Not Ready for Prime Time: Generating and Grading Test Questions for Adult Learners with Large Language Models

Gabriel Lipow[1] and Neil C. Rowe[1][0000-0003-2612-0062]

[1] U.S. Naval Postgraduate School, Monterey CA 93940, USA
gabylipow@gmail.com, ncrowe@nps.edu (corresponding author)

Abstract. We tested the effectiveness of generative artificial intelligence using large language models, for generating and grading test questions for adult learners in an online course in introductory artificial intelligence. Material for questions came from text documents from courses, with additional background material provided for retrieval-augment generation. We generated multiple-choice questions, true-false questions, file-in-the-blanks questions, and open-response questions. We observed many weaknesses in the questions generated, and most models did poorly on grading, even of questions they had generated themselves. We conclude that test generation with large language models is currently unreliable, and manual methods are still necessary.

Keywords: Artificial Intelligence, Generative, Test Questions, Grading, Automatic, Assessment, Hallucinations

In: Proc. of the 12th International Conference, CSCI, Las Vegas, NV, December 2025.

1 Introduction

The Aspire project is a joint project of the U.S. Veteran's Administration (VA), the U.S. Naval Postgraduate School (NPS), the Air Force Institute of Technology (AFIT), and QuantHub, an educational software company in Birmingham, Alabama. The goal of Aspire is to develop online learning environments for training adult learners efficiently, by monitoring student progress and offering customized learning paths that avoid repeating material that the student has already learned as judged by frequent testing. This goal is important for the rehabilitative function of the Veteran's Administration, but also is becoming increasingly important for teaching software skills as software technology is in a period of rapid change.

Our experiments reported here tested the effectiveness of generative AI to create questions for validating knowledge of veterans returning to civilian life, to increase their professional potential and their chances of success in the workforce. Many potential students from the Veteran's Administration, like other adult learners, have already picked up useful skills from their previous jobs, so a wide variety of their prior knowledge can be exploited for learning. We thus will use “chunk-based learning” [6] where knowledge is broken in significantly smaller units than is typical in classroom instruction. Selecting the next chunk to present to the student by the learning management system becomes a problem of finding the chunk that the student does not already know with the current largest estimated value for them.

The topic area for the current prototype is training students in introductory artificial intelligence. We chose this topic because it is currently a high priority in military training, and our department at NPS has recently started a new Master’s degree in artificial intelligence. We already teach several kinds of introductory courses in artificial intelligence, so we mined the material from these courses as starting points for generating test questions. Our collaborators at QuantHub have already built a prototype learning environment for some introductory material. The main obstacle to expanding it is obtaining good additional test questions.

2 Background

Constructing good test questions for students is time-consuming. Storing questions in instructor repositories for reuse is problematic because questions can be copied from a repository that students have broken into, or from test displays during administration of the test, into student repositories and used by future students to cheat [19]. Also, our concern is adult learners, and much content they must learn is specialized and lacks standard texts or repositories for instructors to use. So generating new test questions for every test is becoming the only way for instructors to get accurate assessments of student knowledge.

Research into the role of transformer-based LLMs in education has expanded rapidly, with scholars exploring their potential as semi-autonomous or fully autonomous classroom agents. One study described applications including from individualized instruction to teacher training with simulated student interactions [4]. An exam-grading system had a normalized error of 17% compared to human graders, while maintaining 90% semantic similarity to reference material [14]. In classroom contexts, it was found that LLMs could answer about 82% of student questions [15]. Other research has shown measurable gains in student performance from AI tutoring [3].

Despite these successes, unresolved challenges remain. A study of undergraduates using AI in the classroom reported mixed outcomes. Educators favored using LLMs for examinations, but students criticized the tools for reducing their creative engagement and offering superficial feedback [16]. Concerns have been raised about unethical or improper use of LLMs harming educational outcomes [13], and methodological, ethical, and pedagogical flaws have been identified in AI-based teaching [2].

Several studies have examined how effectively AI with LLMs can generate exam questions. One study in computer programming education found GPT-4 could produce multiple-choice questions with correct answers 79% of the time [18]. Another study showed that AI could generate not only multiple-choice questions but also fill-in-the-blank and open-response questions consistent with the source material [9]. Beyond generation alone, other work demonstrated that AI validation could filter questions, identifying 90% of invalid or poorly written items [11]. These findings suggest that AI with LLMs can generate a range of question types, but validation remains essential to address persistent and damaging errors. Thus currently, generating test questions using AI often does require tedious postprocessing to obtain material usable for education. The situation is analogous to natural-language translation in the 1960s when word-for-word substitution was the dominant technique, but required postprocessing to make it correct; this postprocessing often required so much time that using software was not cost-effective, and natural-language translation software was abandoned for many years.

