

Exploratory Study on LLM-Assisted Evaluation of PMBOK-Based Argumentative Essays in a Master's Programme

Anonymous Authors

Affiliation removed for double-blind review

1

Abstract—The integration of professional standards such as the Project Management Body of Knowledge (PMBOK) knowledge areas into graduate-level project management programmes has increased the need for robust and scalable assessment strategies. At the same time, Large Language Models have emerged as potential tools for supporting academic evaluation. This exploratory study examines the feasibility of using assisted assessment to evaluate argumentative essays structured around knowledge areas in a Master's programme. A limited sample of student essays was analysed using predefined evaluation metrics focusing on conceptual integration, argumentative coherence, cross-domain synthesis, and alignment with knowledge area domains. Model-generated evaluations were examined to identify patterns in domain coverage, scoring consistency, and analytical depth. The study does not aim to validate automated grading but to explore methodological viability and identify strengths, limitations, and risks associated with AI-assisted evaluation in professional graduate education. Findings highlight both the potential of LLMs to detect structural alignment with the knowledge areas and the need for human oversight to ensure contextual and critical judgment. This pilot contributes to ongoing discussions on AI-supported assessment in higher education by analyzing essays written by postgraduate students in a PMI-aligned Master's program and comparing dimension-level scoring patterns between human evaluators and GPT-based evaluation.

Index Terms—Large Language Models, AI-assisted assessment, rubric-based evaluation, PMBOK education, engineering education.

I. INTRODUCTION

AI-Assisted Assessment in Higher Education

The rapid development of Large Language Models (LLMs) has significantly influenced teaching, learning, and assessment practices in higher education. Recent research highlights both the potential and the risks of integrating generative Artificial Intelligence (AI) systems into academic evaluation processes [1], [2]. While LLMs have demonstrated strong capabilities in text generation and feedback provision, their reliability in structured assessment contexts remains underexplored [3] [4]. In professional postgraduate programs, where evaluation must balance analytical rigor and disciplinary standards, the question of whether AI can apply validated rubrics consistently and transparently becomes particularly relevant.

Recent research on human–AI collaborative assessment suggests that large language models can complement traditional grading by providing structured feedback and preliminary scoring aligned with rubric-based frameworks. However, concerns remain regarding reliability, bias, and interpretability when AI systems are applied to evaluative tasks traditionally performed by domain experts [1], [5]. Consequently, empirical validation comparing AI-generated scores with professional evaluators remains a critical research priority in engineering and professional education.

Challenges of Rubric-Based Evaluation in Professional Postgraduate Programs

Rubric-based assessment is widely adopted in postgraduate education to enhance transparency, reduce subjectivity, and promote alignment between learning outcomes and evaluation criteria [6]. However, even validated analytical rubrics require trained evaluators, and inter-rater variability remains a persistent challenge, particularly in argumentative writing tasks [7]. In PMI-aligned Master's programs, final essays demand both rhetorical competence and domain-specific reasoning grounded in Project Management Body of Knowledge (PMBOK) standards [8], increasing evaluative complexity. Ensuring consistent scoring across professional evaluators is therefore both a pedagogical and quality assurance concern.

Large Language Models and Analytical Rubric Application

Emerging studies suggest that LLMs may approximate human evaluators when structured prompts and scoring frameworks are provided [9]. Nevertheless, most investigations have focused on general academic writing rather than discipline-specific professional contexts. The ability of LLMs to interpret evaluative criteria [10], [11], assign dimension-level scores, and provide qualitative feedback aligned with validated rubrics remains a developing research area [12]. In particular, little empirical evidence exists regarding whether LLMs can reliably apply multi-dimensional analytical rubrics within structured professional education environments.

Research Gap in Domain-Specific LLM-Based Evaluation

Although AI-assisted grading tools are increasingly discussed in educational innovation literature [13], limited research has examined their application in professional Master's programs aligned with international standards such as those

¹24th LACCEI International Multi-Conference for Engineering, Education, and Technology: "Engineering without Borders: Artificial Intelligence, Knowledge, Innovation, and Alliances for a Future from the Americas". Hybrid Event, Santiago-Chile, July 15 - 17, 2026.

of the Project Management Institute (PMI). The integration of a validated argumentative writing rubric with domain-specific PMBOK essay tasks [14] represents a novel intersection of educational measurement and AI-assisted assessment. There remains a need for exploratory evidence examining the degree of alignment between LLM-generated scores and professional human evaluators in this specialized academic context.

Purpose and Research Questions

This exploratory pilot study aims to examine whether a LLM can reliably apply an existing validated analytical rubric (RUBRIAR [15]) to PMBOK-based argumentative essays in a PMI Master's program. Specifically, the study investigates the degree of correlation between LLM-generated dimension-level scores and those assigned by two professional evaluators. By situating the inquiry within educational measurement innovation [16], this research seeks to provide preliminary empirical evidence regarding the feasibility, reliability, and limitations of LLM-assisted rubric-based assessment in professional postgraduate education. The following questions are formulated:

- 1) To what extent do LLM-generated dimension scores correlate with professional human evaluators when applying a validated analytical rubric?
- 2) Does the LLM systematically overestimate or underestimate essay quality compared to human raters?
- 3) Are certain rubric dimensions more aligned between human and LLM evaluations than others?

