A Prompt-Driven Framework for Reflective Evaluation of Course Alignment Using Large Language Models

Mashael M. Alsulami
Department of Information Technology, College of Computers and Information Technology,
Taif University, Taif 21944, Saudi Arabia

***Abstract*—Large language models (LLMs) such as ChatGPT are gaining attention in educational settings, yet their potential role in supporting course design and academic quality assurance remains underexplored. This study introduces a structured, prompt-driven framework that uses ChatGPT as a reflective tool to help faculty and curriculum designers evaluate the alignment between course learning outcomes (CLOs), assessment methods, and teaching strategies. Grounded in the standards of the National Center for Academic Accreditation and Evaluation (NCAAA) and Bloom's Revised Taxonomy, the system generates targeted, context-aware feedback using structured prompts modeled after official NCAAA forms. To enhance reliability and reduce hallucinations, the framework employs template-based prompt engineering and rule-based cognitive classification. A large-scale analysis was conducted across 56 CLOs to assess internal consistency, followed by expert validation from two academic reviewers who evaluated a sample of AI-generated feedback for accuracy, usefulness, and cognitive alignment. The findings highlight the tool's ability to surface alignment issues and offer constructive recommendations, while also demonstrating its potential as a scalable support system for curriculum review and accreditation readiness, complementing rather than replacing human expertise.**

***Keywords*—Prompt engineering; Course Learning Outcomes (CLOs); Large Language Models (LLMs); curriculum alignment; educational quality assurance**

I. Introduction

Ensuring that course design is pedagogically sound and aligned with learning expectations is a central concern for quality assurance in higher education. Curriculum alignment, linking learning outcomes with teaching strategies and assessments, is widely recognized as a core principle of effective instructional design [1]. However, despite the availability of structured accreditation frameworks, several studies have shown that instructors often struggle with articulating measurable outcomes and ensuring that course activities target the appropriate cognitive levels [2][3]. These challenges become more pressing as institutions are required to demonstrate evidence of alignment for accreditation and continuous improvement.

Recent innovations in educational technology have opened new possibilities for automating parts of the curriculum review process. Artificial intelligence (AI), particularly large language models (LLMs), is being explored for its potential to support outcome-based design by providing reflective feedback and consistency checks [4]. These models can help academic staff navigate regulatory requirements, avoid cognitive mismatches between CLOs and assessments, and generate suggestions based on taxonomic alignment principles such as Bloom's taxonomy. However, such tools are not meant to replace human judgment; instead, they offer a structured augmentation to facilitate deeper reflection and compliance with institutional standards.

Creating well-designed courses that truly align learning outcomes with assessments and teaching strategies is essential for delivering quality education, and for meeting accreditation standards. In Saudi Arabia, the National Center for Academic Accreditation and Evaluation (NCAAA) [5] provides clear frameworks and structured forms to help ensure consistency and excellence across higher education institutions. Still, many instructors and curriculum developers find it challenging to interpret these guidelines, detect misalignments, or clearly justify their course design decisions.

This research presents a new approach that brings artificial intelligence into the heart of course evaluation. We introduce a reflective tool powered by ChatGPT, a cutting-edge large language model, to assist faculty in reviewing and improving how CLOs, assessments, and teaching methods are connected. But unlike typical AI applications that generate open-ended content, our tool follows a structured, prompt-driven framework that mirrors the official NCAAA specification format.

What sets this work apart is its contribution to both practice and process: it offers a reliable, transparent, and scalable method for evaluating course alignment. Using Bloom's taxonomy as its foundation, the system identifies the cognitive level of each CLO and checks whether it's well-matched with the stated assessments and teaching strategies. The prompts are carefully designed to guide the model's responses, minimizing hallucination and keeping the feedback focused and consistent.

Rather than replacing human decision-making, this tool augments academic judgment by making the review process more transparent, efficient, and pedagogically sound. It offers a practical contribution to educational quality assurance practices, particularly in accreditation-driven contexts, and opens new directions for human-AI collaboration in curriculum evaluation.

II. Related Work

The integration of artificial intelligence (AI) into higher education has drawn significant attention, particularly with

the emergence of large language models (LLMs) such as ChatGPT. These tools have sparked interest as assistive agents in instructional design, assessment, and academic support. This section reviews existing research on the use of LLMs in educational quality assurance, with a focus on their applications in curriculum evaluation, limitations, and reflective learning support.

