DOI:10.32703/2617-9040-2025-46-11

UDC 004.8

Oleksii Nesterenko*[1]

[1]Postgraduate student, Automation and Computer-Integrated Transport Technologies Department, National Transport University, M. Omelyanovicha-Pavlenko St., 1, Kyiv, 01010, Ukraine. ORCID: https://orcid.org/0009-0008-7249-3189.

***Corresponding author:** nesterenko_ob@gsuite.duit.edu.ua.

Methodological Aspects and Models for Assessing the Effectiveness of Artificial Intelligence in Project Management

Rapid integration of artificial intelligence (AI) into project management offers significant potential to improve productivity through data automation, performance monitoring and schedule optimization. However, challenges such as "effective inefficiency" and the variability of AI model output complicate the assessment of its effectiveness. This article analyses the methodological aspects of evaluating AI effectiveness in project management, classifies existing methods (benchmarks, explainable AI, mutual information, psychometrics), identifies key challenges (biases, lack of standards, ethical constraints), and proposes novel metrics- indicator of new competency activation (INCA), novelty coefficient in AI-Driven Project Management (NCAPM) and dynamic assessment of transition to new efficiency enabled by AI (DATNE) to measure innovation. The potential of these approaches for transport infrastructure projects is indicated, where AI allows for the creation of fundamentally new opportunities in planning, service forecasting, and resource optimisation. Future directions include hybrid metrics and integration with decision support systems. The study underscores the need for interdisciplinary approaches to adapt AI evaluation to resource constrained project management environments.

Keywords: *model, machine learning, benchmark, performance assessment, methodology, cognitive models, systems analysis, project management, artificial intelligence, decision support system.*

Introduction. Recent studies suggest that, artificial intelligence (AI) demonstrates significant potential to enhance the productivity and added value of project managers by automating data collection and analysis, monitoring performance, and optimizing time planning and scheduling. In the context of project management, this opens opportunities for optimizing project roadmaps, budgeting, performance control, and improving adaptability to changing external conditions [1].

At the same time, the rapid development and implementation of AI is accompanied by a range of risks that may have a reverse effect on the effectiveness of management activities. One such risk is the phenomenon of so-called "effective inefficiency," which manifests in excessive creation or delegation of irrelevant, non-urgent, or unnecessary tasks. This, in turn, can lead to the project manager's attention being scattered, team overload, and a general decline in the quality of management functions [2].

Another notable risk is the variability of results generated by different AI models. Differences in architecture, training algorithms, and training data sources result in substantial differences in predictions or decisions, even for the same management scenarios. An example is the difference between standard large language models and those using the Mixture-of-Experts (MoE) architecture, which allows for scaling without a proportional increase in computational costs. In this architecture, the model has many specialized "experts" (neural networks or parts of them), but only a few are activated per request.

Thus, evaluating the effectiveness of AI solutions cannot be reduced merely to technical specifications or the number of operations per unit of time—a metric describing hardware computational performance (CPU/GPU). Although this metric is important for technical optimization, it does not reflect the real utility or relevance of a model in solving practical project management tasks.

Analysis of the Latest Research and Problem Statement. The existing body of scientific literature demonstrates a growing interest in methodologies for evaluating the effectiveness of artificial intelligence in management and organizational decision-making. Many studies focus on performance-oriented indicators, such as accuracy, processing speed, or cost efficiency, while others emphasize explainability, transparency, and trust in AI-driven systems [3-5].

Therefore, there is a need for comprehensive approaches to evaluating AI effectiveness in project management that combine technical, organizational, and economic criteria. This article aims to analyze current methods of AI efficiency assessment in project management, identify key challenges, and propose perspectives for overcoming them. Special attention is given to large language models. The article first reviews assessment methods, then discusses challenges, and finally suggests promising alternatives adapted to project management under resource constraints.

Analysis of the Latest Research and Problem Statement. The evaluation of artificial intelligence effectiveness has become a distinct and rapidly developing research field at the intersection of computer science, management, and systems analysis. Existing studies offer a wide range of conceptual and methodological approaches; however, they remain fragmented, heterogeneous, and often poorly aligned with the specific requirements of project management practice. In particular, the lack of unified criteria for assessing not only performance improvements, but also managerial impact and capability transformation, complicates both comparative analysis and practical adoption. Against this background, a systematic analysis and classification of existing AI performance evaluation methods is required in order to identify their applicability, limitations, and relevance to project management under conditions of uncertainty and limited human resources.

To clarify this methodological landscape, the present section is structured as a systematic analytical review of existing AI evaluation approaches. The analysis begins with a general classification of performance-based and capability-oriented methods, followed by a detailed examination of benchmarks, explainable artificial intelligence techniques, mutual information analysis, and psychometric approaches. This structure reflects the functional logic of project management practice, moving from technically measurable indicators toward methods that attempt to capture higher-level cognitive and managerial effects. The purpose of such structuring is to identify conceptual limitations and unresolved methodological gaps in existing research, which subsequently motivate the development of original evaluation models presented in the later sections of this article.

Classification of AI performance evaluation methods. With the increasing application of Artificial Intelligence (AI) across various domains particularly in project management, production, and business analytics, there is a growing need to establish reliable and reproducible methods for evaluating its effectiveness. How- ever, this evaluation process remains a complex and multidimensional challenge. This complexity arises not only from the wide variety of AI solution types—ranging from automated agents to generative models - but also from the lack of universal metrics that would be relevant across all use cases.

In today's project management context, the effectiveness of AI cannot be reduced to questions of technical integration alone. Of critical importance is AI's impact on key managerial functions: whether it truly accelerates processes, enhances decision-making quality, reduces cognitive load on project managers, and delivers measurable economic benefits in both the short and long term. Answering these questions is vital for informed decision-making about AI integration into business processes. Yet without proper performance evaluation methods, these remain mere assumptions or subjective impressions. Therefore, the following sections of this paper provide an in-depth analysis of these questions and justify the need to adapt existing metrics to the specific challenges of project management under conditions of limited human resources.

One of the first steps toward effective evaluation is the formulation of clear, data- oriented objectives. The assessment methodology should be based on Key Performance Indicators (KPIs) that reflect the strategic priorities of the organization. Even so, identifying the actual impact of AI often entails a "chicken-and-egg" paradox: high-quality data is needed to deploy AI, but AI's influence manifests precisely through improvements in data quality and availability.

The type of AI system also increases evaluation complexity. For instance, the effective- ness of predictive maintenance solutions can typically be assessed using straightforward metrics such as failure rate reduction. In contrast, generative AI systems used for employee training or organizational knowledge retention produce outcomes that are harder to quantify and require more sophisticated approaches to data collection and interpretation.

Contemporary approaches to AI evaluation fall into several categories: from performance- oriented methods that focus on achieving specific task outcomes, to capability-oriented methods that attempt to assess the underlying cognitive and functional abilities of AI systems. These are complemented by comprehensive business-oriented frameworks that measure return on investment (ROI) in AI, scalability of solutions, user-friendliness, the creation of fundamentally new capabilities, and the overall impact on workforce productivity [3].