Contributions of this Study

This study contributes to the emerging intersection between AI and professional engineering education by proposing and empirically examining a structured framework for LLM assisted rubric based assessment in project management contexts. Specifically, it integrates human expert evaluation, rubric embedded prompting, and statistical alignment analysis into a coherent methodological pipeline. Moreover, the study provides empirical evidence regarding the degree of agreement between LLM scoring and professional evaluators in PMBOK aligned argumentative essays, while exploring how domain specific knowledge structures may shape evaluative behavior. In doing so, the research seeks to clarify the feasibility and limits of structured LLM application within standards based postgraduate assessment environments.

Structure of the Paper

The remainder of this paper is organized as follows. Section II reviews relevant literature on AI-assisted assessment, rubric-based evaluation, and large language model applications in higher education. Section III describes the study design, essay corpus, human and LLM evaluation procedures, and statistical analysis approach. Section IV presents descriptive statistics, inter-rater agreement, Generative pretrained transformers (GPTs) and human correlation findings, and PMBOK knowledge area analysis. Section V interprets the results in

relation to educational measurement innovation and domain-specific assessment challenges. Finally, Section VI summarizes the main findings, discusses limitations, and outlines directions for future research.

II. RELATED WORK

This section reviews prior research on automated essay scoring, rubric-based assessment, and emerging applications of LLMs in higher education, situating the study within contemporary debates on reliability, validity, and AI-assisted evaluation practices.

Automated Essay Scoring and AI in Higher Education

Automated Essay Scoring (AES) systems have evolved considerably over the past decade, transitioning from rule-based and feature-engineered approaches to deep learning and transformer-based architectures [17]. Recent studies indicate that AI-driven scoring systems can approximate human grading performance in structured writing tasks, particularly when trained on large annotated datasets [18], [19]. However, most AES research has focused on standardized testing environments rather than postgraduate professional education. Moreover, the interpretability and transparency of AI-generated scoring remain ongoing concerns, especially in high-stakes academic assessment contexts.

Early AES systems relied primarily on handcrafted linguistic features and regression-based scoring models. With the introduction of deep learning architectures, particularly transformer-based models, scoring systems increasingly rely on contextual semantic representations capable of capturing discourse-level features of writing quality. More recently, LLM have demonstrated the ability to perform zero-shot or few-shot rubric-based evaluation without task-specific training datasets, opening new possibilities for AI-assisted educational assessment [20].

Large Language Models for Academic Feedback and Evaluation

With the emergence of large language models (LLMs), research has increasingly explored their use for generating formative feedback, grading support, and rubric-based evaluation [20]. Preliminary evidence suggests that LLMs can provide coherent qualitative feedback aligned with explicit criteria when prompts are carefully structured [21]. Nevertheless, questions remain regarding scoring stability, reproducibility across runs, and sensitivity to prompt formulation [22]. While LLMs demonstrate strong natural language understanding, empirical validation of their alignment with validated analytical rubrics remains limited.

Rubric-Based Assessment and Reliability Challenges

Analytical rubrics are widely used in higher education to promote fairness, transparency, and criterion-based grading [6]. Despite their theoretical strengths, inter-rater variability continues to pose challenges, particularly in argumentative and discursive tasks. Studies in educational measurement emphasize the importance of reliability testing and construct validity

when applying rubrics in professional contexts [23]. The potential of AI systems to support or augment rubric-based assessment has been discussed, but systematic comparisons between LLM scoring and expert professional evaluators are still emerging [5].

LLM Alignment with Domain-Specific Standards

Most current research on AI-assisted grading focuses on general academic writing rather than discipline-specific professional frameworks. In fields such as engineering, law, and business education, evaluation often requires alignment with established professional standards [24]. In project management education, argumentative essays must integrate PMBOK knowledge areas and structured decision reasoning, adding a domain-specific dimension to evaluation. To date, little empirical research has examined whether LLMs can consistently apply multi-dimensional rubrics within professional standards-based postgraduate programs [25].

Research Gap

Although AI-assisted assessment is rapidly expanding, there is limited exploratory evidence examining the reliability of LLM-based rubric application in PMI-aligned Master’s programs. Existing literature has not sufficiently addressed the intersection between validated analytical writing rubrics, professional standards frameworks [26] such as PMBOK, and LLM-based scoring comparison against domain experts. This study seeks to address this gap by empirically examining whether GPT can approximate professional evaluators when applying a validated multi-dimensional rubric to postgraduate PMBOK-based argumentative essays.

Unlike traditional automated essay scoring systems that rely on feature engineering or supervised training pipelines, the present study evaluates a zero-shot, rubric-embedded prompting strategy within a domain-specific professional framework. This distinction positions the work within emerging LLM-based interpretive evaluation paradigms rather than conventional predictive grading models.

III. METHODOLOGY

This section describes the study design, participants, evaluation procedures, prompt configuration, and statistical analysis. It outlines how human and GPT scores were collected, structured, and compared. The methodological framework was designed to examine alignment patterns while ensuring transparency, reproducibility, and conceptual rigor within a PMI-aligned postgraduate assessment context.

LLM-Assisted Evaluation Pipeline

Figure 1 illustrates the experimental workflow used to compare human evaluators with LLM-assisted rubric scoring. The process begins with the collection of student argumentative essays developed within a PMBOK-aligned academic context. Prior to analysis, all essays were anonymized to ensure confidentiality and eliminate potential evaluator bias. The anonymized corpus was then assessed independently by

two professional evaluators using the RUBRIAR analytical rubric. In parallel, the same essays were submitted to the GPT model through a rubric-embedded prompt designed to replicate standardized evaluation conditions. The model generated structured outputs containing dimension-level scores and brief justifications in JSON format, enabling systematic data extraction. Subsequently, human and GPT scores were aggregated and analyzed using descriptive statistics and correlation measures. This pipeline enables a controlled comparison between expert human judgment and LLM-assisted evaluation, supporting the investigation of alignment patterns between both assessment approaches.