A. LLMs in Curriculum and Instructional Design

Several studies have investigated how ChatGPT and similar LLMs can support the instructional design process. For instance, Johansson et al. (2024)[6] explored the ability of ChatGPT to independently design higher education field courses, finding that while the model generated coherent learning objectives and activities, it often failed to account for contextual factors and discipline-specific pedagogy. Similarly, the CDIO 2024 Proceedings [7] include a set of experiments in which ChatGPT was tasked with drafting parts of course specifications. While the model performed well in identifying general structure and terminology, it struggled with generating outcomes that align with the cognitive demands of the discipline.

In line with this, Iftene et al. (2024) [8] emphasized the utility of ChatGPT in supporting educators in the process of formulating learning outcomes. Their findings highlight the model's utility in producing syntactically correct CLOs that often reflect Bloom's Taxonomy, though subject-matter specificity remained a concern.

B. ChatGPT for Assessment Evaluation and Feedback

The potential of ChatGPT in generating or evaluating assessments has also been documented. In a recent study by Van Dis et al. (2024)[9], ChatGPT was tested for its ability to assess the quality of student writing and exams. While it demonstrated promising consistency, its performance deteriorated on open-ended questions where nuanced human judgment was needed. Likewise, Sahlgren et al. (2023)[10] conducted controlled evaluations and concluded that while LLMs can surface generic feedback for rubric-based grading, they are less effective in subjective judgment or when factual correctness is crucial.

C. Limitations and Hallucination Risks

Several works have acknowledged inherent risks in relying on LLMs for educational judgment. Hallucination, generating incorrect or fabricated content, remains a key challenge. Salam. (2023) [11] specifically caution that ChatGPT often invents plausible but inaccurate content, especially in contexts involving institutional policies or accreditation standards. To mitigate this, recent approaches suggest structured prompt engineering [12], limiting model creativity by embedding scaffolding within the input prompt, and validating outputs against expert reviews.

Our approach,relying on structured, template-based prompts that mirror the NCAAA course specification format,builds upon this literature by explicitly controlling the scope of generative output. Additionally, our system incorporates Bloom's taxonomy verb parsing, a method aligned with existing literature [13], to establish a foundation for cognitive demand classification before invoking ChatGPT. This layered methodology aims to reduce drift and hallucination by constraining the model's interpretive boundaries.

D. Reflective Tools and Human-AI Collaboration

While many studies focus on automation, our research contributes to a growing body of work that views LLMs as reflective tools—not replacements for human judgment but catalysts for deeper reflection and review. This aligns with the philosophy of human-AI collaboration articulated by Luo et al. (2023) [14], who argue for maintaining human oversight in contexts where contextual, ethical, or pedagogical nuances are critical.

Our reflective framework differs from fully automated assessment systems by situating the model as a reviewer-in-dialogue rather than an oracle. It provides feedback on the alignment between learning outcomes, teaching strategies, and assessments. Such granular critique, when presented through a dashboard or structured feedback format, facilitates human deliberation and supports transparency in course design decisions.

In summary, our research addresses several limitations found in prior work: (1) it avoids open-ended generation by using grounded prompts and taxonomy-driven structure, (2) it evaluates assessment alignment with cognitive levels rather than generating them, and (3) it augments human judgment rather than substituting it. These features position our reflective tool as a practical and scalable solution for academic quality assurance, particularly in contexts governed by formal standards such as those of the NCAAA in Saudi Arabia.

III. Methodology

This study adopted a computational content analysis approach to evaluate the alignment between Course Learning Outcomes (CLOs) and their corresponding assessment methods within course specification documents. The methodology integrated automated text extraction, verb-level classification [15] using Bloom's Revised Taxonomy [16], and interpretation via a large language model (LLM), specifically OpenAI's GPT-3.5. As illustrated in Fig. 1, the process begins with inputting course specifications in PDF format, followed by text extraction. Extracted verbs are mapped to cognitive levels in Bloom's taxonomy, which guides prompt generation for ChatGPT-based analysis. The model then evaluates the alignment between CLOs and assessment and teaching methods and strategies, providing feedback and suggestions, which are compiled into structured outputs for review.

A. Data Collection

The dataset used in this study consisted of 56 annotated CLO instances extracted from course specifications of a Master's program in Artificial Intelligence. Course specifications were retrieved from publicly available sources on the websites of Saudi universities. Each entry includes structured metadata such as course title, CLO code and description, teaching strategy, assessment method, and model-generated feedback. Data ingestion was performed, where batch extraction of course content was automated. The dataset covered multiple CLOs per course and preserved contextual elements such as program title and level, ensuring validity across hierarchical program structures.

