MIX: Jurnal Ilmiah Manajemen

Management Scientific Journal
ISSN (Online): 2460-5328, ISSN (Print): 2088-1231
https://publikasi.mercubuana.ac.id/index.php/jurnal_Mix

Comparative Study of Artificial Intelligence Accuracy on Project Team Member Assessment

Poltak Evencus Hutajulu[1)*]; Andreas Rumata Simanjuntak[2)]; Toba Sastrawan Manik[3)]; R Reza El Akbar[4)]

[1)] poltakevencus@kemenperin.go.id, Industrial Chemical Technology Polytechnic, Indonesia
[2)] andreas16@ptki.ac.id, Industrial Chemical Technology Polytechnic, Indonesia
[3)] tobasastrawanmanik@ptki.ac.id, Industrial Chemical Technology Polytechnic, Indonesia
[4)] reza@unsil.ac.id, Siliwangi University, Indonesia
**) Corresponding Author*

ABSTRACT

Objectives*: This study compares five AI platforms (ChatGPT 5.2, Claude Sonnet 4.5, Gemini 3 Pro, GLM 4.7, DeepSeek V3.2) with manual HR assessment in team performance evaluation, examining accuracy and predictive validity by correlating with program success rates.*
Methodology*: Using a within-subjects design, 32 team members managing 619 locations were evaluated by six methods via standardized prompts. Analysis included comparative tests, accuracy metrics, and correlation between performance scores and graduation rates.*
Finding*: All AI platforms scored higher than the HR baseline. Correlation analysis revealed substantial differences in predictive validity. Claude Sonnet 4.5 demonstrated the strongest correlation with graduation rates (ρ = 0.70), followed by GLM 4.7 (ρ = 0.67), accounting for approximately 50-55% of the outcome variability. The HR baseline unexpectedly showed a negative correlation with program success (ρ = -0.49), indicating misalignment between the evaluation criteria and actual effectiveness. DeepSeek V3.2 most closely matched HR scores but showed only moderate predictive validity, while ChatGPT 5.2 showed no meaningful relationship with outcomes.*
Conclusion*: AI platforms differ substantially in predictive validity. Organizations should prioritize outcome-based validation over human agreement when selecting AI for HR decisions and consider recalibrating manual evaluation systems.*

Keywords*: Artificial Intelligence; Performance Assessment; Prediction Accuracy; Comparative Method.*

Submitted: 19-02-2026 Revised: 20-07-2026 Accepted: 22-07-2026

Article Doi:
http://dx.doi.org/10.22441/jurnal_mix.2026.v16i2.014

p-ISSN: 2088-1231
e-ISSN: 2460-5328

INTRODUCTION

Digital transformation has driven the adoption of artificial intelligence (AI) as a decision-support system in various managerial contexts, including human resource evaluation and selection (Dwivedi et al., 2021). The development of generative AI based on large language models (LLMs) such as ChatGPT, Claude, Gemini, and similar platforms open new opportunities for more efficient, scalable performance appraisal automation (Chui et al., 2023). In the education and training sector, AI shows significant potential as a tool in optimizing instructor allocation, competency evaluation, and project team performance assessment *(Chien & Chen, 2020)*. However, differences in mechanisms and characteristics across AI platforms can lead to inconsistent recommendations if not evaluated comparatively (Liang et al., 2023). Research by (Sánchez-Monedero et al., 2020) found that variations in AI output can reach a difference of up to 30% in the evaluation of candidates with similar profiles, while *(Lambrecht & Tucker, 2019)* uncovered the instability of AI algorithm recommendations when faced with the multidimensional decision context that is prevalent in project team performance assessments.

The gap between theoretical expectations and empirical reality is increasingly evident when organizations adopt multiple AI platforms without an adequate understanding of each system's consistency and accuracy in real operations. Most previous studies have focused on AI technical performance, such as prediction accuracy or processing efficiency (Zhang et al., 2022), while comparative studies systematically evaluating the consistency of recommendations across AI platforms using real organizational data remain relatively limited. Based on the previous explanation, the term "accuracy" in this study refers to agreement with the HR baseline. The importance of organizational context-based calibration mechanisms to ensure that AI recommendations align with company performance values and standards (Benbya et al., 2021). However, previous studies have generally relied on simulation data or benchmark datasets that do not fully represent the complexity of managerial dynamics. In addition, the issue of algorithmic bias stemming from non-representative training data has the potential to result in discriminatory decisions in human performance assessment, making comparative validation a methodological urgency before AI is operationally implemented in personnel decision-making systems (Mehrabi et al., 2021).

The assessment of each team member prior to the implementation of project activities is based on several assessments that serve as references, such as communication skills, work experience, and technical skills, which are fundamental competencies in forming an effective project team (Özkan & Yücel, 2024; Ribeiro et al., 2021; Sampaio et al., 2021). In some organizations, projects are carried out by having the personnel department conduct assessments of communication skills, work experience, and technical skills. These three assessments are sufficient to form a team to complete project activities (Chen et al., 2019; Podgórska & Pichlak, 2019). The project organizations in this study were categorized as training service providers. Individual assessments at the end of the project activities were conducted using the variable of project success rate, which was measured by the pass rate of participants based on the Kirkpatrick evaluation model, which emphasizes the measurement of results as an indicator of the effectiveness of training programs and team member performance assessments from the personnel department that adopted a competency-based performance assessment framework (Paul et al., 2024; Sudarmika et al., 2023; Urbancová et al., 2021). This assessment compared personnel assessments with assessments of the five most popular artificial intelligences.

p-ISSN: 2088-1231
e-ISSN: 2460-5328

Artificial Intelligence assessments can achieve a high degree of accuracy when the platform is trained on sufficient, high-quality data; conversely, if the training data is inadequate, the platform's output is likely to exhibit bias and inaccuracies. Nevertheless, AI can still generate quality output without specific training data provided the prompt is clear and incorporates the essential elements of effective prompting, demonstrating that both the input data and the precision of the prompt are indispensable components of the interaction with AI. HR assessments of project teams tend to be subjective when conducted without established guidelines or proper evaluation matrices, whereas AI-based HR evaluations tend to be more objective, as AI assesses without emotional influence. Since project outcomes are shaped by numerous factors such as team competence and training methods, predicting team competence is a unique and compelling area of study, leading to the hypothesis that there is a significant difference between AI-driven assessments and those conducted by humans (specifically HR personnel).

