This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: URL: https://ieeexplore.ieee.org/document/XXXXXXXX, 2025 International Conference on IT Innovation and Knowledge Discovery (ITIKD)

Developing a Framework for Using Large Language Models for Viva Assessments in Higher Education

Sara Alaswad
College of Information Technology
Ahlia University
Manama, Bahrain
salaswad@ahlia.edu.bh

Tatiana Kalganova
Brunel University of London
London, UK
tatiana.kalganova@brunel.ac.uk

Wasan Awad
College of Information Technology
Ahlia University
Manama, Bahrain
wawad@ahlia.edu.bh

***Abstract*— This paper presents a comprehensive framework for evaluating Large Language Models (LLMs) based on educational performance areas and established evaluation metrics. The study bridges the gap between traditional academic assessment criteria and modern AI evaluation techniques, aligning metrics such as coherence, relevance, completeness, and creativity with performance areas like problem definition, methodology, and product outcomes. Drawing insights from experimental results, the framework highlights the top 10 evaluation metrics frequently observed and emphasizes their significance in assessing AI-generated responses. A critical analysis identifies limitations in the initial framework proposed by ChatGPT, leading to refined strategies for more comprehensive evaluation. The refined framework addresses limitations of subjectivity, overlapping criteria, and weighting mechanisms, offering a dynamic evaluation model for both technical and educational contexts. The findings contribute to advancing interdisciplinary evaluation methodologies and offer valuable insights for educators, researchers, and developers in optimizing LLM applications for educational purposes.**

Keywords—LLM, higher education, VIVA, academic assessment, evaluation metrics, ChatGPT.

I. INTRODUCTION

The integration of large language models (LLMs) such as ChatGPT into various facets of education has sparked considerable academic interest and debate. As these advanced technologies increasingly influence how knowledge is disseminated and assessed, scholars are examining their potential to reshape traditional practices [1]. Recent studies have explored the transformative impact of LLMs across a spectrum of educational domains, from grading systems to critical thinking evaluation, and their broader usage in higher education.

Fagbohun et al. address the potential of LLMs to redefine educational assessments by challenging conventional grading paradigms [2]. Similarly, Tang et al investigate the use of ChatGPT in evaluating critical thinking, employing peer feedback analysis as a lens to measure its efficacy against established classification systems [3]. These studies emphasize the evolving capabilities of LLMs in fostering innovative approaches to evaluation and skill measurement.

In higher education, Yigci et al. provide a comprehensive review of the applications and challenges associated with LLMs, particularly ChatGPT, within university settings [4]. Their work underscores both the opportunities presented by these technologies and the critical concerns regarding their ethical and practical implications. Extending this discussion, Tayan et al. (2024) highlight the necessity of re-evaluating educational frameworks, particularly in technology courses, to accommodate the advent of AI-powered tools [5].

Partially funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission-EU. Neither the European Union nor the granting authority can be held responsible for them.

The role of LLMs in educational evaluation is further explored by Pillera, who discusses the limitations and opportunities of AI in supporting evaluation processes through a dialogue-based investigation with ChatGPT [6]. Collectively, these studies reveal a growing interest in the integration of LLMs into education, driven by their ability to address long-standing challenges while introducing new ethical and pedagogical considerations.

The unique context of VIVA assessment in higher education demands a broader set of evaluation criteria. These include project and human-centered metrics like project performance areas and students' skills. By refining these metrics, researchers and educators can better leverage LLMs to enhance VIVA experiences, address individual learning needs, and prepare students for a technologically advanced future.

This paper builds on the existing body of research by synthesizing insights from these diverse studies, providing a critical overview of the transformative role of LLMs in educational settings. Specifically, it aims to develop a comprehensive framework for using LLMs for Viva Assessments in Higher Education, focusing on their potential to enhance assessment practices while addressing associated challenge.

II. RELATED WORK

The integration of large language models (LLMs) into education has been studied across diverse applications, highlighting their potential and limitations. This section reviews recent advancements, methodologies, and findings relevant to evaluating LLMs in educational settings.

Meissner et al. (2024) introduced ItemForge, an automated tool for generating competence-based e-assessment items in higher education mathematics. Using GPT-3.5 and GPT-4, the study reviewed 240 generated items with input from three mathematical experts. The tool demonstrated proficiency in creating high-quality and concept-aligned items, though issues with incomplete or inaccurate solutions underscored the need for further refinement [7].