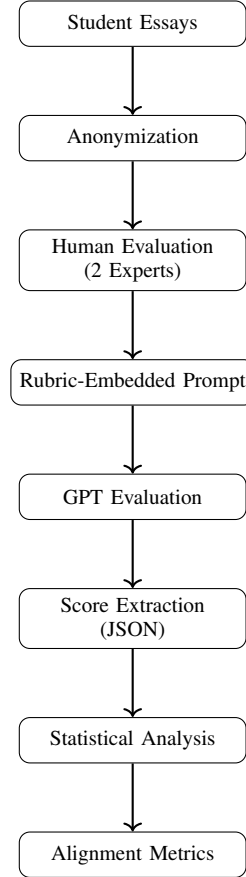


Fig. 1. Experimental workflow for comparing human evaluators and LLM-assisted rubric scoring.

This pipeline ensures methodological separation between human expert judgment and AI-assisted scoring, allowing controlled comparison of evaluative behavior.

Study Design and Context

This study adopts an exploratory pilot design aimed at examining the feasibility of LLM assisted evaluation in a professional Master’s program aligned with PMI standards. Specifically, the study investigates whether GPT can reliably apply an existing validated analytical rubric (RUBRIAR) to argumentative essays developed within a PMBOK-based academic context.

The research is positioned within the domain of educational measurement innovation, seeking preliminary empirical evidence regarding the alignment between human expert evaluation and AI-assisted rubric-based assessment in postgraduate professional education.

Participants and Essay Corpus

The corpus comprises 10 final argumentative essays produced by students enrolled in a PMI aligned Master's program. These essays constituted the final graded assignment and were explicitly structured around PMBOK knowledge areas. All participants belonged to the same academic cohort, ensuring academic homogeneity. Essays were anonymized prior to analysis to preserve confidentiality and comply with ethical standards. Students were required to articulate a clear thesis linked to a selected PMBOK knowledge area and to develop structured arguments grounded in project management principles. Additionally, they were expected to incorporate considerations of trade offs, risks, and stakeholder implications, while maintaining precise professional terminology and formal academic conventions. Therefore, the task was intentionally designed to integrate argumentative competence with domain specific alignment. This design enables the study to examine how structured professional reasoning can be consistently evaluated within a standards based framework. This study is designed as a pilot methodological exploration rather than a statistical validation study.

Human Evaluation Procedure

Two professional evaluators with expertise in project management and postgraduate assessment independently assessed the essays using the validated RUBRIAR analytical rubric [15]. This instrument evaluates four macro dimensions, Genre Adequacy (AGDE), Rhetorical Organization (ORGE), Cohesion and Coherence (RCC), and Normative Accuracy (ANL), each operationalized through explicit criteria. Specifically, each dimension was scored on a 0–5 numerical scale to ensure standardized quantification. Evaluators conducted the grading process independently and remained blind to each other's scores, thereby minimizing potential influence or convergence effects. Subsequently, the mean of both evaluators' scores was calculated for each dimension and used as the human reference benchmark. This procedure establishes a reliability baseline against which LLM generated scores can be systematically compared within a structured analytical framework.

A. LLM-Based Evaluation Procedure

GPT was employed to evaluate the same anonymized essay corpus under identical task conditions. The evaluation was conducted using the GPT-4 architecture through the OpenAI API. The temperature parameter was fixed at 0 to reduce output variability and ensure deterministic scoring behavior. Each essay was evaluated in a single run using a standardized prompt containing the full RUBRIAR rubric and structured JSON output instructions. The full RUBRIAR analytical rubric, including explicit dimension descriptions and scoring

criteria, was embedded within the prompt to guide the model's evaluation process [15]. Specifically, GPT was instructed to assign a numeric score from 0 to 5 for each macro dimension and to generate qualitative feedback at the subdimension level. Additionally, the model was required to justify each assigned score in direct relation to the rubric criteria. By structuring the prompt in this manner, the study sought to approximate standardized human evaluation conditions while examining the model's capacity for criterion aligned judgment. This approach enables a systematic comparison between expert assessment and LLM based scoring within a controlled evaluative framework. To assess prompt sensitivity, a subset of essays was re-evaluated across multiple runs. Moderate intra-model variability was observed, reinforcing the importance of structured prompt design. Although mean differences remained within acceptable exploratory margins, the presence of variability underscores the necessity of structured prompt design and stability testing in educational applications.

Corpus of PMBOK-Based Argumentative Essays

The essay corpus was structured around the ten PMBOK Knowledge Areas, ensuring domain-specific alignment with professional project management standards. Each essay required students to ground their argumentative reasoning within one selected Knowledge Area, thereby integrating rhetorical competence with formal disciplinary concepts. Table I summarizes the Knowledge Areas and their corresponding definitions as defined in the PMBOK 6th Edition framework.

B. Alignment of RUBRIAR Dimensions with PMBOK-Based Argumentative Essays

The evaluation framework adopted in this study is grounded in the validated RUBRIAR analytical rubric for argumentative academic essays [15]. Although originally developed for general academic writing, RUBRIAR offers a multidimensional structure that can be systematically adapted to domain specific contexts. In this study, the rubric is reinterpreted to assess essays situated within the PMBOK knowledge areas framework. PMBOK based essays extend beyond generic persuasion, requiring structured reasoning aligned with project management standards, articulation of process interdependencies, and evidence informed decision making. Therefore, adapting RUBRIAR to this professional context enables a rigorous evaluative model that integrates discursive competence with standards based disciplinary alignment. This alignment allows the research to examine whether structured professional reasoning can be assessed consistently under a validated rubric framework. Table II summarizes this conceptual correspondence.