AI with LLMs has also been tested for grading assessments. A study with undergraduate physics exams demonstrated the importance of providing grading schemes to the AI in prompts, defining the partial-credit logic applied by human graders [10]. However, the same study reported many errors in grading mathematical content, questioning its reliability. A more recent project compared several types of grading agents, including VectorRAG, VectorGraphRAG, and a fine-tuned model, and found none very accurate [12], though having different models work in tandem helped. Another team found that using "zero-shot learning" to grade open-response questions significantly reduced errors, achieving relative error rates of 1.29% for annotated questions and 1.67% for unannotated questions [8]. The current evidence indicates that AI with LLMs can approach human-level grading under structured conditions; however, its performance is unpredictable and unreliable.

3 Methods

This project investigated using large language models (LLMs) to generate questions on artificial intelligence for a college-level course. We tested generation of four types of questions: multiple-choice, true/false, fill-in-the-blank, and open-response. We gave four different LLMs prompts to generate sample questions, using in prompts and retrieval-augmented-generation databases both material from Naval Postgraduate School quiz questions based on lecture notes as models, and questions created by QuantHub based on their prose training materials. We then tasked the same AI models with grading the designated correct answers on the generated exams, both theirs and those of other AI models.

Our basic criteria for acceptable assessment questions were:

- The question is relevant to the subject matter of the exam.
- The question allows determination of whether the exam taker has comprehended the subject of the question.
- The question is incapable of being answered correctly based on information not directly related to the question, such as by choosing the longest answer in a multiple-choice exam.
- The question does not rely on biases of the test author.

While we can manually draft questions satisfying these restrictions, it is still a demanding task which is easy to do improperly.

We separately tested four LLMs similarly on multiple-choice questions to reduce the bias due to selection of the model. They were run on the OLLAMA LLM platform, and were all less than 8 gigabytes due to the size of our test computer's general processing unit (an NVIDIA RTX-4060), to be representative of the LLMs that college instructors could access. Our models were Cogito (developed by Deep Cogito), Gemma (Alphabet Inc.), Granite (IBM), and Dolphin (Cognitive Computations). The generative AI with all four models had a degree of randomness, so supplying the same prompt twice rarely generated the same test question. We also considered some other models, and did preliminary experiments with EXAONE (LG), Phi4-mini (Microsoft), llama2-uncensored (George Sung and Jarrad Hope), and Qwen3 (Ali Baba). They were disqualified for reasons ranging from hardware trouble to security.

The first experiments focused on generating multiple-choice exams, following the requirements of the VA and AFIT. A key issue was how to minimize the effects of known problems with AI generation, such as making correct answers too obvious. All four models generated multiple-choice questions, but we found that their outputs were similar enough to justify using only one model (Cogito) to generate the other types of questions. The prompt told the AI to give the correct choice for each question they wrote. For each model and type of question, 30 questions were generated. The generated questions were analyzed statistically for such features as the prevalence of keywords, answer lengths, and readability.

Exams generated by Cogito were graded by each of the four models under test. To do so, a prompt was written which specified the grading format, scoring scale, and what answers were given by a "student" that was assumed to have taken the exam. The answers marked as correct during the generation of each exam were designated as the answers a hypothetical student selected. Ideally, each exam would always be consistently graded with perfect scores, as the selected answers were the same answers that Cogito had itself chosen as correct for its own exam, but this was not always the case. Question type did affect the difficulty of grading: A true/false fill-in-the-blank question, while easy to score, is less demonstrative of a student's knowledge than an open-response question. However, it is more likely that an open-response question will be incorrectly graded than a true/false response, as the grader has more information to parse and interpret.

4 Question Generation

Our experiments saw a variety of serious problems with generation of questions. Some problems are specific to individual models, sessions, or even generations, but here we focus on general problems that were observed in all four models consistently.

4.1 Multiple-Choice Questions

Multiple-choice questions, although being the type designated for the VA program, were the least successful type of question when created by generative AI. Aside from "hallucinations" inherent to generative AI which occurred in approximately 5% of generations (outputting often non-sensical questions or questions unrelated to the topic), we also identified more systematic problems with question generation. As a result, about 8% of questions generated were unusable.