Unlike some researchers [4] who express profound concerns about the potential negative consequences and existential risks associated with AI use, the authors of this paper do not fully share such apprehensions. Instead, they identify the key risk as ineffective AI implementation. This view is supported by the Massachusetts Institute of Technology (MIT) 2025 report on the state of AI in business, which states: "Despite investments of $30–40 billion in generative AI, 95% of organizations report no significant returns on these investments" [5]. The absence of adequate quantitative assessment methods for AI integration leads many organizations to invest in these technologies without receiving the feedback necessary to refine their strategies. This in turn hampers rational evaluation and comparative analysis of projects, complicating their effective management and improvement. At the same time in recent years, the adoption of artificial intelligence across various business functions has accelerated significantly. According to McKinsey Global Surveys, the percentage of organizations using AI in at least one function surged from 55% in early 2023 to 78% by the end of 2024. This rapid growth is particularly evident in IT, marketing, and service operations. The IT function alone experienced a nine-point increase in reported AI adoption within six months - from 27% to 36% - highlighting the growing reliance on AI-driven tools for automation, risk prediction, and resource optimization.

Figure (1) illustrates the dynamics of AI adoption across different industries between 2020 and 2024, emphasizing the expanding role of AI in project management and strategic operations.

The key objective of this section is to systematize existing groups of AI performance evaluation methods, classify them, and analyze their relevance to project management tasks in resource-constrained environments.

Traditional and Capability-Based Approaches to Evaluating AI Efficiency. Traditional and capability-based approaches to evaluating the effectiveness of artificial intelligence (AI) in project management encompass methods that originated during the early phases of AI development or were adapted from adjacent disciplines. These approaches aim to assess technical performance, economic viability, and organizational impact.

Performance-based evaluation frameworks focus on measuring how well AI systems execute predefined tasks, such as scheduling, resource allocation, or risk prediction. In contrast, capability-based assessments examine the broader cognitive competencies of AI models, including reasoning, adaptability, and learning capacity. Both approaches have been applied to support project management functions through automation and decision support.

However, these methodologies exhibit notable limitations. Performance metrics are often static and insufficiently responsive to the rapidly evolving, uncertain, and resource- constrained environments characteristic of modern project management. Moreover, traditional indicators—such as task completion time or computational throughput - may not reflect the actual managerial value or practical utility of an AI solution in a project setting. Historically, these methods emerged during the development of

foundational AI theories, when the emphasis was placed on benchmarking hardware capabilities (e.g., CPU/GPU performance) or evaluating the return on investment (ROI) of AI-driven systems. At that time, less attention was paid to the complex reasoning abilities and generalization potential of modern AI architectures.

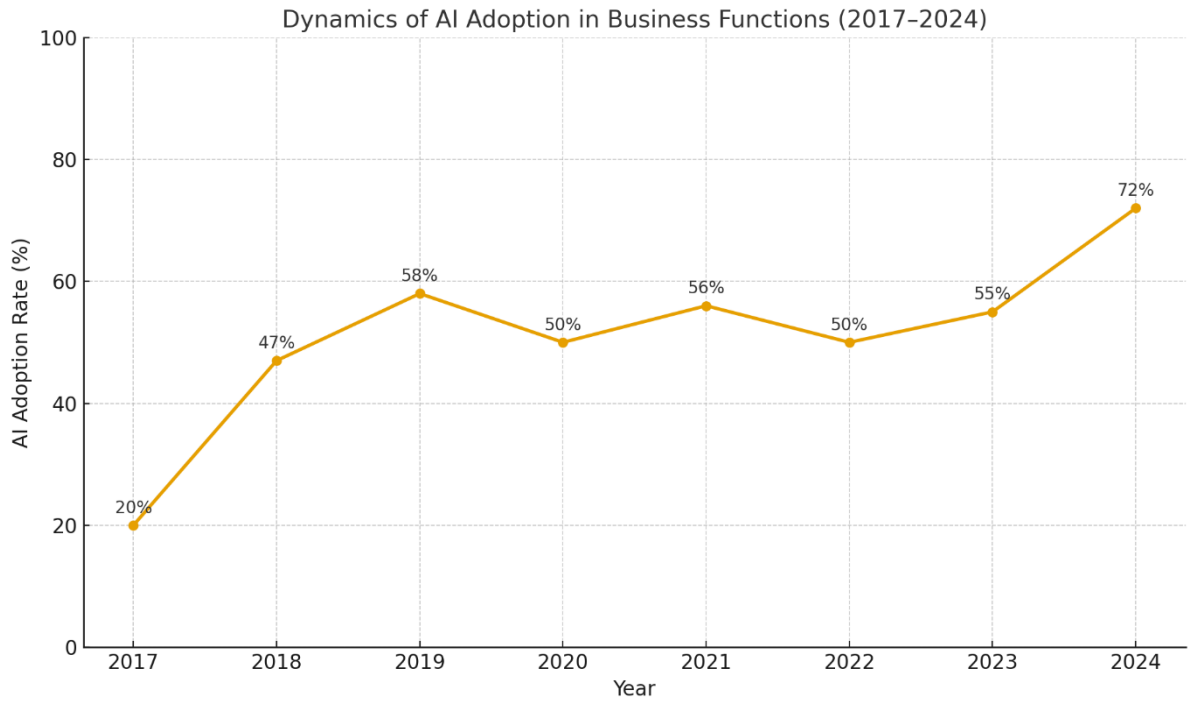

Fig.1. **Share of organizations using AI in at least one business function, 2018–2024. Source: McKinsey & Company (2025).**

Consequently, while these legacy evaluation methods are still prevalent in industry and academia, they often fall short in addressing the demands of contemporary project management, especially under dynamic conditions and tight resource constraints. The following sections review the key methods within this category, examine their application to project environments, and outline the critical limitations that motivate the exploration of more adaptive, context-aware evaluation strategies presented later in this paper.

Benchmarks. In the context of artificial intelligence (AI) performance evaluation, benchmarks refer to standardized sets of tests, tasks, or metrics used to measure system characteristics such as accuracy, execution speed, data processing capacity, and reliability under controlled conditions. Benchmarks serve as a core tool in traditional evaluation methods, enabling consistent comparison of different AI models, assessment of their capabilities, and identification of limitations. However, it is important to note that benchmarks do not always reflect the complexity of real-world project management environments. Within project management (PM), benchmarks are commonly used to evaluate AI effectiveness in task specific areas such as automated scheduling, risk forecasting, or resource optimization.

One prominent benchmark is SuperGLUE (Super General Language Understanding Evaluation), developed as a more challenging successor to GLUE (General Language Understanding Evaluation). SuperGLUE was introduced to address the limitations of GLUE, where AI models had already surpassed the performance of non-expert humans. It consists of a suite of complex tasks aimed at testing language understanding capabilities, with a particular focus on low-resource learning scenarios. While GLUE reached or exceeded human-level performance, SuperGLUE focuses on tasks that remain difficult for algorithms but are solvable by educated humans.