Theoretical Foundations of the RUBRIAR–PMBOK Alignment

Within PMBOK aligned essays, Genre Adequacy requires arguments grounded in project management standards, distinguishing personal interpretation from normative guidance. Thus, AGDE captures disciplined reasoning aligned with professional criteria. Rhetorical Organization reflects structured

TABLE I
PMBOK KNOWLEDGE AREAS AND THEIR FORMAL DESCRIPTIONS,
ALIGNED WITH THE PMBOK 6TH EDITION FRAMEWORK.

Knowledge Area	Description
Integration Management	Includes the processes and activities required to identify, define, combine, unify, and coordinate the various project management processes and activities within the Project Management Process Groups to ensure overall project coherence and alignment with strategic objectives.
Scope Management	Encompasses the processes required to ensure that the project includes all the work necessary, and only the work necessary, to complete the project successfully by defining, validating, and controlling scope boundaries.
Schedule Management	Consists of the processes required to manage the timely completion of the project through defining activities, sequencing tasks, estimating durations, developing schedules, and controlling schedule performance.
Cost Management	Involves the processes required to plan, estimate, budget, finance, fund, manage, and control project costs so that the project can be completed within the approved budget baseline.
Quality Management	Includes the processes for incorporating the organization's quality policy regarding planning, managing, and controlling project and product quality requirements in order to meet stakeholder expectations.
Resource Management	Encompasses the processes to identify, acquire, and manage the physical and team resources needed for successful project completion, including team development and performance management.
Communication Management	Consists of the processes required to ensure timely and appropriate planning, collection, creation, distribution, storage, retrieval, management, monitoring, and disposition of project information.
Risk Management	Includes the processes of conducting risk management planning, identification, analysis, response planning, implementation, and monitoring to increase the probability and impact of positive events and reduce the probability and impact of negative events.
Procurement Management	Involves the processes necessary to purchase or acquire products, services, or results needed from outside the project team, including contract management and change control.
Stakeholder Management	Includes the processes required to identify stakeholders, analyze their expectations and impact, and develop appropriate management strategies to effectively engage stakeholders in project decisions and execution.

project logic, including thesis clarity, justified argumentation, and consideration of trade offs among scope, cost, schedule, risk, and stakeholders. Accordingly, ORGE assesses whether reasoning mirrors professional decision processes. Because PMBOK emphasizes systemic integration, Cohesion and Coherence is interpreted as systems level articulation, examining logical connections across phases and management processes. Similarly, Normative Accuracy extends beyond grammar to conceptual precision in defining baselines, risk responses, and stakeholder strategies. Collectively, these dimensions establish a multidimensional evaluation model integrating discursive competence with domain standards, thereby supporting analytical scoring and reducing subjectivity. This framework enables the study to explore whether LLMs can consistently assess argumentative rigor and adherence to professional project management principles.

TABLE II
ALIGNMENT BETWEEN RUBRIAR DIMENSIONS AND PMBOK-BASED
ARGUMENTATIVE ESSAYS

RUBRIAR Dimension	Original Focus	Operational Alignment in PMBOK Essays
Genre Adequacy (AGDE)	Discursive adequacy to the argumentative academic essay; persuasive intention; differentiation of authorial voice and sources.	Conceptual alignment with PMBOK knowledge areas; correct use of project management terminology; explicit argumentative positioning grounded in PMBOK principles and standards.
Rhetorical Organization (ORGE)	Functional rhetorical structure: thesis, argumentation, counterargument, conclusion, references.	Structured reasoning based on PMBOK processes; coherent thesis development referencing knowledge areas; inclusion of counterarguments related to project trade-offs and risk scenarios.
Cohesion and Coherence (RCC)	Logical progression, textual cohesion, discourse markers, lexical precision.	Logical sequencing of project management concepts; coherent integration across knowledge areas; systemic articulation of planning, execution, monitoring, and control processes.
Normative Accuracy (ANL)	Orthographic and grammatical correctness.	Technical precision in defining PMBOK concepts; consistent terminology usage; formal academic citation standards aligned with professional writing conventions.

Prompt Design and Model Configuration

The prompt was deliberately structured to replicate standardized academic evaluation conditions. It positioned GPT as an expert postgraduate evaluator, embedded the complete RUBRIAR rubric, and provided explicit scoring instructions together with the anonymized student essay. Specifically, the rubric text was fully included to constrain interpretation and ensure criterion aligned judgment. Additionally, clear instructions were provided to standardize numeric scoring and qualitative justification across dimensions. By consolidating role specification, evaluative criteria, and task input within a single prompt, the design sought to approximate controlled human assessment procedures. Thus, the prompt architecture was intended to minimize ambiguity and enhance reproducibility, enabling systematic comparison between expert evaluators and LLM based scoring within a structured analytical framework.

To ensure transparency and reproducibility, the full

RUBRIAR rubric was embedded directly within the GPT prompt. The model was positioned as an expert postgraduate evaluator and instructed to assign numeric scores and provide brief justifications aligned strictly with rubric criteria. The prompt constrained the output format to structured JSON in order to standardize data extraction and minimize variability in responses.

