BOARD #103: Work-in-progress: Evaluating Course Learning Objectives with Generative AI using SMART criteria

Mr. Ahmed Ashraf Butt, University of Oklahoma

Dr. Ahmed Ashraf Butt is an Assistant Professor at the University of Oklahoma. He recently completed his Ph.D. in the School of Engineering Education at Purdue University and pursued post-doctoral training at the School of Computer Science, Carnegie Mellon University (CMU). He has cultivated a multidisciplinary research portfolio bridging learning sciences, Human-Computer Interaction (HCI), and engineering education. His primary research focuses on designing and developing educational technologies that facilitate various aspects of student learning, such as engagement. Additionally, he is interested in designing instructional interventions and exploring their relationship with first-year engineering (FYE) students' learning aspects, including motivation and learning strategies. Prior to his time at Purdue, Dr. Butt worked as a lecturer at the University of Lahore, Pakistan, and has been associated with the software industry in various capacities.

Dr. Saira Anwar, Texas A&M University

Saira Anwar is an Assistant Professor at the Department of Multidisciplinary Engineering, Texas A and M University, College Station. She received her Ph.D. in Engineering Education from the School of Engineering Education, Purdue University, USA. The Department of Energy, National Science Foundation, and industry sponsors fund her research. Her research potential and the implication of her work are recognized through national and international awards, including the 2023 NSTA/NARST Research Worth Reading award for her publication in the Journal of Research in Science Teaching, 2023 New Faculty Fellow award by IEEE ASEE Frontiers in Education Conference, 2022 Apprentice Faculty Grant award by the ERM Division, ASEE, and 2020 outstanding researcher award by the School of Engineering Education, Purdue University. Dr. Anwar has over 20 years of teaching experience at various national and international universities, including the Texas A and M University - USA, University of Florida - USA, and Forman Christian College University - Pakistan. She also received outstanding teacher awards in 2013 and 2006. Also, she received the "President of Pakistan Merit and Talent Scholarship" for her undergraduate studies.

Dr. Asefeh Kardgar, Texas A&M University

Asefeh Kardgar is a researcher at Texas A&M University.

Work-in-progress: Evaluating Course Learning Objectives with Generative AI using SMART criteria

Introduction

In recent years, integrating artificial intelligence (AI) in education has gained significant attention, particularly with the emergence of Large Language Models (LLMs). The LLM models, in general, are trained on a large corpus of data to produce human-like responses in text, images, or other sources [1]. Their ability to generate human-like responses has made them an invaluable tool in education, particularly for automating and enhancing various educational tasks [2-4]. One key area that significantly impacts LLMs' ability to understand and process the text and generate responses is based on the prompt language used to instruct LLMs [5]. Due to their capability, researchers have used LLMs to perform various educational tasks, from simple conversation (e.g., [6]) to complex text analyses (e.g., [7]). For instance, in a study [8], the authors showed that the LLM model is an effective tool for creating engaging content for students. Similarly, another study [9] demonstrated its ability to evaluate students' essays like human graders.

As LLMs continue to transform various aspects of education, one critical area where their potential remains largely unexplored is their ability to help in curriculum design, particularly in evaluating learning objectives (LOs) [10]. Writing and implementing good LOs are essential for aligning the three aspects of curriculum design: course content, assessments, and instructional strategies [11]. The LOs provide a framework for students and instructors to ensure that learning outcomes are clearly defined and achievable. These basic building blocks can guide towards the sequence that builds upon each other and help describe how students will construct their understanding of the course material [11]. On the instructor's end, it allows students to be taught according to guiding principles. Also, it allows the examination of students' course progress with directed assessments and instructional strategies. On the student end, it lays out a clear plan of action to develop a complex understanding of the course's content.

Despite the importance, creating effective LOs can be challenging and time-consuming for educators, often requiring significant expertise in instructional design [12]. In this context, the SMART (specific, measurable, Achievable, relevant, and time-bound) criteria ensure that LOs are well-defined and assessable [13]. The SMART criteria suggest that meaningful learning objectives should be Specific (clear and focused), Measurable (trackable progress), Achievable (realistic to attain), Relevant (aligned with course goal), and Time-bound (within a set timeframe). These criteria ensure that LOs are clear, focused, and aligned with desired course learning outcomes, a key to an effective curriculum. However, instructors, particularly new faculty, often struggle to draft and evaluate the learning objective [12].

