# AssessMate: Revolutionizing Assessment Design with AI

Yiliu Pan, Mingmei Zhang, Tzu-Yun Huang, Luojia Chen, & Kanghua Qiu

AI in Education | Assessment design | learning engineering | LLM

*In the era of Generative Artificial Intelligence (GenAI), traditional assessment methods, such as written tasks and multiple-choice questions, are increasingly susceptible to AI-generated responses, raising concerns about academic integrity. While AI detection tools have been proposed, they remain limited and often biased. Instead of focusing on detection, there is a growing shift toward redesigning assessments to integrate AI constructively. Learning Engineering (LE) provides a structured, iterative approach to designing, implementing, and evaluating technology-enhanced educational solutions. Our study applies an LE framework to the development of AssessMate, an AI-driven system that*

*automates assessment design while ensuring alignment with educational objectives. By embedding principles of human-centered design, systematic evaluation, and pedagogical alignment, AssessMate supports educators in creating more authentic and robust assessments. This paper discusses the co-design process, user evaluations, and future refinements to enhance the system's effectiveness within the Learning Engineering paradigm.*

# Introduction

The rapid rise of Generative Artificial Intelligence (GenAI) tools, like GPT-4, is transforming traditional assessment practices in education, raising concerns about academic integrity. Large Language Models (LLMs) can generate sophisticated responses to assessments, challenging the core principle of students independently demonstrating knowledge (Brown et al., 2020). Efforts to address this, such as AI detection tools, have proven flawed, often misidentifying work by non-native English speakers and failing to reliably detect GenAI misuse (Liang et al., 2023).

Given these limitations, educators are shifting from detection toward redesigning assessments to integrate GenAI constructively. GenAI can enhance learning through feedback, question generation, and personalized support, positioning LLMs as valuable educational tools, such as tutors or learning peers (Zawacki-Richter et al., 2019; Perkins & Roe, 2023). Universities are increasingly encouraging assessments that promote creativity, critical thinking, and real-world problem-solving, with many developing guidelines for incorporating AI tools into higher education (Perkins & Roe, 2023).

However, integrating GenAI into assessments raises challenges, particularly in aligning AI-generated assessments with educational goals and learning outcomes. Effective assessments must measure specific competencies, knowledge, and skills, aligning with cognitive frameworks like Bloom's taxonomy (Biggs, 2003). Ensuring that these assessments validate student progress requires significant instructor input.

Recognizing the challenges, this study explores a co-design methodology to develop assessment tools that align with the evolving educational landscape. We conducted user sessions with instructors to collaboratively design assessment tools, gaining insights into the essential features and elements required for effective assessment design. These co-design sessions revealed key components necessary for creating assessments that foster

creativity, critical thinking, and application of knowledge while integrating GenAI constructively. Based on these insights, we developed a system that automates the assessment design process, leveraging AI to align assessments with specific learning outcomes and educational goals. The system aims to support instructors in creating assessments that not only measure knowledge but also validate progression toward learning objectives, addressing challenges associated with traditional and AI-integrated assessments.

# The Platform: AssessMate

At the core of the AssessMate tool is its ability to automate the alignment of learning objectives with Bloom's taxonomy, enhancing assessment design through systematic evaluation. Bloom's taxonomy was chosen because it provides a well-established hierarchical framework for categorizing educational goals, ensuring that assessments measure a range of cognitive skills from basic recall to complex problem-solving (Anderson & Krathwohl, 2001). The taxonomy aligns with Learning Engineering (LE) principles, as it supports structured, evidence-based instructional design, ensuring that assessments promote deep learning and higher-order thinking (Goodell & Kolodner, 2022).

# System Functionality and Architecture

AssessMate operates by integrating the ABCD model (Audience, Behavior, Condition, Degree) to assess the clarity of instructor-defined learning objectives before mapping them to Bloom's taxonomy. This process enables the system to recommend assessment types that match cognitive demands, ensuring that generated tasks measure intended learning outcomes while fostering higher-order thinking (Biggs, 2003). The system employs Constructive Alignment to ensure that AI-generated assessments adhere to pedagogical best practices (Kessler et al., 2022).

