Features of the practical use of LLM for generating quiz

Oleh Ilarionov[1*], Hanna Krasovska[1†], Iryna Domanetska[1†], Olena Fedusenko[1†]

[1] *Taras Shevchenko National University of Kyiv, Volodymyrs'ka str. 64/13, Kyiv, 01601, Ukraine*

Abstract

The article explores the possibilities of using large-scale language models (LLMs) such as GPT, Claude, Copilot and Gemini for automated test task generation in the field of education. The ability of these models to generate different types of tasks, including multiple choice, open-ended and fill-in-the-blank, as well as their compliance with educational standards and cognitive levels according to Bloom's taxonomy is assessed. A comparative analysis of the quality of the generated tests in terms of complexity, structure and adaptability is carried out. The limitations of the models for generating tasks of higher cognitive levels are identified and recommendations for their integration into educational platforms are given. The results of the study can improve the process of assessing students' knowledge and promote the development of adaptive learning.

Keywords.

LLM, test generation, Bloom's taxonomy, educational standards, test automation

1. Introduction

Thanks to their ability to create new content based on existing data, generative artificial intelligence models (GAMs) are opening up new opportunities in many industries, from business to art, science and education, and allowing for the automation of sometimes routine tasks, speeding up processes that used to require a lot of time and creative effort.

The impact of generative AI is becoming increasingly visible in the educational environment. Generative AI tools can be used for a variety of educational purposes, making the educational process more individualised, adaptive and efficient, providing access to education for a wider range of people, including those with disabilities.

Already, GMIS are helping teachers generate multimodal teaching materials, adjust lesson plans, select relevant literature, and generate various tasks, scenarios, or simulations that help students develop analytical and research skills.

It should be noted that one of the key elements of the educational process is the quality control of students' knowledge and skills, as this allows not only to assess the level of learning but also to identify gaps in learning and improve teaching methods. The control provides feedback between teachers and students, encouraging the latter to learn more actively and develop themselves.

In recent years, one of the most common methods of assessing students' knowledge in modern education has been the test form of control, which has a number of significant advantages: objectivity of assessment, speed and convenience of testing, coverage of a large amount of material, variety of task formats, possibility of analysing statistics, transparency and clarity, standardisation of assessment, adaptability, etc. However, testing, although an effective control method, has its limitations: tests cannot always adequately assess the depth of understanding of the material or

Information Technology and Implementation (IT&I-2024), November 20-21, 2024, Kyiv, Ukraine

* Corresponding author.

† These authors contributed equally.

oleg.ilarionov@knu.ua (O. Ilarionov); hanna.krasovska@knu.ua (H. Krasovska); irinadomanetskaya@gmail.com (I. Domanetska); elvenff@gmail.com (O. Fedusenko)

0000-0002-7435-3533 (O. Ilarionov); 0000-0003-1986-6130 (H. Krasovska); 0000-0002-8629-9933 (I. Domanetska); 0000-0002-5782-5922 (O. Fedusenko)

practical skills of students. Therefore, it is important that test tasks are well thought out, as their quality directly affects the results of control.

Recently, special attention has been paid to LLMs (Large Language Models), which are designed to process and generate texts and can solve various tasks: translation, text creation, emotion analysis, answering questions, etc., and have great potential in the field of test task automation. They can greatly simplify the work of teachers, create adaptive, diverse and personalised tests, which improves the quality of student knowledge control. LLMs can create tasks of various formats, such as closed-ended questions with one or more correct answers; matching questions; fill-in-the-blank tasks; open-ended questions; logical thinking and analysis tests. LLMs can not only generate questions, but also provide detailed explanations for correct and incorrect answers, which allows students to better understand their mistakes and improves their learning.

However, the question arises as to how effectively different LLMs generate such test tasks and whether the tests they create meet generally accepted pedagogical standards.

2. Literature review

2.1. Opportunities for LLM in education

Large-scale language models (LLMs) offer great prospects for improving the efficiency of the educational process, in particular through the automation of tasks that previously required significant time and intellectual resources. Research shows that LLMs can adapt to different learning contexts by automating the development of test tasks, personalising educational materials, and improving access to knowledge for students with different learning needs[5].

The GPT, Claude, Copilot and Gemini models provide the ability to generate both simple and complex test items, including multiple-choice, matching and open-ended questions. Studies of GPT-4 have shown that this model has a high ability to adapt the complexity of tasks, which allows teachers to create questions of different cognitive levels, according to Bloom's taxonomy [5,6]. Claude, on the other hand, demonstrates strengths in ethical and safe content generation, which is especially important in educational environments focused on preventing bias and harmful materials[5,6].