The benchmark encompasses various formats - including classification, question-answering, causal reasoning, and disambiguation - and employs a unified evaluation system based on average scores across tasks

$$S_{core} = \frac{1}{N} \cdot \sum_{i=1}^{N} M_i, \quad (1)$$

where N is the number of tasks;
M_i is the normalized model result for task i, measured by the appropriate metric.

SuperGLUE includes eight tasks. BoolQ is a yes/no question-answering task where models must find an answer in a short text excerpt, typically taken from Wikipedia, based on a search engine user query. Commitment Bank (CB) evaluates the model's ability to determine the degree to which the author believes in the truth of a subordinate clause; this task is essentially a variant of textual implication with three classes - entailment, contradiction, and neutrality. Choice of Plausible Alternatives (COPA) is built on causal relationships: the system is given a premise sentence and two possible alternatives for cause or effect, and it must choose the more logical one. MultiRC is a reading comprehension task where each question can have multiple correct answers; the model must not only identify them but also combine facts from different sentences in the text. ReCoRD presents news articles in a Cloze-style task where one entity is masked, and the correct variant must be selected from several possible candidates; the complexity lies in the fact that the entity can be expressed in the text in several different forms. Particular attention in SuperGLUE is paid to lexical and semantic polysemy. The Word-in-Context (WiC) task involves determining whether a word has the same meaning in two different contexts; it tests the model's ability to distinguish lexical polysemy. Even more challenging is the Winograd Schema Challenge (WSC) task, which requires the model to correctly interpret a pronoun in a sentence, relying not only on syntactic structure but also on common sense. Additional sets AX-b and AX-g are used for diagnosing the quality of implication task performance.

In addition to the main tasks, SuperGLUE includes a diagnostic dataset that allows analysis of model strengths and weaknesses. It includes examples aimed at testing logical operators, coreferences, temporal relationships, semantic roles, and polysemy. As a result, the benchmark not only determines the average performance level but also helps understand in which specific aspects of language understanding the model has limitations [6].

Despite its popularity in the scientific community, SuperGLUE as a benchmark does not provide insight into how effective an AI model can be for business. The reason lies in the nature of this method: it measures model performance in narrowly defined academic tasks that reduce to classification, text-based answer search, or implication determination. Such tasks do not reflect the complexities of real business processes.

Firstly, SuperGLUE tasks are artificial and limited in context. For example, choosing the correct pronoun in the Winograd Schema or determining word meaning in WiC has no direct analogy in business scenarios such as supply chain management, sales forecasting, or user data work. Thus, a high score on the benchmark does not mean the model can generate value in practical conditions.

Secondly, SuperGLUE does not account for economic metrics - implementation costs, integration speed, user training expenses, or return on investment. In business, key indicators include ROI, TCO, or reductions in operational costs, rather than the percentage of correct answers on a test sample. A model may perform excellently on benchmark tasks but be too expensive to operate or overly complex to integrate.

Thirdly, SuperGLUE lacks the dynamism characteristic of the market. Business environments change constantly, with data that is noisy, incomplete, or contradictory. In contrast, benchmark tasks are static and relatively "clean" creating a gap between academic evaluation and how a model behaves in real corporate applications.

Fourthly, SuperGLUE does not test a model's ability to work with domain-specific information. For example, companies in pharmaceuticals, agriculture, or finance have their own unique data, unlike

Wikipedia articles or news texts. High results on general tasks do not guarantee that a model can process complex legal documents, scientific patents, or technical specifications.

In conclusion, SuperGLUE evaluates the linguistic abilities of models but does not answer the question of how much economic value they can create. Businesses expect AI to deliver concrete benefits - faster decision-making, cost reduction, resource optimization, and profit growth—rather than abstract accuracy. Therefore, for corporate applications, specialized metrics and benchmarks that directly reflect business efficiency, rather than just linguistic skills, are needed.

The performance of modern large language models (LLMs) is commonly assessed across multiple dimensions, including reasoning, agentic tasks, and coding abilities. Figure 2 illustrates comparative results for 15 leading models across 12 benchmark suites, such as MMLU-Pro, AIME 24, SciCode, SWE-Bench, TAU-Bench, and others. Each model is evaluated by aggregating results in the following categories:

- agentic: Tasks that require planning, goal execution, and multi-step reasoning across benchmarks like TAU-Bench and BFCL V3;
- reasoning: Logic, mathematical problem-solving, and comprehension, as measured by MMLU-Pro, AIME 24, and LCB;
- coding: Software engineering challenges, including benchmark tasks like SWE-Bench and Terminal-Bench.

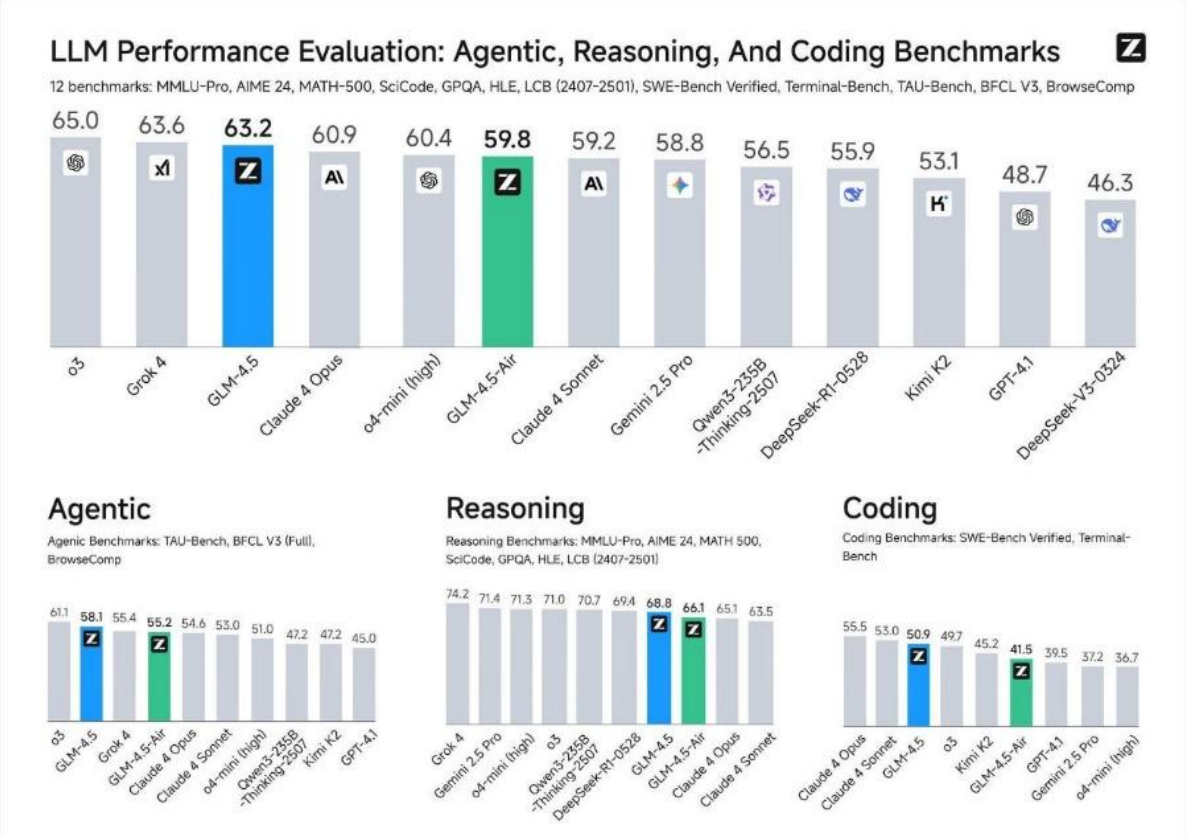

Fig.2. **LLM Performance Evaluation: Agentic, Reasoning, And Coding Benchmarks Source: https://z.ai/blog/glm-4.5**

