Challenges and opportunities of using an AI chatbot for learning assessment

Laura V. Florez

Department of Plant and Environmental Sciences
University of Copenhagen

Introduction

The concept of constructive alignment has largely influenced pedagogy in the last decades, placing strong emphasis on how the strategy for learning assessment should be coherent and hold a clear connecting thread to what the student is meant to learn (Biggs & Tang, 2011). Oral exams have the potential to cover a broad range of learning levels through open questions. Provided a skilled examiner, there is opportunity to flexibly and adaptively explore the students' capabilities from factual knowledge and understanding all the way to deeper analysis and application abilities. Oral exams are also a useful experience to cultivate discussion skills and a valuable interaction between teacher/examiner and student. As for all exam formats, there are also trade-offs, including larger time investment, biases of verbal communication and student stress as a potential barrier for reliable assessment. On another end, highly efficient assessment methods like multiple choice questionnaires do not impose these challenges and are frequently used for large courses. While carefully designed questionnaires can be powerful for assessment, these are comparatively limited in assessing deep learning.

As hopefully many other educators, I am of the opinion that "the future of education appears not as a battleground between technology and tradition, but as a fertile ground for synergy" (Owoseni et al., 2024). In the last few years, the rapid development and expansion in the use of AI and large language models (LLMs) has sparked an interest in making sensible use of this technology to potentiate educational strategies (e.g. Chen et al. 2023). Thus, I was curious whether and how it could be useful for learning assessment. Specifically, I explored whether and how it would be possible to use an AI chatbot to implement a different assessment strategy that

possibly mitigates some of the issues in the exam formats mentioned above. Evidently, such an approach comes with new challenges, and a deep understanding of the strengths and weaknesses of certain types of assessment is central to make the best possible use of an exam (Leth Andersen et al., 2015).

To explore if and how an AI chatbot can be used to develop and carry out learning assessment, I did a pilot experiment linked to a teaching session. After the session, the students had the chance to both answer a multiple-choice questionnaire (MCQ) and have a discussion with a chatbot, which asked questions on the topic in a similar manner to an oral exam. I asked how they perceived the chatbot format, and evaluated if I could use the material from the chatbot conversation to assess whether the students learned what I expected. The aim was to evaluate whether it was similar, better or worse than the MCQ and how. Also, to get first insights on how this assessment format would change both the student's and the teacher's experience in relation to an exam.

Methodology

Assessment formats and general set-up

The experiment was embedded in a teaching session within the course "Animal and plant diversity" for BSc students in the Natural Resources program of the University of Copenhagen. I played a role as teaching support and was responsible for the content and activities of a few sessions but was not involved in the overall course structure and final course assessment design. This course is generally taught in Danish, but there were some classes in English, including the one relevant to this project. This session consisted of a lecture and a group activity on the topic of "Microbial partners as hidden players in animal and plant diversity". In the last part of the session, the students answered a quiz composed of two sections, each covering a subtopic from that session (not the whole course). The format of the first section was a multiple-choice questionnaire (MCQ) and the second section was a chatbot-based discussion. I designed two versions of the quiz, swapping the format (MCQ or chatbot) for each subtopic. The questions were not identical, as they were adapted to each evaluation format. Both versions of the quiz are included in Appendix I.

A total of 22 students participated in the experiment. 12 students took the first version of the quiz (Q1) and 10 students the second (Q2), such that the performance in both formats could be compared independent of the subtopic and for different individuals. 7 of the participants completed only one of the two assessment formats in the quiz, yet their scores were also included in the analysis. The quiz was available for the students in the corresponding course module on Absalon (online course platform at the University of Copenhagen), as a practice quiz (not graded). The last item on the form provided a link to the chatbot, which was created as described in the corresponding section below. The students were asked to copy their discussion or share a link to the conversation with the chatbot, which was uploaded to Absalon as an assignment.

The students were not formally assessed on this activity, and they were informed about this as well as about the purpose of the experiment before they responded the quiz.

