BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **24**, No 3
Sofia • 2024 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2024-0027

Exploring the Efficacy of GenAI in Grading SQL Query Tasks: A Case Study

Thair Hamtini, Abdelbaset J. Assaf

The University of Jordan, Amman, Jordan
E-mails: thamtini@ju.edu.jo ab.assaf@ju.edu.jo

Abstract: *Numerous techniques, including problem-solving, seeking clarification, and creating questions, have been employed to utilize generative Artificial Intelligence* (*AI*) *in education. This study investigates the possibility of using Generate AI* (*GenAI*) *to grade Structured Query Language* (*SQL*) *queries automatically. Three models were used which are ChatGPT, Gemini, and Copilot. The study uses an experimental approach to assess how well the models perform in evaluating student responses by comparing the models' accuracy with those of human experts. The results showed that despite some inconsistencies, GenAI holds great promise for streamlining. Thus, further research is required in light of inconsistent GenAI performance. If these issues were resolved, GenAI can be utilized in education. However, human oversight and ethical issues must always come first.*

Keywords**:** *Grading SQL, ChatGPT, Gemini, Copilot, Automation.*

1. Introduction

Due to limited resources and full class schedules, teachers frequently find it difficult to provide feedback on assignments like grading SQL queries. The opportunity for individualized coaching may be limited by this drawn-out process. Generative Artificial Intelligence (GenAI) offers a promising solution because it can speed up grading and save teachers precious time. GenAI can be used to create sample or exam questions with little to no experience in Artificial Intelligence (AI) [1]. Furthermore, it can carry out various tasks as suggested by L u c k i n et al. [2] who believe that AI has the power and the potential to change how education is assessed by providing personalised feedback based on each students' needs. This will improve the students' learning experience and it is especially helpful in classroom settings where students don't get much one-on-one time. Since grading task is time consuming, using GenAI to automate grading tasks will allow teachers to concentrate on developing curricula and giving each student individualized support.

Grading computer programming and SQL questions is considered a time-consuming process. Especially when a certain question can be answered in many different ways. This process requires resources that aren't available for all schools.

Employing GenAI to manage the grading process will help schools overcome this obstacle. Messer et al. [3] through their investigation, demonstrated that AI tools are proficient in grading computer programming questions. Grading SQL queries using GenAI shares a common base. Once grading tasks are automated, teachers will have more time to improve their teaching materials, skills and their own professional development. This research contributes to the expanding field of AI in education by examining how ChatGPT [4], Gemini [5] and Copilot [6] perform in grading SQL query tasks. Our goal is to identify methods for evaluating students work in a database course. This includes examining the connection between educational assessment techniques and AI. As Luckin and Cukurova [7] highlight, AI integration in education holds the potential to enhance learning outcomes and promote personalized instruction.

This paper looks into how GenAI could improve the grading process of SQL queries. The study is driven by the question: "How can GenAI be effectively utilized for assessing students' solutions to SQL query tasks?" The research attempts to take a close look at how AI technologies might change the way SQL query evaluation is conducted. The purpose of this study is to clarify potential obstacles to the implementation of AI grading systems in educational settings and to demonstrate the potential benefits of these tools.

2. Related work

AI has been involved in almost everything in our lives, including educational aspects. In this section, we investigate how evaluation methods for tasks such as SQL queries could be carried out by GenAI. Previous Studies on the use of GenAI in education concentrating on its potential for automating assessments of programming and related skills are reviewed. Through an analysis of its advantages, drawbacks, and specialties, this section seeks to lay the groundwork for understanding how well GenAI grades students work in SQL queries.

2.1. Benefits of GenAI in higher education

GenAI holds the potential to significantly enhance higher education by offering personalized learning experiences represented by individualized feedback based on each students' needs. Yan et al. [8] critical evaluation emphasizes the revolutionary impact of GenAI in assessment, demonstrating how important it is to revisit assessment principles and rethink assessment practices in higher education. Using GenAI is becoming a daily routine for educators. Thus, Chan's [9] highlights the importance of giving university faculty members and students AI literacy training, so they can use GenAI in the classroom effectively. Further, Chan asks for additional efforts to create more comprehensive and focused policy documents on AI in the classroom.