Bloom's taxonomy was selected due to its widespread adoption in instructional design, offering a structured approach for aligning assessments with intended learning objectives (Thai et al., 2022). The taxonomy provides clear cognitive categories—Remember, Understand, Apply, Analyze, Evaluate, and Create—which help define assessment complexity and expected student learning outcomes. This structured approach aligns with Learning Engineering principles by ensuring that assessment tools support learning objectives at various cognitive levels and encourage evidence-based decision-making in assessment design (Goodell et al., 2023).

The system employs prompt engineering and utilizes the fine-tuned Assistant API to ensure the accurate classification of learning objectives and the generation of appropriate assessments. The process begins with instructors inputting learning objectives to ensure they are clear and complete. This evaluation step verifies that the learning objectives are well-defined and ready for alignment with Bloom's taxonomy.

Once the objectives pass the ABCD evaluation, the system uses prompt engineering techniques to interact with the Assistant API, aligning the learning objectives with the appropriate cognitive levels of Bloom's taxonomy, ranging from basic recall to advanced

skills like evaluation and creation. Four annotators independently annotated learning objectives from a university in Australia, achieving a Cohen's Kappa of 0.65 after two rounds of annotation, indicating a moderate level of agreement. Subsequently, 30 selected examples of learning objectives and their corresponding Bloom's taxonomy classifications were integrated into the knowledge base of the LLMs. To validate the alignment process, we conducted an experiment involving 117 learning objectives, comparing Bloom's taxonomy levels assigned by the Assistant API with those labeled by human experts. The experiment aimed to evaluate the accuracy of the Assistant API in classifying the cognitive levels of learning objectives according to Bloom's taxonomy, providing a rigorous assessment of the system's performance in aligning instructional goals with appropriate cognitive demands. The detailed comparison is presented in Table 1 below.

**Table 1**

*Comparison between human-labeled data and the LLM labeled data*

| Cognitive Level | Accuracy | F1 Score | Cohen's Kappa | AUC |
|---|---|---|---|---|
| remember | 0.98 | 0.67 | 0.66 | 0.99 |
| understand | 0.93 | 0.81 | 0.77 | 0.96 |
| apply | 0.97 | 0.97 | 0.93 | 0.96 |
| analyze | 0.94 | 0.82 | 0.79 | 0.84 |
| evaluate | 0.97 | 0.95 | 0.93 | 0.97 |
| create | 1.00 | 1.00 | 1.00 | 1.00 |

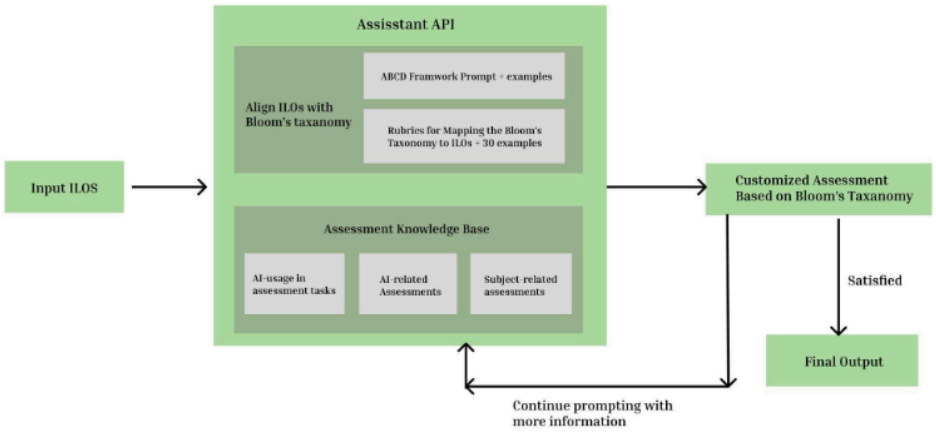
Following alignment, the system's assessment recommendation module uses the Assistant API to generate tailored assessment suggestions that correspond to the cognitive level of each objective. For example, objectives aligned with the "remembering" level might result in multiple-choice questions, while those aligned with "creating" may suggest project-based tasks. These recommendations are designed to align with the best pedagogical practices for each cognitive level.

