Model
 MCQ - Multiple-Choice Question
 NLP - Natural Language Processing
 PDF - Portable Document Format
 PLC - Programmable Logic Controller
 SCADA - Supervisory Control and Data Acquisition
 UML - Unified Modeling Language
 OITE - Orthopaedic In-Training Examination

Appendix A

Table 8. The examples of the MCQ associated with PLC used in this study.

In Thai	Translated in English
12. ข้อใดหมายถึงภาษาที่ใช้กับ PLC ก. ภาษาบูลีน ข. ภาษาแลดเดอร์ ค. ภาษาบล็อก ง. ถูกทุกข้อ	12. Which of the following refers to the languages used with PLCs? a. Boolean language b. Ladder language c. Block language d. All the above
18. ข้อใดคือลักษณะของ PLC ชนิดโมดูล ก. แหล่งจ่ายไฟฟ้าจะรวมอยู่ในชุดเดียวกันทั้งหมด ข. ส่วนประกอบทั้งหมดสามารถถอดแยกออกจากกันได้ ค. ฟังก์ชันในการทำงานจะทำให้มากกว่า PCL ชนิดโมดูล ง. ไม่มีข้อถูก	18. Which of the following is a characteristic of modular-type PLCs? a. The power supply is integrated into a single unit. b. All components can be detached and separated. c. Its functions are more advanced than modular-type PLCs. d. None of the above
23. ข้อดีของระบบ PLC เมื่อเทียบกับระบบรีเลย์ ก. การบำรุงรักษาบ่อย ข. ลดภาระเรื่องกระแสไฟฟ้า ค. เปลี่ยนแปลงวงจรได้ง่าย ง. เป็นระบบที่ควบคุมได้ง่าย	23. What is the advantage of a PLC system compared to a relay system? a. Requires less maintenance b. Reduces current load issues c. Easy to modify the circuit d. Easy to control the system
31. การลบข้อมูลทำโดยนำชิปไปฉายแสงอัลตราไวโอเล็ต เป็นวิธีการลบข้อมูล ROM ชนิดใด ก. PROM ข. EPROM ค. EEPROM ง. ROM ทุกชนิดสามารถทำได้	31. Erasing data by exposing the chip to ultraviolet light is the method used for which type of ROM? a. PROM b. EPROM c. EEPROM d. All types of ROM
37. เอาต์พุตแบบใดสามารถขับโหลดได้ทั้งแบบไฟฟ้ากระแสตรงและไฟฟ้ากระแสสลับ ก. เอาต์พุตแบบทั่วไป ข. เอาต์พุตแบบทรานซิสเตอร์ ค. เอาต์พุตแบบโซลิดสเตทรีเลย์ ง. เอาต์พุตแบบรีเลย์	37. Which type of output can drive both DC and AC loads? a. General output b. Transistor output c. Solid-state relay output d. Relay output

Table 8. (Continued)

In Thai	Translated in English
42. ข้อใดคือข้อดีของอินพุตแบบไฟฟ้ากระแสสลับ ก. ลดปัญหาแรงดันตกคร่อม ข. ปรับระดับแรงดันได้ง่าย ค. สัญญาณไม่สับสน ง. ตอบสนองเร็ว	42. Which of the following is an advantage of AC input? a. Reduces voltage drop problems b. Voltage level is easy to adjust c. Signal does not fluctuate d. Fast response
52. ข้อใดคืออุปกรณ์อินพุตของ PLC ก. Encoder ข. 7 - Segment ค. Solid State Relay Output ง. Source Type	52. Which of the following is an input device of a PLC? a. Encoder b. 7-segment display c. Solid-state relay output d. Source type
70. เลข 480 ₁₀ มีค่าเท่าใดในเลขฐานสิบหก ก. 2A1 ₁₆ ข. 4A3 ₁₆ ค. 1E0 ₁₆ ง. 1B2 ₁₆	70. What is the hexadecimal equivalent of 480 ₁₀ ? a. 2A1 ₁₆ b. 4A3 ₁₆ c. 1E0 ₁₆ d. 1B2 ₁₆
96. เมื่อเราจำเป็นต้องเลือกใช้งาน PLC ควรคำนึงถึงสิ่งใด ก. จำนวนอินพุต ข. จำนวนเอาต์พุต ค. อุปกรณ์ต่อขยาย ง. ถูกทุกข้อ	96. When selecting a PLC for use, which of the following should be considered? a. Number of inputs b. Number of outputs c. Expansion devices d. All the above
99. DM6144 - DM6599 คือคำสั่งใด ก. ใช้บันทึกปัญหา PLC ข. ใช้เก็บค่าตัวเลขทั้งฐานสิบและฐานสิบหก ค. อ่านได้อย่างเดียว ง. ใช้เซตค่าของ PLC	99. DM6144 – DM6599 refers to which instruction? a. Used to record PLC problems b. Used to store decimal and hexadecimal values c. Read-only d. Used to set PLC values