Chatbot creation

I used the AI chatbot platform Poe (Platform for Open Exploration) to create two bots using the GPT-4o model. These were named "MicroPartner_Q1" and "MicroPartner_Q2", corresponding to each quiz version. A set of four questions (one set for each chatbot/quiz version) was included in the prompt used to create the bots, and was a baseline for the bots. The prompt was otherwise the same for both bots and is included in Appendix II. The knowledge base, i.e. the source material for the bots, consisted of three documents:

1. Background literature: a document synthesizing relevant information related to the questions in the bot prompts.
2. The full list of questions (for both quizzes) with corresponding answers, which I prepared previously.
3. The transcripts of videos that were part of a group activity in the teaching session, describing relevant examples/biological systems that illustrate the topic.

Evaluation and data analysis

The MCQ section of the quiz was automatically evaluated on Absalon and the chatbot discussion was evaluated manually using the rubric

shown in Appendix III. The maximum score for each section was 8 points. The time used to evaluate chatbot discussions was recorded individually.

The scores from the two assessment formats (MCQ and chatbot) were compared using a generalized linear mixed model using a Poisson distribution family, including format as a fixed effect and both student id and quiz version (Q1 or Q2) as random effects. Likelihood ratio tests against corresponding null models were applied to evaluate the goodness of fit considering each factor. All analyses were carried out in R version 4.4.1.

Results and Discussion

Practical insights and considerations of the chatbot-based assessment

While designing and testing the chatbot, as well as during the actual experiment with the students, several aspects were relevant to familiarize myself with this technology and explore its shortcomings and potential. First, it was surprisingly challenging to find the right balance between an automated list of questions and a discussion partner, so that the student would have an interactive experience similar to an oral exam. The bots tended to make the discussions very lengthy, including extended explanations and even answers throughout the conversation with the students. My approach was to add a rule in the prompt stating that no answers to the questions should be provided and to include a defined set of questions that the bot should make sure to ask throughout the discussion. While this mitigated the problem partially, there was still a strong tendency towards long comments with extensive detail and it remained somewhat unpredictable if the bot would sometimes provide the answer to the student. This could easily confound formative and summative assessment and could be undesirable if the latter is intended, as would be for an oral exam. It is possible that a different type of chatbot (e.g. not based on generative artificial intelligence) could be more appropriate and is worth considering for the future.

Another pitfall of the chatbot was the excessively positive and complimenting style, despite trying to avoid this by emphasizing it in the prompt. The bot’s response occasionally gave the impression that

the student's answer was correct, although it was partially or fully wrong. This might reinforce some misunderstandings in the students. Also, the students' answers sometimes led the bot to integrate information that is mistakenly linked or brought out of context. These are significant issues of the method for an educational purpose and might be overcome with sufficient tailoring of the bot prompt and the bot's knowledge base (source materials), or the use of an alternative model. Also, encouraging the students to provide more elaborate responses complementing short or single-word answers could be a useful improvement on the current chatbot prompt (Appendix II).

Despite these potential shortcomings, some discussions with the chatbot flowed very well. Notably, the quality of the exchange seemed to depend on the accuracy and style of answers provided by the student, and it is possible that English language and written communication skills influence the outcome. This could compromise equal treatment (reliability) in the exam, yet oral exams have arguably similar challenges. The student's personality, body language and verbal communication skills might in fact interfere with an oral assessment (Davis & Karunathilake, 2005). Such limitations underline the advantage of including diverse assessment formats and the importance of being aware of these potential biases.

Quantitative comparison between assessment formats

A formal comparison between the results obtained by the students in the MCQ and chatbot assessment formats showed that the scores were generally higher with the MCQ method (glmm, likelihood ratio test, $p < 0.001$) (Fig. 1). The individual student, but not the quiz version (i.e. the topics of the questions) also influenced the scores. These quantitative differences were not fully surprising but rather expected after reading the chatbot discussions. The fact that I could more easily assess deeper learning levels in the chatbot method, likely had an impact on how challenging it was to get a higher grade. Adding to this effect, it is possible that —as non-native speakers— the need to use English actively when discussing with the chatbot instead of passively in the MCQ, introduced an extra challenge for some students to fully show their understanding. This potential compromise in the validity of the assessment could be overcome by allowing the bot to operate in

different languages, which is currently feasible. In fact, one of the students in the experiments switched to Danish in the discussion with the bot. Given sufficient quality in automated translations, this could be a viable and diversity-embracing approach.