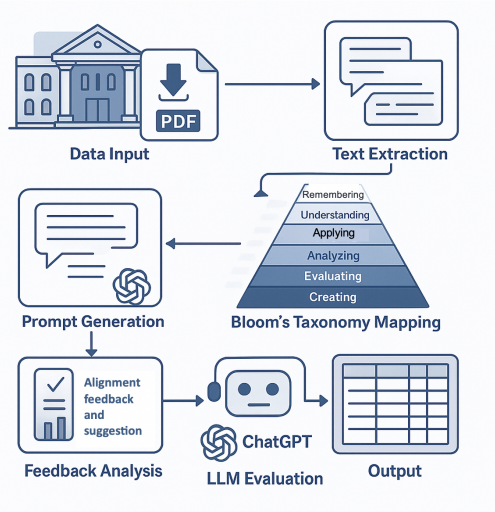

Fig. 1. Computational workflow for evaluating CLO-assessment alignment using Bloom's taxonomy and ChatGPT.

B. CLO Identification and Filtering

To ensure analytical rigor and consistency with national academic standards, only CLO entries that followed a standardized numerical format (e.g., 1.1, 2.2, 3.1) were included in the dataset. This structure directly reflects the official course specification template mandated by the National Center for Academic Accreditation and Evaluation (NCAAA), as specified in the official guidelines provided by the NCAAA [5]. Nearly all Saudi universities follow this core specification model as part of the broader transformation of higher education quality and accreditation in Saudi Arabia [17].

To avoid misinterpretation, the extraction process was designed to filter out non-actionable section headers such as "Knowledge," "Skills," or "Values," unless they were immediately followed by a clearly numbered sub-outcome. This rule-based filtering, implemented using regular expressions, minimized false positives and helped ensure that only assessable, action-oriented CLOs were analyzed.

For each valid CLO, the corresponding teaching strategy and assessment method were also extracted, in accordance with the tripartite structure outlined by the NCAAA. Assessment methods were categorized as either direct (e.g., quizzes, exams, case studies) or indirect (e.g., student surveys, course evaluations), enabling a detailed mapping between learning outcomes and how they are taught and assessed [5]. This structured alignment forms the backbone of outcome-based education and was critical to our study's analysis.

C. Cognitive Level Classification

To determine the cognitive demand embedded in each Course Learning Outcome (CLO), a verb-level classification algorithm was employed, grounded in Bloom's Revised Taxonomy [18]. This taxonomy remains a foundational framework in curriculum design and assessment, especially in outcome-based education models implemented by accreditation bodies such as the NCAAA [19]. The algorithm matched verbs within each CLO to a curated and expanded list of action verbs categorized across the six hierarchical levels of Bloom's cognitive domain:

- Remembering: list, define, recall
- Understanding: describe, explain, summarize
- Applying: implement, demonstrate, solve
- Analyzing: differentiate, compare, examine
- Evaluating: assess, critique, justify
- Creating: design, construct, develop

Each CLO text was scanned for one or more of these verbs, and the corresponding cognitive level was automatically assigned based on the match. This automated process was not only efficient but also allowed for objective categorization, ensuring consistency across all entries. Importantly, this form of classification provides a quantifiable lens through which CLOs can be evaluated for cognitive rigor, a key requirement in accredited program design.

Notably, approximately 86% of the CLOs analyzed contained clearly identifiable action verbs that could be confidently mapped to a Bloom's level. The remaining 14% were flagged as "unknown" due to abstract or vague phrasing, or due to formatting inconsistencies that hindered proper parsing (e.g., missing or malformed sentence structures). These entries were reviewed separately to avoid skewing the cognitive alignment analysis.

By integrating this classification step, the study enabled a structured comparison between intended learning outcomes and their respective assessment strategies. This alignment checking based on cognitive complexity, serves as a diagnostic tool for academic programs seeking to enhance the coherence and integrity of their curriculum.

D. AI-Assisted Evaluation and Hallucination Mitigation

Each CLO–assessment pair was evaluated using OpenAI's ChatGPT (GPT-3.5), which served as a reflective assistant rather than a final judge. To guide the model's reasoning and reduce the chances of irrelevant or inconsistent responses (commonly known as hallucinations), we designed a structured prompt template. This template included key elements such as:

- Program and course title
- CLO description
- Assessment method
- Teaching strategy
- The extracted Bloom's verb and its associated cognitive level

These components were chosen to provide the model with enough context to make informed and consistent evaluations. Recognizing the limitations of large language models, especially their tendency to fabricate or deviate from input, we implemented several best practices to keep the responses grounded and trustworthy.