Automation of the creation of training materials and tests using LLM reduces the workload of teachers, freeing up time to work on individual student support. Another important aspect is the ability to create adaptive tasks that adjust to the level of knowledge and learning pace of each participant. This helps to increase motivation to learn, as students receive immediate feedback and can identify gaps in their knowledge at the early stages of learning[7].

Despite these advantages, the issue of integrating LLMs into learning platforms remains relevant. Teachers need to learn how to properly formulate queries to the models to ensure the relevance of the results obtained. Researchers also draw attention to the limitations of free versions of LLMs, which may restrict their use in educational institutions, especially when processing large amounts of textual data or graphical content[5,8]. However, the prospects for the development of these technologies, in particular in terms of improving the accuracy and reliability of models, open up new horizons for innovation in education.

Thus, the use of LLMs in education allows for an integrated approach to learning, combining the automation of routine processes with the increased individualisation of educational experience. This helps to create conditions for more effective knowledge control and the development of students' analytical skills, which is critical for modern education.

A query optimisation algorithm was used to ensure the relevance and quality of the received tasks. It included several key stages: defining the goal, forming a role for the LLM, checking for key details in the request, and eliminating ambiguities before sending it (Figure 1) [9, 10].

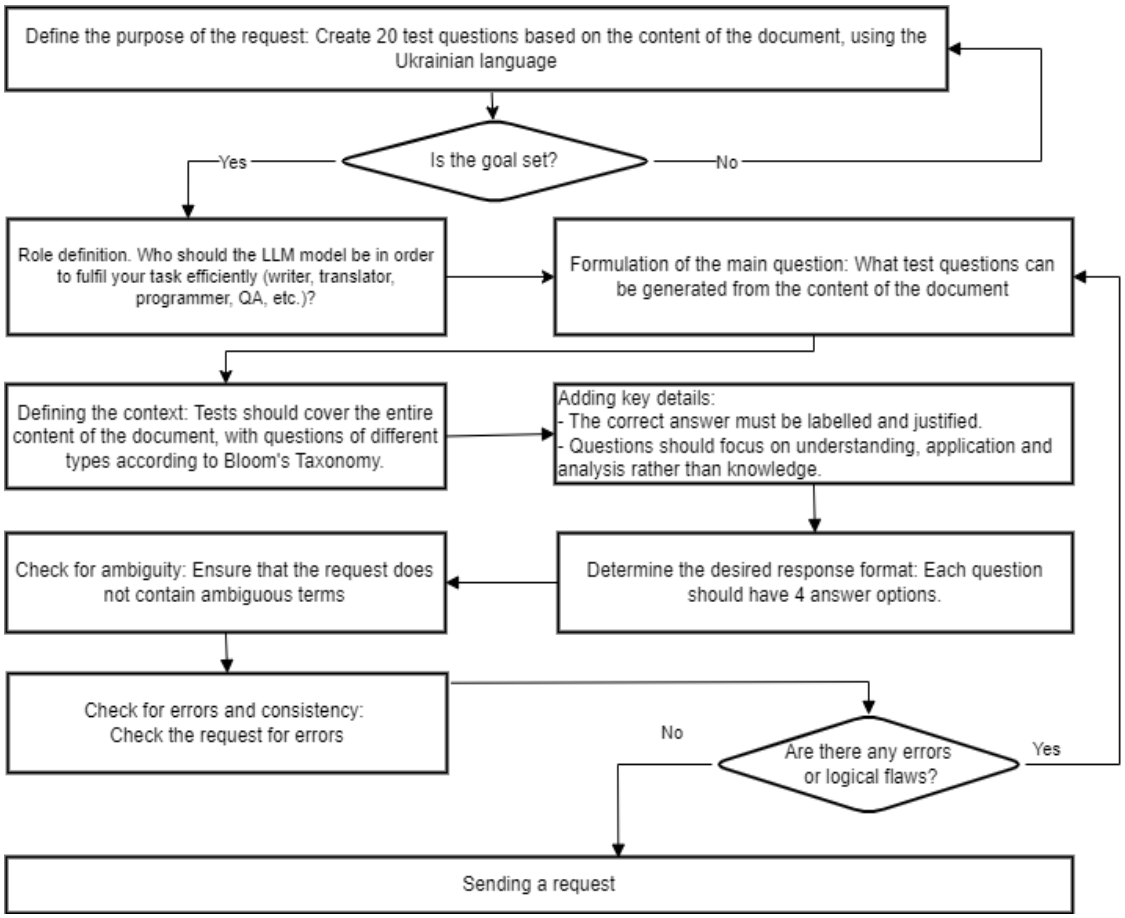

Figure 1: Algorithm for generating an optimal query for large language models

2.2. Analysing the correspondence of tests to Bloom's taxonomy

Bloom's Taxonomy is a widely recognised tool for assessing the level of cognitive complexity of learning tasks, and its use in knowledge testing provides structure and consistency in testing different levels of understanding. The taxonomy divides cognitive processes into six levels: memorising, understanding, applying, analysing, evaluating and creating [5]. To assess whether tests generated by large-scale language models (LLMs) meet the standards of Bloom's Taxonomy, it is important to consider how well the test questions cover these cognitive levels and whether they promote critical thinking and analytical skills.

Research on GPT-4 has shown that this model is capable of generating questions that correspond to different levels of Bloom's Taxonomy, including questions that involve basic memorisation of information (e.g., definitions of concepts or terms), as well as more complex analysis and synthesis tasks that require a deeper understanding of the topic[5,7]. For example, GPT-4 can generate questions that require applying knowledge in new contexts, such as problem solving or comparing concepts that belong to the application and analysis level of Bloom's Taxonomy.

The Claude model also demonstrates the ability to generate questions of different cognitive levels, but its focus is mainly on basic-level tasks such as memorisation and comprehension. The analysed tests generated by Claude show a tendency to create questions that require students to reproduce factual information or explain simple concepts, with less attention to tasks that require evaluation or the creation of new solutions[5].