In this evaluation, GLM-4.5 and its lightweight variant GLM-4.5-Air demonstrated competitive performance across all three categories. While GPT-4-o scored highest overall (65.0), GLM-4.5 achieved notable results in reasoning (68.8) and agentic benchmarks (58.1), outperforming some commercial models like Claude Opus and Gemini 2.5 Pro.

Explainable Artificial Intelligence (XAI). Explainable Artificial Intelligence (XAI) represents a set of methods and techniques aimed at ensuring the transparency of AI system decisions, particularly through the interpretation of the internal processes of models that often function as "black boxes". In the context of project management (PM), XAI gains special significance, enabling project managers—who are not necessarily experts in machine learning—to understand the logic behind AI predictions, such as risk assessments or resource optimization. This method enhances trust in automated systems, identifies biases, and ensures compliance with ethical standards, which is critical in dynamic PM environments with limited resources [7].

Among the key XAI techniques are LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), which explain how individual features influence a model's predictions. While XAI provides qualitative evaluation through transparency and comprehensibility of decisions, its ability to quantitatively assess the effectiveness of implementing AI models is limited. XAI focuses on interpreting AI decisions, allowing project managers to evaluate the appropriateness of predictions.

The LIME method creates local approximations of complex models to explain individual predictions, minimizing a loss function

$$argmin_{g \in G} L(f, g, \pi \cdot x) + \Omega(g), \tag{2}$$

where g is the local interpretable model;
G is the set of simple, human-understandable models (e.g., linear regression, decision trees);
L is the loss function measuring how much the predictions of the local model g differ from those of the complex model f in the local neighbourhood of point x, weighted by π·x (the weight of the local region);
$\Omega(g)$ is a penalty for model complexity. In project management, $\Omega(g)$ limits the number of features (e.g., budget, deadlines) used for explanation, ensuring project managers can easily grasp key factors [8].

In PM, LIME can explain why an AI predicts project delays by highlighting the impact of individual factors, such as resource shortages or unclear requirements. For example, a model g explaining the behaviour of f in the local neighbourhood of x might show that a project delay prediction depends 60% on a limited budget.

However, the LIME method has significant drawbacks, including:

- Locality of Explanations: The formula focuses on local approximation (via $\pi \cdot x$), limiting generalization. In project management, where project conditions change (e.g., new requirements), an explanation for one x may not apply to another. For instance, an explanation for one project (with a budget of 100,000) may be irrelevant for another (with a budget of 1 million);

- Instability: Explanations depend on the choice of $\pi \cdot x$ (e.g., kernel width σ and the generated dataset Z). Changes in these parameters can lead to different g values, reducing reliability in project management, where consistency is needed. This can confuse managers if explanations for similar projects vary.

- Computational Complexity: Calculating $L(f, g, \pi \cdot x)$ requires generating many points Z and evaluating $f(z)$ for each, which can be resource-intensive for large models f (e.g., transformers). In project management, where quick decisions are needed, this can be a limitation [9].

- Subjectivity of Penalty $\Omega(g)$**: The choice of $\Omega(g)$ (e.g., limiting the number of features) is subjective and affects interpretability. In project management, managers may need explanations with varying levels of detail, but the formula does not adapt automatically.

- Lack of Quantitative Effectiveness Assessment: The formula does not measure how much implementing AI improves project metrics (e.g., reduced project completion time). It only explains predictions but does not evaluate their impact, for example, through metrics like ROI or KPIs [10].

Similarly, the SHAP method applies game theory to assess the contribution of each feature

$$\varphi_i = \sum_{S \subset N/\{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} \cdot \left[f_x(S \cup \{i\}) - f_x(S) \right], \quad (3)$$

where φ_i is the contribution of feature i;
S is a subset of features;
N is the set of all features;
f_x is the model prediction [11].

In project management, this allows, for example, determining how parameters such as team experience or budget influence project delays.

The process of using XAI to evaluate AI in PM involves selecting a technique (e.g., SHAP for global interpretation), generating explanations based on the model's output data, and interpreting them in the context of project decisions. Tools like SHAP or LIME libraries in Python are integrated with PM platforms, such as Jira, for real-time visualization of explanations.

However, XAI has significant limitations for quantitatively assessing the effectiveness of implementing AI models in project management. Firstly, XAI does not provide direct metrics to measure the impact of AI on key project indicators, such as reduced completion time, budget savings, or productivity improvements. For example, while SHAP can quantitatively assess feature contributions, it does not measure how much implementing AI improves overall project efficiency compared to traditional methods. Additional metrics, which XAI does not generate automatically, are needed, such as the percentage reduction in delays

$$\Delta T = \frac{T_{nonAI} - T_{withAI}}{T_{nonAI}} \cdot 100\%, \quad (4)$$

or resource savings

$$\Delta R = R_{nonAI} - R_{withAI}. \quad (5)$$

Secondly, XAI methods like SHAP require significant computational resources, which can be problematic in PM environments with limited budgets or hardware, especially for real-time applications, such as IoT project monitoring.

Thirdly, the lack of standardized metrics for evaluating the quality of explanations complicates comparing AI effectiveness across different projects or models. For example, there is no universal criterion to quantitatively assess how useful SHAP or LIME explanations are for project managers, limiting their applicability for comparative analysis.

Thus, XAI is a valuable tool for qualitative evaluation of AI in project management, providing transparency and comprehensibility of decisions. However, its limitations in quantitative effectiveness assessment, high computational complexity, explanation instability, and lack of standardization indicate the need to combine XAI with traditional PM metrics, such as ROI or KPIs, for a comprehensive evaluation of AI impact. Further research should focus on developing standardized approaches to integrating XAI with PM platforms and creating quantitative metrics that evaluate not only explanations but also the overall effectiveness of AI in project management.