First, we used prompt confinement, a strategy where prompts are written in a tightly scoped and consistent format.

Research has shown that this approach helps reduce factual drift and variability in model output [12], [20]. Second, we employed template-based reasoning, where the model was asked to reflect using logic similar to rubrics commonly used by educators [21]. Rather than asking the model to "know" the right answer, it was asked to reason based on the information provided.

Importantly, the model was not expected to draw on external or domain-specific knowledge. Instead, it was asked to evaluate only what was explicitly included in the prompt. After generation, responses were stored and reviewed for clarity, internal logic, and coherence. A portion of them were manually inspected to ensure that the feedback generated was interpretable and relevant.

It's worth emphasizing that the role of ChatGPT in this process is supportive, not authoritative. It serves as a first-pass evaluator, surfacing potential misalignments and offering justifications or improvement suggestions that can help curriculum designers reflect more deeply on the alignment between learning outcomes and assessment strategies. Human judgment remains central to the final decision-making process.

By combining structured prompting with thoughtful review, this approach enables scalable, replicable, and explainable evaluation of CLO–assessment alignment. It also shows promise for future integration into broader curriculum quality assurance systems suggesting the potential of large language models for educational auditing and curriculum evaluation workflows [22].

IV. Results

This section presents the results of a two-part evaluation of the AI-powered reflective tool designed to review the alignment between Course Learning Outcomes (CLOs), assessment methods, teaching strategies and cognitive complexity. The first part focuses on the tool's internal consistency, analyzing how well it identifies Bloom's taxonomy verbs, assigns cognitive levels, and interprets alignment across 56 CLOs. The second part examines how human experts perceive the tool's feedback. Two academic quality experts independently reviewed a sample of CLOs and rated the accuracy, usefulness, and cognitive alignment of the generated responses. Together, these findings offer a well-rounded view of how the tool performs in both automated checks and human-led evaluations. Section VI(A) explores patterns in Bloom's verbs, distributions of cognitive levels, and how frequently the tool flagged alignment issues. Section VI(B) highlights the level of agreement between experts and the AI tool, shedding light on where the model's evaluations aligned with expert judgment, and where improvements may be needed.

A. Self-Consistency Checks

To evaluate the internal consistency of ChatGPT's reflective evaluations across 56 Course Learning Outcomes (CLOs), a comprehensive analysis was performed on three key components: the Bloom's action verbs extracted from each CLO, the associated cognitive levels based on Bloom's Revised Taxonomy, and the nature of the corresponding assessment tasks. This triangulation aimed to determine whether the model's feedback logically corresponded with the expected learning complexity and pedagogical intent.

The CLOs were automatically parsed to detect the primary action verb and its associated Bloom's cognitive level. As illustrated in Fig. 2, the most commonly identified verbs included discuss, explain, analyze, and justify, spanning cognitive levels 2 (Understanding) to 5 (Evaluating). This distribution aligns with the higher-order thinking skills typically emphasized in postgraduate-level courses. A small number of CLOs were categorized under the "unknown" label, indicating either missing or ambiguous verbs that did not match the taxonomy database, highlighting potential issues in CLO phrasing or document formatting.

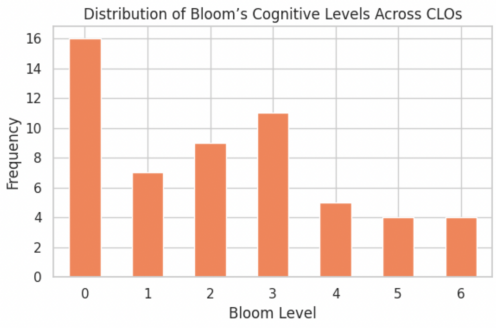

Fig. 2. Distribution of bloom's cognitive levels across CLOs.

Each ChatGPT-generated feedback response was then classified into one of three categories: Aligned, Misaligned, or Suggested. Feedback labeled as Aligned indicated that the model found no major inconsistencies between the CLO, its assessments, and teaching strategies. Misaligned denoted explicit concerns or mismatches raised by the model. Suggested included responses that, while not flagging critical misalignments, recommended improvements to strengthen alignment or clarity.

B. Illustrative Examples of Alignment and Misalignment

Table I presents two anonymized, representative CLO–assessment cases from our dataset, one *Aligned* and one *Misaligned*. For each case, we report the CLO text, Bloom level, assessment/teaching methods, the model's alignment label, and an abridged rationale highlighting specific alignment cues (e.g., verb–task cognitive match/mismatch, direct vs. indirect assessment, authenticity of performance tasks).