Bridging the gap in the literature, this work-in-progress study explores the capability of Generative Pre-trained Transformer (GPT) version 4, an LLM model, to assess LOs based on the SMART criteria. We hypothesize that a well-prompted LLM can efficiently evaluate LOs, saving instructors time. Specifically, this study was informed by the research question: How well does GPT evaluation align with human experts when evaluating course learning objectives using the SMART criteria?

Literature Review

Advancements in AI technology, particularly the development of Large Language Models (LLMs), offer promising new tools that can revolutionize higher education in various ways, such as providing constructive feedback, designing assessments, grading, tailored curricula, and personalized guidance. Due to its abilities, LLMs can help the instructor design an effective and aligned curriculum. One area where LLMs' potential can be very advantageous could be assisting in writing clear and specific learning objectives. Also, LLMs can assist in assessing the learning objectives by providing feedback to improve them. Prior literature highlights the importance of LOs in curriculum design for STEM (Science, Technology, Engineering, and Mathematics) education. More specifically, Los articulates expected learning outcomes, essential for guiding instructional strategies and assessment methods [14]. Effective learning objectives ensure students' educational experiences are directly aligned with the desired outcomes, particularly crucial in STEM fields where concepts and skills are complex and foundational for future learning [15].

Prior literature suggests that LLMs can be used to assess or entirely generate the learning objective. For instance, Sridhar et al. [10] explore using LLM to generate the learning objective and found that LLM was able to generate LOs based on their established criteria. However, studies also discussed the need for further refinement of this process. As the nature of the prompt bounds the LLMs' ability to perform, it is important to provide effective evaluation guidelines for effective judgment. To make the contextual and effective prompt, it is essential to provide clear evaluation criteria [16]. SMART is one such criterion that can be used to evaluate course learning objectives [17]. Prior literature dates back to the 1940s and 1950s when specific and measurable goals were discussed in engineering and educational publications [18]. These criteria ensure that objectives are well-defined, achievable within a specific timeframe, and aligned with the course goals [17]. Implementing SMART objectives facilitates assessment and supports pedagogical strategies to meet these ends, optimizing student engagement and learning outcomes [19 - 20].

With this study, we aim to describe that the systematic use of SMART criteria in crafting learning objectives and supporting advanced AI technologies can significantly improve the quality and effectiveness of curriculum design in STEM education. Educators can enhance instructional clarity and student achievement by integrating structured, objective frameworks

with innovative technological tools, preparing learners effectively for their academic and professional futures.

Methodology

This study employs a quantitative, correlation research design approach to evaluate the effectiveness of LLMs in assessing the LOs from STEM courses using the SMART framework.

Data Collection

We collected 30 LOs from a publicly available syllabus of courses. The selection criteria for the LOs were that they belong to a STEM course, and the syllabus should have a separate section for learning objectives. The collected learning objectives cover programming, engineering design, and database systems courses.

Evaluation Criteria

We used the SMART as the criteria to evaluate the quality of LOs, which both the LLM model and experts used. The SMART rubric is an evaluation criterion that assesses learning objectives based on five key criteria. **Specific**: The objective should clearly describe what students can do. It should have an action or outcome for the students to achieve. **Measurable**: The objective should include measurable indicators of success. This means identifying how students' progress or achievement will be tracked or assessed. **Achievable**: The objective should be realistic and achievable, given the students' existing knowledge level and the time available for the course. **Relevant**: The objective should be meaningful and connected to the student's overall goals, career aspirations, or the course's long-term objectives. **Time-bound**: The objective should have a clear timeline or deadline, specifying when students are expected to complete or achieve the objective.

An example of a LO for an introductory class that teaches to program using MATLAB would be *By the end of the semester, students will be able to write MATLAB functions that solve linear equations using matrix operations*. The LO is ***specific*** as it focuses on clear tasks, i.e., writing MATLAB functions to solve linear equations. It is ***measurable*** and can be assessed through coding assignments or exams on matrix operation. The objective is ***achievable***, given the foundational nature of the course, ensuring it is realistic for students to achieve. It is ***relevant*** and aligns directly with the course content and goals. Finally, the objective is ***time-bound***, with a clear deadline for completion set by the end of the semester.