This study aims to evaluate the consistency, accuracy, and comparative validity for project team member performance assessment in the context of training project, consisting of manual human resources division (HR) assessment or as baseline and five generative AI platforms. How do artificial intelligence assessments differ overall from human assessments, and based on these evaluations, which artificial intelligence comes closest to human judgment? These are among the intriguing questions addressed by this study. Generative AI: In this research, use ChatGPT 5.2, Claude Sonnet 4.5, Gemini 3 Pro, GLM 4.7, and DeepSeek V3.2, with platform selection informed by the Artificial Analysis Leaderboard. (Intelligence, 2025). The work experience variable has a positive and significant influence on Competency (Purnomo et al., 2023). To ensure a fair comparison, the Research applied input standardization using structured prompts based on interview results, so that all AI platforms received identical information on individual profiles, experience, and project achievements related to the program objective. This approach allows the differences in performance scores generated to reflect the inherent characteristics of each AI platform assessment. Integrating the dimensions of competency-based assessment, which include communication skills, work experience, technical skills, and project team member performance assessment (Aguinis, 2019).

The Research design used a within-subjects approach with a repeated-measures ANOVA, in which 32 members of the same team were assessed repeatedly using five AI platforms. This design was chosen because it can control individual variability by treating each subject as its own control, while increasing statistical power by reducing variance errors (Field, 2018). The analysis of the AI system's accuracy was carried out by comparing baseline assessment of human resources division using a composite score, average Mean Absolute Error (MAE), and average Pearson correlation. Furthermore, this study not only assesses the suitability of AI with human assessment, but also tests external validity through correlation analysis between performance scores and project success measured from participants' graduation rates at each handle locus using Spearman correlation, thus generating empirical evidence regarding the practical relevance of AI recommendations to outcomes of project performance (Botvinick et al., 2020).

The main contribution of this study is to provide empirical evidence on the variability and consistency of AI recommendations in a real managerial context, using operational data from the Sakata Innovation Center Foundation's Training Organizing Institute (LPD), involving 32 team members and 619 locations. The study each site represented by a designated participant who acted on behalf of their educational institution and took part in a series of training projects. The training sessions were conducted at 22 different locations across four cities and regencies

p-ISSN: 2088-1231
e-ISSN: 2460-5328

in the Eastern Priangan region of West Java, Indonesia. Unlike previous studies that rarely involved more than two AI platforms in a single study design, this study compares five current generative AI platforms simultaneously with validated baseline methods, yielding a more robust and comprehensive comparison. The findings of this study are expected to have practical implications for the development of AI governance in project organizations, especially by informing AI platform selection criteria aligned with operational needs, and by providing methodological guidance for future Research evaluating artificial intelligence-based decision support systems.

The comparative validity of AI assessments serves as a reference point, as AI can provide evaluations that either align with or diverge from HR assessments regarding predicted team competencies and performance; these AI predictions are based on experience, education, and other parameters found in curriculum vitae each team members. Ultimately, this information influences team composition decisions, thereby supporting greater effectiveness in the context of training projects. Furthermore, a comprehensive review is essential when selecting an AI platform to ensure accurate assessments and to serve as a preliminary basis for evaluating team performance.

LITERATURE REVIEW

The assessment criteria for project team members differ across organizations. However, three variables are usually the main assessment criteria, namely communication skills, which are defined as the ability to convey information, expectations, and ideas effectively, including verbal, written, and non-verbal communication to create a common understanding in achieving project objectives *(Özkan & Yücel, 2024; Siddiqui et al., 2024).* Work experience, which represents the knowledge and skills acquired through direct practice and involvement in previous projects relevant to the tasks to be undertaken (Chen et al., 2019; Stewart, 2019), and technical skills that include the ability to implement schedules, assess technical risks, use technology effectively, and adapt to the tools, techniques, and methods necessary for project success *(Özkan & Yücel, 2024).* The personnel division usually evaluates three variables to assess team members before implementing a project, using interviews and portfolio or curriculum vitae assessments (Brook Fischer, 2024; Derry Holt, 2024).

Communication skills are essential for project team members. This skill can be defined as the ability to convey information, expectations, and ideas effectively, through verbal, written, and nonverbal communication, to create a common understanding among team members and achieve project goals *(Özkan & Yücel, 2024; Siddiqui et al., 2024).* The project organization in this study stated that this variable was the most important priority. Work experience is the second assessment conducted to evaluate team members. Work experience can be defined as the process of developing knowledge and skills in work methods for employees through their involvement in carrying out their work (Foster, 2020; Rosalia et al., 2018). Technical skills are an important part of carrying out project activities. These technical abilities are the foundation for a project's success. In this study, technical skills are defined as the ability or knowledge to use specific techniques in performing specific jobs or tasks, including the ability to implement schedules, assess technical risks, use technology effectively, and adapt to the tools, techniques, and methods necessary for project success *(Özkan & Yücel, 2024; PPM School of Management, 2025).*

