Generative AI: A High-Performing Assistant in Examination Design

Winn Wing-Yiu Chow
The University of Melbourne

Generative AI is rapidly advancing and holds great promise for transforming teaching and learning methodologies. This study evaluates the performance of large language models in assessing examination questions based on four key quality criteria: cognitive complexity, alignment with learning outcomes, content quality, and language appropriateness. Preliminary findings show that ChatGPT 3.5 performs comparably to state-of-the-art methods in classifying questions by cognitive complexity, as defined by Krathwohl's revised Bloom's taxonomy. Designing high-quality examinations requires careful consideration, including diverse question types to assess various levels of understanding and alignment with learning outcomes. Many university educators face challenges in creating effective examinations due to insufficient training. This paper outlines our work-in-progress in evaluating generative AI, specifically large language models, as an assistant in examination design. In addition to excelling in classifying questions by cognitive complexity, ChatGPT 3.5 demonstrated promising initial results across other criteria. In conclusion, generative AI has substantial potential to assist university educators in enhancing the overall quality of examination design.

Keywords: Generative AI, Large language model, ChatGPT, Examination design, Bloom's taxonomy

Introduction

The advancements in generative AI are remarkable and continue to accelerate. We are currently witnessing the emergence of powerful generative AI tools such as ChatGPT and Gemini. These tools not only excel in comprehending and processing extensive amounts of text data but also in generating text that closely resembles human-written content. For instance, ChatGPT 3.5 debuted in November 2022, followed swiftly by ChatGPT 4.0 in March 2023, showcasing the swift pace of innovation in this space. In a recent study by Meyer et al. (2024), comparing the performance of ChatGPT 4.0 to that of ChatGPT 3.5 on the written German medical licensing examination, a substantial improvement of 27% was noted with ChatGPT 4.0. Despite being in its infancy, generative AI shows great potential and is expected to contribute to meaningful changes in higher education, particularly in teaching and learning methodologies.

Assessment is crucial in teaching and learning, acting as a tool to gauge students' comprehension, progress, and mastery of learning materials. Additionally, it provides valuable feedback to both students and educators, informing instructional strategies and fostering improvement. Examinations are commonly employed as a key final assessment method to evaluate students' attainment of learning outcomes. However, it is worth noting that many university educators may lack extensive training in crafting high-quality examination questions (Arya et al., 2022). Constructing effective test items involves numerous considerations, as outlined by Clay (2001), including cognitive complexity, alignment with learning outcomes, content quality, and language appropriateness. For instance, cognitive complexity refers to the different levels of learning, from basic remembering and understanding to advanced evaluating and creating, based on Krathwohl's (2002) revised Bloom's taxonomy. Language appropriateness concerns whether the language is clear and suitable for the test item and for the students. As a result, crafting high-quality examinations necessitates the pedagogical expertise of university educators, and additional assistance is often needed to support them in this endeavor.

Given the remarkable text comprehension and analysis capabilities of generative AI, we aim to investigate the feasibility of leveraging this technology to create a system for assisting in examination design. This paper reports the work-in-progress of this exploration. Within this study, we are evaluating how generative AI, such as ChatGPT, can support university educators in crafting higher education examinations that adhere to the quality criteria outlined by Clay (2001). Serving as an assistant, generative AI shows potential in reviewing and analysing the drafted examination questions, offering suggestions and feedback to educators to improve the overall quality of examination design.

Background

Criteria for developing effective examination questions

Clay (2001) outlines several key criteria to consider when crafting effective examination questions. In short, examination questions should encompass (1) Cognitive Complexity - measuring an appropriate level of student knowledge, (2) Alignment with Learning Outcomes - assessing the achievement of learning outcomes, (3) Content Quality - measuring important aspects of the subject, and (4) Language Appropriateness - being written clearly and without ambiguity. Similar criteria are also suggested in other studies, such as Bramley et al. (2019). Consequently, this study adopted these four key quality criteria to evaluate how generative AI can support examination design.