Lyu et al. (2024) conducted a semester-long field study evaluating CodeTutor, an LLM-powered assistant for introductory computer science education. The study involved 50 students and revealed significant improvements in final scores among those using CodeTutor. Despite its success in aiding programming tasks, such as syntax comprehension and debugging, its limited impact on fostering critical thinking and a decline in student engagement with the tool over time suggest room for improvement in its design and implementation [8].

Moore et al. (2024) explored the comparative effectiveness of LLM-driven, crowdsourced, and expert rubric applications for evaluating multiple-choice (MCQs) and

short-answer questions (SAQs) across six educational domains. Their study highlighted GPT-4's high reliability among LLMs, though challenges like prompt sensitivity and inherent biases in both human and AI evaluations pointed to the complexity of ensuring objective assessments [9].

Abeysinghe and Circi (2024) introduced EdTalk, a framework combining automated, human, and LLM-based evaluations. Their findings emphasized the benefits of factor-based evaluation in identifying areas for improvement in LLM applications, strengthening the argument for human oversight in critical contexts. However, the high cost of human evaluation remains a significant limitation [10].

Weissburg et al. (2024) investigated biases in LLMs used for personalized education. By analyzing 17,000 educational explanations across diverse topics, the study exposed biases in content generation linked to demographic attributes, including race, gender, and income. These findings underscore the need for more inclusive datasets and a nuanced understanding of LLM deployment in education [11].

Shankar et al. (2024) presented EvalGen, a mixed-initiative framework aligning LLM-generated evaluation criteria with human preferences. Although participants appreciated EvalGen's utility in creating initial assertions, the study's limited scope—focused on medical and product evaluation pipelines—highlighted the need for further iteration and real-world deployment to validate its broader applicability [12].

These studies collectively underscore the growing role of LLMs in educational assessments. They demonstrate their potential in automating evaluation tasks and enhancing learning outcomes, while also drawing attention to challenges such as biases, engagement dynamics, and the importance of human oversight in high-stakes scenarios. Building on the findings of prior research, this study makes the following key contribution:

- A novel framework that maps LLM-generated evaluation metrics to educational performance areas, offering a structured approach to integrating AI-Models into VIVA evaluations.

III. METHODOLOGY

This paper employs an experiment-based methodology to systematically investigate how ChatGPT evaluates its own responses. Given the nature of the research—examining the role of LLMs in educational assessments—this approach provides a clear, quantifiable understanding of the metrics that ChatGPT considers most relevant, offering insights into both AI-driven evaluation methods and their potential alignment with human assessment frameworks.

A. Setting Experiment

The primary objective of this experiment is to investigate the evaluation metrics employed by ChatGPT by examining its responses to a set of predefined questions and analyzing the metrics it uses for self-evaluation.

The experiment consisted of four stages: data collection, results cleaning, metric mapping, and result analysis.

1) Data Collection

Five different questions were identified based on their thematic categories, each falling into distinct subject areas. Questions were posed to ChatGPT, each asked five times to ensure variability and comprehensiveness. For each generated response, ChatGPT was subsequently asked to evaluate its own answer, providing a grade out of 5 and stating the evaluation metrics used for this evaluation. This process was repeated for each question-response pair, resulting in a total of 785 metrics.

2) Results Cleaning

Upon collecting the generated metrics, a cleaning process was undertaken. Initially, 82 unique metrics were identified. This set included variations due to typographical distinctions, , such as differences in capitalization (e.g., Language fluency vs. Language Fluency) or variations in the order of combined metrics when using the "and" operand (e.g., Clarity and Coherence vs. Coherence and Clarity). These inconsistencies were resolved, reducing the number of distinct metrics to 75.

3) Metric Mapping

To further refine the list of metrics, ChatGPT was tasked with defining all 75 metrics. Based on these definitions, a mapping exercise was conducted to group similar metrics based on semantic similarity. This process resulted in the consolidation of the metrics into 34 distinct categories.

4) Result Analysis

The final stage involved analyzing the refined metrics. Fig 1 shows the frequency of each metric's appearance to determine the most commonly used evaluation metrics. Fig 2 shows the top ten metrics identified, along with their respective frequencies.