You are an expert evaluator in postgraduate project management education. Evaluate the essay using the RUBRIAR rubric.

For each dimension | AGDE, ORGE, RCC, ANL | assign a score from 0 to 5 and provide a brief justification aligned with the rubric criteria.

Return the output in JSON format with score and justification for each dimension. Apply the rubric strictly and objectively.

Essay: [Anonymized text]

Implementation Details

The GPT-based evaluation was implemented through the OpenAI API using the GPT-4 architecture. Essays were submitted individually through structured prompts containing the full RUBRIAR rubric. The API configuration included temperature = 0 to ensure deterministic scoring behaviour, maximum token limit sufficient to accommodate full essay inputs, and standardized JSON output formatting. All prompts and evaluation scripts were executed through a Python-based pipeline to enable automated extraction of dimension-level scores and justifications. This configuration ensured consistent evaluation conditions across the essay corpus.

Scoring Structure and Data Extraction

For each essay, dimension level scores from two professional evaluators were collected, alongside their averaged human reference score and corresponding GPT dimension scores. In addition, GPT generated qualitative feedback at the subdimension level. While all quantitative analyses relied exclusively on numeric dimension scores to ensure statistical comparability, qualitative comments were examined descriptively to explore conceptual alignment with RUBRIAR criteria. Thus, the dataset enabled both structured numerical comparison and interpretive assessment of evaluative coherence. Importantly, this dual extraction strategy was designed to investigate not only score proximity but also whether GPT feedback reflected the underlying construct definitions embedded in the rubric. In doing so, the study sought to examine alignment patterns at both metric and conceptual levels within a controlled evaluative framework.

Statistical Analysis

Given the exploratory design and limited sample size, non-parametric statistical methods were applied. Spearman's rank correlation coefficient was calculated to examine the relationship between GPT dimension scores and mean human scores.

Additionally, mean absolute differences between GPT and human scores were computed to identify potential systematic overestimation or underestimation patterns. Descriptive statistics were used to examine variance across dimensions.

Exploratory Reliability Assessment

To assess agreement trends, dimension-level score distributions were compared visually and numerically. Although not intended as a full psychometric validation study, the analysis aimed to determine whether GPT scoring patterns approximate human evaluator judgments within acceptable exploratory thresholds.

Potential scoring bias and dimension-specific discrepancies were identified and interpreted cautiously, given the pilot nature of the study.

Ethical Considerations

All essays were anonymized prior to analysis. The study did not involve experimental manipulation of student grades; GPT scores were used exclusively for research purposes and did not affect official academic outcomes. The research adheres to institutional ethical standards for educational data analysis.

IV. RESULTS

This section presents the quantitative and qualitative findings derived from the comparative analysis of human and GPT evaluations. Descriptive statistics, agreement measures, and domain-level patterns are reported to examine scoring alignment across rubric dimensions and PMBOK Knowledge Areas within the exploratory pilot design.

Descriptive Statistics of Human and GPT Scores

This subsection reports descriptive statistics comparing scores assigned by human evaluators and GPT across the four RUBRIAR dimensions. Specifically, Table III presents the mean scores and standard deviations for each dimension, enabling direct comparison of central tendency and dispersion between evaluative sources. By examining these patterns, the analysis seeks to determine whether GPT scoring behavior approximates expert judgment at the dimension level. Thus, descriptive statistics serve as a preliminary indicator of alignment before more detailed inferential analysis. This step establishes the empirical foundation for assessing whether LLM based evaluation can reflect structured human scoring patterns within a standards aligned framework.

TABLE III
DESCRIPTIVE STATISTICS AND SCORE DIFFERENCES BETWEEN HUMAN AND GPT EVALUATIONS (N=10)

Dimension	Human Mean	GPT Mean	Mean Diff
AGDE	3.5	3.9	+0.3
ORGE	4.0	3.9	-0.1
RCC	4.0	4.2	+0.1
ANL	4.5	4.2	-0.2
Overall	4.0	4.1	+0.1

To visually illustrate dimension-level alignment patterns, Figure 2 presents a comparison of mean scores between human evaluators and GPT.

GPT scores exhibited a slight tendency toward overestimation in selected dimensions, while remaining closely aligned with human evaluative patterns overall. Importantly, variation

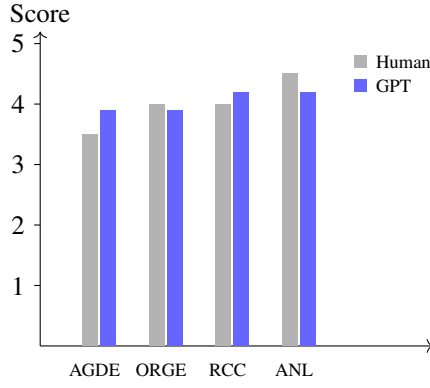


Fig. 2. Comparison of mean dimension-level scores between human evaluators and GPT.

across dimensions did not indicate systematic inflation or deflation, suggesting that GPT scoring behavior approximates the general structure of expert judgment. Thus, differences appear moderate and dimension specific rather than uniformly biased.

Dimension-Level Analysis

This subsection explores alignment patterns across individual RUBRIAR dimensions to identify areas of stronger or weaker agreement between GPT and professional evaluators. Overall, agreement was strongest in ANL, indicating that GPT demonstrated consistent performance in identifying grammatical accuracy and technical precision within PMBOK terminology. Similarly, high alignment in RCC suggests that the model reliably assessed logical progression and conceptual coherence across project management arguments. In contrast, comparatively lower agreement in AGDE reflects greater variability in evaluating genre positioning, authorial voice, and standards aligned argumentative framing. Qualitatively, this pattern suggests that GPT performs more consistently in structurally observable dimensions, such as language accuracy and cohesion, while exhibiting increased sensitivity and interpretive variability in dimensions requiring nuanced judgment of disciplinary stance and persuasive adequacy.