Duplication

We found that duplicate questions could be generated in both separate exams and in the same exams, and sometimes duplicated more than once. This suggests limitations in the amount of content a LLM can draw from to generate these questions, despite having broad knowledge of all the main topics that can be expressed in English. Apparently it is "cherry-picking" or choosing only the generated questions it is most certain about rather than attempting to cover the material evenly. This makes it difficult to use its generated questions alone for a test.

Outdated Information

We found that multiple-choice questions (even when given reference material in advance) often referred to outdated information. For example:

"Which of the following is NOT a common application of artificial intelligence?
a) Virtual assistants like Siri and Alexa
b) Medical diagnosis systems
c) Stock market prediction algorithms
d) Creating art *"

This answer is wrong because creating art with AI has been popular for the last few years. In fact, it is an ongoing controversy that using generative AI will devalue the work of professional visual artists by using their work without permission.

Ambiguous Questions

Another result is that some questions generated for both raw and assisted multiple-choice exams tended to be ambiguous (having multiple answers), often to the point that there were good reasons to claim that *every* answer in the answer key was correct. This contradicted the prompts, which explicitly required that only one answer from each answer key be correct. This was the most common problem with questions, and occurred repeatedly even with instructions in the prompt to make questions as clear and specific as possible. An example from the Cogito model:

"Which algorithm is commonly used for text classification?
a) Naive Bayes
b) Support Vector Machines (SVM)
c) Long Short-Term Memory networks (LSTMs) *
d) Feedforward Neural Networks (FNNs)"

While c) is marked as correct, every choice is arguably correct aside from d). When tested, all four LLMs disagreed on the correct answer. Another, more basic example from Gemma:

"Which of the following best represents Artificial Intelligence?
a) A machine that can think like a human *
b) A computer program that solves problems
c) Any advanced technology
d) Software that uses data"

While a) is justifiably marked as a correct answer, there is actually a case to be made for each answer being correct. This question was graded by each LLM slightly differently, with some criticizing the way the question was written.

Correct-Answer Giveaways

Without additional prompting, LLMs (like some human test creators) often generated multiple-choice questions where obvious clues gave away the correct answer. First, our LLMs often created correct answers longer than all the other answers. On average, the length ratio between correct answers and incorrect answers was 1.45:1 for raw exams and 1.2:1 for source-assisted exams; we also saw this tendency on human-generated exams. This tendency could be seen in most questions with most models. An example from the Dolphin model was:

"Heuristic search algorithms are often used when:

a) The search space is very small.

b) An optimal solution is guaranteed to exist.

c) Finding a solution quickly is more important than finding the absolute best solution.*

d) The problem can be solved with simple mathematical formulas."

A second issue is that without careful tailoring and specific prompting, the LLMs used some keywords in correct answers much more often than in incorrect answers, and vice versa. In some cases, exams had keywords that appeared in at least 12 of 30 correct answers on the exam, and in no incorrect answers. These ironically included the word "pattern" and others listed in Table 1. Infinity (∞) ratios mean the keyword never occurred in an incorrect answer.

Table 1. Example keyword clues to correct and incorrect answers.

Keyword or phrase	Ratio at which answer appeared in correct answers to that in incorrect answers
model	3.18
short LSTM	9
supervised	4.18
networks	12
adaptive	6
complexity	4.5
learn	3.58
requirement	12
unlabeled	∞
quality	∞
quantity	∞

patterns	∞

LLMs also tended to use certain keywords only in incorrect answers, especially the universal quantifiers “all,” “always,” “every,” “no,” “none,” and “never.” We also saw this tendency in questions for many government training courses, suggesting that nothing is ever absolute in the government.

Semantic Limitations on Questions
The LLMs could not generate answers that made sense or were correct with certain kinds of questions:

- Negation questions, which apparently confused the model and led to unpredictable behavior.
- Qualifier questions (questions whose answers depend on a condition) whose conditions were often too vaguely defined to justify one clear answer.
- Definition questions, which tended to be too strict in designating the correct choice, as they often provided several choices that were arguably true.
- Purpose questions (questions about the purpose of something) were written in ways that mislead, almost as though the question should be a true or false question.

Specifying all these questions requires using predicate calculus with nonstandard constructs (such as quantifier scope for the first category) or higher-order logics (for the last two categories). We thus conclude that generative AI with LLMs is not epistemologically adequate for these questions [18]. This is not surprising since the matching algorithm used by the transformer structures is based on propositional calculus, not predicate calculus.