The development of large language models (LLMs) based on the transformer architecture has encouraged the use of artificial intelligence to automate human resource performance

p-ISSN: 2088-1231
e-ISSN: 2460-5328

assessment by improving natural language understanding and reasoning skills (Brown et al., 2020) Various generative AI platforms, such as ChatGPT, Claude, Gemini, GLM, and DeepSeek, adopt different architectural approaches but fundamentally rely on self-attention and contextual embedding mechanisms for processing textual information (Touvron et al., 2023; Zeng et al., 2023). The selection of the five platforms in this study is based on the Artificial Intelligence Index, which evaluates model performance based on output quality, speed, and cost efficiency (Intelligence, 2025). ChatGPT 5.2 and Claude Opus obtained the highest scores. However, Claude Sonnet 4.5 was chosen as a representation of the Claude ecosystem because of the balance between reasoning capabilities, cost, and latency for operational implementation (Anthropic, 2024). Gemini 3 Pro represents Google's multimodal capabilities and enterprise integration. At the same time, GLM 4.7 and DeepSeek V3.2 were chosen to represent architectural diversity as well as geographic contexts through an efficient bilingual and mixture-of-experts open-source approach (Zeng et al., 2023). The combination of models with an Intelligence Score range of 41–51 allows for a comprehensive representation of the contemporary generative AI landscape. It enhances the generalization of findings in diverse organizational contexts.

This study adopts a Competency-Based Performance Management (CBPM) framework that integrates individual competency assessments and actual performance outcomes to produce a comprehensive and predictive evaluation of organizational success (Aguinis, 2019). The variables of communication ability, work experience, and technical skills were chosen because they were consistently proven to be strong predictors of job performance, while staffing performance assessments were used to integrate behavioral dimensions and outcomes in the context of multi-location work (Ones et al., 2021). The operational definition of variables refers to the instructional communication literature, work experience based on the relevance and complexity of the task, and technical competencies observed through performance and problem-solving (Ford et al., 2018; Hargie, 2021). The within-subjects design with Repeated Measures ANOVA was chosen for its ability to control individual differences, improve estimation accuracy, and strengthen statistical power over between-subjects designs (Maxwell et al., 2018). Each subject was assessed by all AI platforms so that the difference in scores reflected the effect of the AI output, not individual characteristics (Field, 2018b). The sphericity assumption test is carried out using Mauchly's test, with Greenhouse–Geisser or Huynh–Feldt corrections applied when assumptions are violated to prevent Type I inflation errors (Haverkamp & Beauducel, 2017). Follow-up analysis was performed through paired-sample t-tests with Bonferroni correction. At the same time, effect magnitude was reported using partial eta squared and Cohen's d to assess the practical significance of the score difference (Tabachnick & Fidell, 2019).

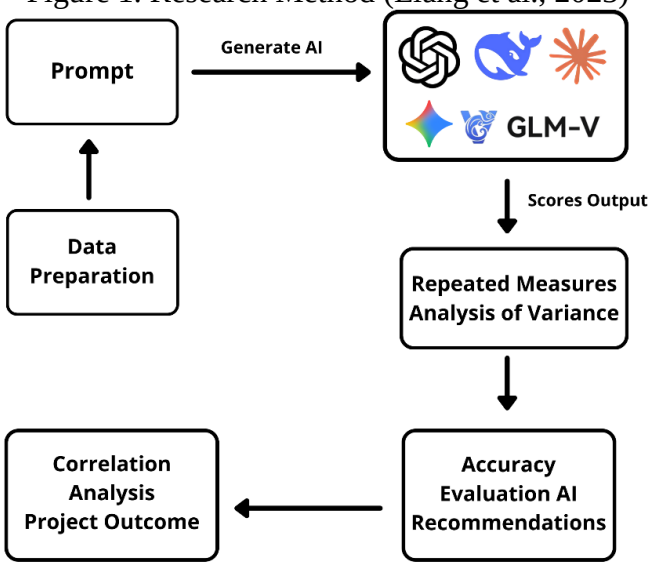
The validity of performance appraisal in this study is focused on criterion-related validity, namely the ability of evaluation methods to predict relevant organizational outcomes *(Cascio & Aguinis, 2019)*. The participant's graduation rate is used as a criterion variable because it represents the success of the program objectively and in line with the results-level evaluation approach in training evaluation *(Kirkpatrick & Kirkpatrick, 2016)*. Spearman correlation is used as a key indicator of external validity due to its non-parametric and robust nature to outliers as well as non-linear relationships, with Pearson correlation reported as a complement to linear relationships. This approach asserts that conformity with human judgment does not always guarantee predictive validity, given that human judgments are prone to bias and do not necessarily reflect actual performance (Kahneman et al., 2021).

p-ISSN: 2088-1231
e-ISSN: 2460-5328

This study is not positioned as an attempt to replace human judgment but rather as an analytical reference on how AI can support HR in evaluating team performance; its findings directly relate to essential issues of AI ethics, explainability, and governance (Tambe, Cappelli, & Yakubovich, 2019). First, the finding that all five platforms systematically rated team members higher than the HR baseline, combined with their near-zero agreement with HR regarding communication skills, highlights risks associated with algorithmic bias and fairness; AI systems may disadvantage individuals whose strengths are interpersonal in nature or difficult to infer from data (Köchling & Wehner, 2020). Consequently, fairness cannot be assumed; it must be explicitly assessed. Second, while the study presents the scores generated by these platforms, it does not reveal the underlying penalties applied. From an Explainable AI (XAI) perspective, high predictive validity is necessary but insufficient, as HR managers require transparent, interpretable justifications before acting on algorithmic recommendations (Adadi & Berrada, 2018; Barredo Arrieta et al., 2020). Third, the substantial differences observed among platforms along with the divergence between HR agreement and predictive validity for actual outcomes underscore the need for AI governance. Models should be validated, audited, and periodically re-audited against actual performance results, while maintaining human involvement in the process (a "human-in-the-loop" approach) (Charlwood & Guenole, 2022). Overall, these results reposition AI as a complementary, continuously validated decision-support tool that reinforces, rather than replaces, the HR function, while offering practical guidance for managers on selecting and integrating AI platforms for team performance assessment.