LLM Evaluation

We used Generative Pre-trained Transformers (GPT) version 4 as an LLM model to evaluate the

learning objective, as it is one of the most used and robust models for Natural Language Processing (NLP) tasks [1]. GPT belongs to the family of neural networks, which is based on the transformer architecture trained on a large corpus of data and is considered a key AI advancement [21]. It can mimic human-like responses such as text, images, music, etc. To interact with GPT, we provide a prompt that describes the task, and then the model generates a response based on its pre-trained knowledge. In our case, prompts were designed to guide the LLM in evaluating each learning objective according to the SMART rubric. Also, the prompt included the course description, the learning objective, and specific instructions to assess each objective against the five SMART criteria. The prompts also clarified that responses should be binary (Yes/No) for each criterion.

Human Evaluation

We asked two experts with experience in curriculum design for an independent evaluation of the Learning objectives based on the SMART framework. For evaluation, we gave them 15 LOs each by randomly selecting from the collected LOs for them to evaluate independently. Each expert evaluated each LO with a yes or no for each SMART if they met the criteria, similar to the LLM evaluation. Also, we asked them to follow the same criteria definition as discussed in the Evaluation Criteria section to ensure consistency in both evaluations. Experts were asked not to discuss their evaluation with each other to maintain independence in evaluation.

Data Analysis

For the analysis, we ensured that the same set of LOs across the LLM and human evaluators allowed for a more accurate comparison of their assessments. Table 1 shows the distribution of evaluation from the LLM and experts regarding whether the LOs meet the criteria. One key observation is that the LLM, along with Expert 1 and Expert 2, consistently agreed that the LOs were relevant to the course content or goals across 30 LOs. However, the experts largely believed that the LOs did not meet the criteria definitions for all other criteria.

Table 1. Distribution of 'Yes' responses for each smart criterion

SMART Criterion	LLM (%)	Expert 1 (%)	Expert 2 (%)
Specific	80	20.00	60
Measurable	60	33.33	66.66
Achievable	96.6	53.33	73.33
Relevant	100	80.0	100
Time-bound	3.33	73.33	100

We calculated the percent agreement between the evaluations done by the LLM and the experts for each SMART criterion, as shown in Table 2. Percent agreement was calculated for each criterion. We compare whether the LLM and the human annotator agreed on the same decision for each LO, i.e., both marked the criterion as qualified or not. The percentage reflects the proportion of agreements over the total number of learning outcomes for each criterion. We interpret the agreement as suggested various authors, and it is considered acceptable if there is a

75% to 90% agreement [22 - 23]. While looking at Table 2, it is notable that both Expert 1 and Expert 2 showed strong agreement with the LLM on the "Relevant" criterion, which was aligned with the findings in Table 1.
Similarly, the LLM evaluation was aligned more closely with Expert 1 on "Specific" (80%) and "Achievable" (90%) and showed lower agreement with Expert 2, especially for "Specific" (20%) and "Measurable" (33.33%). We did not see a strong agreement for the other criteria and may need to revisit these criteria.

Table 2. Percent Agreement between the LLM and human evaluators on each SMART criterion

SMART Criterion	LLM vs Expert 1	LLM vs Expert 2
Specific	80	20.00
Measurable	60	33.33
Achievable	96.6	53.33
Relevant	100	80.0
Time-bound	3.33	73.33

We then conducted Cohen's Kappa analysis to understand better the agreement between the evaluation done by the LLM and the human evaluator. The results of the analysis are shown in Table 3. The result was interesting as even though we found strong agreement in the evaluation done by LLM and experts, as shown in Table 3, the agreement was not statistically meaningful or consistent, particularly when adjusted for chance, as shown in Table 3.

Table 3. Cohen's Kappa between the LLM and human evaluators on each SMART criterion

SMART Criterion	LLM vs Expert 1	LLM vs Expert 2
Specific	0.22	-0.12
Measurable	0.4	0.47
Achievable	0.15	0.0
Relevant	0.0	0.0
Time-bound	0.035	0.0

Discussion

This work-in-progress study is part of a larger project that aims to facilitate the instructor in designing an effective and student-centered curriculum. In this regard, this study investigates the ability of LLMs to evaluate the LOs with a minimum context (e.g., course description) based on the SMART criteria. For that, we collected publicly available LOs from different STEM courses. Then, we used GPT-4 (i.e., LLM model) to evaluate the LOs using a prompt and asked two experts, with Expert 1 evaluating the first 15 LOs and Expert 2 evaluating the remaining 15 LOs. To inform our study, we used percent agreement and Cohen's Kappa to assess the alignment between the LLM and experts' evaluation of the LOs.