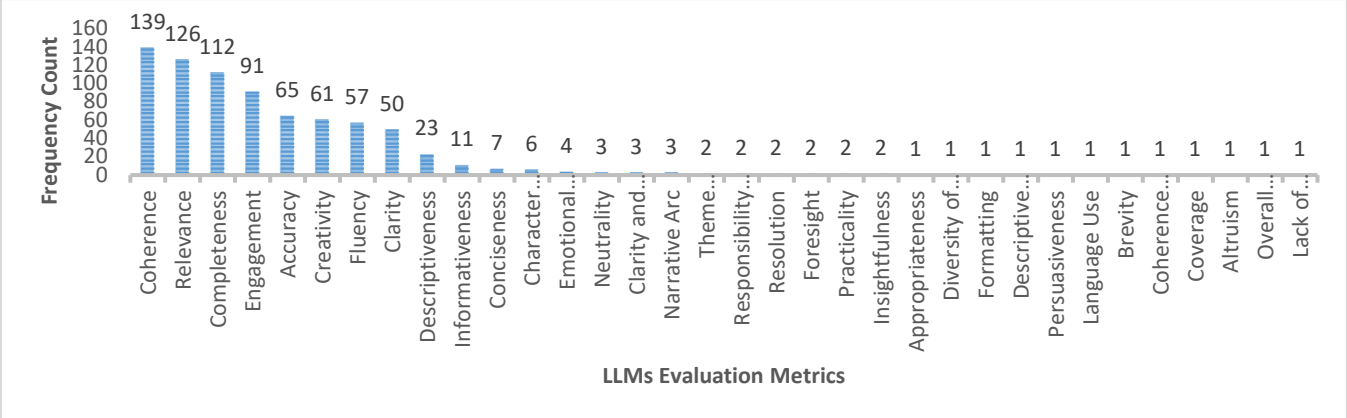

Fig. 1. LLM Evaluation Metrics derived from experimental results, with the corresponding frequency count for each metric.

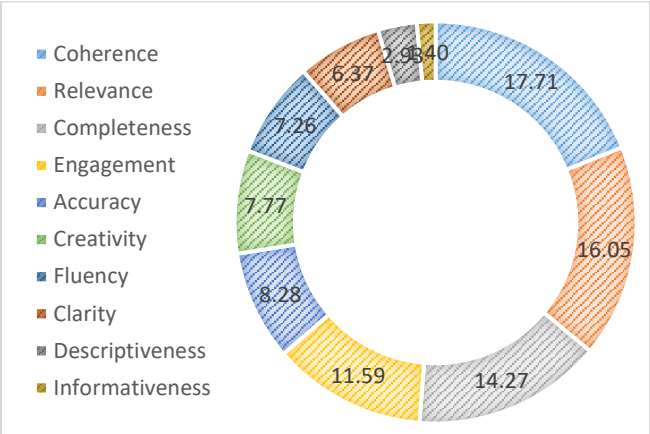

Fig. 2. The top 10 LLM evaluation metrics identified from the experiment showing the frequency percentage for each. The frequency was calculated based on the frequency count for each metric, normalized as a percentage of the total metric count.

This experiment provides insights into the evaluation metrics utilized by ChatGPT, highlighting key criteria such as Coherence, Relevance, Completeness, Engagement, Accuracy, Creativity, Fluency, Clarity, Informativeness, and Descriptiveness.

B. Proposed Evaluation Framework Using ChatGPT

A rubric is a structured framework used for assessment, providing clear criteria for evaluating performance across multiple dimensions. It typically includes defined categories, performance levels, and specific descriptors to ensure consistent and objective grading.

For the final project, the rubric table for the final project serves as a comprehensive evaluation framework to assess various aspects of a student's work. It categorizes performance into ten distinct areas as shown in TABLE I. Each area is graded on a four-tier scale:

1) *Exceeds Standards:* Exceptional performance, significantly above expectations.
2) *Meets Standards:* Satisfactory performance, meeting expected criteria.
3) *Partially Meets Standards:* Adequate but needs improvement in certain areas.
4) *Fails to Meet Standards:* Insufficient performance, lacking key elements.

The rubric provides a holistic scoring mechanism, ensuring objective evaluation of both technical and presentation aspects of the final project.