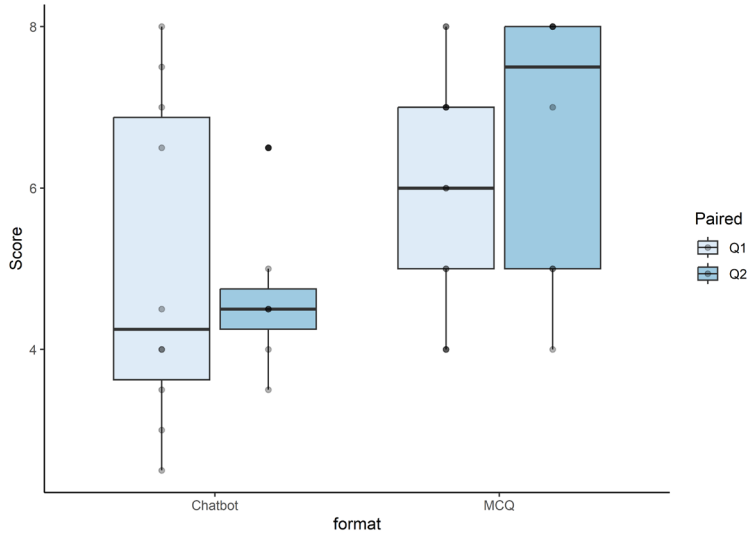

Fig. 1. Scores obtained by a total of 22 students shown for each assessment format. 12 students took version 1 (Q1) and 10 students' version 2 (Q2) of the quiz. Both Q1 and Q2 included an MCQ and a chatbot-based section.

Directly visualizing the frequencies of the scores across the group of students according to assessment format shows that in the MCQ method, there is a more homogeneous distribution than in the chatbot scores (Fig. 2a). 28% of the students (6/21) obtained the highest grade and none got a score below 4 points. Contrastingly, in the chatbot discussion only 5% (1/17) of the students achieved a maximum score, and the majority of the group had a score below 4.5. These differences did not depend on the quiz version, that is, the topic of the questions did not influence this distribution (Fig. 2b).

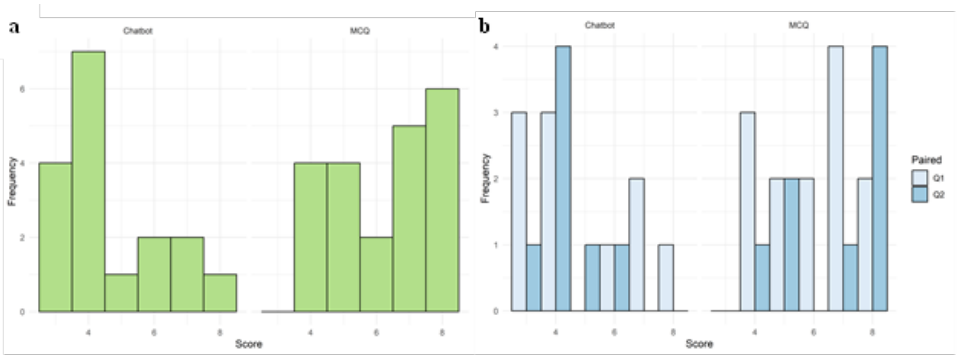

Fig. 2. Frequency distributions (histograms) of scores obtained by the students shown for (a) each assessment format and (b) distinguishing both quiz versions within each assessment format.

This experiment was designed so that the topics, length, and time at which the different assessment formats were applied was as comparable as possible. While I find the qualitative comparison between the outcomes more insightful and relevant (discussed below), these quantitative differences likely reflect intrinsic challenges of MCQ as an assessment method. The results suggest that the teaching session was overall quite successful in conveying factual knowledge, getting the students to analyze it and apply it to concrete situations, which were skills evaluated through the MCQ. The synthesis level, which was very important to achieve higher scores in the discussion with the chatbot, was weaker and more variable across the students. This was not detected through the MCQ, while the chatbot allowed the assessment of a broader range of abilities. This is central to the design of assessments, since these should not only be able to show that specific learning objectives have been achieved, but also to uncover potential weak spots (Leth Andersen et al., 2015).