Mutual Information (MI) and Model Inspection are methods for evaluating the internal workings of AI systems, aimed at analysing the information that a model actually captures and uses for its predictions. In the context of project management (PM), these methods allow for a deep understanding of how AI processes data, for example, how internal model representations reflect dependencies between project features (budget, deadlines, risks), contributing to the identification of inefficiencies and optimization of decisions. MI measures the dependency between input data and the model's internal states, while model inspection involves analysing layers, activations, and weights, providing a detailed overview of AI functioning [12]. The essence of MI lies in the quantitative assessment of

interdependence between variables, for example, between input features (X) and the model's internal representations (Y). The MI formula is based on information theory

$$I(X,Y) = \sum_{x,y} p(x,y) \cdot log \frac{p(x,y)}{p(x) \cdot p(y)}, \quad (6)$$

where $p(x, y)$ is the joint probability distribution;
$p(x)$ and $p(y)$ are the marginal distributions [13].

In project management, MI can analyse how much an AI model captures dependencies between project features, for example, how budget data influences layer activations for risk forecasting. Model inspection complements MI through the direct analysis of internal components: neural network layers, neuron activations, and connection weights. For instance, visualization techniques like t-SNE (for dimensionality reduction) allow one to see how a model clusters project data, while weight analysis (e.g., via gradient methods) identifies key features. In project management, this can be used to verify whether a model effectively processes data from tools like Jira, for example, how layer activations reflect the impact of team size on delay estimates.

Despite their advantages, MI and model inspection methods have significant limitations for evaluating AI effectiveness in PM.

Firstly, high computational complexity: calculating MI for large models with N features has exponential complexity due to the need to estimate joint distributions, making it unsuitable for real-time evaluation in resource- constrained project environments.

Secondly, the need for high technical expertise: interpreting MI results or activation visualizations requires machine learning expertise, which may be unavailable to project managers, leading to misapplications.

Thirdly, instability with dynamic data: MI depends on data quality, and thus MI values may vary, com-plicating comparative analysis of models.

Fourthly, limited scalability: for deep models with millions of parameters, inspecting layers (e.g., weight analysis) becomes practically impossible without specialized hardware, reducing its usefulness in projects with limited budgets. Finally, the lack of standardization: there are no unified metrics for assessing the quality of MI or inspection, making them subjective and less suitable for quantitative comparison of AI model effectiveness across different scenarios.

Thus, MI and model inspection provide a deep understanding of AI's internal workings, which is valuable for optimizing models in project management. However, their limitations in computational complexity, need for expertise, and instability in dynamic environments indicate the necessity of combining them with other methods, such as benchmarks or XAI, for a comprehensive effectiveness evaluation.

Psychometrics, traditionally used to assess human cognitive abilities such as intelligence, memory, or abstract thinking, is adapted for evaluating AI systems through analogy with human tests. In the context of project management, psychometrics allows the assessment of AI's "cognitive abilities" for example, its capacity for analysing causal relationships, solving complex planning tasks, or generating creative solutions for resource optimization. This method goes beyond standard performance metrics, focusing on evaluating the general intellectual capabilities of AI, which is particularly valuable in dynamic environments requiring adaptability to changing conditions. However, psychometrics faces challenges in formalization and the risk of anthropomorphism, which limits its practical application. The essence of psychometrics in AI evaluation lies in applying tests similar to those used for humans, such as IQ tests, analogies, causal reasoning, or linguistic creativity assessments. In project management, this may include evaluating an AI's ability to analyse dependencies between project factors (e.g., how budget changes affect deadlines) or generate innovative resource allocation strategies. For example, an analogy test can check whether an AI can map the relationship "budget/risk" to other pairs like "resources/productivity," reflecting its abstract thinking capacity. A causal reasoning test can assess how

an AI determines that a resource shortage causes project delays. Formally, psychometric evaluation can be represented through an assessment function:

$$S = f\left(T, R\right), \tag{7}$$

where S is the psychometric score;
T is a set of tests (e.g., on analogies or memory);
R is the AI's responses, compared to benchmark or human results [14].

Methods for Measuring Fundamentally New Capabilities Created by AI. A number of metrics and frameworks have been specifically designed to assess fundamentally new capabilities enabled by artificial intelligence (AI), where the baseline performance was initially zero. Unlike traditional percentage-based productivity improvements - which require a non-zero baseline for comparison - these metrics focus on the emergence of new skills, processes, or functions that were previously unavailable. They are often classified as indirect or business-oriented metrics and typically combine quantitative indicators with qualitative assessments, such as maturity levels.

The following analysis draws from both academic and business literature, with a focus on project management and managerial productivity enhancement.

Maturity-Based AI Metrics evaluate an organization's progress in implementing AI by measuring the transition from no capability to new capabilities across a predefined scale (e.g., from level 1 to level 5). These metrics capture the transformative impact of AI, such as enabling software development by non-technical project managers, or the creation of novel project processes like risk prediction or resource optimization. They are crucial for assessing the influence of AI on project-related activities, especially in environments with constrained resources.

Example: The McKinsey AI Maturity Model assesses AI adoption using key performance indicators such as the number of AI-driven use cases deployed and the percentage of revenue generated from AI initiatives. If a project manager previously had no ability to code, this model measures the shift from 0 to a number of real business cases involving AI (e.g., from 0 to 15 projects per year involving automated coding capabilities), with a focus on the pace at which new functionality is created [15].

Innovation and Novelty Creation Metrics. Innovation capacity is a key performance indicator for modern organizations seeking long- term competitiveness in dynamic markets. A critical aspect of evaluating AI implementation in project management involves measuring innovation outcomes directly attributable to AI technologies. These metrics aim to quantify an organization's ability to generate entirely new products, services, or processes - that is, radical innovations.

Several types of metrics are used to measure innovation performance, including:

Quantitative metrics:

- number of patent applications (patent activity);
- number of new products or services launched;
- number of scientific publications, technical reports, or prototypes;
- share of R&D budget allocated to experimental AI initiatives.

Qualitative metrics:

- radicalness of innovation (i.e., departure from existing technologies);
- ability of AI to autonomously generate novel ideas;
- expert-based assessment of novelty using the Delphi method.

One of the core quantitative metrics is the Innovation Effectiveness Index

$$I_{innovation} = \frac{N_{patents} + N_{new\_products} + N_{scientif\_outputs}}{T}, \tag{8}$$

where $N_{patents}$ is the number of patents filed during the evaluation period;
$N_{new\_products}$ is the number of novel product/service launches;

$N_{scientific\_outputs}$ is the count of relevant technical or academic outputs.

Generative AI models (e.g., LLMs, GANs, diffusion models) act as innovation catalysts by:

- supporting idea generation in creative sessions;
- performing automated patent analysis to identify technological gaps;
- simulating and testing innovative concepts in virtual environments;
- generating AI-driven product design prototypes.

In project management, this enables non-intuitive solution generation, especially by less experienced managers. For example, large language models (such as ChatGPT, Claude, or Gemini) can autonomously produce multiple project planning scenarios based on minimal input, thus facilitating novel approaches to resource, time, or budget planning. Example: The number of patents filed as a direct result of AI capabilities (e.g., increasing from 0 to 10 per year) represents the creation of new research potential. In project management, this may include AI-driven generation of planning methodologies for novice project leads. According to a recent study [16], AI-related patent filings are growing at an annual rate exceeding 30%, confirming AI's critical role as an innovation engine.