TABLE I. RUBRIC TABLE FOR THE FINAL PROJECT OUTLINING EDUCATIONAL PERFORMANCE AREAS

	Educational Performance Areas	**Description**
1	Problem Definition	Focuses on identifying and analyzing user needs (functional and non-functional) with effective tools. Levels of achievement range from thorough and efficient identification to imprecise or absent problem analysis.
2	Literature Search	Evaluates the ability to conduct a thorough and relevant search of academic, technical, or practical sources to support problem understanding and solution design. Achievement ranges from comprehensive and up-to-date searches to incomplete or irrelevant references.
3	Solution Design	Assesses the creativity, feasibility, and effectiveness of proposed solutions, including methodology and tools used. Ranges from highly innovative and practical designs to generic or incomplete approaches.
4	Result & Analysis	Assesses the clarity and completeness of results, analysis, and testing, along with the quality of conclusions. Ranges from high-quality, inventive analysis to incomplete, unclear results.
5	Solution implementation/Product	Reviews the functionality and logic of the developed product in meeting objectives. Ranges from fully functional, excellent logic to incomplete products with minimal functionality.
6	References & Citation	Evaluates the accuracy, consistency, and ethical use of references and citations according to appropriate academic or industry standards. Ranges from meticulous citation practices to absent or inconsistent references.
7	Documentation & Format	Examines the quality and organization of the report, including the use of sources and adherence to formatting standards. Ranges from error-free, well-structured documentation to poorly constructed or absent documentation.
8	Teamwork	Assesses the collaboration, communication, and contributions of team members. Ranges from highly cohesive, effective teams to dysfunctional or minimally cooperative groups.
9	Organization, Eye Contact & Delivery	Focuses on the logical flow of the presentation, maintaining audience engagement, and clear, confident communication. Ranges from fully engaging and professional delivery to disorganized, inaudible, or disengaging presentations.
10	Time Management & Presentation Skills	Evaluates the speaker's ability to use time effectively, integrate multimedia effectively, and demonstrate mastery of supporting materials. Ranges from well-managed presentations to poorly paced or incomplete ones.

TABLE II. provides the definitions of the top ten LLM Evaluation Metrics acquired from the experiment, detailing the key aspects used to evaluate the quality of language model outputs.

TABLE II. LLMS EVALUATION METRIC

	LLM Evaluation Metric	**Definition**
1	Coherence	The logical and consistent connection of ideas in the text. It measures how well the response flows and whether it makes sense as a whole.
2	Relevance	The degree to which the response addresses the given prompt or question. It evaluates how pertinent the content is to the user's query.
3	Completeness	The extent to which the response fully addresses all aspects of the prompt. It

		checks if the answer covers all necessary points and sub-questions.
4	Engagement	The ability of the response to capture and maintain the reader's interest. It looks at how compelling and captivating the content is.
5	Accuracy	The correctness of the information provided in the response. It evaluates whether the facts, figures, and assertions are true and reliable.
6	Creativity	The originality and inventiveness of the response. It measures how well the answer incorporates unique ideas, novel perspectives, or imaginative concepts.
7	Fluency	The smoothness and natural flow of the text. It looks at how well the sentences are constructed and how easily they read.
8	Clarity	The clearness and ease of understanding of the response. It assesses how well the response communicates its ideas without ambiguity or confusion.
9	Descriptiveness	The level of detail and vividness in the response. It measures how well the response uses sensory details and imagery to enhance the narrative.
10	Informativeness	The extent to which the response provides valuable and useful information. It measures the richness of content and how much the reader can learn from it.

To map the LLMs evaluation metrics with the specified educational performance areas, ChatGPT was asked to align each criterion with the most relevant performance area where it plays a significant role. Fig. 3 shows how these metrics have been mapped:

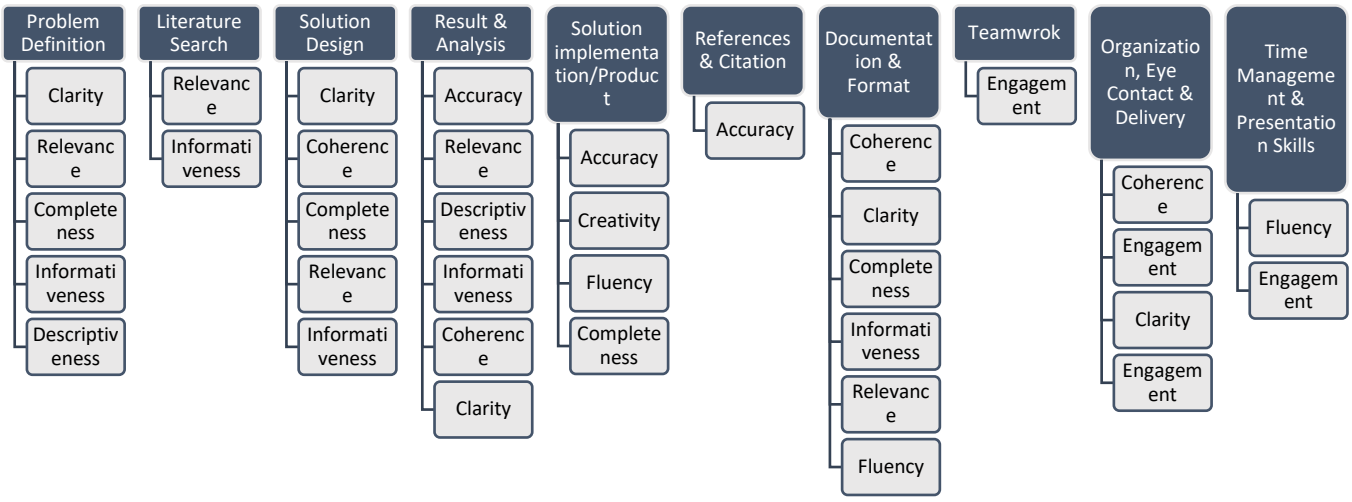

Fig. 3. Proposed Evaluation Framework using ChatGPT offering a structured approach to bridging LLM metrics with human educational evaluations

IV. Discussion

The proposed framework for mapping evaluation criteria (Coherence, Relevance, Completeness, Engagement, Accuracy, Creativity, Fluency, Clarity, Descriptiveness, and Informativeness) to educational performance areas (Problem Definition, Literature Search, Solution Design, Result & Analysis, Solution implementation/Product, Documentation & Format, Teamwork, Organization, Eye Contact & Delivery, and Time Management & Presentation Skills) represents a structured approach to bridging technical metrics with human-centric educational evaluations. This discussion critically examines the framework's strengths, limitations, and potential areas for refinement.

A. Strengths of the Framework

- **Comprehensive Coverage:** The framework aligns a diverse set of evaluation criteria with detailed educational performance areas, ensuring that both technical and pedagogical aspects are assessed comprehensively. For example, Accuracy is tied to multiple performance areas like Solution Design, Evaluation Result & Analysis, and Product, reflecting its importance across phases of a project.
- **Interdisciplinary Applicability:** By incorporating evaluation criteria commonly used in language models (e.g., Coherence, Informativeness) into educational contexts, the framework demonstrates versatility. It can be adapted for IT projects, academic presentations, or interdisciplinary research, making it broadly applicable. For example, in Computer Science, students can develop software solutions, ensuring logical design, functionality, and clear documentation. Engineering projects may involve CAD modeling, stress testing, and prototype development, emphasizing accuracy and teamwork. In Business & Marketing, students can analyze market trends, develop strategic plans, and present data-driven recommendations. Medical students can apply the framework through clinical case studies, focusing on accurate diagnosis, treatment planning, and patient outcome analysis. Similarly, in Humanities, historical research and critical analysis can be assessed through well-structured arguments and source evaluation. By integrating the framework into exams, assignments, and presentations, educators can ensure comprehensive and objective assessments tailored to each discipline.
- **Clarity in Mapping:** The framework establishes clear connections between abstract criteria and practical educational goals. For instance, Fluency is mapped to Time Management & Presentation Skills, emphasizing the importance of smooth and efficient delivery during presentations.
- **Focus on Engagement and Interaction:** Including criteria like Engagement in areas such as Organization, Eye Contact & Delivery, and Teamwork highlights the importance of maintaining

audience interest and interaction, which are critical in educational settings.

B. Limitations of the Framework

- **Overlapping Criteria:** Certain criteria, such as Clarity and Coherence, overlap significantly in their application across performance areas. While these overlaps reflect the interconnected nature of educational tasks, they may impact the precision of evaluation if not clearly distinguished.
- **Subjectivity in Interpretation:** Criteria like Creativity and Engagement are inherently subjective, and their evaluation can vary widely depending on the evaluator's perspective. This variability could lead to inconsistencies when applied across different contexts or evaluators.
- **Lack of Weighting Mechanism:** The framework treats all criteria equally, without specifying their relative importance in different educational performance areas. For example, Accuracy might be more critical in Problem Definition or Result & Analysis than Creativity, but this distinction is not explicitly addressed.
- **Limited Focus on Feedback Mechanisms:** While the framework excels at mapping evaluation metrics to performance areas, it does not emphasize feedback loops or iterative improvement. For instance, how the insights from an assessment could guide learners in refining their work is not discussed.
- **Potential for Overgeneralization:** The framework assumes equal applicability of all criteria across educational performance areas. However, certain criteria (e.g., Descriptiveness) may hold less relevance in specific contexts, such as Organization, Eye Contact & Delivery, leading to potential mismatches.

C. Refined Framework and Solutions

To enhance its effectiveness, the initial framework was first refined to address its identified limitations, providing more precision, flexibility, and applicability for both educational performance and LLM evaluation contexts. TABLE III. underscores the suggested improvements for each of the limitations of the initial framework.

