Large Language Models for Learner Assistance in Massive Open Online Courses: Challenges, Tools, and Approaches

Jesus-Angel del-Hoyo-Gabaldon[1], Eva Garcia-Lopez[1], Antonio Garcia-Cabot[1,*], David de-Fitero-Dominguez[1], Mary-Ellen Wiltrout[2], Jessica Sandland[2] and Ana Bell[2]

[1] Universidad de Alcalá, Ctra. Madrid-Barcelona km 33.6, 28805 Alcalá de Henares, Spain

[2] Massachusetts Institute of Technology, Massachusetts Avenue 77, 02139 MA, Cambridge, USA

Abstract

Artificial Intelligence has undergone a significant revolution in recent years. The emergence and subsequent development of the Transformers architecture led to extensive research resulting in large language models (LLMs). These systems power widely used applications, such as ChatGPT, which is based on LLMs fine-tuned with human instructions to enhance their performance. Evidence shows that they surpass the results of previous models (BERT/GPT/T5 families) in terms of outcomes, even when they have less complex configurations. In addition, online courses and Massive Open Online Courses (MOOCs) experience a well-known issue: high dropout rates. Scholars aim to tackle this problem by introducing innovative systems and alternatives to enhance students' learning experiences and prevent course abandonment. One option is to complement these courses with LLMs, which can incorporate chatbots or comparable systems to facilitate learning and engagement. The present proposal focuses on developing an automated system wherein a chatbot automatically produces questions related to course content. Then, learners will receive feedback on their answers through semantic similarity mechanisms, indicating the specific content they need to review.

1. Introduction

One of the most compelling features of Massive Open Online Courses (MOOCs) is their ability to reach an extremely diverse set of learners. MOOC learners come from a wide range of demographic backgrounds, and they represent a wide variety of ages, educational backgrounds, and employment levels [1].

Furthermore, the diversity of MOOC learners is not limited to demographics. The motivations that learners have for participating in MOOCs can differ greatly. Perdue identifies a variety of motivations that drive MOOC participation, including curiosity and exploration, skill acquisition, and the desire to connect with others [2]. This range of motivations leads to

---

[*] Corresponding author.

✉ jesus.hoyo@edu.uah.es (J. A. del-Hoyo-Gabaldon); eva.garcial@uah.es (E. Garcia-Lopez); a.garciac@uah.es (A. Garcia-Cabot); david.fitero@uah.es (D. de-Fitero-Dominguez); mew27@mit.edu (M. E. Wiltrout); jgsandla@mit.edu (J. Sandland); anabell@mit.edu (A. Bell)

🆔 0000-0002-7598-3289 (E. Garcia-Lopez); 0000-0002-0298-3237 (A. Garcia-Cabot); 0000-0002-4647-4282 (M. E. Wiltrout)

learners interacting with MOOCs in ways which range from sampling specific course content of interest to engaging with all of the course material and earning a certificate of completion.

This diversity of participants provides a unique opportunity for instructors to engage learners who differ substantially in background and motivation from the students in their university classrooms; however, it also provides a unique challenge. Meeting the needs of a group of learners with vastly different backgrounds, preparation, and motivations is challenging within the context of a single course.

Many MOOCs, including those considered in this paper, aim to provide a university-level learning experience for course participants, including challenging formative and summative assessments. However, not all learners enter a course with the background and skills necessary to successfully complete these challenging assessments. Indeed, both Gütl et al. and Onah, Sinclair and Boyatt have identified course difficulty and lack of learner preparation as one factor motivating learner dropout from MOOCs [3], [4].

This study proposes leveraging large language models (LLMs) to help bridge this gap between learners' current levels of knowledge and the knowledge required to successfully engage with the course assignments. Additionally, we aim to employ LLMs to create additional content for learners who do not aim to complete all the assessments in the course, but rather choose to engage with the course content in a less formal way.