An important characteristic of LLM-generated tests is their ability to adapt the level of difficulty to different levels of student knowledge. For example, the GPT and Copilot models can generate adaptive tests that match both basic and advanced cognitive levels. This allows teachers to create tests that gradually increase the level of difficulty, starting with simple memorisation questions and ending with analysis and evaluation tasks that require deeper processing of the material[7].

It is important to note, however, that the ability of models to generate questions at the highest cognitive levels, such as creating new concepts or evaluating solutions, is still limited. For example, only a few models, such as GPT-4, are able to effectively formulate tasks that include elements of synthesis and critical evaluation, while other models, such as Claude and Gemini, mainly focus on lower cognitive level tasks[6,7].

Thus, the analysis of the correspondence of LLM-generated tests to Bloom's taxonomy shows that these models are able to cover different cognitive levels, but the level of complexity and variety of tasks vary depending on the specific model. The use of LLMs to create tests opens up prospects for flexible adaptation of learning tasks, which contributes to improving the quality of student assessment.

2.3 Limitations of models in task generation

Despite its considerable potential, the use of large language models (LLMs) for generating test items has a number of limitations that should be taken into account when implementing them in the educational process. These limitations can affect the quality of the created tests, as well as the effectiveness of their use in different learning contexts. The main challenges relate to both technical aspects of the models and pedagogical limitations.

1. Limitation of context size and task types. LLMs, such as GPT-4, Claude or Gemini, process text within a predefined context size, which limits their ability to generate tasks based on large learning materials. For example, when dealing with long lecture materials, the model may lose relevant details or create incomplete questions that do not cover all the necessary information[6]. Some models, such as Copilot and Gemini, further limit text processing in the free versions, forcing users to manually copy content into queries, which reduces their usability[5].

2. Problems with the validity and relevance of questions. Generated tasks do not always fully meet the learning objectives and may not cover the full range of competencies required by the educational programme. Such tasks often have reduced construct validity, as models can formulate questions without taking into account a deep understanding of the subject area or course specifics[7]. In addition, the possibility of randomly guessing the correct answer is especially relevant for multiple-choice tasks if the answer options are not sufficiently differentiated or plausible.

3. Lack of specialisation and domain knowledge. LLMs are able to generate tasks from various disciplines, but their performance may decrease in specific subject areas that require expert knowledge. For example, when creating tasks in programming or medicine, models can generate questions that contain inaccuracies or do not match the level of difficulty of the course[8]. Claude and Copilot demonstrate limited ability to create high cognitive level tasks, such as evaluation and new solution creation, which may reduce their effectiveness for advanced courses.