TABLE III. INITIAL FRAMEWORK LIMITATIONS AND SUGGESTED IMRPOVMENTS

	Limitation	Improvement	Example Implementation
1	Addressing Overlapping Criteria	Distinct Definitions and Scope for Each Criterion	When mapping these criteria to Problem Definition, focus on "Coherence" for logical structuring of objectives, "Clarity" for articulation of the problem, and "Relevance" for aligning the problem to educational goals.
2	Mitigating Subjectivity	Quantifiable Rubrics for Subjective Criteria	During a presentation, audience participation can be tracked using real-time polling or feedback tools, ensuring engagement levels are measurable.
3	Introducing Weighting Mechanisms	Prioritize Criteria by Performance Area	When evaluating an IT project methodology, prioritize Accuracy and Clarity over Creativity, ensuring technical soundness remains central.
4	Embedding Feedback Mechanisms	Feedback Loops for Iterative Improvement	After initial evaluation, a reviewer could provide a scorecard with targeted suggestions, encouraging iterative refinement of the project.
5	Adapting to Context-Specific Needs	Contextual Flexibility in Mapping	A presentation in computer science might value Accuracy in results more, while a design project would prioritize Creativity in outcomes.
6	Enhancing Real-Time Evaluation	Real-Time Tools for Objective Measurement	During a live presentation, an AI tool could provide instant feedback on pacing, volume, and audience attention.
7	Integration of Dynamic Reporting	Generate Automated Performance Reports	After evaluating a student's presentation, provide a report showing scores for Clarity, Engagement, and Coherence alongside specific suggestions for improvement.

1) Addressing Overlapping Criteria

- **Coherence:** Logical flow and connection between ideas, ensuring smooth transitions.
- **Clarity:** Simplicity and understandability of the language and concepts.
- **Relevance:** Pertinence of content to the stated objectives or questions.
- **Completeness:** Inclusion of all necessary elements or information for thorough understanding.

2) Mitigating Subjectivity

Develop scoring rubrics for subjective criteria such as Creativity and Engagement. For example:

- **Creativity:** Evaluate based on originality (25%), applicability (25%), innovation (25%), and user-centricity (25%).
- **Engagement:** Assess via interactive elements (30%), audience response (30%), and sustained attention (40%).

3) Introducing Weighting Mechanisms

Assign weights to criteria based on the significance of each in a given performance area:

- **Problem Definition:** Clarity (40%), Relevance (30%), Completeness (20%), Coherence (10%).
- **Solution Design:** Accuracy (40%), Clarity (30%), Relevance (20%), Fluency (10%).

4) Embedding Feedback Mechanisms

Integrate structured feedback cycles for each performance area:

- **Problem Definition:** Provide feedback on unclear objectives and suggest refinements.
- **Solution implementation/Product:** Highlight areas of improvement in usability, scalability, or innovation.
- **Documentation & Format:** Suggest edits for structural improvements or incomplete sections.

5) Adapting to Context-Specific Needs

Tailor criteria to fit specific domains:

- **For IT projects:** Emphasize Accuracy, Relevance, and Scalability in Problem Definition and Solution Design.

- **For creative fields:** Highlight Creativity and Descriptiveness in Product and Documentation.

6) Enhancing Real-Time Evaluation

Use technology to monitor criteria such as:

- **Engagement:** Eye-tracking, polling tools, or sentiment analysis during a presentation.
- **Time Management:** Automated timers or pacing tools to ensure adherence to allotted time.
- **Clarity:** Speech analysis software to detect filler words or unclear articulation.

7) Integration of Dynamic Reporting

After evaluation, generate a report that breaks down scores by performance area and provides actionable insights:

- Highlight strengths and areas for improvement.
- Include visualizations (e.g., graphs, radar charts) for easy interpretation.

Figure 4 presents the refined framework, updated to address the limitations identified in the initial design.