Instructors review the suggested assessments and have the flexibility to make modifications, allowing them to tailor the generated tasks to their specific course needs. This flexibility ensures that instructors retain control over the assessment design process while benefiting from the automation provided by the tool.

A key feature of AssessMate is its ability to incorporate Generative AI into the assessment design process. By leveraging prompt engineering, instructors can choose to create AI-integrated assessments, such as those requiring students to engage with AI-generated content or opt for AI-resistant assessments that mitigate the risks of academic dishonesty. This capability allows the tool to adapt to the evolving technological landscape and maintain academic integrity.

**Figure 1**

*General User Flow of AssessMate Tool*




# Study Design

The research employed a co-design methodology to collaboratively develop an assessment design tool tailored to the needs of instructors in the context of Generative Artificial Intelligence (GenAI). This approach involved iterative, user-centered design sessions that actively engaged instructors from diverse academic disciplines in the development process. Participants were recruited from a university in Australia who were either professors or instructors who had previous experience in teaching or instructional design.

# Participants

A total of 10 instructors participated in the co-design sessions, representing various disciplines, including humanities, sciences, and medical school. Participants were selected to ensure a diverse range of perspectives on assessment design, particularly concerning the integration of AI tools in educational settings.

# Procedure

The co-design session went through three rounds of interaction. In the first phase, instructors participated in focus group discussions to identify key challenges in current assessment practices and to brainstorm potential features and elements that would be beneficial in an AI-integrated assessment tool. Participants were asked to outline the specific pain points they encountered when designing assessments, particularly concerning aligning assessments with intended learning outcomes. After the first phase, the insights were used to develop initial prototypes of the assessment design tool. These prototypes included customizable assessment templates, AI-driven alignment suggestions based on Bloom’s taxonomy, and features that allowed instructors to generate assessment tasks tailored to specific learning outcomes. In the final phase, participants were invited to interact

with the prototypes and provide feedback through hands-on sessions. Instructors tested the tool by creating sample assessments, after which they engaged in structured interviews and focus group discussions to evaluate the tool's usability, feature relevance, and overall alignment with their needs.

# Evaluation

The co-design sessions lasted for months and went through 3 rounds of interaction, the developed assessment tool was evaluated through qualitative methods, including semi-structured interviews and focus groups. The evaluation aimed to assess the tool's effectiveness, usability, and alignment with learning outcomes from the instructors' perspectives. The evaluation involved 10 instructors who had participated in the co-design sessions and 2 additional instructors who were new to the tool. This mix ensured both continuity in feedback and fresh insights from those unfamiliar with the tool's development process. Data were collected through individual interviews and focus groups, each lasting approximately 20-30 minutes. During these sessions, participants were asked to use the tool to design assessments for their courses and to reflect on their experiences. Key questions focused on the perceived alignment of generated assessments with learning outcomes, the ease of use of the tool, and areas for further enhancement.

To ensure methodological rigor, AssessMate was evaluated using a mixed-methods approach, incorporating both qualitative and quantitative measures:

1. Alignment Accuracy: We conducted a comparative analysis of 117 learning objectives, evaluating AI-generated classifications against expert-labeled Bloom's taxonomy levels. The analysis yielded 93% agreement, indicating strong reliability.
2. Usability & Instructor Satisfaction: We collected Likert-scale survey responses measuring ease of use, perceived effectiveness, and instructor confidence in AI-generated assessments.
3. Pedagogical Impact: Instructors provided qualitative reflections on how AssessMate influenced their assessment design practices, instructional flexibility, and ability to integrate AI ethically.

Data collection spanned three rounds of evaluation, ensuring that both experienced and new users provided feedback. Qualitative data were thematically analyzed, following an inductive coding approach to identify emerging themes related to tool effectiveness, AI alignment, and potential refinements.

# Results

The user testing sessions provided valuable insights into the effectiveness, usability, and areas for improvement of the AssessMate tool. The feedback was categorized into three main areas: positive attitudes toward the tool, the likelihood of future use, and suggestions for enhancements.