Criterion 1: Cognitive complexity

Examination questions should be crafted to assess the appropriate levels of student learning, ranging from basic remembering and understanding to advanced evaluating and creating. A well-constructed examination should align closely with the subject’s intended learning outcomes, considering the specified level of cognitive complexity. For instance, in a Digital Business Analysis subject, where students are expected to apply modeling techniques to comprehend complex business contexts, assessments should emphasize their application rather than simple recall. Bloom’s taxonomy offers a widely recognized framework for classifying examination questions based on cognitive complexity. However, research suggests that university educators often struggle to classify questions accurately using this taxonomy (Karpen & Welch, 2016). Automated classification presents a potential solution to this challenge. Various machine learning techniques, such as support vector machine and convolutional neural network, have been explored for the classification of examination questions using Bloom’s taxonomy, achieving an F1 score of around 0.897 (Mohammed & Omar, 2020; Valentine & Oliveira, 2023). In these studies, word embedding emerges as a crucial technique for capturing the semantic meaning of words in examination questions. Given that generative AI commonly incorporates word embedding and possesses exceptional text comprehension and analysis capabilities, it holds promise for effective classification. However, the application of generative AI for classification is yet to be investigated.

Criterion 2: Alignment with learning outcomes

Learning outcomes define the specific, observable skills and knowledge that students are expected to demonstrate as a result of their learning. Examinations, being a key measure of learning, should assess the achievement of these learning outcomes. Thus, there should be a clear alignment between examination questions and learning outcomes. In previous studies, automated classification was utilized to classify learning outcomes based on Bloom’s taxonomy with an accuracy of approximately 87% (Shaikh et al., 2021). However, to our knowledge, no study has evaluated an automated mapping of examination questions to learning outcomes. Therefore, it is of significant research interest to evaluate the effectiveness of generative AI in performing this task to provide university educators feedback on the alignment and examination of observable skills and knowledge.

Criterion 3: Content quality

Examinations should adequately cover the syllabus and include all important topics within the subject matter. However, manually evaluating the breadth of topic coverage in an examination is time-consuming. Recent studies have shown promising results in automating this assessment using natural language processing techniques (Premathilaka et al., 2020). With its advanced language processing capabilities, generative AI is anticipated to effectively analyse examination questions, offering insights into topic coverage.

Criterion 4: Language appropriateness

Clear instructions are crucial because students must accurately understand what is being assessed to effectively demonstrate their understanding of the subject matter. Recent studies consistently highlight the language proficiency of generative AI. For instance, in a large-scale study comparing human-written essays to those generated by ChatGPT, experts rated ChatGPT essays as higher in quality (Herbold et al., 2023). Additionally, another study demonstrated ChatGPT's effectiveness as a tool for formal English language learning (Shaikh et al., 2023). Therefore, generative AI has promising potential to also function as a language-checking tool, enhancing clarity and correcting grammatical or spelling errors in examination questions.

Generative AI

This study examined a particular kind of generative AI called a large language model (LLM), with ChatGPT being one of the most well-known examples of such a model. The size of an LLM is usually measured by the number of parameters it holds. For instance, ChatGPT 3.5 has 175 billion parameters, while ChatGPT 4.0 escalates to an impressive 1 trillion parameters. Operating these massive models demands substantial computational resources, rendering them impractical for standard local server hosting. Moreover, since these commercial models are proprietary, submitting data for their processing raises significant data privacy concerns. Recently, smaller-scale LLMs like Flan-T5 (Chung et al., 2024) have emerged as viable alternatives. Some versions of Flan-T5 can be hosted on standard local servers and are licensed under the Apache License 2.0. Despite their reduced size, they have demonstrated notable performance compared to their larger counterparts.

Data, methodology, and initial results

Criterion 1: Cognitive complexity

Our study began with examining the effectiveness of LLMs in categorizing exam questions based on cognitive complexity, as defined by Krathwohl's (2002) revised Bloom's taxonomy. To achieve this, we evaluated the classification performance of multiple LLM models using a well-established dataset of examination questions by Yahya et al. (2012), expertly labeled with Bloom's taxonomy levels. The dataset was selected due to its common adoption as a benchmark in the literature (Mohammed & Omar, 2020; Valentine & Oliveira, 2023). This dataset comprised 600 questions, with 100 questions representing each cognitive level. We tested both state-of-the-art LLMs, such as ChatGPT 3.5, and smaller versions of LLMs, as outlined in Table 1.