Qualitative evaluation of the chatbot-based format and potential for formative assessment

Both the MCQ and the chatbot discussion were useful for me as an examiner to evaluate the intended learning outcomes, but the quality of the information was different. I got a better understanding of the individual student's biases and deeper learning outcomes when reading their answers to the chatbot although the MCQ allowed to cover more content. While the difference in breadth might seem obvious given the characteristics of each method, I intended to cover

similar amounts of content with each section of the quiz. However the students usually provided short responses in the chatbot discussion, and the bot compensated with long answers. It was therefore the bot, not the student, covering the topic broadly. In an oral exam, the examinator can (and should) have the control to adapt to the student and ask questions such as “Can you elaborate on that”. As suggested above, it should be possible to improve the bot to do this. However, I believe that an experienced human examiner will easily outperform the bot in abilities like the constant adaptation to the student, asking the appropriate question at the right time and mindfully following the progression of the exam. Such skills are key to making the most of an oral exam (Leth Andersen et al., 2015) and we can only speculate if or when large language models will come truly close to this kind of perceptive interaction. Interestingly, this might also be a concern from the student’s perspective. During the experiment, one of them asked the chatbot “How do you think I did? And how do you think you did as the role of the interviewer?” The bot’s answer was very positive and unnatural, especially hard to believe or relate to. This led me to the question whether we can currently (or ever) make a bot that can answer these questions in a truly useful way and emphasized the value of a teacher maintaining the role of a reflective and self-critical examiner.

Whether a chatbot will be able to guide a student through a personalized discussion meant for formal assessment might not be the most crucial question. Instead, we should ask if this is desirable. There is growing interest in using AI-powered discussion partners for learning, and tools like Khanmigo (https://www.khanmigo.ai/teachers) or GeniousTutor AI (https://geniustutor.ai/?via=topaitools) have become popular for students from elementary school up to higher education levels. These can indeed be suitable and powerful strategies to engage students and promote self-paced learning. However, assessment should demand high involvement of the teacher as, independent of the methodology, it is a measure of learning success and is key to guarantee that the very essence of teaching is in place. Automatizing or outsourcing this responsibility is risky, and tools like AI should be used in a carefully supervised manner to maintain the overview. This might be less crucial during learning activities, where the student’s independence can come with multiple advantages. Luckily, currently available AI-

based educational toolboxes like those mentioned above are not only promising for student learning, but can also equip teachers with question generators, discussion prompts or automatic creation of multiple-choice evaluations (e.g. (https://www.khanmigo.ai/teachers) that can complement and enhance assessment strategies.

The students' experience

The experiences of the students providing oral feedback on the experiment with the chatbot were varied but generally positive. While some said that they would have liked to have more depth in the answers and comments from the chatbot, others said that the length and depth of answers was fine. This might result from different needs from the students, but also from a degree of unpredictability in the chatbot's style that can be problematic for a reliable assessment. In terms of how comfortable they felt in the discussion with a bot, one student said, "it was refreshing" and another mentioned that it was encouraging to have very positive and praising answers. I find this particularly interesting, since stress can be a strong limitation for many students during oral exams and including chatbot discussions among exam formats could, even unconsciously, benefit these students. A chatbot might not meet everyone's needs in this sense though, as another student confessed to the bot that "This [the conversation] was painstakingly awkward…"

Some students commented that it was nice to discuss with the bot. However, they did point out that this rather resembled a learning tool and many said that they did learn something while interacting with the chatbot. As discussed above, this was not the original intention when designing the bot and is not recommended for oral exams (Willum Johansen, 2023), but it underlines that the approach can be useful for formative assessment.

The examiner's experience

The most demanding task as an examiner for the chatbot format was to design a bot that fulfilled the needs for the discussion. This might be a significant time investment in the beginning, especially considering the multiple rounds of testing and improving the prompt, defining the source material and identifying the most appropriate model for this purpose. However, once having designed the bot, the

examiner's role has some advantages compared to an oral exam. It took me on average 5.5 minutes to evaluate discussions that lasted approximately twice as long. I estimate that revising these written discussions would be significantly faster compared to carrying out oral exams for multiple students. While I would likely prefer the direct interaction component of an oral exam, it was interesting and insightful to read the discussions. This approach might offer a good compromise for large courses in which oral exams are too time consuming, while still improving the assessment of deep learning in comparison to an MCQ.