4. Ethical and methodological limitations. Another important issue is the risk of bias in the questions, as LLMs are trained on large amounts of data that may contain cultural or gender stereotypes. Claude has built-in mechanisms to minimise such risks, but it is not always possible to completely avoid bias[7]. In addition, the tasks created by LLMs may not take into account the individual needs of students with special educational needs, which can limit the accessibility and fairness of testing.

5. Technical limitations and integration. Integration of LLMs into learning platforms can also be difficult due to technical limitations, such as limited access to free versions of models or the complexity of setting up APIs for automatic task generation. Teachers may need to undergo additional training to generate high-quality queries, which increases the complexity of using these tools in everyday practice[6, 8].

Thus, the limitations of LLM in generating tasks are related to both their technical characteristics and pedagogical aspects. Nevertheless, the development of new algorithms and the improvement of models open up opportunities to overcome these challenges in the future.

3. Objective.

General-purpose LMMs were chosen for the study: GPT-4 (OpenAI - chatgpt.com); Claude (Anthropic - claude.ai); Copilot (Microsoft - copilot.cloud.microsoft) and Gemini (Google DeepMind - gemini.google.com), which are available in free versions. These models were chosen due to their availability, popularity, and ability to generate different types of tasks (e.g., multiple choice questions, open-ended questions, matching tasks). This decision allowed us to focus on the possibilities of using models that do not require additional hardware or payment costs, which is an essential factor for the widespread introduction of such tools in educational institutions. The use of available models allows us to evaluate their potential for automating the creation of test tasks without the need for significant investment in resources.

4. Methodology

4.1 Research models and tools

General-purpose LMMs in their free versions were chosen for the study: GPT-4 (OpenAI - chatgpt.com); Claude (Anthropic - claude.ai); Copilot (Microsoft - copilot.cloud.microsoft) and Gemini (Google DeepMind - gemini.google.com). These models were chosen for their availability, popularity, and ability to generate different types of tasks. This allowed us to focus on analysing the possibilities of using the models and assessing their potential for automated test task creation without the need for monthly fees or investments in additional hardware or other resources, which is an essential factor for the widespread introduction of such tools in the educational process of educational institutions.

4.2 Selecting learning content for test task generation

To generate test tasks, a fragment of lecture material in Ukrainian on the discipline "Technology of creating software products" intended for 2nd year students of the speciality "Computer Science" was used. The text document had the following parameters:

- Characters without spaces: 9710;
- Word count: 1523;
- Format: PDF, size 361 KB.

The content covered the basic concepts of UML use case diagrams, which provides sufficient depth to test the models' ability to generate questions at different cognitive levels.

4.3. The procedure for creating and optimising queries

To ensure the relevance and quality of the tasks received, we used a query optimisation algorithm (Figure 1). The following key steps were taken into account before sending the request:

- Objective: To receive 20 test questions that cover the entire content of the lecture and correspond to Bloom's taxonomy.
- Formation of a role for LLM: The models were set up as a "virtual teacher", able to explain the material and create questions based on the reading.
- Checking for clarity and consistency: The request was checked for logical errors and ambiguities before being sent.

The procedure for creating and optimising queries was formed on the basis of the algorithm (Fig. 1) to obtain relevant test tasks for the downloaded lecture fragment. An example of an optimised query is shown in Figure 2.

Figure 2: Optimised query for test case generation

4.4. Assessing the quality of generated tasks

Several criteria were used to assess the quality of the tasks:

1. Compliance with Bloom's Taxonomy: We assessed whether the questions corresponded to different cognitive levels, from memorisation to creation.
2. Structure and clarity: The clarity of wording and the presence of explanations for correct and incorrect answers were analysed.
3. Variety of task types: The ability of the models to generate questions of different formats (multiple choice, open-ended, etc.) was compared.
4. Completion time: It was taken into account how quickly students could complete the test within the given time.
5. Validity and discriminative power: Expert analysis was used to check the extent to which the tasks meet the educational objectives and discriminate between students with different levels of knowledge.

4.5. Collecting and comparing data

For each model, 20 test questions were generated based on the same training content. The generated questions were compared by the following parameters:

- Average length of the question and justification (in characters);
- The median number of words in a question;
- The level of difficulty of the questions according to Bloom's taxonomy.