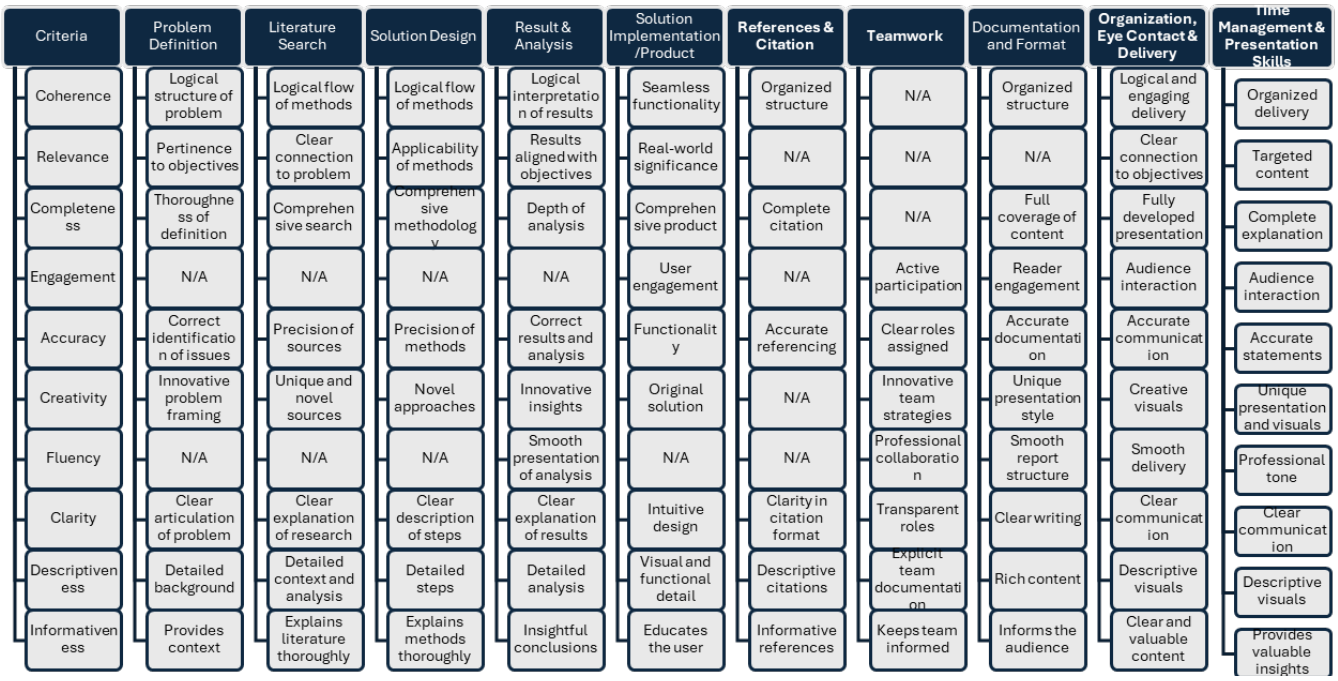

Fig. 4. Refined Framework to Address Identified Limitations

D. *Strategies for Training Educators to Use LLMs in Assessments*

While AI-driven assessment tools enhance efficiency and objectivity, human oversight remains crucial to prevent over-reliance on automated evaluations. Educators must critically review AI-generated assessments to ensure contextual accuracy, fairness, and alignment with learning objectives. AI may struggle with nuanced reasoning, creativity, and ethical considerations, requiring human judgment to validate results. Moreover, instructors play a key role in interpreting student responses, providing personalized feedback, and addressing potential biases in AI-generated evaluations. By integrating AI with human expertise, the assessment process remains balanced, ensuring that technology complements rather than replaces the educator's role in fostering critical thinking and holistic learning.

1) Workshops and Hands-on Training

- Organize interactive workshops to familiarize educators with LLM functionalities, applications, and best practices.
- Provide hands-on exercises where educators use LLMs for grading, feedback generation, and question formulation.

2) Guided Practice with Real Assessments

- Allow educators to experiment with LLM-generated assessments using actual student work.
- Compare AI-generated feedback with human evaluations to identify strengths and limitations.

3) Ethics and Bias Awareness Training

- Educate teachers on AI biases, ethical considerations, and responsible use of LLMs in assessments.
- Develop guidelines for validating AI-generated content and ensuring fairness.

4) Customizable AI Integration

- Train educators on adjusting AI parameters to align with specific course objectives and assessment criteria.
- Teach them how to fine-tune prompts for discipline-specific assessments.

5) Human-AI Collaboration Framework

- Develop protocols for human oversight to review, validate, and refine AI-generated assessments.
- Encourage a blended approach where AI assists but does not replace human judgment.

6) Peer Learning and Knowledge Sharing

- Establish educator communities for sharing experiences, best practices, and challenges in AI-assisted assessments.
- Facilitate mentorship programs where experienced AI users support new adopters.

7) Continuous Support and Evaluation

- Provide ongoing technical support and access to AI literacy resources.
- Periodically evaluate AI's impact on assessment quality and refine training accordingly.