Conclusions and outlook

Designing a chatbot for learning assessment has significant challenges, particularly in refining the prompts to ensure effective interaction for the desired purpose and in guaranteeing reliability and validity of the assessment. While the chatbot based discussions tested in this pilot study were useful to assess learning outcomes to some extent, they currently seem more feasible and appropriate as learning tools, with promising potential for formative assessment. I look forward to adapt this chatbot as a discussion partner using the reflections from this project. It will be useful to include it as an activity for reinforcing concepts after the session, or even as part of a flipped-classroom program in other courses. Asking the students to share their discussions can still be useful for the teacher to get a close look at their understanding, rather than for summative assessment.

Despite their own set of challenges and advantages, AI chatbots expand the array of assessment methods available, providing examiners with more diverse options and offering students varied ways of being assessed. It's crucial to keep teacher supervision and closely monitor assessment activities to ensure quality, relevance, and most importantly, to maintain the perceptiveness that a teacher —and not a bot— can offer. Looking ahead, it seems sensible to keep pace with advancements in AI tools for education and integrate them into a diverse toolbox, while understanding and remaining critical about their limitations.

References

Biggs, J., & Tang, C. (2011). Constructively aligned teaching and assessment. In *Teaching For Quality Learning At University* (Fourth, pp. 95–110). McGraw-Hill Education (UK).

Chen, Y., Jensen, S., Albert, L. J., Gupta, S., & Lee, T. (2023). Artificial Intelligence (AI) Student Assistants in the Classroom: Designing Chatbots to Support Student Success. *Information Systems Frontiers*, *25*(1), 161–182. https://doi.org/10.1007/s10796-022-10291-4

Davis, M. H., & Karunathilake, I. (2005). The place of the oral examination in today's assessment systems. *Medical Teacher*, *27*(4), 294–297. https://doi.org/10.1080/01421590500126437

Leth Andersen, H., Dahl, B., & Tofteskov, J. (2015). Assessment and exams. In L. Rienecker, P. Stray Jørgensen, G. Holten Ingerlsev, & J. Dolin (Eds.), *University Teaching and Learning* (pp. 369–407). Samfundslitteratur. https://www.academicbooks.dk/da/content/university-teaching-and-learning-0

Owoseni, A., Kolade, O., & Egbetokun, A. (2024). Applications of Generative AI in Summative Assessment. In A. Owoseni, O. Kolade, & A. Egbetokun (Eds.), *Generative AI in Higher Education: Innovation Strategies for Teaching and Learning* (pp. 97–122). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-60179-8_4

Willum Johansen, M. (2023, 2024). *Oral exams in practice. Video material for University Pegadogy Course*. Department of Science Education, University of Copenhagen.

Appendix 1

Quiz 1

1. According to the endosymbiotic theory, bacteria were relevant to the evolution of eukaryotic cells because they
    a. Protected the ancestors of eukaryotic cells from antagonists
    b. Conferred metabolic advantages and became mitochondria and chloroplasts
    c. Are the ancestors of the nucleus

2. The microbiota of an animal can include
    a. Bacteria and fungi
    b. Archaea
    c. Protists
    d. All are correct

3. You are studying the microbiota of a bird species in Denmark and find that 80% of the individuals you have sampled have the same bacterium in their uropygial gland and they are healthy. You isolate this bacterium and confirm that it can grow in pure culture in the lab. You can conclude that this bacterium is likely:
    a. A pathogen
    b. An obligate symbiont
    c. A facultative symbiont

4. A horizontal symbiont transmission route means that a microbial symbiont is acquired from
    a. The environment or unrelated hosts
    b. Members of the same colony
    c. The mother

5. A vertical symbiont transmission route favors
    a. Flexible acquisition of novel symbionts when environment changes
    b. Increased dependence between the partners
    c. That symbionts become harmful

6. A bacteriocyte is
    a. A symbiont that can fix nitrogen
    b. A bacterial cell
    c. A host cell that contains bacterial symbionts

7. In which type of animals are intracellular symbionts particularly common?
    a. Vertebrates
    b. Amphibians
    c. Insects
    d. Reptiles

8. Microbial communities associated to plants are usually more diverse
    a. Belowground
    b. Aboveground
    c. Within the plant cells

9. Please follow the link to the next section of the quiz, which is an interactive chatbot. Once you are done, use the second link to submit the record of your discussion as a word or text file. Thanks!
   QuizBot: MicrobePartner_Q1