Chen, Chen and Heffernan [10] conducted a study where students received customized math tutoring from a conversational agent based on the generative model ChatGPT. The results showed that the conversational agent adjust its explanation based on students' misconceptions and it is customised according to

the learner comprehension. This shows part of GenAI potential in improving learning. Further, the study by Kim, Park and Lee [11] used ChatGPT to mark essays written by high school students. The model was trained on a human-judged essay dataset. The results showed that the model correctly marked the essays with a 0.86 correlation to human grades. Further, the study concludes that the model could recognize important components of well-written essays and offer criticism that was comparable to that of human editors.

Further, Ling and Chiang [12] demonstrated the potential for improved support for students learning programming from an adaptive learning system built on a generative model ChatGPT, which led to better performance in programming assignments. Their research showed that the model was capable of understanding students' knowledge and consequently altered the difficulty of the problems it produced. Overall, GenAI has the potential to be a powerful tool for improving teaching and learning by providing customized tutoring for each individual, automated essay grading, language translation, and adaptive and interactive learning.

2.2. Challenges and considerations in using GenAI for grading student assignments

Although GenAI presents advantages in terms of effectiveness and the possibility, of expansion, its implementation for grading student assignments requires careful consideration of many challenges. For instance, Rudolph, Tan and Tan [13] raise concerns about the possibility of GenAI to accidentally propagating misinformation, especially in summarizing or generating creative text formats. This demonstrated the necessity for educators to remain vigilant about the accuracy and source of information within student assignments. Chan [9] also highlights the absence of rules and the existence of situations, in evaluating practices with GenAI. This highlights the significance of creating protocols in universities to guarantee the efficient application of GenAI, for grading and assessment purposes.

Another challenge arises from GenAI's comprehension of the topics it evaluates. Machine learning models learn from data patterns. Thus, GenAI may struggle to grasp the core concepts students are studying. This limitation can hinder the effectiveness of providing explanations and feedback to address a student's requirements and misunderstandings. A teaching system based on a generative model, for instance, was unable to provide explanations that addressed students' particular misconceptions, as demonstrated by the study of Wang, Chen and Heffernan [14]. Because of this, GenAI can be used in addition to human oversight rather than as a replacement.

The training data of generative models is critical to their operation. Therefore, biases in the training data will cause the results to be biased, creating even another issue. For instance, if a model is trained on a dataset of essays that are primarily written by students from a certain demographic, it may not be able to accurately grade essays written by students from other demographics. This shows that in order to use GenAI for grading, the data which the model used for training must be carefully selected.

Although GenAI holds promise for automatic grading, its use requires caution. One concern is the limitations in how GenAI evaluates student work which could

potentially lead to spread of misinformation. Another is that GenAI might not be able to provide customized feedback for each individual if it does not fully understand the concepts that students are learning. Finally, biases in the models' training set of data may result in unfair grading policies that disadvantage particular student populations.

3. Methodology

This study followed an exploratory approach to investigate how GenAI can be applied in education. An experiment was conducted in which, three different GenAI models: ChatGPT, Gemini, and Copilot were fed by students' solutions to SQL query tasks. The goal was to evaluate these models' accuracy efficacy and consistency in marking these questions. This methodology will review and explore the capabilities of GenAI models in assessing student solutions, the possible advantages of integrating GenAI into education, and the potential drawbacks of using GenAI in education. While also reflecting on the implications for educational practices. The study was driven by the research question: How can GenAI be effectively utilized for assessing students' solutions to SQL query tasks? The results add to the literature on AI in education, which is primarily focused on grading technical assignments like SQL queries.