The percent agreement result showed that LLM strongly agrees with expert evaluations when measuring the "Relevant" criterion. LLM has a good agreement when measuring “Achievable” with Expert 1, but it was much lower with Expert 2, highlighting a disparity between the evaluations. Similarly, for other criteria, such as "Specific" and “Measurable,” the agreement was notably lower with expert 2 as compared to expert 1. A mixed result was shown regarding the agreement between human experts and LLM evaluation of the LOs. Furthermore, we used Cohen's Kappa to understand the relationship between the evaluation done by LLM and experts. The rationale for using Cohen's Kappa for the analysis was that while percent agreement offers a simple measure of agreement, it does not account for the possibility of agreement occurring by chance [24].
Therefore, Cohen's Kappa can better help us understand the strength and significance of the observed agreement. The result of Cohen Kappa values is similar to the percent agreements, showing moderate agreement for "Measurable" criteria between the evaluation of LOs done by both LLM and experts. However, all the other criteria have zero kappa values, showing no agreement beyond chance or a random pattern of agreement/disagreement [25], and small kappa values showing fair or slight agreements in the evaluations.

The mixed result of using LLM in education is consistent with the literature, where several studies have shown that LLM facilitates content creation [3] and evaluation [26]. At the same time, some studies found that LLM could generate sensible yet wrong responses, especially when subjective criteria are involved [27]. Furthermore, as this work-in-progress study, these mixed results in agreements may be due to the inherently subjective nature of SMART criteria used for LO evaluations, requiring evaluators to understand the course content and goal holistically. Therefore, there is potential to explore LLM's potential in assessing the LOs further by refining the LLM's contextual understanding and developing well-defined criteria. Overall, the result suggests that while LLM could effectively assess a certain part of learning objectives, further refinement of GPT with more contextual information is needed.

Limitations and future directions

As this is a work-in-progress study, several limits must be considered. These limits also provide an opportunity for future research on this topic. The first limitation was the limited LOs dataset (n= 30), which limited the generalizability of the findings. Therefore, one future direction for this study is to collect a large and more diverse dataset to improve the finding of generalizability. The second limitation was that we used a general-purpose GPT-4 model without any fine-tuned human annotator. Fine-tuning GPT-4 with human-annotated LO evaluation based on the SMART criteria may improve the LLM's performance. The third limitation was that although we used the SMART criteria, its criteria needed to be refined and evaluated by educational experts. This process will help us to design better guidelines for evaluating learning objectives. Lastly, we only used 1 LLM model (i.e., GPT-4) to evaluate LOs. Therefore, exploring the efficacy of other LLM models and comparing their ability to assess LOs is necessary.

References:

[1] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individ. Differ.*, vol. 103, p. 102274, Apr. 2023, doi: 10.1016/j.lindif.2023.102274.

[2] A. Goslen, Y. J. Kim, J. Rowe, and J. Lester, "LLM-Based student plan generation for adaptive scaffolding in game-based learning environments," *Int. J. Artif. Intell. Educ.*, Jul. 2024, doi: 10.1007/s40593-024-00421-1.

[3] M. M. Rashid *et al.*, "Humanizing AI in Education: A Readability Comparison of LLM and Human-Created Educational Content," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 68, no. 1, pp. 596–603, Sep. 2024, doi: 10.1177/10711813241261689.

[4] S. Avogadri and D. Russo, "On opportunities and challenges of large language models and gpt for problem solving and TRIZ education," in *World Conference of AI-Powered Innovation and Inventive Design*, D. Cavallucci, S. Brad, and P. Livotov, Eds., Cham: Springer Nature Switzerland, 2025, pp. 193–204.

[5] S. K. K. Santu and D. Feng, "TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks." 2023. [Online]. Available: https://arxiv.org/abs/2305.11430

[6] Y. Zhuang, Y. Yu, K. Wang, H. Sun, and C. Zhang, "ToolQA: A Dataset for LLM Question Answering with External Tools," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Curran Associates, Inc., 2023, pp. 50117–50143. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/9cb2a7495900f8b602cb10159246a016-Paper-Datasets_and_Benchmarks.pdf

[7] M. Elaraby, Y. Zhong, D. Litman, A. A. Butt, and M. Menekse, "ReflectSumm: A Benchmark for Course Reflection Summarization," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 13819–13846.

[8] S. Al Faraby, A. Romadhony, and Adiwijaya, "Analysis of LLMs for educational question classification and generation," *Comput. Educ. Artif. Intell.*, vol. 7, p. 100298, Dec. 2024, doi: 10.1016/j.caeai.2024.100298.