As shown in Fig. 3, a majority of responses were classified as Aligned, suggesting that many CLOs were appropriately constructed and assessed. However, a notable portion of feedback was marked as Misaligned or included improvement Suggestions, underscoring recurrent challenges in articulating learning outcomes and aligning them effectively with assessments and pedagogy.

These findings underscore the potential of structured LLM-based tools as formative quality assurance mechanisms. The distribution of Bloom's levels offers insight into instructional emphasis across the curriculum, while the alignment feedback highlights areas requiring further improvement in outcome

TABLE I. TWO COMPACT EXAMPLES: (*Left*) ALIGNED, (*Right*) MISALIGNED. TEXTS ARE ANONYMIZED AND LIGHTLY ABRIDGED

Field	Case A: Aligned	Case B: Misaligned
CLO	(example) "Describe core concepts of X and apply them to simple scenarios."	(example) "State definitions of X."
Bloom	Apply (L3)	Remember (L1)
Assess/Teach	Direct: exam, quiz, project; Indirect: survey / Lectures, guided exercises	Direct: quiz, exam; Indirect: survey / Lectures
Label	Aligned	Misaligned
Rationale	Tasks include a project and applied items that explicitly require L3 "apply", matching the CLO verb and expected level; a mix of direct/indirect evidence is present.	CLO targets L1 "remember" (state), yet assessment emphasizes higher-demand tasks (exam items requiring application/analysis) without corresponding CLOs; survey adds only indirect evidence, causing a verb–task mismatch.

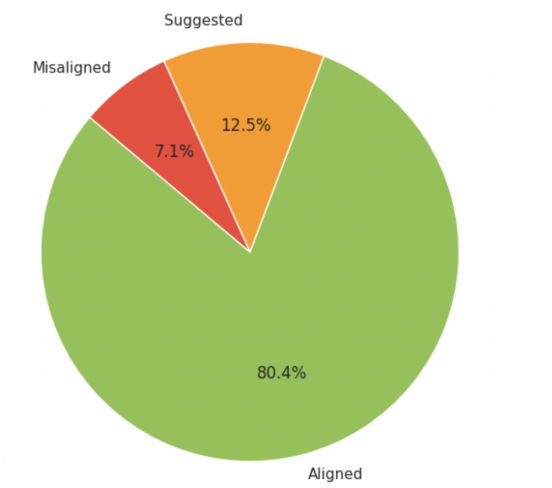

Fig. 3. Distribution of CLO–Assessment alignment outcomes.

construction and strategy selection. For curriculum designers and quality officers, these visualizations serve as practical dashboards, enabling them to detect trends, monitor inconsistencies, and prioritize reviews of misaligned CLOs. Crucially, this approach complements, rather than replaces, expert judgment by enhancing consistency, scalability, and transparency in curriculum evaluation workflows.

C. Human Expert Evaluation

To ensure the external validity of the AI-assisted evaluation and to provide a qualitative benchmark, a stratified sample of 12 Course Learning Outcomes (CLOs), representing approximately 20% of the full dataset, was independently reviewed by two academic quality experts. These experts were faculty members with substantial experience in course specification, learning outcomes design, and curriculum alignment in higher education.

Each expert received the original CLO statement, the associated assessment method, and the feedback generated by the AI-powered reflective tool (ChatGPT). They were asked to evaluate the AI's responses using a structured rubric built around three key dimensions, informed by best practices in AI-supported assessment [23], [24]:

- Accuracy: Whether the feedback correctly assessed the alignment between the CLO and the assessment.
- Usefulness: Whether the feedback provided practical value for improving CLO or course design.
- Cognitive Alignment: Whether the feedback appropriately matched the CLO's Bloom's level with its assessment method.

Ratings were recorded on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree), allowing for both descriptive and inter-rater agreement analysis.

Across the sample, the descriptive results showed strong overall agreement between the two experts, with most mean scores falling between 3 and 5. This suggests that the AI-generated feedback was generally perceived as accurate and useful. To statistically assess inter-rater consistency, Cohen's Kappa was calculated for each rubric dimension as shown in Table II. The results confirmed a moderate to substantial level of agreement, consistent with recent work highlighting the interpretability and limitations of large language models in educational tasks [25].

TABLE II. INTER-RATER AGREEMENT BASED ON COHEN'S KAPPA

Dimension	Kappa Score	Interpretation
Accuracy	0.41	Moderate agreement
Usefulness	0.12	Slight agreement
Cognitive Alignment	0.24	Fair agreement

This expert validation step reinforces the potential of LLMs to function as reflective aids in curriculum evaluation, provided that their outputs are critically reviewed by human experts, as part of a hybrid human–AI workflow.