V. Conclusion

In this study, we examined the transformative potential of large language models (LLMs) across a wide range of educational applications, including grading systems, critical thinking assessment, and broader implementation in higher education. In contrast to existing frameworks mentioned in the literature, which often focus on specific aspects of LLM evaluation or rely heavily on human oversight, our proposed framework introduces a structured approach that directly maps LLM evaluation metrics to educational performance areas, specifically in the context of VIVA assessments.

The framework proposed and refined in this paper provides a structured, multidimensional approach to aligning large language model (LLM) evaluation metrics with key educational performance areas. Initially robust in its integration of technical and pedagogical perspectives, the framework's subsequent iterations address critical limitations such as overlapping criteria, the absence of weighting mechanisms, and the lack of feedback loops. These enhancements make the framework not only more precise but also adaptable to diverse educational contexts and AI-driven outputs.

By embedding quantitative rubrics, introducing context-specific adaptation, and ensuring real-time evaluative capabilities, the refined framework bridges the gap between technical metrics and human-centric evaluations. Its emphasis on actionable insights ensures relevance across traditional and AI-enabled educational environments. Future work will focus on extending this evaluation framework to viva assessments, leveraging AI models to evaluate students' performance in real-time. This advancement aims to ensure consistency, objectivity, and depth in oral assessments, further demonstrating the versatility and applicability of AI-driven educational tools.

References

[1] W. Awad and J. Moosa, "Implications of AI Chatbots in Education: Challenges and Solution," *J Stat Appl Probab*, vol. 13, no. 2, pp. 611–622, Mar. 2024, doi: 10.18576/jsap/130203.

[2] O. Fagbohun, N. P. Iduwe, M. Abdullahi, A. Ifaturoti, and O. M. Nwanna, "Beyond Traditional Assessment: Exploring the Impact of Large Language Models on Grading Practices," *Journal of Artificial Intelligence, Machine Learning and Data Science*, vol. 2, no. 1, pp. 1–8, Feb. 2024, doi: 10.51219/jaimld/oluwole-fagbohun/19.

[3] T. Tang, J. Sha, Y. Zhao, S. Wang, Z. Wang, and S. Shen, "Unveiling the efficacy of ChatGPT in evaluating critical thinking skills through peer feedback analysis: Leveraging existing classification criteria," *Think Skills Creat*, vol. 53, Sep. 2024, doi: 10.1016/j.tsc.2024.101607.

[4] D. Yigci, M. Eryilmaz, A. K. Yetisen, S. Tasoglu, and A. Ozcan, "Large Language Model-Based Chatbots in Higher Education," 2024, *John Wiley and Sons Inc*. doi: 10.1002/aisy.202400429.

[5] O. Tayan, A. Hassan, K. Khankan, and S. Askool, "Considerations for adapting higher education technology courses for AI large language models: A critical review of the impact of ChatGPT," *Machine Learning with Applications*, vol. 15, p. 100513, Mar. 2024, doi: 10.1016/j.mlwa.2023.100513.

[6] G. C. Pillera, "In dialogue with ChatGPT on the potential and limitations of AI for evaluation in education," 2023, doi: 10.7346/PO-012023-36.

[7] R. Meissner *et al.*, "LLM-generated competence-based e-assessment items for higher education mathematics: methodology and evaluation," *Front Educ (Lausanne)*, vol. 9, 2024, doi: 10.3389/feduc.2024.1427502.

[8] W. Lyu, Y. Wang, T. R. Chung, Y. Sun, and Y. Zhang, "Evaluating the Effectiveness of LLMs in Introductory Computer Science Education: A Semester-Long Field Study," in *L@S 2024 - Proceedings of the 11th ACM Conference on Learning @ Scale*, Association for Computing Machinery, Inc, Jul. 2024, pp. 63–74. doi: 10.1145/3657604.3662036.

[9] S. Moore, N. Bier, and J. Stamper, "Assessing Educational Quality: Comparative Analysis of Crowdsourced, Expert, and AI-Driven Rubric Applications," 2024. [Online]. Available: www.aaai.org

[10] B. Abeysinghe and R. Circi, "The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches," Jun. 2024, [Online]. Available: http://arxiv.org/abs/2406.03339

[11] I. Weissburg, S. Anand, S. Levy, and H. Jeong, "LLMs are Biased Teachers: Evaluating LLM Bias in Personalized Education," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.14012

[12] S. Shankar, J. D. Zamfirescu-Pereira, B. Hartmann, A. G. Parameswaran, and I. Arawjo, "Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences," Apr. 2024, [Online]. Available: http://arxiv.org/abs/2404.12272