The emergence of the Transformer architecture [5] gave rise to transformer-based LLMs, which had a profound impact on the field of Natural Language Processing (NLP). Currently, the most promising approach is to align them with human instructions [6]. Consequently, research has been conducted to investigate the impact of instruction-based LLMs in learning. A recent example demonstrates that a chatbot powered by GPT-4 enhanced the performance of adopters on an exam in an online coding class [7]. However, the engagement of learners was reduced, suggesting that a different approach to integration may be required. Other examples illustrate other uses in MOOCs, such as a system to validate peer-assigned essay scores [8] or a GPT-4-based system for automatically grading writing assessments [9]. These systems would be significantly enhanced by the incorporation of effective feedback for learners. Recent evidence has demonstrated the feasibility of providing constructive feedback to learners through ChatGPT [10]. Consequently, this study proposes a novel approach for learner assessment and assistance in MOOCs.

The remaining sections are structured as follows: Section 2 presents an initial pipeline design for the automatic generation of test questions, and that is focused on their subsequent evaluation to provide valuable feedback for learners. Section 3 contains a conclusion of the proposal.

2. Pipeline design

The primary objective of the proposed system is to minimize the impact of the aforementioned issues, with particular emphasis on engagement. As previously stated, LLM-powered solutions have the potential to enhance learning outcomes in MOOCs. However, it is crucial to integrate them carefully to avoid any adverse effects on engagement. To achieve this, learners will have access to tailored AI-generated questions and feedback that will assist them in determining the specific contents they should review. In general, this process would be laborious when undertaken manually. However, it is anticipated that AI will reduce human effort. The

procedure for obtaining a robust system is illustrated in **Figure 1**, which outlines the following steps:

1. The initial stage of the process involves the implementation of a heuristic, rule-based system. This first iteration is constructed upon a pipeline that is capable of automatically generating multiple-choice questions (MQC-AI), and which is explained in further detail in Subsection 2.1. Based on the content of a course, MCQ-AI will generate a series of questions, which will then be presented to the learner for response, and by comparing their answers with the ground truth, basic AI-generated recommendations will be provided for each failed question. As a global evaluation, learners will be presented with an aggregation of individual recommendations, in addition to a recommendation of the course contents that they should review.

2. The second iteration will attempt to enhance the aforementioned system for automated feedback generation by capitalizing on the human knowledge provided by real instructors through playtesting. In a controlled environment, instructors will be presented with real examples of learner responses, which they will then provide with tailored feedback. Subsequently, the human-generated feedback will be used to fine-tune an LLM to provide a global evaluation from an aggregation of individual feedback of failed questions. Furthermore, learners will be provided with recommendations regarding the course contents that they are required to review.

3. The third step is focused on text input problems (TIP). As in the previous iterations, the LLM responsible for providing feedback will be enhanced with human knowledge through playtesting. However, it will now be used to also provide feedback for text input problems. In the previous iterations, multiple-choice questions were generated with the assistance of an LLM (MCQ-AI), but this step involves a different LLM, specifically fine-tuned to generate text input problems (TIP-AI). This model will generate question-answer pairs and will be explained in greater detail in Subsection 2.2. Then, embedding representation features can be used to evaluate answers through comparison using semantic similarity. Finally, the AI model trained for recommendation generation can be used to provide a global evaluation to learners.