Table 1
Large Language Models Tested

Large Language Model	Size	License	Testing Platform
ChatGPT 3.5	175 billion parameters	Commercial	Online, accessed by API
Flan-T5-base	250 million parameters	Apache License 2.0	Local server
Flan-T5-large	780 million parameters	Apache License 2.0	Local server
Flan-T5-xl	3 billion parameters	Apache License 2.0	Google Colab

Following the common testing practice, we divided the dataset into a training set (90%) and a test set (10%), employing stratified sampling to ensure a proportional representation of each cognitive level in both sets. The technique used for training the LLMs was fine-tuning (Chung et al., 2024). The average results of the accuracy and F1-score attained by the tested LLMs after 3 repetitions are detailed in Table 2.

Table 2
LLMs' Performance on Classifying Cognitive Complexity

	ChatGPT 3.5	Flan-T5-base	Flan-T5-large	Flan-T5-xl
Accuracy	88.3%	73.5%	82.2%	86.9%
F1-score	0.88	0.73	0.82	0.87

The results demonstrated that LLMs effectively classified examination questions into various cognitive levels, with ChatGPT 3.5's accuracy reaching as high as 88.3% and an F1-score of 0.88. This performance was noteworthy, although it fell slightly short of the best-reported F1 score of 0.897 by Mohammed & Omar (2020) on the same dataset. As expected, larger LLMs tended to perform better. Surprisingly, even the smaller Flan-T5-large model, around 175 times smaller than ChatGPT 3.5, achieved only 6.1% lower accuracy, while Flan-T5-xl demonstrated comparable performance. Given its capability for local hosting to ensure examination confidentiality, Flan-T5-xl emerged as a very promising choice. Unfortunately, ChatGPT 4.0 has not been released yet for fine-tuning. We may include testing of other advanced and popular LLMs in the future.

Furthermore, LLMs possess a distinct advantage over traditional machine learning classifiers. In addition to providing the classified cognitive complexity level, LLMs can offer justification for the classification. For instance, when asked about why the question "*Distinguish between micro and macro economics"* was at the analysis level, ChatGPT 3.5 offered the following explanation: "*In Bloom's taxonomy, the analysis level involves breaking down information into parts and understanding the relationship between those parts. When asked to distinguish between micro and macroeconomics, students are required to analyze the characteristics, principles, and scope of each field. This involves identifying key differences such as the level of aggregation (individual vs. aggregate), the focus (specific markets vs. the economy as a whole), and the tools used for analysis (microeconomic theories vs. macroeconomic theories). By comparing and contrasting these aspects, students demonstrate analytical thinking skills, which align with the analysis level of Bloom's taxonomy.*"

Criterion 2: Alignment with learning outcomes

Our upcoming investigation will focus on assessing how effectively LLMs can map examination questions to learning outcomes. We will assess LLMs using a designated dataset and adopt an unsupervised approach. Our goal is to evaluate LLMs' ability to perform mapping learning outcomes spanning various subject matters.

Criterion 3: Content quality and Criterion 4: Language appropriateness

We will also evaluate LLMs based on criteria 3 and 4 using a designated dataset in an unsupervised manner. Initial experimentation with ChatGPT 3.5 yielded promising results. For example, when ChatGPT 3.5 was given the task of mapping the question "*List TWO tools that you can use to represent requirements and define designs from the user's perspective*" into one of four key topics: (1) Business analysis planning, (2) Strategy analysis, (3) Requirement analysis and design, or (4) Solution evaluations, it correctly identified the topic as (3) Requirement analysis and design. Furthermore, in our preliminary testing, ChatGPT 3.5 demonstrated strong performance in correcting grammatical and spelling errors and enhancing the clarity of examination questions.