4.6. Analysis of reliability and practical limitations

After the tasks were generated, an expert analysis was conducted to ensure that they met the learning objectives and were clear to students. The usability of each model in the educational process was also taken into account. Particular attention was paid to technical limitations, such as the amount of text to be processed and the ability to integrate models into existing testing platforms.

5. Data and analysis

5.1. Comparative analysis of model capabilities

To compare the capabilities of the GPT-4 (chatgpt.com), Claude, Copilot and Gemini models, 20 test tasks were generated on the basis of the same training content. The models were evaluated according to the following parameters: the number of aspects of the topic covered, compliance with Bloom's taxonomy, the variety of task formats, and the level of question complexity.

Table 1 summarises the ability of the models to process text and graphic information in the free mode. As you can see, only GPT-4 and Claude can process uploaded text files, while Copilot and Gemini require manual text input.

Table 1.
Types of data processed free of charge

Model	Text data	Image
GPT-4	+	+
Claude	+	+
Copilot	-	+
Gemini	-	+

5.2. Variety of test item formats

Table 2 shows a comparison of the types of test tasks that can generate models without specialised queries. Of the 10 possible formats, GPT-4 supports all of them, while Copilot and Gemini demonstrate limited functionality.

Table 2.
Types of tasks available for generation

Format of tasks	GPT-4	Claude	Copilot	Gemini
Multiple choices	+	+	+	+
Several correct answers	+	+	+	+
True/False	+	+	+	-
Open answer	+	-	+	+
The challenge of compliance	+	+	+	+
Sequence tasks	+	-	+	+
Filling in the blanks	+	+	+	-
Crosswords	+	-	-	-
Essay.	+	+	+	-
Graphic tasks	+	-	-	-

5.3. Assessment of the quality and complexity of tasks

A comparison of the quality of questions generated by GPT-4 and Claude revealed a difference in the level of difficulty and depth of topic coverage. GPT-4 demonstrated a tendency to create higher cognitive level questions, including analysis and synthesis, while Claude focuses on memorisation and comprehension. The following parameters were chosen to assess the quality of the generated tasks: the average length of the question and explanation were used as indicators of the structuredness and level of detail of the answers, which affects the clarity and completeness of the information provided to students; and the median value of the number of words in the question was chosen to assess the conciseness and clarity of the wording. The details of the selected parameters are presented in Table 3 below.

Table 3.
Statistical analysis of task characteristics

Parameter	GPT-4	Claude	Δ
Average question length (characters)	85	110	-25
Average length of justification (characters)	120	150	-30
Median number of words per question	15	18	-3

Figure 3 shows an example of a GPT-4-generated test question that demonstrates the structure of the question and the explanation for the answer.

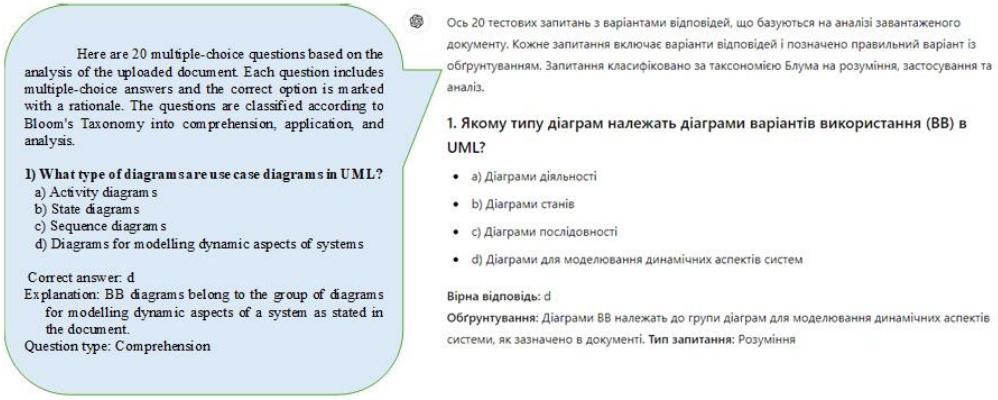

Figure 3: An example of generating a test case for a promt in GPT

Figure 4 shows an example of a test question created by Claude that illustrates a different approach to question generation and level of detail.