METHOD

Figure 1. Research Method (Liang et al., 2023)



The study follows a systematic flow consisting of five main stages, as shown in Figure 1, which is designed to ensure consistency, objectivity, and comparative validity in the performance evaluation of project team members using multiple AI platforms. The initial stage of data preparation involves collecting and preparing data through structured interviews with 32 project team members.

The project team consisted of 32 members, representing the entire KKA facilitator team at LPD SIC rather than a sample. LPD SIC was designated as such by a decision of the

p-ISSN: 2088-1231
e-ISSN: 2460-5328

Indonesian Ministry of Education, Culture, Research, and Technology. This research focuses on analyzing the outputs generated by each AI model to determine their accuracy and alignment with HR assessments. A census sampling technique (sampling the entire population) was employed; thus, the figure of 32 represents the full scope of the units under study rather than a limitation. Although the number of subjects was relatively small (n = 32), this does not undermine the findings. Small samples carry the risk of failing to detect a genuine effect, rather than producing a false positive; consequently, if an effect is found to be significant despite a small sample size, it indicates a particularly robust effect. Despite the small sample size, differences between platforms remained significant in the RM-ANOVA ($p < 0.001$) with large effect sizes ($\eta^2G = 0.15–0.47$), and correlations with actual outcomes were strong for Claude Sonnet 4.5 ($r = 0.74$) and GLM 4.7 ($r = 0.72$). Small samples risk missing genuine effects rather than generating false significance; since these effects were detected as strong and significant, the findings are not compromised by the sample size (Lakens, 2022; Cohen, 1992). To further substantiate the results, the study includes effect size reporting along with 95% confidence intervals and Bayes factors and a bootstrap robustness test, while explicitly acknowledging that the study's scope (limited to a single organization) presents an opportunity for future cross-organizational replication.

The interview uses four key questions designed to dig into comprehensive information on how to deliver Material to participants with diverse backgrounds, professional experience relevant to the role in the program, the application of technical skills to help participants achieve targets, and the management of work responsibilities, including the number of loci addressed. The data collected included demographic information in the form of names, education, workload, the number of shade locations, program outcomes in the form of participants' graduation rates, and transcripts of interview answers. The data is then tabulated in a structured format using the pandas DataFrame to facilitate integration with the template prompt at a later stage. In addition, the internally validated manual HR assessment baseline data is also compiled as the benchmark standard for AI accuracy evaluation.

When the data is ready, a structured prompt that is standardized for all AI platforms is developed according to the principles of effective prompt engineering (White et al., 2023). The prompt protocol employed comprises ten constituent elements: context (providing background information to the AI); instructions (defining the task the AI must perform); input data (specifically, files such as curriculum vitae team member and others); examples of expected results; technical parameters; output format; language style; constraints; the AI's role; and the iteration process specified at the end of the prompt. Standardized structured prompts are consistently generated into five generative AI platforms such as ChatGPT 5.2, Claude Sonnet 4.5, Gemini 3 Pro, GLM 4.7 (GLM-V), and DeepSeek V3.2. Each AI platform receives an identical prompt that contains the same information for each subject. Hence, the difference in scores generated reflects the inherent characteristics of the AI algorithm, not the variation in the quality of the data input. Outputs from each platform are extracted in a score format for communication ability, work experience, technical skills, and staffing performance assessments. The extraction process is carried out using regular expressions through the re library in Python to ensure accurate and consistent parsing (Vallat, 2018). The extraction results are then tabulated in a pandas DataFrame with a dataframe structure.

Furthermore, an inferential analysis was carried out using ANOVA Repeated Measures to test whether there was a difference in average performance scores between the six evaluation methods (five AI platforms and one human resources division assessment). All analysis was done using Python (Jupyter Notebook). Prior to the ANOVA test, the normality assumption

p-ISSN: 2088-1231
e-ISSN: 2460-5328

was checked using the Shapiro–Wilk test, while the sphericity assumption was tested using Mauchly's test. If the sphericity assumption is violated ($p < 0.05$), the Greenhouse–Geisser correction is applied to the degrees of freedom to obtain a more conservative F estimate. The RM-ANOVA model is composed with composite score as a dependent variable, evaluation method (6 levels) as a within-subjects factor, and SubjectID as a subject identifier. The null hypothesis tested is expressed in Equation (1) with a significance level of $\alpha = 0.05$.

At the same time, the magnitude of the effect is reported using Generalized eta squared (η^2G) with the interpretation criteria in Equation (2). If the omnibus test shows significant results, the analysis is continued with post hoc pairwise comparisons using paired-sample t-tests with Bonferroni correction, as shown in Equation (3). The magnitude of the effect for each pair comparison is calculated using Cohen's d in Equation (4) with the interpretation in Equation (5) (Norouzian & Plonsky, 2018; Tabachnick & Fidell, 2019b).

$$H^0: \mu_{HR} = \mu_{ChatGPT} = \mu_{Claude} = \mu_{gemini} = \mu_{GLM} = \mu_{DeepSeek} \quad \text{..... (1)}$$

$$\eta^2G = 0{,}01(small), \eta^2G = 0{,}06\ (medium), \eta^2G = 0{,}14\ (high) \quad \text{..... (2)}$$

$$adjusted = 0{,}05/15 = 0{,}0033 \quad \text{..... (3)}$$

$$d = \frac{(M^1 - M^2)}{SD\ pooled} \quad \text{..... (4)}$$

d = 0,2 (Small), d = 0,5 (medium), d = 0,8 (high)..... (5)

The evaluation of the accuracy and validity of AI platform recommendations was carried out quantitatively to assess the suitability of AI assessments with the staffing (HR) baseline and its relevance to the success of the real project. The accuracy of AI scores on HR was analyzed using composite score agreement as a measure of categorical consistency (Pedregosa et al., 2011). as well as Mean Absolute Error (MAE) to measure the absolute deviation of AI scores from HR assessments, where lower MAE scores indicate higher prediction accuracy. In addition, the relational consistency between AI scores and HR baseline was evaluated using Pearson correlation to assess the strength and direction of linear relationships, with a significant positive correlation ($p < 0.05$) indicating alignment of scoring patterns (Virtanen et al., 2020).