Participants generally expressed high satisfaction with AssessMate, particularly regarding the quality and relevance of the assessments generated. The tool was praised for its ability

to understand and align with course learning objectives, enhancing instructors' confidence in its utility for assessment design.

Quality of Generated Assessments: Participants consistently reported that the assessments produced by AssessMate were of high quality, aligning well with the course objectives and demonstrating a nuanced understanding of the instructional goals. This feedback reflects the tool's effectiveness in generating diverse, relevant assessment tasks that support learning outcomes. One participant noted, "The tool really understands the course objectives well, and the assessments generated are of high quality".

Integration of Bloom's Taxonomy: The integration of Bloom's taxonomy was identified as a significant strength of AssessMate. Instructors valued the tool's ability to ensure that assessment tasks were aligned with specific cognitive levels, supporting the creation of pedagogically sound assessments. A participant commented, "Bloom's taxonomy integration is helpful and ensures that the assessments are aligned with the intended cognitive levels".

Diverse Assessment Types: Several users were impressed with the variety of assessment types generated by the tool, noting that it effectively fostered creativity and critical thinking. This diversity was seen as a key factor in making assessments more engaging and reflective of real-world skills. One participant stated, "I was impressed with the diverse assessment generated," underscoring the tool's capacity to produce varied and stimulating assessment formats. Moreover, feedback indicated a strong likelihood of AssessMate being adopted more broadly within educational settings. Participants expressed high interest in using the tool for future course assessments and were inclined to recommend it to colleagues.

Potential for Wider Adoption: Participants viewed AssessMate as a valuable tool for enhancing the assessment design process and were optimistic about its broader applicability in educational contexts. The tool's efficiency in generating aligned assessments was particularly noted as a key driver of this positive reception. A participant remarked, "I would definitely use AssessMate for my future course assessments and would encourage my colleagues to try it as well". Also, participants found the tool easy to navigate and appreciated its ability to streamline the assessment design process, reducing the time and effort required to create high-quality, aligned assessments.

The feedback from user testing highlighted several areas for improvement in The evaluation of AssessMate highlighted critical areas for enhancement, particularly regarding guidance on AI usage, transparency of recommendations, and alignment of assessments with student capabilities. Instructors expressed uncertainty about the appropriate level of AI involvement for students, underscoring the need for clearer guidance and consistent institutional policies on AI integration in coursework. Participants also emphasized the importance of understanding the rationale behind the tool's recommendations, suggesting that transparent explanations would allow instructors to engage more deeply with AI-generated suggestions and make informed decisions. Feedback further indicated that the specificity of prompts significantly influences the quality of AI-generated assessments, with instructors favoring the tool's role as assistive support rather than a fully automated solution. Additionally, the assessment difficulty levels were often found to be too advanced for first-year students, highlighting the need for the tool to include adjustable difficulty settings that align

assessments with the appropriate cognitive levels of students, particularly in entry-level courses. Addressing these areas is essential for refining AssessMate's functionality and ensuring its effective integration into educational practices.

# Discussion

The results of this study highlight both the strengths and areas for improvement in AssessMate, emphasizing the role of user-centered design in developing AI-driven assessment tools. The positive feedback on assessment quality, alignment with learning objectives, and integration of Bloom's taxonomy validates the tool's ability to transform assessment practices within a Learning Engineering (LE) framework. However, several areas require refinement to enhance usability, build instructor trust, and ensure alignment with institutional policies and best practices in AI-assisted education.

AssessMate's development aligns with LE principles by incorporating a systematic design process, iterative prototyping, and evidence-based evaluation (Goodell & Kolodner, 2022). The co-design methodology reflects the human-centered focus of LE, ensuring that the tool meets real-world instructor needs while maintaining pedagogical control. Future work should further integrate LE best practices in adaptive learning, data-driven assessment design, and real-time feedback mechanisms (Kessler et al., 2022). Strengthening these connections will help position AssessMate as a leading tool in AI-assisted educational technology.