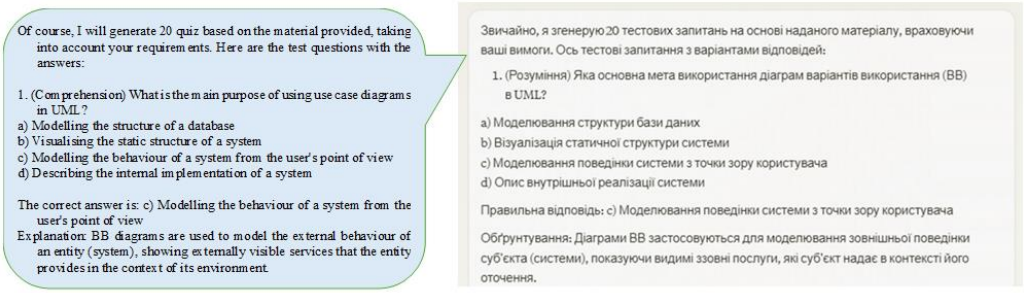

Figure 4: An example of generating a test case for a manufacturing task in Claude

5.4. Correspondence of tasks to Bloom's taxonomy

The analysis of the generated tasks showed that GPT-4 covers all levels of Bloom's Taxonomy, including the highest levels - evaluation and creation. Claude, on the other hand, mostly generates questions on the basic levels (knowledge and understanding). This shows that GPT-4 is more flexible in generating tasks for different learning contexts (Figure 5)

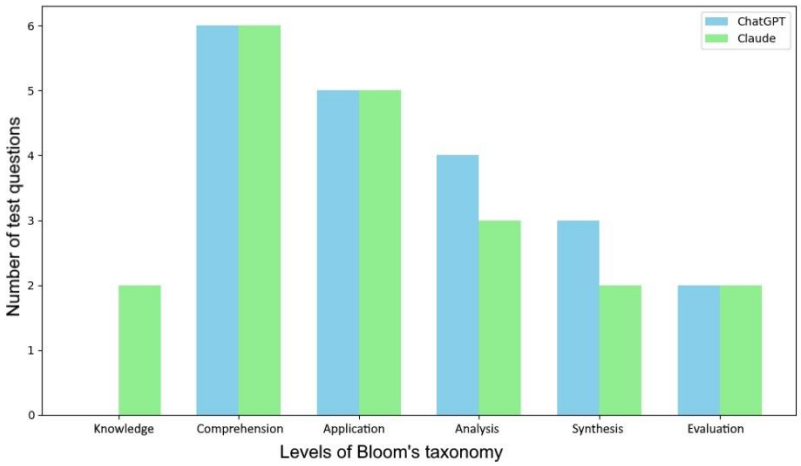

Figure 5: Graphs comparing question difficulty (a) and coverage of use case aspects for the generated GPT and Claude tests

5.5. Additional features of the models

GPT-4 provides more advanced functionality by allowing you to add justifications to your answers. Claude is less detailed in explaining correct and incorrect options, which may limit its effectiveness for training purposes that require in-depth feedback.

5.6. Analysing the validity of generated tasks

A comprehensive approach was used to assess the validity of the tasks, covering several key aspects: content validity, construct validity, clarity of wording, relevance of answer options, and absence of ambiguity.

Each criterion was assessed by experts on a five-point scale (1 - very low validity, 5 - very high validity). Leading academic staff with many years of teaching experience in computer science disciplines and practical experience in applying object-oriented analysis and design using UML were involved as experts in the quality assessment of the generated tasks. The scores for the GPT-4 and Claude models are presented in Table 4.

Table 4.
Assessing the validity of tasks generated by GPT-4 and Claude

Criterion	GPT-4	Claude	Δ
Content validity	4.5	4.5	0
Construct validity	4.0	4.5	+0.5
Clarity of wording	4.5	4.5	0
No ambiguity	4.0	4.5	+0.5
Relevance of answer options	4.5	4.5	0
Cognitive level (according to Bloom)	3.5	4.0	+0.5
Average score	4.17	4.42	+0.25

The analysis showed that the Claude model received higher scores for construct validity and lack of ambiguity, which indicates the clarity and relevance of the questions. At the same time, the GPT-4 demonstrated flexibility in generating tasks at different cognitive levels, although some of them may have minor ambiguities.

6. Conclusion

This study demonstrates the high potential of large-scale language models (LLMs) for automating the creation of test tasks in the educational process. However, the results also revealed a difference in the capabilities and effectiveness of different models. GPT-4 and Claude have shown high performance in generating tasks, but each of them has its own advantages and limitations that affect their application.