3.1. Materials

Students were provided with three populated tables to answer questions using JOIN operations in SQL:

Table 1. Students (student_id (PK), student_name, major, GPA)

Table 2. Courses (course_id (PK), course_name, credits)

Table 3. student_courses (student_id (PK), course_id (PK))

They were then asked to answer the following questions using JOINs:

1. Write a SQL query to retrieve the names of all students enrolled in the "Introduction to SQL" course.

2. Write a SQL query to retrieve all course names and find the total number of students enrolled in each course.

To ensure consistency in grading, we provided each of the ChatGPT, Gemini, and Copilot with the correct answers and established a detailed marking criterion for each question as follows:

3.2. Marking criteria for Question 1 (Total Marks: 5)

- 1 mark for selecting the correct attribute (e.g., student_name) to retrieve (0 marks for incorrect attribute selection).
- 1 mark for retrieving data from the correct tables (students, courses, student_courses) (0 marks for incorrect or incomplete tables).
- 1 mark for using the correct WHERE clause to filter data (0 marks for incorrect or missing WHERE clause).
- 1 mark for creating the correct INNER JOIN between students and student_courses tables on student_id (0 marks for incorrect join type, table names, or join condition).

- 1 mark for joining the results from the previous step with the Courses table on course_id (0 marks for incorrect join type, table names, or join condition).

3.3. Marking criteria for Question 2 (total marks: 5)

- 1 mark for selecting the correct attribute (e.g., course_name) to retrieve (0 marks for incorrect attribute selection).
- 1 mark for retrieving data from the correct tables (courses, student_courses) (0 marks for incorrect or incomplete tables).
- 1 mark for using the aggregate function COUNT (0 marks for missing or incorrect aggregate function).
- 1 mark for using the GROUP BY clause to group data (0 marks for missing or incorrect GROUP BY clause).
- 1 mark for using a LEFT JOIN between courses and student_courses tables on course_id (0 marks for incorrect join type, table names, or join condition).

4. Results

This study compared the effectiveness of ChatGPT, Gemini, and Copilot in grading student responses to SQL queries. The focus was on two key aspects:

- **Accuracy.** How well did the models' grades match the marks awarded by human experts?
- **Consistency.** Did the models provide consistent grades for the same student responses across different assessments?

4.1. Initial challenges and revised approach

We initially present all the information at once (tables, questions, answers, marking criteria, student responses) to the models. However, the outputs from all three models (ChatGPT, Gemini, and Copilot) revealed inaccuracies, indicating they couldn't understand everything presented together. Notably, Copilot even missed grading three questions entirely.

To address these challenges and ensure comprehensive understanding, we adopted the following step-by-step communication strategy:

1. Tables and data. The models first received the tables with data and responded with sample queries to demonstrate their understanding of the data structure.

2. Questions. Next, they were presented with the two SQL query questions and answered them with their own SQL queries.

3. Answers. Following this, we provided the correct answers to both questions, which all models validated successfully.

4. Marking criteria. Finally, the marking criteria were introduced, and all models consistently scored the correct answers perfectly (10/10). This confirmed their grasp of the specific evaluation criteria.

This breakdown clarifies the initial struggles and highlights the revised approach that ensured the models understood each element individually before tackling student responses.

4.2. Student response evaluation and analysis

We presented the student responses to all three models (ChatGPT, Gemini, and Copilot). Tables 1, 2 and 3 summarize the performance of all three models compared to the human-assigned marks.

Table 1. Student 1 marking

Question	Part	Mark	ChatGPT	Gemini	Copilot
1	1	1	1	1	1
	2	1	1	1	1
	3	1	1	1	1
	4	1	1	1	1
	5	1	1	1	1
Out of 5		5	5	5	5
2	1	1	1	1	1
	2	1	1	1	1
	3	1	1	1	1
	4	1	1	1	1
	5	0	0	1	1
Out of 5		4	4	5	5
Total		9	9	10	10

The tables detail student number, question number, part number, and individual marks awarded by human experts and each GenAI model. This allows for a detailed analysis of their performance on specific student queries, such as selecting the correct attributes, using appropriate joins, and achieving the desired results.