References

- [1] M. Koçak, A. K. Oğuz, and Z. Akçalı, "The role of artificial intelligence in medical education: an evaluation of Large Language Models (LLMs) on the Turkish Medical Specialty Training Entrance Exam," *BMC Med. Educ.*, vol. 25, p. 609, 2025. doi: 10.1186/s12909-025-07148-0
- [2] F. Guillen-Grima et al., "Evaluating the Efficacy of ChatGPT in Navigating the Spanish Medical Residency Entrance Examination (MIR): Promising Horizons for AI in Clinical Medicine," *Clin. Pract.*, vol. 13, no. 6, pp. 1460–1487, 2023. doi: 10.3390/clinpract13060130
- [3] P. A. Massey, C. Montgomery, and A. S. Zhang, "Comparison of ChatGPT-3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations," *J. Am. Acad. Orthop. Surg.*, vol. 31, no. 23, pp. 1173–1179, 2023. doi: 10.5435/JAAOS-D-23-00396
- [4] C. J. Chen, V. K. Bilollikar, D. VanNest, J. Raphael, and G. Shaffer, "Artificial intelligence in orthopaedic education: A comparative analysis of ChatGPT and Bing AI's Orthopaedic In - Training Examination performance," *Medicine Advances*, vol. 2, no. 3, pp. 284–290, 2024. doi: 10.1002/med4.77
- [5] S. Stankovski, G. Ostojić, S. Tegeltija, M. Stanojević, M. Babić, and X. Zhang, "Generative AI Applications and Tools in Engineering Education," 2024 23rd International Symposium INFOTEH-JAHORINA (INFOTEH), pp. 1–4, 2024. doi: 10.1109/INFOTEH60418.2024.10495941.
- [6] S. M. Ismail, D. R. Rahul, I. Patra, and E. Rezvani, "The impact of formative assessment on learners' achievement and self-regulation," *Language Testing in Asia*, vol. 12, p. 40, 2022. doi: 10.1186/s40468-022-00191-4.
- [7] B. C. Wesolowski, "Validity, reliability, and fairness in classroom tests," in *Applying Model Cornerstone Assessments in K–12 Music: A Research-Supported Approach*, F. Burrack and K. A. Parkes, Eds., Lanham, MD: Rowman & Littlefield, 2018, ch. 5, pp. 83–102.
- [8] G. T. L. Brown, "Products, processes, psychology, and technology: Quo vadis educational assessment?" *Frontiers in Education*, vol. 7, p. 1060633, 2022. doi: 10.3389/educ.2022.1060633.
- [9] C. H. Aguayo-Hernández, A. Sánchez Guerrero, and P. Vázquez-Villegas, "The learning assessment process in higher education: A grounded theory approach," *Education Sciences*, vol. 14, no. 9, p. 984, 2024. doi: 10.3390/educsci14090984.
- [10] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, 2019. doi: 10.18653/v1/p19-1493.
- [11] C. J. Chen, V. K. Bilollikar, D. VanNest, J. Raphael, and G. Shaffer, "Artificial intelligence in orthopaedic education: A comparative analysis of ChatGPT and Bing AI's Orthopaedic In-Training Examination performance," *Medicine Advances*, vol. 2, no. 3, pp. 284–290, 2024. doi: 10.1002/med4.77.
- [12] L. Dolinar, E. Calcina, and E. Novak, "Evaluating Open-Source Large Language Models for Synthetic Non-English Medical Data Generation Using Prompt-Based Techniques," *Informatica*, vol. 49, no. 27, pp. 125–134, 2025. doi: 10.31449/inf.v49i27.7763