4. The fourth stage of this process involves the generation of mathematical problems using an LLM. LLMs demonstrate robust arithmetic capabilities and logical reasoning, yet they have limited abilities in mathematical and abstract reasoning, and fails on graduate-level problems [11], [12]. While these issues present a significant challenge, some examples have shown neural network capabilities in solving, explaining, and generating university-level mathematical problems [13]. A similar approach can be employed to fine-tune on code an LLM that has been pretrained on text. This can be enhanced with general scientific knowledge [14] and well-known structures on how to solve mathematical problems [15], as well as with prompt engineering techniques, such as Chain-of-Thought (CoT), which has demonstrated a favorable performance for mathematical problem solving and complex reasoning [16]. The resulting LLM is expected to generate university-level mathematical problems of different fields (e.g., mathematics, physics, chemistry, and programming) with a well-explained solution that can be used to provide recommendations to learners.
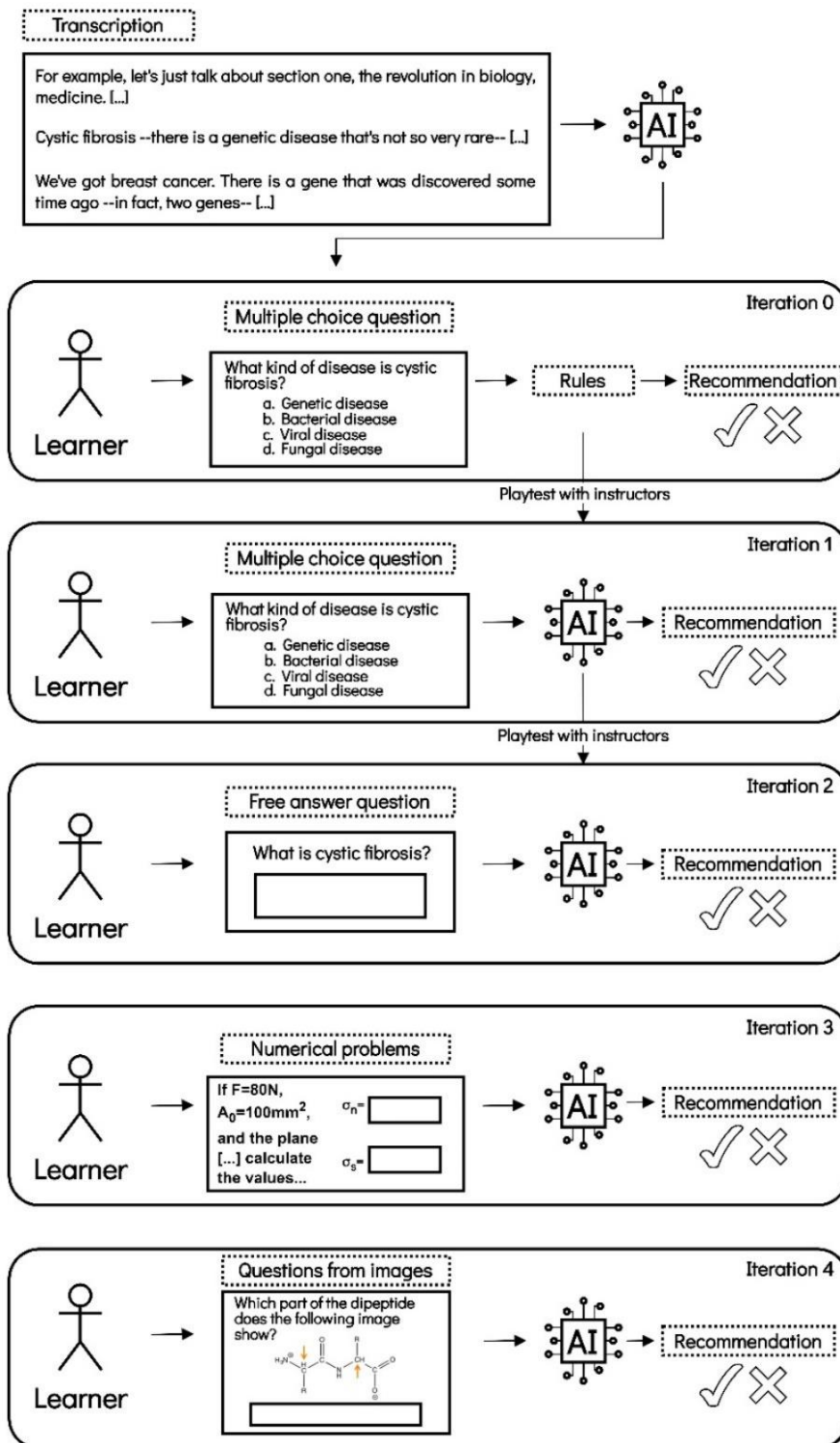
Figure 1: Procedure for creating a system for Question Generation and subsequent automatic correction. Learners will be presented with AI-generated questions. Once they have answered, a customized feedback message will be generated. The models responsible for evaluation and feedback generation will be fine-tuned using real data obtained from playtesting.