Furthermore, correlation analysis project outcome with the external validity of the assessment method was tested through the Spearman correlation (ρ) between the performance score and the graduation rate of participants at each handle locus as an indicator of real project success, considering its non-parametric and robust nature to non-normality and non-linear relationships. Significant positive correlations indicate that the assessment method has predictive capabilities that are practically relevant to program outcomes, with the interpretation of relationship strength referring to weak, medium, and strong classifications. All analyses were conducted at a significance level of $\alpha = 0.05$ with a 95% confidence interval and were transparently documented to ensure repeatability and reliability of the study results (Kluyver et al., 2016).

p-ISSN: 2088-1231
e-ISSN: 2460-5328

RESULTS AND DISCUSSION

Results

This study analyzed data from 32 team members managing 619 coding and AI program locations. The predominance of the elementary (50%) and junior secondary (36.36%) levels indicates a primary focus on basic education. Participant variability (CV = 24.73%) with a mean of 28.14 ± 6.96 reflects moderate heterogeneity in managerial complexity. Descriptive statistics demonstrate a consistent pattern across the three assessment variables. Generative AI platforms produced higher mean scores and greater scoring consistency (SD = 1.01–3.08) than the HR baseline (SD = 2.17–2.46). In communication competence, Gemini 3 Pro achieved the highest mean (M = 92.22, SD = 1.01), followed by GLM 4.7 (M = 91.53, SD = 1.32) and ChatGPT 5.2 (M = 91.03, SD = 1.45). In contrast, the HR baseline recorded the lowest score (M = 88.06, SD = 2.17), with nearly twice the variability of the AI-based assessments. For work experience, Gemini 3 Pro (M = 90.97, SD = 2.90) and GLM 4.7 (M = 90.84, SD = 3.08) again outperformed other methods, with overall dispersion ranging from SD = 2.33 to 3.08, indicating greater individual heterogeneity. A similar pattern emerged in technical skills, with GLM 4.7 scoring highest (M = 91.94, SD = 2.72), followed by Gemini 3 Pro (M = 91.69, SD = 1.99) and ChatGPT 5.2 (M = 91.31, SD = 2.55). DeepSeek V3.2 consistently generated the lowest scores among AI systems, yet remained above the HR baseline. Overall, AI-based evaluations exhibited higher central tendency and comparatively more stable scoring patterns than the conventional HR assessment.

In the testing phase, Repeated Measures ANOVA showed significant differences between platforms in all variables ($p < 0.001$) with moderate to large effect sizes ($\eta^2G = 0.15–0.47$). Communication ability showed the strongest effect, $F(5,155) = 45.85$, $\eta^2G = 0.47$. The Greenhouse-Geisser correction-maintained significance despite the violation of the sphericity assumption, as shown in Table 1.

Table 1. Repeated Measures ANOVA

Variabel	SS	df	MS	F	p-value (uncorr)	p-value (GG-corr)	η^2G	ε	Sphericity	W	p (Mauchly)
Communication Skills											
Platforms	337,81	5	67,56	45,85	< 0,001	< 0,001	0,47	0,53	FALSE	0,11	< 0,001
Error	228,37	155	1,47	—	—	—	—	—	—	—	—
Work Experience											
Platforms	252,60	5	50,52	36,70	< 0,001	< 0,001	0,15	0,71	FALSE	0,30	0,001
Error	213,40	155	1,38	—	—	—	—	—	—	—	—
Technical Skills											
Platforms	345,78	5	69,16	53,85	< 0,001	< 0,001	0,22	0,67	FALSE	0,34	0,005
Error	199,06	155	1,28	—	—	—	—	—	—	—	—

Post hoc analysis with Bonferroni correction identified 11 of 15 different platform pairs as significant ($\alpha = 0.05$). All AIs were significantly different from the HR baseline ($p < 0.001$, $d = 0.77–1.46$), confirming the tendency for AIs to give higher ratings. DeepSeek V3.2 vs GLM 4.7 ($d = -1.07$) and DeepSeek V3.2 vs Gemini 3 Pro ($d = -1.12$) showed the largest differences between AIs.

The robustness of the findings is supported by multiple layers of verification accompanying the analysis. Assumption violations were explicitly tested: the Shapiro-Wilk test

p-ISSN: 2088-1231
e-ISSN: 2460-5328

revealed that 12 of the 18 combinations were not normally distributed, and Mauchly's test confirmed a violation of sphericity ($p < 0.05$), necessitating the use of the Greenhouse-Geisser correction for the RM-ANOVA results; even after this conservative correction ($\varepsilon = 0.53–0.71$), all three variables remained significant ($p < 0.001$), indicating that the conclusions do not hinge on the violated assumptions. Correlations were also calculated using two methods: Pearson (parametric) and Spearman (rank-based non-parametric), yielding consistent patterns (e.g., Claude Sonnet 4.5: $r = 0.74/\rho = 0.70$; GLM 4.7: $r = 0.72/\rho = 0.67$; HR: $-0.49/-0.49$), demonstrating that the findings were not driven by outliers or the shape of the distribution. Furthermore, post-hoc tests employed the Bonferroni correction to control for inflation due to multiple comparisons and were reinforced by a very large Bayes Factor ($BF10 > 10^9$), evidence independent of the frequentist framework, while the cross-platform accuracy rankings remained identical across all error and fit metrics (MAE, RMSE, and R^2). Overall, this consistency across methods, assumptions, and metrics demonstrates the robustness of the study's conclusions.