- [13] J. P. d. O. Hansen, S. Hansson, and M. Koenig, "Teaching programmable logic controllers: a hands-on and skill-building approach," 2024 IEEE 7th International Conference on Industrial Cyber-Physical Systems (ICPS), pp. 1–6, 2024. doi: 10.1109/ICPS59941.2024.10640000.
- [14] J. Lasso, G. Banks, K. Wing, T. Lybeck, and J. Shovic, "Flipping the Switch: Leveraging Pinball Machines to Teach PLC Programming and Control Systems," 2025 Intermountain Engineering, Technology and Computing (IETC), pp. 1-6, 2025, doi: 10.1109/IETC64455.2025.11039489.
- [15] M. M. Sakanti, A. Efendi, A. Surjanto, M. A. Falahuddin, N. Khakim, and C. B. D. Kuncoro, "Development of PLC Trainer Kit as a Teaching Aid for Electric Motor Practice," 2023 10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), pp. 467-474, 2023. doi: 10.1109/EECSI59885.2023.10295587.
- [16] S. Lu, L. Wang, T. Yang, and Y. Wu, "An Engineer-Oriented Teaching Practice of Industrial Process Control Systems," 2025 IEEE Global Engineering Education Conference (EDUCON), London, United Kingdom, pp. 1–9, 2025. doi: 10.1109/EDUCON62633.2025.11016402.
- [17] P. Kim, V. Chhoeum, S. Heng, and H. Yahoui, "Accreditation of PLC course in Royal University of Phnom Penh to meet Industry 4.0," 2023 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), pp. 360–363, 2023. doi: 10.1109/ECTIDAMTCON57770.2023.10139494.
- [18] S. Pratumsuwan and W. Pongaen, "An embedded PLC development for teaching in mechatronics education," International Journal of Engineering Education, vol. 27, no. 3, pp. 567–574, 2011.
- [19] V. Bencheva, S. Rübmann, M. Tanev, and D. Iliev, "Capacity development through intensive training of higher education teachers for implementation of PLC curricula in South-East Asia," International Journal of Engineering Pedagogy (iJEP), vol. 14, no. 2, pp. 34–49, 2024.
- [20] M. A. Mohd Khairudin, N. A. Ahmad, and M. S. A. Rahman, "Design of portable industrial automation education training kit compatible for IR 4.0," International Journal of Engineering and Technology, vol. 8, no. 1.9, pp. 221–226, 2019.
- [21] K. Kheowsakul, P. Senajit, and S. Maneechay, "Designing and building a basic programming logic controller (PLC) kit for teaching in the programmable logic controller course," Journal of Technical Education and Training, vol. 14, no. 2, pp. 45–54, 2022.
- [22] R. Sakanti, A. H. Putra, and C. Y. Lin, "Development of PLC trainer kit as a teaching aid for electric motor practice," International Journal of Engineering Research and Technology, vol. 12, no. 5, pp. 620–628, 2023.
- [23] B. Vogel-Heuser, C. Legat, and D. Rinker, "Evaluation of a UML-based versus an IEC 61131-3-based software engineering approach for teaching PLC programming," IEEE Transactions on Education, vol. 56, no. 3, pp. 360–366, Aug. 2013.
- [24] B. Vogel-Heuser, C. Legat, and D. Rinker, "Usability evaluation on teaching and applying model-driven object-oriented approaches for PLC software," IEEE Transactions on Industrial Informatics, vol. 8, no. 4, pp. 898–907, Nov. 2012.
- [25] R. DeGuglielmo, M. A. Dempsey, and J. S. Marshall, "Introduce ladder logic and programmable logic controller (PLC) modules for engineering curriculum integration," ASEE Annual Conference and Exposition Proceedings, pp. 1–12, 2020.
- [26] N. Abdullah, M. R. Ismail, and A. A. Rahman, "Programmable logic controller (PLC) automation control learning kit programmed using the CX-programmer software," International Journal of Interactive Mobile Technologies (iJIM), vol. 15, no. 7, pp. 144–156, 2021.
- [27] S. Chumchuen, T. Pankan, and P. Akatimagool, "Experimental-based learning package of automatic control systems combining programmable logic controller and cloud SCADA system," International Journal of Interactive Learning Environments, 2025, doi: (as listed in article).
- [28] Q. T. Tran, H. Yahoui, and R. Siauve, "An interactive approach to teach automation in the training of the industry 4.0," International Journal of Engineering Pedagogy (iJEP), vol. 9, no. 3, pp. 50–62, 2019.
- [29] L. Meng and Y. Dong, "Teaching reform of PLC course based on outcome-based education in smart manufacturing context," International Journal of Emerging Technologies in Learning (iJET), vol. 16, no. 18, pp. 135–148, 2021.
- [30] P. Khieovongphachanh, S. Phommahaxay, and B. Vongphachanh, "Implementation of curricula in Industry 4.0 context: Faculty of Engineering, National University of Laos," International Journal of Engineering Pedagogy (iJEP), vol. 13, no. 1, pp. 89–101, 2023.
- [31] B. Mesnard et al., "Artificial Intelligence: Ready To Pass the European Board Examinations in Urology?," Eur. Urol. Open Sci., vol. 60, pp. 44–46, 2024. doi: 10.1016/j.euros.2024.01.002
- [32] M. Mousavi, S. Shaheei, J. M. Harley, et al., "Performance of generative pre-trained transformers (GPTs) in Certification Examination of the College of Family Physicians of Canada," Fam. Med. Com. Health, vol. 12: e002626, 2024. doi: 10.1136/fmch-2023-002626
- [33] N. Kadoya et al., "Assessing knowledge about medical physics in language-generative AI with large language model: using the medical physicist exam," Radiol. Phys. Technol., vol. 17, pp. 929–937, 2024. doi: 10.1007/s12194-024-00838-2
- [34] A. Sood, N. Mansoor, C. Memmi, M. Lynch, and J. Lynch, "Generative pretrained transformer-4, an artificial intelligence text predictive model, has a high capability for passing novel written radiology exam questions," Int. J. Comput. Assist. Radiol. Surg., vol. 19, pp. 645–653, 2024. doi: 10.1007/s11548-024-03071-9
- [35] B. Sorenson and K. Hanson, "Identifying Generative Artificial Intelligence Chatbot Use on Multiple-Choice, General Chemistry Exams Using Rasch Analysis," J. Chem. Educ., vol. 101, pp. 3216–3223, 2024. doi: 10.1021/acs.jchemed.4c00165
- [36] J. Metsämuuronen, "Generalized Cohen's d for multiple means and polytomous settings," Applied Psychological Measurement, 2026, doi: 10.1177/01466216261416025.