One key challenge identified in the study is the need for greater transparency in AI decision-making. Instructors expressed uncertainty regarding how AI-generated assessments were mapped to Bloom's taxonomy and learning objectives. Future iterations of AssessMate should incorporate explanatory AI features, providing clear, detailed justifications for assessment recommendations. This aligns with explainable AI (XAI) techniques, which help users understand and trust AI-generated outputs. Additionally, the development of institutional guidelines for AI-assisted assessment design will be critical to ensuring that instructors and students engage with AI tools ethically and effectively. Universities adopting AssessMate should establish best-practice policies that define appropriate AI use in assessment to maintain academic integrity.

Another challenge raised by participants is the misalignment of assessment difficulty with student capabilities, particularly for introductory-level learners. Future versions of AssessMate should incorporate adaptive learning principles to personalize assessments based on student proficiency. By leveraging student learning profiles and performance history, the system can generate assessments at appropriate difficulty levels, ensuring that students are neither overwhelmed nor under-challenged. Competency-based tracking can further refine assessment recommendations by progressively increasing complexity as students demonstrate mastery. Additionally, providing greater instructor control over difficulty settings will help educators fine-tune assessments to their specific course requirements.

To refine AssessMate's impact, the design and development process must remain iterative, incorporating ongoing instructor feedback and real-world classroom testing. Expanding user testing across multiple institutions and disciplines will help validate the tool's effectiveness,

scalability, and pedagogical adaptability. Future research will also explore automated assessment analytics, leveraging AI-driven insights to track student engagement, assessment reliability, and alignment with learning outcomes. Additionally, integrating AssessMate with learning management systems (LMSs) such as Canvas and Moodle will facilitate seamless adoption and usability, further streamlining the instructor workflow.

By embedding Learning Engineering principles into the continued development of AssessMate, the tool can evolve to better serve educators while ensuring that AI remains a supportive, transparent, and pedagogically sound element of assessment design. These refinements will position AssessMate as a key contributor to the growing field of evidence-based AI applications in education, advancing the integration of AI-driven assessment while maintaining best practices in learning design and educational technology.

# References

Altbach, P. G., Reisberg, L., & Rumbley, L. E. (2009). Trends in global higher education: Tracking an academic revolution. United Nations Educational, Scientific and Cultural Organization.

Biggs, J. (2003). Teaching for quality learning at university. The Society for Research into Higher Education & Open University Press.

Black, P. J., & Wiliam, D. (2018). Classroom assessment and pedagogy. ASSESSMENT IN EDUCATION, 25(3). Advance online publication. https://doi.org/10.1080/0969594X.2018.1441807

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). Assessment for learning: Putting it into practice. Open University Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Goodell, J., Kessler, A., & Schatz, S. (2023). Learning Engineering at a Glance. Journal of Military Learning.

Goodell, J., & Kolodner, J. (Eds.). (2022). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. Taylor & Francis.

Kessler, A., Craig, S. D., Goodell, J., Kurzweil, D., & Greenwald, S. (2022). Learning Engineering is A Process. In J Goodell & J. Kolodner (Eds.), Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. (pp. 29-46). Routledge.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7).

Lindsay, B. D. (2023). Lilly Library at Wabash College: AI in the Classroom: AI-Proof Assignments.

Perkins, M., & Roe, J. (2023). Decoding Academic Integrity Policies: A Corpus Linguistics Investigation of AI and Other Technological Threats. Higher Education Policy. https://doi.org/10.1057/s41307- 023-00323-2

Thai, K. P., Craig, S. D., Goodell, J., Lis., J., Schoenherr, J. R., & Kolodner J. (2022). Learning Engineering is Human-Centered. In J. Goodell & J. Kolodner (Eds.), Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. (pp. 83-124). Routledge.

Yan, D., Fauss, M., Hao, J., & Cui, W. (2023). Detection of AI-generated essays in writing assessments. Psychological Test and Assessment Modeling, 65(1), 125-144.

Zawacki-Richter, O., Marín, V.I., Bond, M. et al (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators?. Int J Educ Technol High Educ 16, 39. https://doi.org/10.1186/s41239-019-0171-0

**Yiliu Pan**

**Mingmei Zhang**

**Tzu-Yun Huang**

**Luojia Chen**

**Kanghua Qiu**