Conclusion

Examinations remain the predominant method of final assessment in higher education. However, university educators frequently encounter challenges in crafting effective examinations due to a lack of support and guidance in examination design. This paper presents our work-in-progress in investigating the potential use of generative AI, particularly large language models, to assist in the design of examinations. Specifically, we are examining how well large language models evaluate examination questions based on four key quality criteria. Initial findings indicate that these models excelled in assessing the cognitive complexity levels of examination questions using Bloom's taxonomy. Notably, ChatGPT 3.5 achieved performance comparable to the state-of-the-art and demonstrated a unique ability to provide justifications for its classifications as feedback to university educators. Additionally, ChatGPT 3.5 demonstrated promising initial results across the other criteria. These findings suggest that generative AI has significant potential to become a valuable tool in supporting university educators in examination design.

References

Arya, A., Chan, A. D. C., Hellemans, K., & Hornsby, D. J. (2022, June 2). *Faculty lack training programs in pedagogical skills*. University Affairs. https://universityaffairs.ca/opinion/in-my-opinion/faculty-lack-training-programs-in-pedagogical-skills/

Bramley, T., Crisp, V., & Shaw, S. (2019). The art of test construction: Can you make a good Physics exam by selecting questions from a bank? *Research Matters*, *27*. https://doi.org/10.17863/CAM.100385

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2024). Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, *25*(70), 1–53. https://jmlr.org/papers/v25/23-0870.html

Clay, B. (2001). *Is this a Trick Question?: A Short Guide to Writing Effective Test Questions*. Kansas Curriculum Center. https://www.k-state.edu/ksde/alp/resources/Handout-Module6.pdf

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, *13*(1), 18617. https://doi.org/10.1038/s41598-023-45644-9

Karpen, S. C., & Welch, A. C. (2016). Assessing the inter-rater reliability and accuracy of pharmacy faculty's Bloom's Taxonomy classifications. *Currents in Pharmacy Teaching and Learning*, *8*(6), 885–888. https://doi.org/10.1016/j.cptl.2016.08.003

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Meyer, A., Riese, J., & Streichert, T. (2024). Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study. *JMIR Medical Education*, *10*, e50965. https://doi.org/10.2196/50965

Mohammed, M., & Omar, N. (2020). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. *PLOS ONE*, *15*(3), e0230442. https://doi.org/10.1371/journal.pone.0230442

Premathilaka, Y. T., Banujan, K., & Kumara, B. T. G. S. (2020). Ontology-Based Approach to Determine the Coverage of Examination Papers. *2020 International Conference on Decision Aid Sciences and Application (DASA)*, 613–617. https://doi.org/10.1109/DASA51403.2020.9317235

Shaikh, S., Daudpotta, S. M., & Imran, A. S. (2021). Bloom's Learning Outcomes' Automatic Classification Using LSTM and Pretrained Word Embeddings. *IEEE Access*. https://doi.org/10.1109/access.2021.3106443

Shaikh, S., Yayilgan, S. Y., Klimova, B., & Pikhart, M. (2023). Assessing the Usability of ChatGPT for Formal English Language Learning. *European Journal of Investigation in Health, Psychology and Education*, *13*(9), 1937–1960. https://doi.org/10.3390/ejihpe13090140

Valentine, A., & Oliveira, E. (2023). Creating a software application to help university educators to reflect on the cognitive complexity of their exam questions, using Bloom's Taxonomy and automated classification. *ASCILITE 2023 Conference Proceedings: People, Partnerships and Pedagogies*. https://doi.org/10.14742/apubs.2023.613

Chow, W. (2024). Generative AI: A High-performing Assistant in Examination Design. In T. Cochrane, V. Narayan, E. Bone, C. Deneen, M. Saligari, K. Tregloan, & R. Vanderburg. (Eds.), *Navigating the Terrain: Emerging frontiers in learning spaces, pedagogies, and technologies*. Proceedings ASCILITE 2024. Melbourne (pp. 330-335). https://doi.org/ 10.14742/apubs.2024.730

Note: All published papers are refereed, having undergone a double-blind peer-review process.

ASCILITE 2024

Navigating the Terrain:

Emerging Frontiers in Learning Spaces, Pedagogies, and Technologies