Authors' Profiles



Amnaj Prajong is an Assistant Professor in the Faculty of Agricultural Technology and Industrial Technology, Nakhon Sawan Rajabhat University, NSRU. He received B.S.Tech.Ed. in Electrical Engineering from KMUTNB in 2006. He received M.Eng. in Electronic Engineering from Technology Mahanakorn University in 2010. His research interests include sensor and transducer, automation, programmable logic controller, IoT, IIoT, mobile networks, analog circuit design, wireless communications, and AI.



Therdpong Daengsi, Ph.D., is an Associate Professor in the Faculty of Engineering at Rajamangala University of Technology Phra Nakhon (RMUTP). He received his B.Eng. in Electrical Engineering from KMUTNB in 1997. He later obtained an M.Sc. in Information and Communication Technology from Assumption University in 2008. In 2012, he earned a Ph.D. in Information Technology from KMUTNB. With 19 years of experience in the telecommunications sector, he also worked as an independent academic for a short period before becoming a lecturer. His research interests include QoS/QoE, mobile networks, multimedia communications, cybersecurity, and artificial intelligence.

How to cite this paper

Amnaj Prajong, Therdpong Daengsi, "Comparative Performance Analysis of Generative AI Applications in PLC: An Industrial Electrical Engineering Subject", International Journal of Modern Education and Computer Science(IJMECS), Vol.18, No.4, pp. 25-39, 2026. DOI:10.5815/ijmecs.2026.04.02