5. The final iteration is focused on the problem of Visual Question Generation (VQG), which is described as the task of asking a natural and engaging question when shown an image using AI models [17]. Following a similar approach, image-to-text models will be used to generate questions for learning environments. One example is the novel LLaVA-1.5, a Large Multimodal Model (LMM). LLaVA-1.5 has demonstrated efficacy in the field of Visual Question Answering [18], and it is anticipated that it will perform similarly well in the Visual Question Generation task. Then, as in previous iterations, the Feedback Generation system will be used to compare learners' answers with the descriptions of the images.

The following subsections provide a detailed description of the Multiple-Choice Question generation system (MCQ-AI), the Text Input Problem generation system (TIP-AI), and the Feedback Generation system (FG-AI).

2.1. Multiple-Choice Question generation system (MCQ-AI)

MCQ-AI comprises a series of Deep Learning models, arranged in sequence, which are used to generate multiple-choice questions. At the beginning, T5 models were employed to solve the problem in English [19], and then they were subsequently replaced by mT5 models to implement the pipeline in Spanish [20].

These LLMs, designated as instruction-based models, demonstrate superior performance when compared to previous models (e.g., BERT/GPT/T5 families) in terms of outcomes, even when they have less complex configurations [6]. The ability of these models to replace the ones used in the pipeline will be studied accordingly. The initial pipeline operates as follows:

1. The course content will be divided into different paragraphs. Initially, they were obtained using *sent_tokenize*, a library for dividing text into sentences. However, this approach is not optimal. The issue will be addressed by means of semantic chunking, which entails dividing a text into sentences, comparing the semantic similarity (e.g., cosine similarity) of each sentence with the others, and then grouping sentences with the most similar embeddings together. As a result, meaningful chunks are obtained.
2. From each paragraph, answers are extracted using a T5 model (T5-AE). Subsequently, a second T5 model (T5-QG) is employed to generate the corresponding questions based on the same paragraph and the related answers. Finally, a third T5 model (T5-DG) is utilized to generate the distractors, i.e., the incorrect answers in multiple-choice questions.

Although the pipeline is effective, instruction-based LLMs have demonstrated superior performance, suggesting that their use will result in a simplification and improvement of the original pipeline. To fine-tune models such as GPT-4 [21] or LLaMA 2 [22], different datasets will be employed, including SQuAD [23], which is focused on enhancing learners' reading comprehension, and HotpotQA [24], to improve reasoning across course contents.

In this preliminary study, we were granted access to LLaMA 3 [25]. The efficacy of this LLM can be assessed by examining **Figure 2**, where a prototype of the MCQ system is presented.

Context

Text to analyze

PROFESSOR: So I want you to find the gene now. It's not the gene that rescues arginine biosynthesis in yeast. It's not the gene that encodes beta globin where you know that protein very well, and you can purify that protein and make antibodies. Again, it's that it's Huntington's disease, let's say. Huntington's disease. That's a dominantly inherited disorder. It leads to brain degeneration typically in the fifth decade of life. Dominantly inherited-- dad might be affected here. You might have a bunch of children here. And on average, what fraction of the children will be affected? Half of the children will be affected. And let's cause some of them to be affected here. There's no sex preference there. There we go. About half of the children in that family are affected. Or maybe It's cystic fibrosis we're talking about. Cystic fibrosis is a recessive disease. It causes cysts in the pancreas. That's how it got its name. But it also causes the tremendous buildup of mucus in the lungs which leads to breathing problems, which is one of the serious hallmarks of cystic fibrosis. And because it's a recessive disease, the parents are

Or

Choose a SRT file

> ☁ Drag and drop file here
> Limit 200MB per file
> **Browse files**

`Generate questions!`

2386 tokens.