Fig. 4 provides a side-by-side comparison of ratings from each expert across all CLOs and criteria. While the scores generally converged, certain items, particularly within the "Usefulness" dimension, showed divergence, highlighting the need for domain-specific refinement in feedback generation. The results show that the reflective tool provides reasonably consistent evaluations of CLO-alignment, particularly in identifying surface-level alignment mismatches. The tool's strength lies in its structured analysis based on Bloom's Taxonomy and its ability to assist with systematic curriculum review. However, lower inter-rater agreement on "Usefulness" suggests that some feedback lacked sufficient context or actionable depth. This emphasizes that while AI tools can support accreditation workflows, human oversight remains essential, especially when interpreting pedagogical intent or discipline-specific nuance. This expert validation underscores the tool's role as a decision-support system, offering scalable consistency while leaving room for human educational judgment.

V. DISCUSSION

The findings from this study underscore the promising role of structured, prompt-based large language models (LLMs)

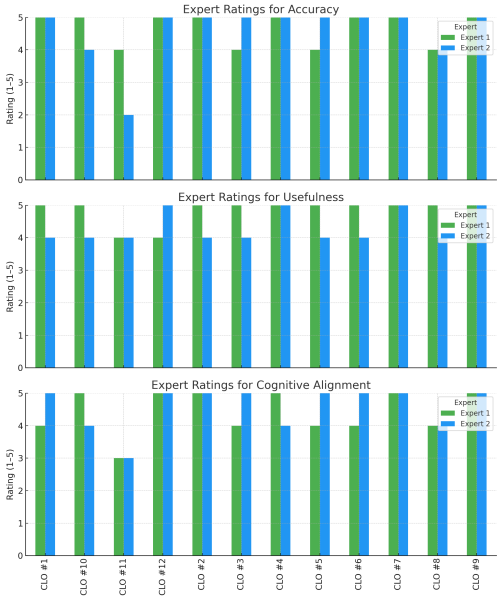

Fig. 4. Comparison of expert ratings for each CLO across three evaluation dimensions: accuracy, usefulness, and cognitive alignment.

in supporting curriculum evaluation tasks that require both cognitive interpretation and alignment with institutional standards. By focusing on Course Learning Outcomes (CLOs), their associated assessments, and pedagogical strategies, the reflective tool demonstrated an ability to surface misalignments and provide constructive, taxonomy-informed feedback.

A key insight emerging from the self-consistency analysis is the model's sensitivity to cognitive distinctions embedded in Bloom's Revised Taxonomy. The high prevalence of verbs corresponding to the "Understanding" through "Evaluating" levels aligns well with the nature of postgraduate course design, where deeper cognitive engagement is expected [26]. However, the relatively low frequency of "Creating" level verbs may suggest either an underrepresentation of higher-order synthesis tasks in course documentation or conservative formulation of CLOs that avoid explicitly committing to evaluative or generative assessments. This observation resonates with prior literature indicating that course developers often shy away from cognitive verbs associated with advanced competencies due to either misinterpretation or fear of overpromising learning deliverables [27], [13].

The model's classification of CLO-assessment relationships into "Aligned," "Misaligned," or "Suggested" reflects an operational advancement over prior implementations of LLMs in educational contexts, where feedback tends to be generic or contextually shallow [28], [29]. The structured prompt engineering and taxonomy-based scaffolding used in this study likely contributed to this improvement by narrowing the model's interpretive scope, a technique echoed in recent AI reliability research [12], [30]. These findings affirm that template-based confinement not only mitigates hallucinations but also enhances the model's ability to reason reflectively based on provided input rather than drawing from its latent knowledge alone.

Notably, the human expert evaluation results further validate the model's potential as a pedagogical aid. While moderate agreement on "Accuracy" and "Cognitive Alignment" suggests foundational reliability, the limited consensus on "Usefulness" points to an important limitation. Specifically, the model-generated feedback may lack domain-specific depth or contextual nuance, especially when CLOs involve interdisciplinary goals or implicit institutional constraints. This aligns with the findings of Luo et al. [31] and Iftene et al. [8], who caution that while LLMs can serve as reflective collaborators, they are not yet replacements for domain expertise. The divergences in expert ratings highlight the necessity of maintaining human oversight, particularly for evaluating the pedagogical appropriateness and practical feasibility of suggested improvements.