Quiz 2

1. For an animal, living in symbiosis with a microorganism can be beneficial because
   a. Bacteria and fungi can produce various compounds that animals can't
   b. There is no cost in maintaining microbial symbionts
   c. Microbes and animals never* compete for the same resources

   *In future versions: replace for "don't"

2. If you find an obligate intracellular bacterial symbiont in a cicada, which feeds on plant sap, it is likely that the bacterium can provide its host with
   a. Sugars
   b. Digestive enzymes
   c. Essential amino acids

3. Leguminous plants usually benefit from *Rhizobium* bacteria because they
   a. Avoid the formation of root nodules
   b. Transform atmospheric nitrogen into ammonia
   c. Detoxify ammonia

4. The protection of eggs by symbiotic bacteria is
   a. Often mediated by bioactive compounds
   b. Restricted to insects
   c. Restricted to terrestrial animals

5. Animals can benefit from microbes to maintain a plant-based diet because
   a. Animals lack enzymes to break down complex polymers like pectin and cellulose
   b. Some microbes can detoxify harmful compounds produced by plants
   c. Both are true

6. Defensive symbiosis in insects living underground can involve
   a. Preservation of food or immobile life stages
   b. Difficulty to acquire new symbionts
   c. Complete loss of the insect's immune system
   d. All are true

7. Chemosynthetic bacterial symbionts of marine invertebrates can broaden the ecological niche of their hosts because:
   a. They produce compounds that fend off pathogens and other natural enemies
   b. They allow the host to exploit inorganic matter for energy and colonize habitats without sunlight
   c. They have chloroplasts

8. Which of the following is an example of how microbial symbionts can constrain diversification in their hosts?
    a. By enabling plants to access essential nutrients like phosphorus and nitrogen more efficiently
    b. By producing antimicrobial compounds that protect plants from pathogens
    c. By creating obligate relationships that limit the host's ability to adapt to new environments
    d. By helping animals detoxify environmental toxins

9. Please follow the link to the next section of the quiz, which is an interactive chatbot. Once you are done, use the second link to submit the record of your discussion as a word or text file. Thanks! QuizBot: MicrobePartner_Q2

Appendix II

Chatbot prompt

Context
You are a quiz bot. You will lead a short discussion with a student to assess their knowledge on the impact of beneficial microbial partners in animal or plant hosts at a BSc level. You will be provided retrieved documents from the course associated to this discussion. You should make sure to ask the questions listed below [under "### Questions"]. After I respond to each, create one new question based on my answer. Do not be overly praising. Please kindly remind me in the beginning and end of the discussion that I should copy and paste our conversation into a word document once the session is finished.
Rules for the discussion

- Wait until I reply to engage in discussion
- Do not provide the answers to the questions
- The full conversation should be in English.

Questions
MicroPartner_Q1:

- Explain why it can be advantageous for an animal or plant to associate with a microbial symbiont or outsource specific functions.
- Name one example in which symbiotic bacteria protect an immature life stage of an animal and briefly explain the mechanism.
- How can microbial symbionts favor animal and plant diversification?
- Can microbial symbionts also constrain diversification? [if the answer is affirmative, ask how]

MicroPartner_Q2:

- Why does the endosymbiotic theory highlight the relevance of bacteria in the evolution of eukaryotic cells?
- What is the difference between obligate and facultative symbiosis?
- Discuss the advantages and disadvantages of horizontal transmission of microbial partners in animal hosts.
- What are the implications of a vertical transmission route

Appendix III

Rubric used to evaluate each answer provided by the student in the chatbot discussion.

Criteria	**Description**
Accuracy	Concepts and facts integrated in the answer are correct
Breadth of Knowledge	Demonstrates comprehensive understanding of the topic.
Examples (if applicable)	Provides relevant examples that effectively illustrate the topic.
Depth of Arguments (if applicable)	Arguments are well-developed, logical, and demonstrate deep understanding.

Scoring:

- **2 Points:** Meets all criteria effectively.
- **1.5 Points:** Meets most criteria, with minor gaps.
- **1 Point:** Meets some criteria, but lacks depth or clarity.
- **0.5 Points:** Meets few criteria, with significant shortcomings.
- **0 Points:** Does not meet criteria.