Such innovation metrics should be treated not only as reporting indicators, but also as strategic signals of digital transformation effectiveness. Their dynamic monitoring via internal dashboards enhances an organization's adaptability within a VUCA environment.

Metrics for Human Capital Development Driven by AI. AI deployment in project management should be evaluated not only through technical and financial indicators, but also through its impact on human capital transformation. Human capital metrics reflect qualitative and quantitative shifts in staff skills, knowledge, and productivity - particularly when baseline digital or AI competence was previously lacking. A common approach involves administering digital literacy questionnaires or assess- ments to track growth in knowledge and confidence when interacting with AI systems.

For instance, a Digital AI Literacy Index (L) can be calculated as

$$L = \frac{N_{qualifield}}{N_{total}} \cdot 100\%, \tag{9}$$

where $N_{qualified}$ is the number of employees demonstrating at least basic AI system understanding;
N_{total} is the total number of participants assessed.

Another valuable metric is the share of staff who completed AI-related training or certification within a defined period (e.g., annually). This serves as an indicator of the effectiveness of internal personnel development programs or external educational initiatives.

In project-oriented environments, such metrics help evaluate the extent to which AI can compensate for shortages in skilled labour by automating routine tasks and offering decision-making support.

Table 1 provides a comparative overview of key AI evaluation approaches across various domains of project management impact.

Table 1. **Overview of AI Evaluation Methods and Their Applications in Project Management**

Method Category	Example Metric	Application Area	Evaluation Focu
Technical Effectiveness	Accuracy, Latency, F1-Score	Predictive tools	System performance
Business Efficiency	ROI, NPV, Time savings	Strategic and operational decisions	Financial impact
Innovation Impact	Patents, Novelty index	R&D, product pipeline enhancement	New value creation
Human Capital Development	AI training rate, Skill index	Workforce AI adoption	Capability building
Organizational Maturity	AI maturity models	Implementation roadmaps	Transformation stage

The Purpose and Objectives of the Research. The purpose of this research is twofold and reflects a transition from analytical assessment to methodological development in the evaluation of artificial intelligence effectiveness in project management.

The first objective of the study is to conduct a structured analytical review of existing methods for evaluating the effectiveness of artificial intelligence in project management. This includes the classification and critical examination of performance-based, interpretability-oriented, and maturity-based evaluation approaches, as well as the identification of their conceptual and methodological limitations when applied to dynamic and resource-constrained project environments.

The second objective of the study is to develop original methodological approaches for assessing artificial intelligence effectiveness in "zero-to-one" scenarios, where AI enables fundamentally new managerial capabilities that were previously unavailable. These approaches are aimed at capturing AI-induced transformations in decision-making, task feasibility, and project execution, which cannot be adequately evaluated using traditional productivity or maturity metrics.

Novel Approaches to Measuring the Effectiveness of AI Application in Project Management. Based on a critical analysis of existing approaches to measuring the effectiveness of artificial intelligence (AI) implementation in management practices, particularly in project management, three original evaluation methods are proposed. These methods have not been adequately described in contemporary scientific literature. Most existing studies focus on classical metrics such as Key Performance Indicators (KPIs), Return on Investment (ROI), or technology adoption maturity models. In contrast, the proposed approaches emphasize under-explored aspects, including: the dynamics of transitions from a baseline to an innovative level ("zero-to-one transitions"), the interaction between artificial intelligence and project management methodologies (notably aligned with the Project Management Institute standards), and the use of explainable AI models alongside ethical decision-making components. Each method includes a description, formula/algorithm, application, and potential for empirical validation. These methods can benefit organizations aiming to integrate intelligent tools into management processes - not only to enhance productivity but also to develop new training approaches for personnel, automate decision-making, and ensure transparency in human-algorithm interactions under conditions of uncertainty. It is worth noting that these methods do not duplicate each other but form a comprehensive evaluation system.

Indicator of New Competency Activation (INCA) in Project Management. The INCA index introduces a novel metric focused not on the efficiency of performing existing tasks but on measuring the emergence of new competencies and functions arising from AI application. This distinguishes it from classical approaches - such as KPIs, ROI, or maturity models - which assess only productivity or the maturity level of technology adoption. The scientific novelty lies in three aspects:

- **Conceptual**: The introduction of the "zero-to-one transition" concept in the context of project management, i.e., the shift from the absence of a capability to its emergence through AI.

- **Methodological**: A formula is proposed that combines the number of new tasks, their impact on project outcomes, and their explainability (XAI-score).

- **Analytical**: Integration of XAI (explainable AI) with the PMI PMBOK model allows not only measuring the effect but also explaining the mechanism of its occurrence.

The INCA index is an integral metric calculated as follows

$$INCA = \frac{\sum_{i=1}^{T} \left(N_i \cdot I_i \cdot E_i \right)}{T}, \tag{10}$$

where N_i – the number of new tasks i made possible (ranging from 0 to T);
I_i – impact on the project (scored from 0 to 1, based on PMI scale: impact on timelines, budget, quality);
E_i – explainability (XAI-score, e.g., SHAP value for LLM, ranging from 0 to 1, where 1 indicates full transparency);

T – the total number of tasks in the project.

An example application in an IT project could involve a manager who does not code generating code using AI. INCA would measure this as a transition from 0 to 10 new modules, with an impact on the budget (I_i=0.8) and explainability (E_i=0.7), yielding an overall index for assessing implementation effectiveness.

Thus, the index serves as a dynamic indicator of competency development, which can be empirically validated. To verify its reliability, *A*/*B* testing is recommended: one group of managers works with AI support, while another does not. Comparing the mean INCA values between groups, followed by a statistical test (e.g., Student's t-test), would quantitatively demonstrate AI's impact.

INCA can be integrated into project management maturity assessment systems or internal PMO dashboards. It enables the identification of competencies activated by AI (rather than merely improved) and the assessment of the relationship between new skills and reduced risks or costs. Additionally, this method can be used to develop competency maps, where INCA acts as an indicator of training, innovativeness, and digital maturity. Compared to traditional approaches, INCA aligns more closely with evolutionary metrics of organizational intelligence development, making it suitable for strategic analysis.

Novelty Coefficient in AI-Driven Project Management (NCAPM). The NCAPM is an original metric for evaluating the innovative potential of AI in project management through a combination of qualitative (expert novelty assessment) and quantitative (scenario simulation) metrics. The metric is based on the integration of generative AI for modelling "what-if" scenarios in project management (e.g., real-time risk forecasting). This approach is absent in existing KPIs or innovation speed metrics, which typically do not account for ethical dilemmas associated with novel solutions. The coefficient is calculated as follows

$$NCAMP = \frac{Q_n + S_n - B_n}{3}, \tag{11}$$

where Q_n – qualitative novelty assessment (expert score from 0 to 10);
S_n – simulation assessment (number of unique scenarios generated by LLM, ranging from 0 to n, with uniqueness verified via NLP analysis);
B_n – correction for ethical ambiguities (ranging from 0 to 1, based on the NIST AI Risk Framework, for ethical novelty).