Analysis by PMBOK Knowledge Area

This subsection examines whether alignment between GPT and professional evaluators differs across the PMBOK knowledge areas represented in the essay corpus. Table IV summarizes the distribution of essays by Knowledge Area, reflecting natural variation in student topic selection within the Master's program. Because participation was not experimentally assigned, group sizes per area remain limited. Nevertheless, analyzing alignment at this level allows the study to explore potential domain specific variability in LLM scoring patterns. Thus, the subsection seeks to determine whether GPT demonstrates consistent evaluative behavior across distinct PMBOK contexts or whether alignment varies depending on disciplinary focus.

TABLE IV
DISTRIBUTION OF ESSAYS BY PMBOK KNOWLEDGE AREA (N=10)

Knowledge Area	Number of Essays
Integration Management	1
Scope Management	1
Schedule Management	1
Cost Management	1
Quality Management	1
Resource Management	1
Communication Management	1
Risk Management	1
Procurement Management	1
Stakeholder Management	1
Total	10

Table V summarizes the mean human and GPT scores across knowledge areas, along with mean score differences.

TABLE V
MEAN HUMAN AND GPT SCORES BY PMBOK KNOWLEDGE AREA

Knowledge Area	Human	GPT	Difference
Integration	4.0	3.6	-0.4
Scope	3.5	4.2	+0.7
Schedule	3.5	3.4	-0.1
Cost	4.5	4.4	-0.1
Quality	4.5	4.2	-0.3
Resource	4.0	4.0	0.0
Communication	4.0	4.4	+0.4
Risk	4.5	3.6	-0.9
Procurement	3.5	4.2	+0.7
Stakeholder	4.0	3.6	-0.4

Across PMBOK Knowledge Areas, alignment between GPT and professional evaluators varied in magnitude rather than direction, revealing domain specific sensitivity in scoring behavior. While several areas exhibited minimal differences, indicating close approximation of expert judgment, others showed more pronounced deviations. In particular, GPT assigned higher scores in domains such as Scope, Procurement, and Communication, suggesting a relative tendency toward generosity in areas emphasizing structured articulation and stakeholder framing. Conversely, lower GPT scores in Integration, Risk, and Stakeholder Management indicate greater divergence in domains requiring complex systemic reasoning or nuanced evaluative judgment. Importantly, no uniform inflation pattern is observed across all Knowledge Areas. Given the limited number of essays per domain, these findings remain exploratory. Thus, the results suggest that GPT's rubric application is broadly stable, yet potentially influenced by the cognitive demands and conceptual density characteristic of specific PMBOK areas.

Summary of Findings

This subsection synthesizes quantitative and qualitative results to provide an integrated interpretation of alignment patterns between GPT and professional evaluators. Overall, GPT demonstrated moderate to strong correspondence with expert scoring across RUBRIAR dimensions. Agreement was most

pronounced in structurally observable domains, particularly those related to technical precision and textual coherence, while comparatively lower alignment emerged in dimensions requiring nuanced evaluative judgment and argumentative positioning. Importantly, although minor directional differences were detected in selected areas, no systematic or extreme scoring bias was observed. Discrepancies remained within acceptable exploratory bounds relative to human inter rater variability. Taken together, the findings support the preliminary feasibility of LLM assisted rubric based assessment in a PMI aligned postgraduate context, while also highlighting dimension specific sensitivity that warrants further investigation. Inter rater reliability between the two professional evaluators was strong, as reflected by a Spearman correlation coefficient of $\rho = 0.87$. This high level of association indicates substantial consistency in the application of the RUBRIAR dimensions across essays. The result supports the internal stability of the human evaluation process and establishes a robust benchmark for subsequent comparison with GPT based scoring patterns within the exploratory framework.

V. DISCUSSION

This section interprets the empirical findings in relation to educational measurement theory, rubric-based assessment, and emerging research on LLM-assisted evaluation. It examines alignment trends, domain sensitivity, and practical implications, while acknowledging methodological limitations and outlining directions for future research in postgraduate professional education contexts.

Interpretation of Alignment Patterns

The findings indicate that GPT demonstrates moderate to strong alignment with professional evaluators across RUBRIAR dimensions, suggesting that LLM-based scoring can approximate structured human judgment in postgraduate project management assessment. Stronger agreement in technically observable dimensions, such as Normative Accuracy and Cohesion and Coherence, supports the view that LLMs perform reliably when evaluating formal linguistic and structural features. Conversely, comparatively lower alignment in Genre Adequacy suggests that dimensions requiring interpretive judgment, disciplinary positioning, and nuanced argumentative framing remain more challenging for automated evaluation.

Domain-Specific Sensitivity Across PMBOK Knowledge Areas

Variation across PMBOK Knowledge Areas suggests that GPT scoring behavior is not uniformly stable across domains. Areas emphasizing structured articulation and explicit procedural reasoning exhibited closer approximation to human judgment, whereas domains requiring systemic integration or complex risk reasoning displayed greater divergence. These patterns indicate that domain complexity and conceptual density may influence LLM scoring consistency. However, given the exploratory sample size, these observations should be interpreted cautiously.