Questions

What type of inheritance pattern is characteristic of Huntington's disease?

○ Autosomal recessive
○ X-linked
○ Dominant
○ Recessive

What percentage of children of an affected parent are likely to inherit the disease in the case of Huntington's disease?

○ 75%
○ 25%
○ 50%
○ 100%

What is the primary cause of cystic fibrosis?

○ Respiratory infections
○ Genetic mutation
○ Pancreatic insufficiency
○ Buildup of mucus in the lungs

How often do children of unaffected parents inherit cystic fibrosis?

○ 25%
○ 10%

Figure 2: A preliminary model of the MCQ system, implemented with LLaMA 3.

2.2. Text Input Problem generation system (TIP-AI)

TIP-AI is similar to the MCQ-AI system previously described. In this case, the system is focused on Text Input Problem generation. The TIP-AI system structure is as follows:

1. The initial step is analogous to MCQ-AI Step 1.
2. An LLM will generate a series of questions related to each semantic chunk. Subsequently, the real answer (semantic chunk) will be stored in a vector representation (embeddings), along with each question generated for the chunk. This results in the creation of different question-answer pairs. Embeddings allow comparisons between the real answer and learners' answers, which is required to determine whether an answer is correct or, alternatively, whether the learner needs further feedback.

To achieve this, instruction-based LLMs fine-tuned with datasets such as TQA [26] will be employed. TQA contains a variety of materials, suitable for the text input problem generation task, as well as for others, including mathematical and visual question generation tasks.

2.3. Feedback Generation system (FG-AI)

FG-AI is designed to assist users in their learning process and to guide them in reviewing course content where they may have performed less effectively. Upon receipt of an answer from a learner to a question of the types mentioned above, the following process will be initiated:

1. An automated correction system will evaluate the answer and determine its correctness. This system may be rule-based (e.g., MCQ), or it may employ a more complex evaluation method, such as semantic similarity (e.g., TIP). If the answer is not entirely accurate, an LLM will generate tailored feedback based on the failed question. The model will take into consideration the question, context, correct answer, and

distractors, when applicable, as well as the learner's answer. It will then evaluate why the answer is not entirely correct and provide some concept clarification for the learner. An overview of the various potential evaluation and feedback generation frameworks are given below.

a. The rule-based system is relatively straightforward. Thanks to prompt engineering techniques, a prompt can be developed to generate MCQs with a structured format. This allows the question, possible answers and the correct answer to be processed programmatically. The system then displays the question, allowing the learner to interact with it. This is demonstrated in the example of **Figure 2**. Then, the result of the interaction can be compared to the correct answer. If the response is correct, no further feedback is needed. Otherwise, an LLM will receive the context, the question, the correct answer, and the learner's response. Based on this information, the LLM will generate tailored feedback for the incorrect response. **Figure 3** illustrates a preliminary test of this concept.

b. Evaluation and feedback generation for TIP cannot be addressed in the same manner as that of MCQs. In this case, it is not as straightforward to provide an accurate response as in the previous instance, since the same idea can be expressed in a variety of ways with natural language. To address the evaluation and feedback generation problem, three possible solutions can be considered:

   i. **Semantic similarity mechanisms:** as previously stated, semantic similarity mechanisms, such as cosine similarity, can be employed to compare a given answer to the ground truth. Based on the model's confidence, an LLM will generate recommendations for the learner.

   ii. **Template-based recommendations:** Swope et al. propose the creation of templates based on a scale to evaluate learners' answers [27]. Each step involves the examination of different features within the learner's answer to ascertain whether they align with the characteristics in the question. An illustrative example is provided below:

   1: The response provided by the learner is not related to the topic at hand.
   2: The response provided by the learner is partially related to the topic, but lacks some key aspects (key aspect 1, key aspect 2, …, key aspect n).
   3: The response provided by the learner is directly related to the topic and provides a comprehensive overview of the key aspects.

   iii. **FACTSCORE:** FACTSCORE is a novel evaluation method which breaks texts into a series of atomic facts to facilitate comparison [28]. By leveraging the neural mechanisms elucidated in the work, atomic facts can be generated to evaluate learners' answers and provide valuable feedback from those atomic facts.

c. With regard to the mathematical problems and questions generated from images, the current state of the analysis is still in an initial stage. Consequently, further investigation is required to develop a framework that will prove to be of value.
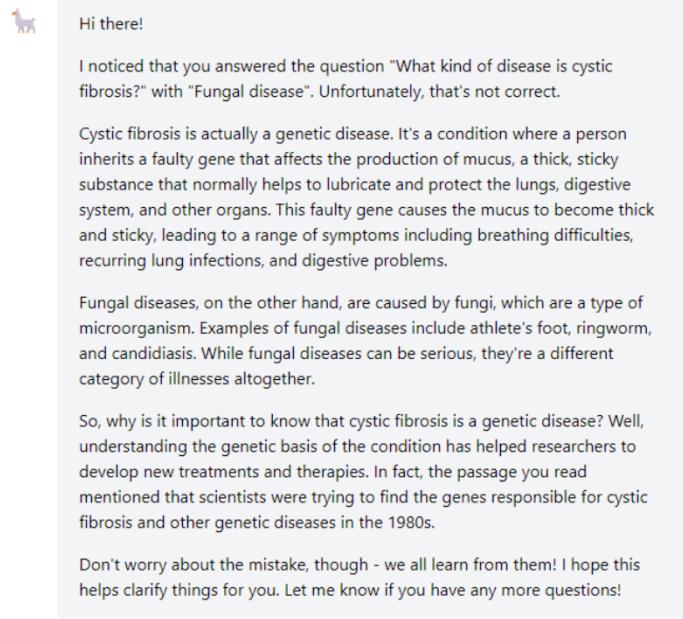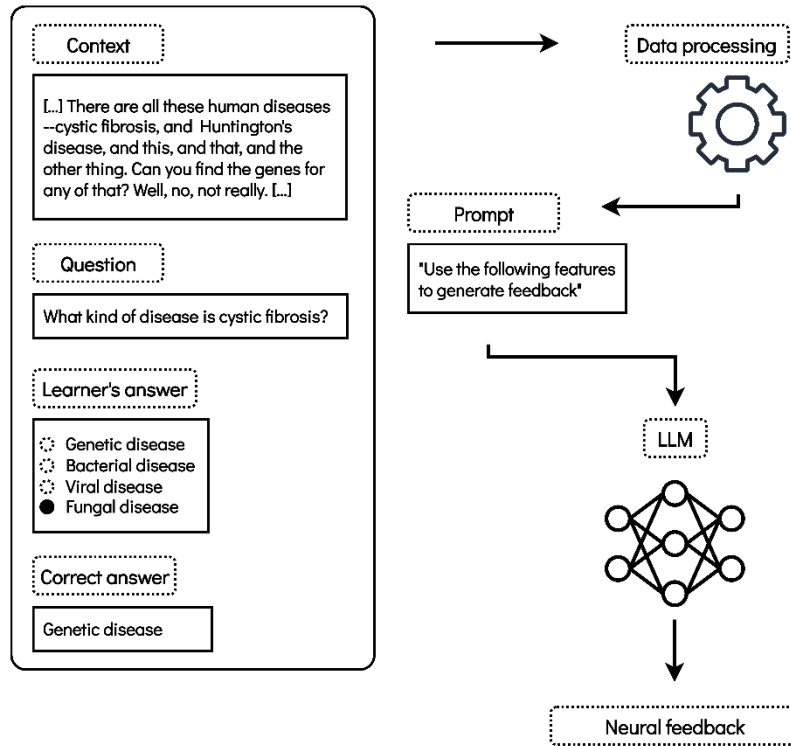
Figure 3: The example illustrates the process of generating feedback based on MCQs. In this case, the data from a real test will be recollected and transformed into a prompt for the generation of valuable feedback. In other cases, the process is analogous but involves alterations in the required data and the manner of its processing, which ultimately yields a prompt suitable for providing feedback. For instance, in the case of TIPs, the data processing step could involve transforming the FACTSCORE into a template that the LLM can use to validate the result.