An example application in a construction project might involve AI creating a new risk simulation capability for non-technical managers. The coefficient would measure novelty as 8/10 (Q_n), with 50 scenarios (S_n=0.9), minus ambiguity (B_n=0.2), yielding a coefficient for assessing implementation effectiveness.

Dynamic Assessment of Transition to New Efficiency En-abled by AI (DATNE). This metric tracks dynamic transitions from a baseline to a new efficiency level in real- time, integrating Internet of Things (IoT) data with project management tools and AI systems. The metric enables real-time monitoring using agent-based AI systems (agents based on large language models), a feature absent in static maturity models or task completion indicators, with a focus on the adaptability of project management to changes, especially during crisis situations.

The coefficient is calculated as follows

$$DANTE = \int \frac{C_t - C_0}{\Delta T} dt, \tag{12}$$

where C_t – capability level at a specific time t, a numerical indicator ranging from 0 to 1, reflecting the ability of a team or individual to perform new task types previously unavailable without AI or digital tools.

This level can be objectively measured using IoT device data, which records actual changes in behaviour or system performance.

The capability level has the following formalized representation

$$C(t) = \frac{Z(t)}{Z_{max}}, \tag{13}$$

where $C(t)$ – capability level at time t;
$Z(t)$ – number of new completed tasks made possible by the intelligent system at time t;
Z_{max} – the maximum possible number of such tasks in the project (defined during the planning phase). -
C_0 – baseline level (0 for a project without AI);
ΔT – observation period (e.g., one week of the project).

An example calculation of the capability level can be provided. Within a project to modernize a production line in a food industry enterprise, the project team implemented a data analysis system based on intelligent sensors. Previously, operators could not identify equipment anomalies in real-time. After implementation - thanks to automatic alerts from the system - the team was able to respond promptly to 9 new event types. The total number of tasks potentially executable with this system was 15. Thus, the capability level one month after implementation can be calculated as $C(t)=9/15=0.6$. This indicates a 60% realization of new capabilities that were impossible before integrating the intelligent module.

The DATNE metric can be applied, for example, in an agile project where a manager gains a new real-time budget adjustment capability through AI. This indicator allows for a quantitative assessment not only of productivity but also of the level of digital transformation and team adaptation to new tools, which is critical in modern project management.

Challenges and Prospects for Evaluating the Effectiveness of AI Application in Project Management. Evaluating the effectiveness of AI application in project management faces numerous challenges, both technical and methodological in nature [9]. One key issue is the lack of established standards. Most existing methods are developed primarily within specific use cases [18], making them difficult to compare or scale to broader contexts, which hinders standardization [4].

Moreover, methodological barriers exist. For instance, many metrics are subjective or lack external validity testing [14]. Anthropomorphism is often observed, where AI systems are attributed human qualities or intentions [21]. Another problem is the presence of “noise” in training datasets, where models exhibit reduced effectiveness due to incorrect or publicly available benchmark data [22]. This is particularly relevant in fields like software development, where it is challenging to work with unstable application programming interfaces, new or external domains, and languages with limited resources [6].

From a technical perspective, difficulties arise from biases in data - measurement, labelling, selection, aggregation, confirmation, and others - which can lead to unfair treatment of different user groups [7]. Challenges also include privacy concerns, such as attempts to extract personal information from statistical models, data poisoning, and adversarial scenarios, necessitating the implementation of differential privacy mechanisms and distributed learning.

Ethical and legal constraints pose additional complexity. Data collection for analysis is often hindered by personal data protection regulations, internal company policies, or national legislative restrictions. This is particularly relevant for metrics requiring continuous real-time monitoring of personnel or user behaviour, as in the case of the dynamic transition assessor.

One of the central challenges in evaluating the effectiveness of AI systems in project management lies in the lack of unified mathematical frameworks that adequately describe the internal logic and decision-making mechanisms of large language models (LLMs). Even foundational elements of transformer-based architectures require complex mathematical justification, involving high-dimensional vector representations and optimization theories such as stochastic gradient descent and regularization techniques. Without a clear mathematical understanding of these systems, the development of reliable evaluation metrics for their application in dynamic, multi-variable environments like project management becomes problematic. This complexity highlights the need for interdisciplinary approaches

that combine project management methodologies with computational models and applied mathematics to ensure explainability, repeatability, and control of AI-driven project decisions [20].

Despite these challenges, there are clear directions for further development. One promising approach is the creation of hybrid metrics that combine technical characteristics (e.g., model interpretability, simulation accuracy, scenario generation efficiency) with classical management criteria (adherence to timelines, budget, quality). Another important direction is the integration of AI into decision support systems, where the model can not only provide recommendations but also self-assess their effectiveness.

The implementation of automated audit mechanisms is also advisable - language agents can analyse project documentation, compare actual results with planned ones, generate reports, and identify patterns in AI usage. Comparative testing of different models under practical project management conditions is also promising, enabling empirical evaluation of each approach's strengths and weaknesses.

Another significant direction involves integration with cognitive sciences to develop reliable psychometric tests and the creation of ethical evaluation frameworks, standardization of bias reduction processes in data, and the implementation of data governance policies. Simultaneously, particular attention is needed for large-context models capable of accounting for complex dependencies in dynamic project management environments, including the use of new architectures that combine algorithmic memory and intelligent information reproduction.

Thus, despite numerous barriers, the development of evaluating the effectiveness of AI application in project management represents a highly promising research direction. Its further advancement requires a multidisciplinary approach, combining project management, artificial intelligence, ethics, psychology, and data engineering.

Particular attention should be given to the potential of these approaches for transport infrastructure projects, where the application of AI enables the emergence of zero-to-one capabilities - functions that did not previously exist in traditional engineering or management systems. For instance, intelligent agents integrated into transport project management platforms can autonomously generate maintenance forecasts for rail tracks or highway pavements using IoT sensor data, optimize the sequencing of construction tasks based on real-time traffic models, and dynamically allocate resources to reduce downtime in rolling-stock operations. These capabilities represent a transition from reactive to predictive management, transforming the way infrastructure projects are planned, maintained, and financed. The integration of metrics such as INCA, NCAPM, and DATNE into these domains would allow quantifying the emergence of such fundamentally new competencies, thereby forming a methodological basis for evaluating innovation efficiency in digitalized transport ecosystems

Conclusions. This study addressed the methodological problem of evaluating the effectiveness of artificial intelligence (AI) in project management, where traditional performance indicators increasingly fail to reflect the actual managerial impact of AI-enabled systems. The results demonstrate that while artificial intelligence has significant potential to enhance planning, budgeting, risk forecasting, and decision-making processes, its effectiveness cannot be adequately assessed without a differentiated methodological framework that accounts for both existing evaluation approaches and emerging AI-driven capabilities.