Importantly, the tool's value lies not in definitive judgment but in its capacity to structure and scale reflective evaluation. In contexts such as those governed by the NCAAA, where alignment and documentation rigor are essential, having a replicable first-pass reviewer that surfaces potential issues can reduce faculty workload and enable earlier intervention in curriculum development cycles. The use of visualizations (e.g., cognitive level distributions and alignment categorizations) enhances interpretability and provides a data-driven dashboard for institutional review, echoing similar recommendations in curriculum analytics literature [32], [33].

Nonetheless, the system is not without limitations. A small percentage of CLOs were unclassifiable due to ambiguous phrasing or formatting inconsistencies, highlighting ongoing challenges in automating quality assurance across diverse documentation standards. Furthermore, while the tool supports alignment evaluation, it does not currently assess whether the assessments themselves are valid measures of cognitive complexity, a task that likely requires deeper qualitative analysis or human triangulation.

VI. Conclusion

This study introduced a structured, prompt-based reflective framework that leverages ChatGPT to evaluate the alignment between Course Learning Outcomes (CLOs), assessment methods, and teaching strategies in line with NCAAA standards. Unlike general-purpose AI applications, our approach focused on targeted prompts grounded in Bloom's Taxonomy and the official course specification structure to generate consistent and relevant feedback.

Through self-consistency checks across 56 CLOs, the tool demonstrated a reasonable ability to detect Bloom's action verbs, classify cognitive levels, and flag potential misalignments in assessment alignment. A majority of the AI-generated feedback was classified as aligned, though a significant portion provided constructive suggestions or identified mismatches highlighting common design issues in course documentation. To validate the tool's reliability, two academic quality experts independently evaluated a stratified sample of CLOs. Their assessments showed moderate inter-rater agreement on accuracy, slight agreement on cognitive alignment, and limited agreement on usefulness suggesting that while the tool provides structured and technically sound evaluations, human oversight remains essential, particularly in interpreting nuanced instructional context.

Overall, this research demonstrates the potential of large language models to act as scalable and structured reflective tools in curriculum quality assurance. Rather than replacing educators, the system is designed to support them offering data-driven insights that can guide course improvement and strengthen institutional readiness for accreditation reviews.

DECLARATION

The author declares that she has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this document.

REFERENCES

[1] J. Biggs and C. Tang, *Teaching for Quality Learning at University*, 4th ed. McGraw-Hill Education, 2011.

[2] O. Al-Diban, H. Alshareef, and M. Saleh, "Analyzing curriculum alignment gaps using bloom's taxonomy in saudi universities," *International Journal of Educational Research*, vol. 118, p. 102234, 2024.

[3] R. Mahmoud, A. Alshammari, and R. Almefleh, "Evaluating learning outcomes for accreditation: A case study in saudi higher education," *Education and Information Technologies*, vol. 29, no. 2, pp. 456–478, 2024.

[4] L. Ha, Y. Chen, and J. Rios, "Reflective course design with llms: Minimizing hallucination and enhancing outcome alignment," *Computers & Education: Artificial Intelligence*, vol. 5, p. 100177, 2024.

[5] National Center for Academic Accreditation and Evaluation (NCAAA), *National Qualifications Framework and Course Specification Guidelines*, Education and Training Evaluation Commission (ETEC), Saudi Arabia, 2021, available at: https://etec.gov.sa/en/productsandservices/NCAAA/Pages/default.aspx.

[6] J. Johansson, D. Broman, and L. Plahn, "Field courses for dummies? to what extent can chatgpt design a higher education field course," *Higher Education*, vol. 87, pp. 1–17, 2024.

[7] V. Authors, "Cdio 2024 proceedings," in *Proceedings of the 20th International CDIO Conference*. CDIO, 2024, https://cdio.org/.

[8] A. Iftene, I. Ciocoiu, and L. Mihăescu, "Language models and higher education: The use of chatgpt in romanian academic settings," *Education Sciences*, vol. 12, no. 541, 2024.

[9] E. A. Van Dis, J. Bollen, and W. Zuidema, "What can chatgpt not do in education? evaluating its effectiveness in assessing educational learning outcomes," *Educational Technology Research and Development*, 2024.

[10] M. Sahlgren, D. Holmer, and F. Olsson, "Educational assessment with llms: Opportunities and risks," *Educational Technology & Society*, vol. 26, no. 2, pp. 22–35, 2023, preprint available from RTH. [Online]. Available: https://rth.se/publications/llms-assessment.pdf

[11] M. Sallam, "Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns," *Healthcare*, vol. 11, no. 6, p. 887, 2023. [Online]. Available: https://www.mdpi.com/2227-9032/11/6/887

[12] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," https://arxiv.org/abs/2102.07350, 2021.