Accuracy evaluation using the HR baseline with MAE, RMSE, and Pearson/Spearman correlation reveals a pattern different from the post hoc findings. DeepSeek V3.2 is closest to the HR baseline (Composite Score = 0.69; MAE = 1.76; RMSE = 2.42), reflecting its conservative assessment character. Claude Sonnet 4.5 ranked second (Composite Score = 0.57; MAE = 2.36; $r = 0.52$) with a good balance between absolute deviation and relational consistency. GLM 4.7 showed the highest correlation with the HR baseline ($r = 0.65$; $\rho = 0.65$), but with greater absolute deviation (MAE = 3.40), while Gemini 3 Pro and ChatGPT 5.2 showed the lowest accuracy compared to conventional HR assessments.

To test the external validity of the assessment method, a correlation analysis was conducted between team performance scores and participant pass rates as objective indicators of project success. The pass rate distribution showed bimodal characteristics, with an average of 95.3% (SD = 3.2%), ranging from 88–100%, and 65.6% of team members achieving a pass rate of ≥95%. Sufficient variability allowed the detection of meaningful correlational relationships, with analysis using Pearson's correlation for linear relationships and Spearman's for monotonic relationships, which are robust to outliers and nonlinearity. The analysis results show dramatic differences in predictive validity across methods (Table 1). Claude Sonnet 4.5 shows the highest external validity, with a strong, highly significant correlation ($r = 0.74$, $\rho = 0.70$, $p < 0.001$), explaining approximately 55% of the variability in participant graduation rates. This platform consistently identified team members who contributed to better program outcomes. GLM 4.7 showed nearly equivalent validity ($r = 0.72$, $\rho = 0.67$, $p < 0.001$) with predictive power of approximately 52% of outcome variability. DeepSeek V3.2, despite achieving the highest accuracy against the HR baseline, showed only a significant Spearman correlation ($\rho = 0.58$, $p = 0.005$) but not a significant Pearson correlation ($r = 0.27$, $p = 0.140$), suggesting a non-linear relationship or the Influence of outliers in its assessment. Gemini 3 Pro showed weak predictive validity with non-robust correlations ($r = 0.41$, $p = 0.020$; $\rho = 0.32$, $p = 0.080$), while ChatGPT 5.2 showed no significant relationship with program outcomes ($r = -0.20$, $p = 0.270$; $\rho = -0.19$, $p = 0.300$).

The next finding shows that the HR baseline has a significant negative correlation with the graduation rate, $r = -0.49$, $\rho = -0.49$, $p < 0.001$, as shown in Table 2. These results indicate a fundamental mismatch between conventional manual evaluation criteria and actual effectiveness in producing program outcomes. In other words, traditional HR assessment systems tend to reflect constructs that differ from or even oppose the competencies empirically associated with project success. These findings underscore the urgency of recalibrating

p-ISSN: 2088-1231
e-ISSN: 2460-5328

organizational evaluation frameworks to better align with outcome-based performance indicators.

Table 2. Correlation Test of Performance Appraisal with Project Success

Platform	Correlation Variables	Pearson_r	P_value Pearson	Spearman_ ρ	P_value Spearman
Claude Sonnet 4.5	Graduation Rate	0.74	0	0.7	0
Deep Seek V3.2	Graduation Rate	0.27	0.14	0.58	0
GLM 4.7	Graduation Rate	0.72	0	0.67	0
Gemini 3 Pro	Graduation Rate	0.41	0.02	0.32	0.08
HR (Baseline)	Graduation Rate	-0.49	0	-0.49	0
Chat GPT 5.2	Graduation Rate	-0.2	0.27	-0.19	0.3

These findings confirm that proximity to the HR baseline assessment does not guarantee predictive validity for actual outcomes. DeepSeek V3.2's closeness to the HR baseline actually resulted in lower predictive validity than Claude Sonnet 4.5, which had greater deviation but the strongest outcome correlation. This shows that AI evaluation systems should be prioritized based on criterion-related validity rather than simply conformity with conventional human assessment practices. Claude Sonnet 4.5 and GLM 4.7 show the highest potential as tools to support managerial decisions based on actual results. At the same time, the negative correlation in the HR baseline underscores the urgency of recalibrating organizational evaluation systems to align with competencies that actually drive project success.

Differences in predictive validity across platforms are not coincidental but can be explained by three mechanisms. First, each model is built using distinct architectures and training data, leading to variations in how they weigh indicators of competence. Models with superior reasoning capabilities are better equipped to process complex interview information into assessments that reflect actual performance differences. Second, training processes reliant on human feedback tend to result in models that assign uniformly high scores; models that operate within an overly narrow scoring range lose their ability to discriminate, thereby reducing predictive validity. This explains why alignment with HR ratings does not guarantee accuracy regarding real-world outcomes. Third, predictive accuracy depends on how well a model's assessment criteria align with the competencies that truly determine program success. Models emphasizing fundamental aspects (such as Claude Sonnet 4.5 and GLM 4.7) show higher correlations with pass rates, whereas models influenced by superficial factors yield low or even negative validity. Thus, variations in predictive validity are predictable consequences of differences in architecture, calibration, and criterion alignment across models, rather than merely isolated empirical findings.