2. Upon completion of a learning unit, every feedback generated during that specific unit for the learner will be gathered and processed by an LLM to identify critical misconceptions. The LLM will then provide a global evaluation and suggest to the learner the content that is required to review. The goal of this process is to increase engagement and improve the learning experience of learners, prompting them to interact with MOOCs.

To operationalize this concept, instruction-based LLMs will be employed, which will be fine-tuned with existing data from several courses. In addition, other sources will be used, such as data obtained through playtesting with real instructors.

3. Conclusions

The enhancement of learners' learning processes in MOOCs is a topic that has been the subject of intense study. Advances in Artificial Intelligence offer the potential for significant improvements, as evidenced by previous examples. However, there is still much room for improvement. The system proposed is expected to enhance existing results and provide a valuable AI-powered tool for various courses, improving engagement and learning experience. Nevertheless, due to the central role of Artificial Intelligence in this solution, the deployment of a system similar to the one proposed in this paper requires careful consideration. It is well-documented that artificial intelligence is prone to hallucination, despite the efforts of researchers to avoid this behavior. This implies that false positives and non-desired outputs may be generated. In addition, to enhance accessibility and usability, learners must have the option of skipping questions generated by this system and also to report incorrect generated recommendations. This will help to ensure that they remain engaged with the learning material, even when they encounter erroneous AI-generated content.

Acknowledgements

References

[1] C. R. Glass, M. S. Shiokawa-Baklan, and A. J. Saltarelli, "Who Takes MOOCs?," *New Dir. Institutional Res.*, vol. 2015, no. 167, pp. 41–55, 2016, doi: 10.1002/ir.20153.

[2] M. Perdue, "Incorporating Learner Perspectives into Course Design," in *2023 IEEE Learning with MOOCS (LWMOOCS)*, Oct. 2023, pp. 1–7. doi: 10.1109/LWMOOCS58322.2023.10306167.

[3] C. Gütl, R. H. Rizzardini, V. Chang, and M. Morales, "Attrition in MOOC: Lessons Learned from Drop-Out Students," in *Learning Technology for Education in Cloud. MOOC and Big Data*, L. Uden, J. Sinclair, Y.-H. Tao, and D. Liberona, Eds., Cham: Springer International Publishing, 2014, pp. 37–48. doi: 10.1007/978-3-319-10671-7_4.

[4] D. F. O. Onah, J. Sinclair, and R. Boyatt, "Dropout Rates of Massive Open Online Courses: Behavioural Patterns," in *EDULEARN14 Proceedings*, in 6th International Conference on Education and New Learning Technologies. Barcelona, Spain: IATED, Jul. 2014, pp. 5825–5834.

[5] A. Vaswani *et al.*, "Attention Is All You Need." arXiv, Dec. 05, 2017. doi: 10.48550/arXiv.1706.03762.

[6] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback." arXiv, Mar. 04, 2022. doi: 10.48550/arXiv.2203.02155.

[7] A. Nie *et al.*, "The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement but Increased Adopters Exam Performances." OSF, Apr. 25, 2024. doi: 10.31219/osf.io/qy8zd.

[8] W. Morris, S. Crossley, L. Holmes, and A. Trumbore, "Using Transformer Language Models to Validate Peer-Assigned Essay Scores in Massive Open Online Courses (MOOCs)," in *LAK23: 13th International Learning Analytics and Knowledge Conference*, in LAK2023. New York, NY, USA: Association for Computing Machinery, Mar. 2023, pp. 315–323. doi: 10.1145/3576050.3576098.