[13] A. Lubbe, E. Marais, and D. Kruger, "Cultivating independent thinkers: The triad of artificial intelligence, bloom's taxonomy and critical thinking in assessment pedagogy," *Education and Information Technologies*, 2025, in Press.

[14] L. Luo, H. Yang, and X. Yu, "Human–ai collaboration in education: Principles and practices," *BMC Medical Education*, vol. 23, no. 1, p. 341, 2023.

[15] H. Kissane, A. Schilling, and P. Krauss, "Probing internal representations of multi-word verbs in large language models," *arXiv preprint arXiv:2502.04789*, 2025. [Online]. Available: https://arxiv.org/abs/2502.04789

[16] S. Nurmatova and M. Altun, "A comprehensive review of bloom's taxonomy integration to enhancing novice efl educators' pedagogical impact," *Arab World English Journal (AWEJ)*, vol. 14, no. 3, pp. 380–388, 2023. [Online]. Available: https://ssrn.com/abstract=4593934

[17] A. Alhazmi, "Higher education reform in saudi arabia: The journey toward vision 2030," *Journal of Education and Learning*, vol. 11, no. 1, pp. 100–110, 2022.

[18] L. W. Anderson and D. R. Krathwohl, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman, 2001.

[19] A. Bakharia, "Can chatgpt teach us about academic writing? large language models as ai writing tutors," *Media Education Research Journal*, vol. 9, no. 1, pp. 1–14, 2023. [Online]. Available: https://doi.org/10.5920/mep.20004

[20] X. Zhou, J. Liu, M. Sun *et al.*, "A survey of hallucination in large language models," *arXiv preprint arXiv:2302.08091*, 2023. [Online]. Available: https://arxiv.org/abs/2302.08091

[21] A. Lubbe, E. Marais, and D. Kruger, "Cultivating independent thinkers: The triad of artificial intelligence, bloom's taxonomy, and critical thinking in assessment pedagogy," *Education and Information Technologies*, vol. 29, no. 1, pp. 1–20, 2024.

[22] E. Kasneci, K. Sessler, H. Krosse, and G. Kasneci, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023.

[23] R. AlAli and Y. Wardat, "Opportunities and challenges of integrating generative artificial intelligence in education," *International Journal of Religion*, vol. 5, no. 7, pp. 784–793, 2024. [Online]. Available: https://doi.org/10.61707/8y29gv34

[24] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "A systematic review of research on artificial intelligence applications in higher education: Learning, teaching and administration," *Educational Technology Research and Development*, vol. 69, no. 2, pp. 733–762, 2021.

[25] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2201.06935*, 2022. [Online]. Available: https://arxiv.org/abs/2201.06935

[26] Y. Luo, T. Liu, P. C.-I. Pang, D. McKay, Z. Chen, G. Buchanan, and S. Chang, "Enhanced bloom's educational taxonomy for fostering information literacy in the era of large language models," *arXiv preprint arXiv:2503.19434*, 2025. [Online]. Available: https://arxiv.org/abs/2503.19434

[27] J. Biggs and C. Tang, *Teaching for Quality Learning at University*, 4th ed. McGraw-Hill Education, 2011.

[28] E. A. Van Dis, J. Bollen, and W. Zuidema, "What can chatgpt not do in education? evaluating its effectiveness in assessing educational learning outcomes," *Educational Technology Research and Development*, 2024.

[29] M. Sahlgren, D. Holmer, and F. Olsson, "Educational assessment with llms: Opportunities and risks," *Educational Technology & Society*, vol. 26, no. 2, pp. 22–35, 2023.

[30] X. Zhou, J. Liu, M. Sun *et al.*, "A survey of hallucination in large language models," *arXiv preprint arXiv:2302.08091*, 2023.

[31] L. Luo, H. Yang, and X. Yu, "Human–ai collaboration in education: Principles and practices," *BMC Medical Education*, vol. 23, no. 1, p. 341, 2023.

[32] O. Al-Diban, H. Alshareef, and M. Saleh, "Analyzing curriculum alignment gaps using bloom's taxonomy in saudi universities," *International Journal of Educational Research*, vol. 118, p. 102234, 2024.

[33] R. Mahmoud, A. Alshammari, and R. Almefleh, "Evaluating learning outcomes for accreditation: A case study in saudi higher education," *Education and Information Technologies*, vol. 29, no. 2, pp. 456–478, 2024.