Human Reliability as a Benchmark

To establish a human reliability reference, agreement between the two professional evaluators was examined using Spearman's rank correlation. The results indicate a strong positive correlation between evaluators, suggesting consistent application of the RUBRIAR rubric across the essay corpus. This agreement provides a baseline against which GPT scoring can be compared. Importantly, observed differences between GPT and human scores generally fall within the range of variability between human evaluators, supporting the interpretation that LLM-assisted scoring approximates structured human judgment in this exploratory context.

Implications for LLM-Assisted Assessment in Professional Education

The results provide preliminary evidence supporting the feasibility of LLM-assisted rubric-based evaluation in PMI-aligned postgraduate programs. Embedding the full rubric within the prompt appears to enhance structured scoring behavior, enabling alignment with domain standards. While LLMs cannot replace expert evaluators, they may serve as decision-support tools, formative feedback assistants, or preliminary screening mechanisms within standards-based professional education.

Implications for Engineering Education Assessment

LLM-assisted evaluation could support scalability in engineering programs with large cohorts while preserving rubric-based rigor. Within engineering education, rubric-based assessment is frequently used to evaluate complex competencies such as design reasoning, systems thinking, and professional communication. The integration of LLM-assisted evaluation could support scalable assessment practices in large engineering programs while preserving structured evaluation criteria. However, responsible deployment requires maintaining human oversight, transparency in AI-assisted scoring processes, and continuous calibration between automated and expert evaluation.

From a practical standpoint, AI-assisted assessment should operate within a supervised review loop, where LLM-generated scores are treated as preliminary analytical inputs rather than final grading decisions. Human evaluators remain responsible for interpretive validation, contextual nuance assessment, and ethical accountability. A hybrid workflow combining automated pre-scoring with targeted human moderation may optimize efficiency while preserving academic integrity.

Ethical Considerations in LLM-Assisted Assessment

The integration of Large Language Models into academic evaluation also raises important ethical considerations related to transparency, algorithmic bias, academic integrity, and the role of human oversight. From a transparency perspective, students and instructors must clearly understand how AI-supported evaluations are generated and how rubric criteria are applied within the model prompting structure. Additionally, LLM-based systems may inherit biases present in their

training data, potentially influencing linguistic or stylistic interpretations of student writing. Although the rubric-embedded prompting strategy used in this study constrains evaluation behavior, the possibility of algorithmic bias cannot be fully excluded and therefore requires ongoing monitoring. Academic integrity considerations are also relevant, as the use of AI in both writing and grading introduces new challenges regarding authorship, originality, and evaluation fairness. Consequently, LLM-assisted assessment should not be considered a replacement for professional evaluators but rather a decision-support mechanism operating under human supervision. Maintaining human oversight ensures that contextual interpretation, disciplinary judgment, and ethical accountability remain central to the evaluation process.

Limitations and Future Research Directions

This study is limited by its small sample size and exploratory design. The number of essays per PMBOK Knowledge Area was restricted, precluding inferential statistical comparisons at the domain level. Additionally, only a single prompt configuration was tested, and multi-run stability analysis was limited. Future research should incorporate larger samples, cross-cohort validation, prompt sensitivity analysis, and experimental manipulation of rubric structure. Investigating calibration strategies and dimension-level weighting may further refine LLM-assisted evaluation reliability. Additionally, prompt sensitivity may influence scoring outcomes, as minor variations in prompt formulation can alter LLM responses. While temperature was fixed to reduce variability, the possibility of latent hallucination or reasoning inconsistencies remains a methodological consideration when applying LLMs in evaluative contexts.

The limited sample size constrains statistical generalizability and prevents inferential domain-level comparisons. The findings should therefore be interpreted as exploratory alignment indicators rather than definitive validation evidence. Future studies should incorporate multi-cohort datasets and cross-institutional validation to evaluate scalability and robustness under heterogeneous academic contexts.

VI. CONCLUSION

This exploratory study examined whether GPT can reliably apply a validated analytical rubric to PMBOK-based argumentative essays within a PMI-aligned Master's program. The findings indicate moderate to strong alignment with professional evaluators across dimensions, with variability primarily observed in higher-order argumentative positioning. Importantly, GPT discrepancies generally remained within the bounds of human inter-rater variability, supporting the preliminary feasibility of structured LLM-assisted rubric evaluation.

The study contributes to emerging research on AI-supported educational measurement by demonstrating that rubric-embedded prompting can yield dimension-level scoring behavior comparable to expert assessment in a professional standards-based context. While not a substitute for human

expertise, LLMs show potential as structured evaluative assistants in postgraduate education.

Beyond its exploratory research contribution, the proposed framework also suggests several potential deployment scenarios within professional postgraduate education. First, LLM-assisted rubric evaluation could be used as a formative feedback system, providing preliminary diagnostic feedback to students before final grading. Second, such systems may function as decision-support tools for teaching assistants or instructors, helping to identify structural strengths or weaknesses in argumentative essays while preserving human judgment as the final authority. Third, LLM-based analysis could support rubric calibration processes, allowing academic programs to monitor scoring consistency across evaluators and cohorts. In this sense, LLM-assisted evaluation should be interpreted not as automated grading but as an augmentation mechanism designed to improve transparency, efficiency, and analytical consistency in rubric-based assessment environments.

Further empirical validation across larger samples and diverse disciplinary contexts is required to determine generalizability and long-term pedagogical implications. However, variability across prompt executions and knowledge domains indicates that LLM-based assessment requires controlled deployment conditions and continued empirical validation before institutional adoption.