The theoretical contribution of this study lies in the observed differences between AI-generated assessments and HR evaluations; the fact that some AI results closely approximate HR assessments demonstrate that different AI systems employ distinct algorithms, leading to varying outputs even when identical prompts are used. These algorithms can be likened to the stages of creating a product, such as brewing coffee, where different methods yield slightly different flavors; this analogy illustrates why AI outputs vary. The study's novelty lies in its focus on a previously unexamined scenario: the assessment of project teams in the training sector. It compares the accuracy of five AI systems against HR evaluations, with accuracy defined as the level of agreement with the HR baseline.

p-ISSN: 2088-1231
e-ISSN: 2460-5328

Discussion

Overall, the results of Repeated Measures ANOVA show that the evaluation method has a significant effect on communication skills, work experience, and technical skills ($p < 0.001$), with a moderate to large effect size ($\eta^2G = 0.15–0.47$), so that the differences between platforms are not only statistically significant but also practically meaningful. The largest effect appeared in communication skills ($F(5.155) = 45.85$; $\eta^2G = 0.47$), followed by technical skills ($F(5.155) = 53.85$; $\eta^2G = 0.22$) and work experience ($F(5.155) = 36.70$; $\eta^2G = 0.15$). Although the sphericity assumption was violated (Mauchly $p < 0.05$), the Greenhouse–Geisser correction ($\varepsilon = 0.53–0.71$) did not alter the significance of the results, confirming the robustness of the findings. Post hoc tests showed that 73.33% of platform pairs (11 out of 15) differed significantly ($\alpha = 0.05$), with all AIs consistently scoring higher than the baseline HR ($d = 0.77–1.46$). However, the difference in scores did not necessarily reflect the quality of the evaluation. DeepSeek V3.2 is indeed the closest to HR in terms of accuracy (MAE = 1.76; RMSE = 2.42; composite score = 0.69), but its predictive validity is limited ($\rho = 0.58$; $p = 0.005$; $r = 0.27$; $p = 0.140$). In contrast, Claude Sonnet 4.5 ($r = 0.74$; $\rho = 0.70$) and GLM 4.7 ($r = 0.72$; $\rho = 0.67$) show the strongest external validity, explaining approximately 55% and 52% of the variability in graduation rates. These findings are crucial because baseline HR was significantly negatively correlated with graduation ($r = −0.49$; $\rho = −0.49$), confirming that the superiority of an evaluation method should be determined by its validity in relation to actual outcomes, not merely its proximity to conventional assessments. Other opportunities for further research could involve various organizations and industries, as well as the use of additional performance indicators and the conduct of cross-validation studies.

This study has several limitations. First, it was conducted within a single organization (LPD SIC) and sector (educational facilitators); although the entire population ($n = 32$) was included, this single-context scope limits generalizability to other organizations and sectors. Second, each AI platform performed the assessment only once; consequently, consistency across repetitions (prompt stability/test-retest) and inter-rater reliability (e.g., ICC) were not measured. Third, the study evaluated AI-generated scores rather than the underlying reasoning, meaning the aspect of explainability was not tested. Fourth, the scores were concentrated within a narrow range near the upper limit (85–95), potentially creating a ceiling effect and limiting the range of variation. Fifth, numerous factors beyond facilitator quality influence the outcome measure (participant pass rate); thus, the observed relationship is associative rather than causal. Finally, the AI results are tied to specific model versions at a specific point in time (e.g., GLM 4.7, ChatGPT 5.2) and are subject to change following model updates.

CONCLUSION

This study found significant differences between the six evaluation methods ($\eta^2G = 0.15–0.47$), with all AI platforms consistently scoring higher than the HR baseline. However, score differences did not always reflect better evaluation quality in terms of actual outcomes. Accuracy against HR and predictive validity were two distinct dimensions: DeepSeek V3.2 was most accurate in replicating HR scores (MAE = 1.76), but had limited predictive validity. Conversely, Claude Sonnet 4.5 and GLM 4.7 showed the highest predictive ability for graduation rates ($r \geq 0.72$), despite deviating more from HR scores. A significant negative correlation between HR and program outcomes ($r = −0.49$) confirms that conventional systems are not suitable as a single gold standard. Therefore, an outcome-based evaluation approach is recommended, with Claude Sonnet 4.5 and GLM 4.7 as the most promising candidates for data-

p-ISSN: 2088-1231
e-ISSN: 2460-5328

driven managerial decision-making. Further research is needed to validate the findings using larger samples, a longitudinal design, and cross-organizational testing.

The findings of this study apply under specific conditions. The recommended outcome-based evaluation approach is applicable only when objective, measurable indicators of operational success (such as pass rates) are available; in contexts lacking clear outcome metrics, this approach is difficult to replicate. The findings are also most relevant to small, limited project teams rather than large-scale populations and to knowledge-based work, such as educational facilitation. Furthermore, the AI's reliability is limited to competencies for which information exists in the data (e.g., work experience and technical skills), whereas interpersonal competencies such as communication skills, which in this study showed almost no correlation with HR assessments, fall outside that scope. Institutional and cultural contexts (specifically the LPD in Indonesia) also serve as boundaries for interpreting the results.

Several future directions could strengthen these findings. First, cross-organizational and cross-sector replication using more diverse samples to test generalizability. Second, testing AI reliability through repeated assessments (test-retest/prompt stability) and calculating inter-rater reliability (ICC) to ensure score stability. Third, integrating explainable AI approaches to assess not only score accuracy but also the transparency of the underlying rationale. Fourth, analyzing fairness across subgroups (e.g., gender, age, tenure) to ensure AI assessments are non-discriminatory. Fifth, conducting longitudinal studies to examine whether individuals predicted by AI to excel actually perform well over time. Finally, exploring hybrid human-AI evaluation models that combine outcome-based AI assessments with managers' contextual considerations.