The first research task - focused on the analysis of existing methods for assessing AI effectiveness - was addressed through a structured classification and critical examination of contemporary evaluation approaches. Traditional methods, including benchmarks, explainable artificial intelligence techniques, mutual information analysis, psychometric assessments, and maturity-based models, were systematized according to their evaluation focus and applicability to project management environments. The analysis showed that benchmarks (e.g., SuperGLUE) provide a standardized basis for comparing model performance but neglect the dynamic, contextual, and managerial nature of project work. Explainability-oriented and capability-based approaches improve transparency and insight into model behavior, yet their practical use is often constrained by high computational requirements, the need for specialized expertise, and limited comparability across projects. Maturity-based models capture organizational adoption dynamics but remain largely descriptive and do not explain how AI alters managerial effectiveness. Across all reviewed approaches, persistent limitations were identified,

including the lack of unified standards, susceptibility to data bias, anthropomorphic interpretations of AI behavior, and ethical and legal constraints, all of which hinder comprehensive quantitative evaluation in project management practice.

The second research task - aimed at developing methodological approaches for evaluating artificial intelligence effectiveness in "zero-to-one" scenarios - was addressed through the formulation of original, innovation-oriented evaluation metrics: INCA, NCAPM, and DATNE. These metrics shift the analytical focus from incremental productivity improvements toward the assessment of AI-enabled transitions from the absence of capability to its emergence. The proposed approaches enable the measurement of newly activated managerial competencies, qualitative novelty in project decision-making, and dynamic changes in effectiveness over time.

Based on the results obtained for both research tasks, the study outlines directions for further development of AI effectiveness evaluation. These include the creation of hybrid metrics combining technical, managerial, and ethical dimensions; the integration of advanced evaluation models into decision support systems; and the use of automated audits based on intelligent agents. Further empirical validation of the proposed methods in real project settings and their alignment with project management standards are identified as necessary steps toward broader practical adoption.

Overall, the study confirms that a comprehensive evaluation of artificial intelligence in project management requires a multidisciplinary approach that integrates computer science, systems analysis, and management science. The findings provide a coherent methodological foundation for assessing both existing AI applications and fundamentally new AI-enabled capabilities, contributing to the sustainable and responsible development of AI-driven project management practices.

REFERENCES

1. Müller, R., Locatelli, G., Holzmann, V., Nilsson, M., & Sagay, T. (2024). Artificial intelligence and project management: Empirical overview, state of the art, and guidelines for future research. *Project Management Journal*, *55*(1), 9-15. https://doi.org/10.1177/87569728231225198.
2. Mills, S., & Spencer, D. A. (2025). Efficient Inefficiency: Organisational challenges of realising economic gains from AI. *Journal of Business Research*, *189*, 115128. https://doi.org/10.1016/j.jbusres.2024.115128.
3. Edwards, J. (2025). How To Measure AI Efficiency and Productivity Gains. In: *InformationWeek, ed. AI and Machine Learning Insights*. Available from:. https://www.informationweek.com/machine-learning-ai/how-to-measure-ai-efficiency-and-productivity-gains.
4. Burden, J. (2024). Evaluating ai evaluation: Perils and prospects. *arXiv preprint arXiv:2407.09221*. https://doi.org/10.48550/arXiv.2407.09221.
5. Challapally, A., & Pease, C. (2025). AI trends and innovations in 2025. In: Artificial Intelligence News, ed. Annual AI Report. Available from:https://www.artificialintelligence-news.com/wp-content/uploads/2025/08/aireport2025.pdf.
6. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, *32*6 1-15. https://doi.org/10.48550/arXiv.1905.00537.
7. Zhang, M., Wang, H., Li, J., & Gao, H. (2020). Learned sketches for frequency estimation. *Information Sciences*, *507*, 365-385. https://doi.org/10.1016/j.ins.2019.08.045.
8. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778.
9. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012.
10. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. https://doi.org/10.48550/arXiv.1702.08608.
11. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*. https://doi.org/10.48550/arXiv.1705.07874.
12. Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), e7. https://distill.pub/2017/feature-visualization.
13. Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*. https://doi.org/10.48550/arXiv.physics/0004057.
14. Hernandez-Orallo, J., 2017. Evaluating intelligence across species and machines. *The Measure of All Minds, ed. Cognitive Science Series. Cambridge, UK: Cambridge University Press*, 50–100. https://doi.org/10.1017/9781316596654.

15. McKinsey & Company. (2023). Generative AI's impact on business in 2023. In: The State of AI Report, ed. Digital Transformation Insights. New York, NY: McKinsey & Company, 10–25. Available from: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-state-of-ai-in-2023.
16. WIPO (2023). *Trends in AI technology development*. Technology Trends 2023: Artificial Intelligence, ed. IP Research Series. Geneva, Switzerland: World Intellectual Property Organization.
17. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, *54*(6), 1-35. https://doi.org/10.1145/3457607.
18. MIT (2025). *Business applications of AI in 2025*. The State of AI in Business 2025, ed. Technology Insights. Cambridge, MA: MIT Press, pp.5–20. Available from: https://www.mit.edu/ai-report-2025.
19. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. https://doi.org/10.48550/arXiv.1702.08608.
20. Nesterenko O., & Kulbovskyi I. (2024). Mathematical framework of transformer-based artificial intelligence architectures in large language models for the development of intelligent agents. *Science and technology today*, *5*(46). http://dx.doi.org/10.52058/2786-6025-2025-5(46)-1860-1872.

Нестеренко О.Б.[1]

[1]Аспірант, Кафедра Автоматизація та комп'ютерно-інтегровані технології транспорту, Національний транспортний університет, вул. Михайла Омеляновича-Павленка, 2, 01010, м. Київ, Україна. ORCID: https://orcid.org/0009-0008-7249-3189.

Методологічні аспекти та моделі оцінки ефективності штучного інтелекту в управлінні проєктами.

Анотація. *Швидка інтеграція штучного інтелекту в управління проєктами пропонує значний потенціал для підвищення продуктивності завдяки автоматизації даних, моніторингу ефективності та оптимізації розкладів. Однак виклики, такі як "ефективна неефективність" та варіативність результатів моделей ШІ, ускладнюють оцінку ефективності. У статті аналізуються методологічні аспекти оцінки ефективності ШІ в управлінні проєктами, класифікуються існуючі методи (бенчмарки, пояснювальний штучний інтелект, взаємна інформація, психометрія), ідентифікуються ключові виклики (упередження, відсутність стандартів, етичні обмеження) та пропонуються нові метрики (ПАНК, КНУПШ, ДОПШ) для вимірювання інновацій. Зазначено потенціал цих підходів для проєктів транспортної інфраструктури, де ШІ дає змогу створювати принципово нові можливості в плануванні, прогнозуванні обслуговування та оптимізації ресурсів. Перспективи включають гібридні метрики та інтеграцію з системами підтримки прийняття рішень. Дослідження підкреслює необхідність міждисциплінарних підходів для адаптації оцінки ШІ до середовищ в управлінні проєктами з обмеженими ресурсами.*

Ключові слова*: Модель, машинне навчання, бенчмарк, оцінка ефективності, методологія, когнітивні моделі, системний аналіз, управління проєктом, штучний інтелект, система підтримки прийняття рішень.*

Дата першого надходження статті до видання 11.09.2025

Дата прийняття до друку статті 24.11.2025