Table 2. Student 2 marking

Question	Part	Mark	ChatGPT	Gemini	Copilot
1	1	1	1	1	1
	2	1	1	1	1
	3	1	1	1	1
	4	0	0	0	1
	5	0	0	0	0
Out of 5		3	3	3	4
2	1	1	1	1	1
	2	1	1	1	1
	3	1	1	1	1
	4	0	1	1	1
	5	1	1	0	1
Out of 5		4	5	4	5
Total		7	8	7	9

Table 3. Student 3 marking

Question	Part	Mark	ChatGPT	Gemini	Copilot
1	1	0	0	0	1
	2	1	1	1	1
	3	0	0	0	1
	4	1	1	1	1
	5	1	1	1	1
Out of 5		3	3	3	5
2	1	1	1	1	1
	2	1	1	1	1
	3	1	1	1	1
	4	1	1	1	1
	5	1	0	0	1
Out of 5		5	4	4	5
Total		8	7	7	10

Looking at Tables 1, 2, and 3, we can see the strengths and weaknesses of each GenAI model. Here are our specific observations from the analysis.

1. Understanding Join Types. Several inconsistencies emerged in how the models handled join types. In Student 1, Question 2, only ChatGPT correctly identified the missing left join. Likewise, Copilot awarded partial marks in Student 2, Question 1 despite the incorrect use of a natural join. These inconsistencies emphasise the significance of interpreting join types accurately, as they are vital for marking students' SQL query solutions correctly.

2. Identifying Missing Clauses. Some models performed weirdly with missing clauses in student responses. In Student 2, Question 2, group by clause was missing, however, both ChatGPT and Copilot incorrectly awarded full marks. Identifying all possible solutions and detecting missing clauses is an essential part in in ensuring the accuracy of SQL query results. It's worth noting that while ChatGPT was marking the student's answers, the model was rewriting the students' responses. For this particular student, when the response was rewritten, a "GROUP BY" clause was added that wasn't there before. This behavior is very weird and cannot be justified.

3. Accuracy in Marking Technically Correct Responses. Many questions can be answered in multiple ways. The models behaved differently in marking technically correct responses with different approaches. For instance, in Student 3, Question 2, a right join (flipped tables) was used to achieve the correct answer. ChatGPT and Gemini considered it wrong because the provided key answer used a left join. Only Copilot marked this part as correct. This shows that technically correct responses can be incorrectly marked if they deviate from the expected criteria.

4. Inconsistencies in Mark Attribution. Some models behaved weirdly in calculating total scores or assigning partial marks. In Student 3, Question 1, Gemini identified one error but awarded a lower total mark than expected. This highlights the potential for discrepancies in scoring, which can affect the overall assessment accuracy.

Well go into further detail regarding these results and their implications for further study in the following section.

5. Discussion

The performance of the GenAI models in grading student SQL queries was examined and both positive and negative aspects were found. While all three models (ChatGPT, Gemini, and Copilot) achieved perfect scores on well-structured responses (e.g., Student 1, Question 1), they showed inconsistent behavior when it came to managing the subtleties and complexity of the student work. (Students 2 & 3).

All models exhibited inconsistencies in grading, particularly with aspects like join (Students 1 & 2) and attribute selection (Student 3, Question 1). Additionally, Copilot demonstrated further inconsistencies across all students, awarding full marks despite missing clauses (Student 2, Question 2) or overlooking incorrect conditions (Student 3, Question 1). These flaws point to a lack of comprehension of the nuances in student responses including questions that aren't fully answered (Student 2,

Question 2), minor errors (Student 3, Question 1), or alternative approaches (Student 3, Question 2).

Despite these limitations, all three models achieved relatively good accuracy in marking some responses. However, instances of assigning incorrect marks, adding irrelevant information, or miscalculating total scores highlight the variability in performance that necessitates further investigation. Fig. 1 shows a line plot that compares the correct assessment with the assessment of the three models used in this case study.