REFERENCES

- [1] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günemann, E. Hüllermeier, *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102274, 2023.
- [2] X. Zhai, "Chatgpt for next generation science learning," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 29, no. 3, pp. 42–46, 2023.
- [3] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang, "What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education," *Smart learning environments*, vol. 10, no. 1, p. 15, 2023.
- [4] V. Nikolovski, D. Trajanov, and I. Chorbev, "Advancing ai in higher education: A comparative study of large language model-based agents for exam question generation, improvement, and evaluation," *Algorithms*, vol. 18, no. 3, p. 144, 2025.
- [5] C. Narreddy, S. Joordens, and S. Prompiengchai, "Harnessing large language models for scalable and effective formative assessment in higher education: A review," *Trends in Higher Education*, vol. 4, no. 4, p. 65, 2025.
- [6] S. M. Brookhart and F. Chen, "The quality and effectiveness of descriptive rubrics," *Educational Review*, vol. 67, no. 3, pp. 343–368, 2015.
- [7] P. L. Cooper, "The assessment of writing ability: A review of research," *ETS Research Report Series*, vol. 1984, no. 1, pp. i–46, 1984.
- [8] A. E. CATANA and A. FLORESCU, "Pmbok standards—a structured literature review," *contexts*, vol. 1, no. 4, p. 5, 2025.
- [9] C. Kooli and N. Yusuf, "Transforming educational assessment: Insights into the use of chatgpt and large language models in grading," *International Journal of Human-Computer Interaction*, vol. 41, no. 5, pp. 3388–3399, 2025.
- [10] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, *et al.*, "A survey on evaluation of large language models," *ACM transactions on intelligent systems and technology*, vol. 15, no. 3, pp. 1–45, 2024.
- [11] S. Baral, A. Botelho, A. Santhanam, A. Gurung, L. Cheng, and N. Hefernan, "Auto-scoring student responses with images in mathematics," *The Proceedings of the 16th International Conference on Educational Data Mining*, 2023.

- [12] T. Lee, H. Tu, C. H. Wong, W. Zheng, Y. Zhou, Y. Mai, J. S. Roberts, M. Yasunaga, H. Yao, C. Xie, *et al.*, “Vhelm: A holistic evaluation of vision language models,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 140632–140666, 2024.
- [13] M. Jukiewicz and M. Wyrwa, “Can chatgpt replace the teacher in assessment? a review of research on the use of large language models in grading and providing feedback,” *Applied Sciences*, vol. 16, no. 2, p. 680, 2026.
- [14] P. Mendonça, J. R. H. Carvalho, and A. C. Oran, “Using prompt engineering to enhance a project-based learning course on project management,” in *2025 IEEE Frontiers in Education Conference (FIE)*, pp. 1–9, IEEE, 2025.
- [15] S. Kloss, M. Tapia-Ladino, and S. Sagredo-Ortiz, “Estrategias de autorrevisión en escritura argumentativa: un estudio con alumnos de pedagogía,” *RLA. Revista de lingüística teórica y aplicada*, vol. 63, no. 1, pp. 103–129, 2025.
- [16] L. A. Baltador, V. Grecu, N. D. Panta, and A. M. Beju, “Design thinking in education: Evaluating the impact on student entrepreneurship competencies,” *Education Sciences*, vol. 14, no. 12, p. 1311, 2024.
- [17] Z. Ke and V. Ng, “Automated essay scoring: A survey of the state of the art,” in *IJCAI*, vol. 19, pp. 6300–6308, 2019.
- [18] M. Sailaja, D. Harini, M. Jhansi, P. Varshith, *et al.*, “Automated essay scoring using advanced transformer-based models for comprehensive assessment evaluation,” *Dasari and Jhansi, Mokirala and Varshith, Poshetty, Automated Essay Scoring Using Advanced Transformer-Based Models for Comprehensive Assessment Evaluation (November 10, 2025)*, 2025.
- [19] A. Pack, A. Barrett, and J. Escalante, “Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability,” *Computers and Education: Artificial Intelligence*, vol. 6, p. 100234, 2024.
- [20] E. Emirtekin, “Large language model-powered automated assessment: a systematic review,” *Applied Sciences*, vol. 15, no. 10, p. 5683, 2025.
- [21] K. Mzwri and M. Turcsányi-Szabo, “The impact of prompt engineering and a generative ai-driven tool on autonomous learning: A case study,” *Education Sciences*, vol. 15, no. 2, p. 199, 2025.
- [22] N. Arabzadeh and C. L. Clarke, “A human-ai comparative analysis of prompt sensitivity in llm-based relevance judgment,” in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2784–2788, 2025.
- [23] A. Yousef and A. Ayyoub, “Rubric development and validation for assessing educational robotics skills,” in *Frontiers in Education*, vol. 9, p. 1496242, Frontiers Media SA, 2024.
- [24] A. Jonsson and G. Svingby, “The use of scoring rubrics: Reliability, validity and educational consequences,” *Educational research review*, vol. 2, no. 2, pp. 130–144, 2007.
- [25] J. Rudolph, S. Tan, and S. Tan, “Chatgpt: Bullshit spewer or the end of traditional assessments in higher education?,” *Journal of applied learning & teaching*, vol. 6, no. 1, pp. 342–363, 2023.
- [26] C. Zhao, “Ai-assisted assessment in higher education: A systematic review,” *Journal of Educational Technology and Innovation*, vol. 6, no. 4, 2024.