The GPT-4 has demonstrated the greatest flexibility in creating tasks of different formats and at different cognitive levels, according to Bloom's Taxonomy. Its ability to generate complex questions, including those requiring analysis and synthesis, makes this model suitable for use in curricula focused on the development of analytical thinking. At the same time, GPT-4 revealed some shortcomings related to possible ambiguities in the questions, which requires additional verification by teachers.

Claude, in turn, received the highest scores for construct validity and clarity of task wording. This indicates its effectiveness in creating questions of basic and medium difficulty. However, this model demonstrated a limited ability to formulate tasks of the highest cognitive levels (synthesis and evaluation), which may reduce its effectiveness for advanced courses.

Copilot and Gemini are less versatile than GPT-4 and Claude, in part because of the limited number of available task formats in the free mode. However, these models can be useful for highly specialised tasks, such as programming testing or visual element integration.

The study also revealed that the correct formulation of queries is an important factor in obtaining relevant answers from models. Teachers need to take into account both the limitations of the models (for example, the amount of text being processed) and the peculiarities of generating questions at different cognitive levels.

Thus, the use of LLMs to create test tasks is a promising area of educational technology development. It is important to note that the study used commonly used models in free versions that do not require specialised hardware. This demonstrates that automated test task creation can be affordable for educational institutions with limited resources, as well as for teachers who want to use modern technologies without additional costs. The choice of a particular model should be based on the purpose of the test and the level of complexity of the tasks.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1]. Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Wenhao Liu, and Caiming Xiong. (2022). Quiz Design Task: Helping Teachers Create Quizzes with Automated Question Generation. In *Findings of the Association for Computational Linguistics: NAACL 2022* (pp. 102-111). Seattle, United States: Association for Computational Linguistics. https://aclanthology.org/2022.findings-naacl.9/.

[2]. Kwan, C.C.L. (2024). Exploring ChatGPT-Generated Assessment Scripts of Probability and Engineering Statistics from Bloom's Taxonomy. In S.K.S. Cheung, F.L. Wang, N. Paoprasert, P. Charnsethikul, K.C. Li, K. Phusavat (Eds.), *Technology in Education. Innovative Practices for the New Normal. ICTE 2023. Communications in Computer and Information Science* (vol. 1974, pp. 275-286). Singapore: Springer. https://doi.org/10.1007/978-981-99-8255-4_24

[3]. Bharatha, A., Ojeh, N., Rabbi, A.M.F., Campbell, M.H., Krishnamurthy, K., Layne-Yarde, R.N., Kumar, A., Springer, D.C.R., Connell, K.L., & Majumder, M.A.A. (2024). Comparing the Performance of ChatGPT-4 and Medical Students on MCQs at Various Levels of Bloom's Taxonomy. *Advances in Medical Education and Practice*, 15, 393-400. https://doi.org/10.2147/AMEP.S457408

[4]. Herrmann-Werner, A., Festl-Wietek, T., Holderried, F., Herschbach, L., Griewatz, J., Masters, K., Zipfel, S., & Mahling, M. (2024). Assessing ChatGPT's Mastery of Bloom's Taxonomy Using Psychosomatic Medicine Exam Questions: A Mixed-Methods Study. *Journal of Medical Internet Research*, 26, e52113. https://doi.org/10.2196/52113

[5]. Aboalela, R.A. (2023). ChatGPT for generating questions and assessments based on accreditations. In *ACITY 13th International Conference on Advances in Computing and Information Technology* (pp. 1-12). https://arxiv.org/abs/2312.00047.

[6]. Agarwal, M., Goswami, A., & Sharma, P. (2023, September 29). Evaluating ChatGPT-3.5 and Claude-2 in Answering and Explaining Conceptual Medical Physiology Multiple-Choice Questions. *Cureus*, 15(9), e46222. https://doi.org/10.7759/cureus.46222

[7]. Brame, C. (2013). Writing good multiple choice test questions. Retrieved [today's date], from https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/

[8]. Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-334. https://doi.org/10.1207/S15324818AME1503_5

[9]. Amatriain, X. (2024). Prompt Design and Engineering: Introduction and Advanced Methods. ArXiv. https://arxiv.org/abs/2401.14423.

[10]. Tran, Andrew & Angelikas, Kenneth & Rama, Egi & Okechukwu, Chiku & Smith, David & Macneil, Stephen. (2023). Generating Multiple Choice Questions for Computing Courses Using Large Language Models. In *IEEE Frontiers in Education Conference* (p.1-8) https://bit.ly/3AE4YOc