The line plot effectively highlights the differences in grading between the correct marks and those assigned by the three models. By comparing the lines, we can see that while ChatGPT and Gemini tend to be closer to the correct marks, Copilot may be more tolerant or inconsistent. This figure makes it easier to see how each model compares to the actual outcomes that were anticipated and highlights how crucial it is to improve these models in order to guarantee impartial and accurate evaluation in learning environments.

The findings align with the previous research about GenAI in education, which suggests that AI can be a valuable tool for educators, but it should not replace human expertise entirely. The study by M o o r e, S h a w o n and J a m i l [15] highlights the potential of AI for automating routine tasks such as grading, which could result in offering more time for educators to provide personalized instruction and feedback.

Despite that AI models can be valuable tools for educators by automating routine tasks such as grading [9, 11, 15], AI models often struggle with the complexities of human-generated content, as observed in our study (e.g., join types, missing clauses). As our findings demonstrate, shortcomings such as inaccurate grading in AI-generated models emphasize the need for a careful approach. Educators must play an active role in the creation and supervision of AI-driven assessment tools, ensuring that they enhance human judgment rather than replace it.

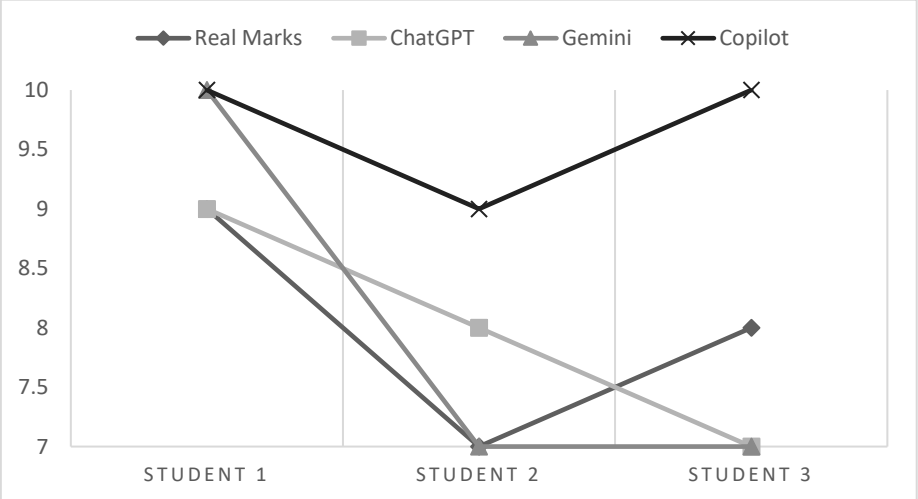

Fig. 1. Comparison of real marks and AI model marks

6. Recommendations for teachers using genai for assessments

Although GenAI models showed potential for improving the efficiency of assessment processes in education such as grading SQL queries, the limitations demand careful implementation. Educators should consider the following points when using GenAI for assessment:

1. Start Small and Focus on Objective Questions: Start with GenAI for clear objective questions and provide a straightforward marking criterion, making use of GenAI strengths in pattern recognition and rule application.

2. Utilize GenAI as a Supportive Tool, not a Replacement: GenAI can be used to free up time by letting it carry out automated tasks.

3. Review GenAI-Generated Marks Carefully: You cannot fully rely on the marks provided by GenAI, especially for complicated questions. Thus, human oversight must be maintained.

4. Provide Clear and Detailed Marking Criteria: GenAI will mark the work based on the training it received. Thus, it must be fed a clear and well-defined marking criterion to maximize its effectiveness. The marking criteria must be comprehensive and should covers unusual ways to answer any particular question.

5. Consider Potential Biases: GenAI results are driven by the datasets it was trained on and bias could appear in many forms. Thus, make use of methods and instruments to encourage fairness and reduce bias in assessments driven by AI.

6. Focus on Overall Learning Outcomes: Learning doesn't always depend on accurate responses. Thus, GenAI might be incorporated to assess a wider variety of learning outcomes for students.

7. Embrace Process-Oriented Learning: Shift the focus from merely analyzing the solutions to the thought and problem-solving processes that produced those solutions, in line with the capabilities of Generative AI models like ChatGPT.