[9] S. Golchin, N. Garuda, C. Impey, and M. Wenger, "Large Language Models As MOOCs Graders." arXiv, Feb. 29, 2024. doi: 10.48550/arXiv.2402.03776.

[10] W. Dai *et al.*, "Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT," in *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, Jul. 2023, pp. 323–325. doi: 10.1109/ICALT58122.2023.00100.

[11] Y. Chang *et al.*, "A Survey on Evaluation of Large Language Models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, p. 39:1-39:45, Mar. 2024, doi: 10.1145/3641289.

[12] S. Frieder *et al.*, "Mathematical Capabilities of ChatGPT," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 27699–27744, Dec. 2023.

[13] I. Drori *et al.*, "A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level," *Proc. Natl. Acad. Sci.*, vol. 119, no. 32, p. e2123433119, Aug. 2022, doi: 10.1073/pnas.2123433119.

[14] R. Taylor *et al.*, "Galactica: A Large Language Model for Science." arXiv, Nov. 16, 2022. doi: 10.48550/arXiv.2211.09085.

[15] D. Hendrycks *et al.*, "Measuring Mathematical Problem Solving With the MATH Dataset." arXiv, Nov. 08, 2021. doi: 10.48550/arXiv.2103.03874.

[16] J. Wei *et al.*, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." arXiv, Jan. 10, 2023. doi: 10.48550/arXiv.2201.11903.

[17] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende, "Generating Natural Questions About an Image," arXiv.org. Accessed: May 09, 2024. [Online]. Available: https://arxiv.org/abs/1603.06059v3

[18] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved Baselines with Visual Instruction Tuning." arXiv, May 15, 2024. doi: 10.48550/arXiv.2310.03744.

[19] R. Rodriguez-Torrealba, E. Garcia-Lopez, and A. Garcia-Cabot, "End-to-End generation of Multiple-Choice questions using Text-to-Text transfer Transformer models," *Expert Syst. Appl.*, vol. 208, p. 118258, Dec. 2022, doi: 10.1016/j.eswa.2022.118258.

[20] D. De-Fitero-Dominguez, E. Garcia-Lopez, A. Garcia-Cabot, J.-A. Del-Hoyo-Gabaldon, and A. Moreno-Cediel, "Distractor Generation Through Text-to-Text Transformer Models," *IEEE Access*, vol. 12, pp. 25580–25589, 2024, doi: 10.1109/ACCESS.2024.3361673.

[21] OpenAI, "GPT-4 Technical Report." arXiv, Mar. 27, 2023. doi: 10.48550/arXiv.2303.08774.

[22] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models." arXiv, Jul. 19, 2023. doi: 10.48550/arXiv.2307.09288.

[23] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," 2016, doi: 10.48550/ARXIV.1606.05250.

[24] Z. Yang *et al.*, "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering." arXiv, Sep. 25, 2018. doi: 10.48550/arXiv.1809.09600.

[25] "Meta Llama 3," Meta Llama. Accessed: May 15, 2024. [Online]. Available: https://llama.meta.com/llama3/

[26] A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi, "Are You Smarter Than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension," *2017 IEEE Conf. Comput. Vis. Pattern Recognit. CVPR*, pp. 5376–5384, Jul. 2017, doi: 10.1109/CVPR.2017.571.

[27] J. Swope, "2024-05-20 Educators WG: AI Powered Assessment - Open edX Community - Open edX Community Wiki." Accessed: Jun. 10, 2024. [Online]. Available: https://openedx.atlassian.net/wiki/spaces/COMM/pages/4246667265/2024-05-20+Educators+WG+AI+Powered+Assessment

[28] S. Min *et al.*, "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation." arXiv, Oct. 11, 2023. doi: 10.48550/arXiv.2305.14251.