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Anthropic. (2024). Claude 4.5 model card. https://www.anthropic.com/claude

Barredo A.A, Díaz-Rodríguez A, Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S. , Gil-Lopez S., Molina D., Benjamins R., Chatila R., Herrera F. (2020). Explainable Artificial Intelligence (XAI): : Concepts, taxonomies, opportunities and challenges toward responsible AI. https://doi.org/10.1016/j.inffus.2019.12.012

Benbya, H., Davenport, T. H., & Pachidi, S. (2021). Special issue editorial: Artificial intelligence in organizations: Current state and future opportunities. MIS Quarterly Executive, 20(4), ix–xxi. https://doi.org/10.17705/2msqe.00054

Brook Fischer. (2024). Candidate evaluation: How to confidently decide who to hire. Homerun. https://www.homerun.co/articles/candidate-evaluations

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Cascio, W. F., & Aguinis, H. (2019). Applied psychology in talent management (8th ed.). SAGE Publications.

p-ISSN: 2088-1231
e-ISSN: 2460-5328

Charlwood, A., & Guenole, N. (2022). Can HR adapt to the paradoxes of artificial intelligence? Human Resource Management Journal, 32(4), 729–742. https://doi.org/10.1111/1748-8583.12433

Chen, T., Fu, M., Liu, R., Xu, X., Zhou, S., & Liu, B. (2019). How do project management competencies change within the project management career model in large Chinese construction companies? International Journal of Project Management, 37(3), 485–500. https://doi.org/10.1016/j.ijproman.2018.12.002

Chien, C.-F., & Chen, L.-F. (2020). Using rough set theory to recruit and retain high-potential talents for semiconductor manufacturing. IEEE Transactions on Semiconductor Manufacturing, 33(4), 529–539.

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023). The economic potential of generative AI: The next productivity frontier. McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

Derry Holt. (2024). Recruitment assessment: In-depth guide for agencies 2024. OneUpSales. https://oneupsales.com/blog/recruitment-assessment

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V, Janssen, M., Jones, P., Kumar, A., Kar, A., Kizgin, H., Kronemann, B., Lal, B., & Williams, M. D. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management, 57, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Foster, B. (2020). Manajemen sumber daya manusia. Alfabeta.

Intelligence, S. H.-C. A. (2025). Artificial Intelligence Index Report 2025. Stanford University.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: A flaw in human judgment. Little, Brown Spark.

Köchling, J., & Wehner, M. (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research, 13(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2016). Kirkpatrick's Four Levels of Training Evaluation. ATD Press.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). Jupyter Notebooks—A publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas, 87–90. https://doi.org/10.3233/978-1-61499-649-1-87

Lakens, D. (2022). Sample size justification. Collabra: Psychology, 8(1), Article 33267. https://doi.org/10.1525/collabra.33267

p-ISSN: 2088-1231
e-ISSN: 2460-5328

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Management Science, 65(7), 2966–2981. https://doi.org/10.1287/mnsc.2018.3093

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Yian, Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., … Koreeda, Y. (2023). Holistic Evaluation of Language Models. http://arxiv.org/abs/2211.09110

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6). https://doi.org/10.1145/3457607

Özkan, N., & Yücel, İ. (2024). Development and validation of the project manager skills scale (PMSS): An empirical approach. Heliyon, 10(2), e24075. https://doi.org/10.1016/j.heliyon.2024.e24075

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

PPM School of Management. (2025). Management skill: Pengertian, fungsi, jenis dan contoh di dunia kerja. https://www.ppmschool.ac.id/management-skill/

Purnomo V.M, F., Perizade Badia & Syapril Yuliani. (2023). The Influence of Training and Work Experience on Performance with Competence as an Intervening Variable in The Head Of The Technical Implementing Unit Pt. Indonesian Railways. International Journal of Social Service and Research (IJSSR), Vol. 03, No. 08, e-ISSN: 2807-8691. https://doi.org/10.46799/ijssr.v3i8.499

Rosalia, D., Mintarti, S., & Fitriadi. (2018). Pengaruh pelatihan dan pengalaman kerja terhadap produktivitas kerja karyawan Jaya Sakti Sentosa. Jurnal Ekonomi Manajemen Akuntansi, 2(2), 1–15.

Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020). What does it mean to "solve" the problem of discrimination in hiring? Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 458–468. https://doi.org/10.1145/3351095.3372849

Siddiqui, J., Kosmala, K., Marszałek, A., & Rudawska, E. (2024). Communication in project team management: Identification of research gaps and direction for future research. European Research Studies Journal, 27(Special Issue), 732–751. https://doi.org/10.35808/ersj/3746

Stewart, G. L. (2019). Team member skill and ability levels, personality traits, background and experience in workplace team performance: A meta-analysis. Team Performance Management, 25(1/2), 3–29.

p-ISSN: 2088-1231
e-ISSN: 2460-5328

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. California Management Review, 61(4), 15–42. https://doi.org/10.1177/0008125619867910

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. http://arxiv.org/abs/2307.09288

Vallat, R. (2018). Pingouin: Statistics in Python. Journal of Open Source Software, 3(31), 1026. https://doi.org/10.21105/joss.01026

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., & Larson, E. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. http://arxiv.org/abs/2302.11382

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., Tam, W. L., Ma, Z., Xue, Y., Zhai, J., Chen, W., Zhang, P., Dong, Y., & Tang, J. (2023). GLM-130B: An Open Bilingual Pre-trained Model. http://arxiv.org/abs/2210.02414

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., & Perrault, R. (2022). The AI Index 2022 Annual Report. Stanford Institute for Human-Centered AI, Stanford University.