8. Engage in Ongoing Evaluation: Continuously review the effectiveness of GenAI in assessments and its impact on student learning, making adjustments as required.

7. Conclusion

The research delved into how GenAI assists in the automated marking of SQL queries. GenAI may not only take over everyday tasks such as marking students' assignments, but it could also tailor its responses to the mistakes students make which might help them learn better. If the hurdles are overcome with further investigation and development GenAI could turn into an essential resource for those teaching.

While GenAI brings about gains in speed it does not always score student answers the same way, which needs more looking into. Researchers in the future need to focus on reducing bias in the data they use for training to make sure evaluations are fair. Incorporating GenAI into educational tools might just make them work better and be more widely accepted. A good way to bring this together could be by using platforms for managing learning to help teachers and students talk to each other more easily. Making the assessment process smoother could also make people feel

more at ease when they use GenAI. Looking into how GenAI works not just for SQL queries but in different situations could help us understand what it's good at and where it struggles.

However, it is crucial to keep in mind that AI should only be applied sparingly and morally with human oversight continuing to be a crucial part of the evaluation process for education. It is critical to ensure that GenAI models' decision-making processes are transparent and simple to understand if educators are to trust and rely on them. Additionally, in the long run checks and updates will be required to ensure the relevance and accuracy of these models.

References

1. Jönsson, A. Prompting for Progression: How Well Can GenAI Create a Sense of Progression in a Set of Multiple-Choice Questions? – Dissertation, KTH Royal Institute of Technology, 2024. **https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-346350**
2. Luckin, R., W. Holmes, M. Griffiths, L. B. Forcier. Intelligence Unleashed: An Argument for AI in Education. – Pearson, 2016.
3. Messer, M., N. C. Brown, M. Kölling, M. Shi. Automated Grading and Feedback Tools for Programming Education: A Systematic Review. – ACM Transactions on Computing Education, Vol. **2**4, 2024, No 1, pp. 1-43.
4. OpenAI. Bard [Large Language Model] 2024.
5. OpenAI. ChatGPT [Large Language Model] 2024.
6. OpenAI. Copilot [Large Language Model] 2024.
7. Luckin, R., M. Cukurova. Designing Educational Technologies in the Age of AI: A Learning Sciences-Driven Approach. – British Journal of Educational Technology, Vol. **50**, 2016, No 6, pp. 2824-2838.
8. Yan, L., L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, D. Gašević. Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review. – British Journal of Educational Technology, Vol. **55**, 2023, No 1, pp. 90-112.
9. Chan, C. K. Y. A Comprehensive AI Policy Education Framework for University Teaching and Learning. – International Journal of Educational Technology in Higher Education, Vol. **20**, 2023, No 38.
10. Chen, Y., Y. Chen, N. Heffernan. Personalized Math Tutoring with a Conversational Agent. – preprint arXiv:2012.1212, 2020.
11. Kim, S., J. Park, H. Lee. Automated Essay Scoring Using a Deep Learning Model. – Journal of Educational Technology Development and Exchange, Vol. **2**, 2019, No 1, pp. 1-17.
12. Ling, H.-C., H.-S. Chiang. Learning Performance in Adaptive Learning Systems: A Case Study of Web Programming Learning Recommendations. Front. Psychol., 2022.
13. Rudolph, J., S. Tan, S. Tan. ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education? – Journal of Applied Learning and Teaching, Vol. **6**, 2023, No 1, pp. 1-22.
14. Wang, W., Y. Chen, N. Heffernan. A Generative Model-Based Tutoring System for Math Word Problems. – Preprint arXiv:2010.04, 2020.
15. Moore, N. C., F. R. Shawon, H. M. Jamil. An Experiment on Leveraging ChatGPT for Online Teaching and Assessment of Database Students. – In: Proc. of International Conference on Teaching, Assessment, and Learning for Engineering, 2023.

Received: 07.08.2024; Second Version: 14.08, Accepted: 21.08.2024 (fast track)