[9] F. Yavuz, Ö. Çelik, and G. Yavaş Çelik, "Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments," *Br. J. Educ. Technol.*, vol. 56, no. 1, pp. 150–166, Jan. 2025, doi: 10.1111/bjet.13494.

[10] P. Sridhar, A. Doyle, A. Agarwal, C. Bogart, J. Savelka, and M. Sakr, "Harnessing LLMs in Curricular Design: Using GPT-4 to Support Authoring of Learning Objectives," presented at the Conference on Education Technology, Tokyo, Japan, Jul. 2023, p. N/A-N/A. [Online]. Available: https://ceur-ws.org/Vol-3487/paper9.pdf

[11] R. A. Streveler, K. A. Smith, and M. Pilotte, "Aligning Course Content, Assessment, and Delivery: Creating a Context for Outcome-Based Education," in *Outcome-Based Science, Technology, Engineering, and Mathematics Education: Innovative Practices*, K. Yusof, N. Azli, A. M. Kosnin, S. Yusof, and Y. Yusof, Eds., IGI Global Scientific Publishing, 2012, pp. 1–26. doi: 10.4018/978-1-4666-1809-1.ch001.

[12] R. M. Harden, "Learning outcomes and instructional objectives: is there a difference?," *Med. Teach.*, vol. 24, no. 2, pp. 151–155, 2002, doi: 10.1080/0142159022020687.

[13] M. Lamm, "Know Where You Are Going! Simple Steps to Writing SMART Learning Objectives." [Online]. Available: https://ctl.jhsph.edu/blog/posts/SMART-learning-objectives/#:~:text=Learning%20objectives%20should%20be%20a,time%2Dbound%20(SMART)

[14] J. B. Biggs and C. Tang, *Teaching For Quality Learning At University*. Maidenhead: McGraw-Hill Education, 2011.

[15] D. Kennedy, *Writing and using learning outcomes: a practical guide*. Cork: University College Cork, 2006. [Online]. Available: https://hdl.handle.net/10468/1613

[16] J. Kim *et al.*, “Which is better? Exploring Prompting Strategy For LLM-based Metrics.” 2023. [Online]. Available: https://arxiv.org/abs/2311.03754
[17] K. B. Lawlor, “Smart Goals: How the Application of Smart Goals can Contribute to Achievement of Student Learning Outcomes,” *Dev. Bus. Simul. Exp. Learn.*, vol. 39, 2012, [Online]. Available: https://api.semanticscholar.org/CorpusID:62121075
[18] M. Morrison, “History of SMART Objectives,” Rapid Business Improvement. [Online]. Available: http://rapidbi.com/management/history-of-smart-objectives/
[19] G. T. Doran, “There’s a S.M.A.R.T. way to write management’s goals and objectives,” *Manage. Rev.*, vol. 70, no. 11, pp. 35–36, 1981.
[20] G. Wiggins and J. McTighe, *Understanding by design*. ASCD, 2005.
[21] S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni, “Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers,” *IEEE Access*, vol. 12, pp. 69812–69837, 2024, doi: 10.1109/ACCESS.2024.3397775.
[22] D. P. Hartmann, “Considerations in the choice of interobserver reliability estimates,” *J. Appl. Behav. Anal.*, vol. 10, no. 1, pp. 103–116, 1977, doi: 10.1901/jaba.1977.10-103.
[23] S. E. Stemler, “A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability,” *Pract. Assess. Res. Eval.*, vol. 9, no. 4, 2004.
[24] M. L. McHugh, “Interrater reliability: the kappa statistic,” *Biochem Med Zagreb*, vol. 22, no. 3, pp. 276–282, 2012.
[25] J. Sim and C. C. Wright, “The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements,” *Phys. Ther.*, vol. 85, no. 3, pp. 257–268, Mar. 2005, doi: 10.1093/ptj/85.3.257.
[26] A. Magooda, D. Litman, A. A. Butt, and M. Menekse, “Improving the quality of students’ written reflections using natural language processing: Model design and classroom evaluation,” in *International Conference on Artificial Intelligence in Education*, Springer, 2022, pp. 519–525.
[27] L. Zhou, W. Schellaert, F. Martínez-Plumed, Y. Moros-Daval, C. Ferri, and J. Hernández-Orallo, “Larger and more instructable language models become less reliable,” *Nature*, vol. 634, no. 8032, pp. 61–68, Oct. 2024, doi: 10.1038/s41586-024-07930-y.