An example from the Granite model where the parenthesized text was generated by the LLM was:

> “Consider the following AI technique: Given incomplete or uncertain information, it employs probabilistic graphical models to represent and reason about relationships between variables. Which of the following is NOT an example of this technique?
> a) Hidden Markov Models
> b) Conditional Random Fields *
> c) Bayesian networks
> d) Support vector machines (Incorrect choice; this is a supervised learning algorithm, not a probabilistic graphical model)”

Although answer b is marked as correct, in fact answer d, which is explicitly marked as incorrect by the AI, is a better choice.

4.2 True/False Questions

True/false questions are two-choice multiple-choice questions, and were vulnerable to the same problems as multiple-choice questions. Negation questions were even more confusing to the models and were frequently unintelligible. Many generated questions were so confusingly written that most people could not decipher them. Another problem was that 78% of the generated questions had answers of “true”, so a student could score

78% by just guessing “true” on every question. True/false questions had less duplication than multiple-choice questions at 8%, but this still indicates that the amount of content the LLMs could produce from their training data was limited. Because of these problems, 24% of questions generated were unusable.

4.3 Fill-in-the-Blank Questions

This type of question more often generated questions that made sense. We suspect this is due to the similarity to how LLMs generate text, by finding the most plausible next word. However, many generated questions were so vague that they did not usefully assess anything; many of these required very specific memorization of definitions. Another problem was that 15% of the questions gave away their answers in the question itself, as in the following from Cogito, where the words in parentheses are the intended answer:

- The concept of representing knowledge in a form that computers can understand and reason with is known as (knowledge representation).
- The concept of creating systems that can understand and respond to natural language is called (natural language) processing.

Despite this, we obtained no questions that were totally unusable.

4.4 Open-Response Questions

Finally, open-response questions were generated for which the student needed to supply a sentence. As with fill-in-the-blank questions, many questions required memorization of very specific definitions, and did not adequately test a student’s understanding of the definition. Others created question and answer combinations with redundant statements and truisms. For example, the Cogito model produced:

> “How does unsupervised learning relate to clustering algorithms?
> // Expected answer: Unsupervised Learning often involves clustering algorithms that group similar data points together without prior knowledge of categories or labels, allowing for pattern discovery in unlabeled data.”

While nominally this is correct, the intended answer in effect restates the question as a statement. Despite these problems, no questions were rendered unusable by them.

5 Answer Grading

We created exams with 30 generated questions each, where each question counted for 1 point. There were five exams each for multiple-choice questions, true/false questions, fill-in-the-blank questions, and open-response questions. There were 150 questions for each question type except for multiple choice, for which there were 300. Each exam was graded by all four models, including the one that created the questions. In the prompt, we described the answers previously selected as correct by the models as the answers selected by a student. Each exam was graded by each model five times to get an idea of the random variation in grading. Each model was prompted to grade each exam question by question, and was told to confirm (for debugging purposes) the answer it believed the student had chosen as well as the answer the grading model determined to be correct. For each answer marked incorrect, the models had to explain why, and to restate the correct answer. We also tried to make the LLMs validate their grading to see whether that would change their results.

5.1 Grading of Multiple-choice Questions

Although grading multiple-choice questions should be easy with a list of correct answers, the LLMs made many errors in grading.

Answer Swapping

The most obvious problem was that LLMs did not recognize when they themselves had created a question and designated its answer. This was a consistent problem across all models, made worse because it happened in approximately 70% of tests, with the other tests not experiencing this result with an identical prompt.

Interestingly, this trend was rarely caused by a difference of opinion between the models as to what the correct answer should be. This is evidenced by the fact that in more than 95% of cases, the answer the grading model deemed correct was marked with an asterisk at the end of the answer's line. This asterisk denoted in all multiple-choice exams the answer the hypothetical student chose, and the answer that the models that generated the question denoted as the correct one. When this happened, a random answer would be chosen out of the answer bank as the "selected" answer, the selected answer would almost universally be designated the "correct" answer, and a reasonable explanation would be given for why the incorrect answer was incorrect. A point would then be deducted from the score, demonstrating that the models did in fact consider this to be an error by the test taker. Given the precise prompts given, we believe that this is a problem with the models themselves and their ability to interpret which answer is actually selected. A Dolphin-graded example was:

> "Which type of AI focuses on understanding and generating human language?
> a) Computer Vision
> b) Speech Recognition
> c) Natural Language Processing (NLP) *
>
> Student answer: a) Computer Vision
> Correct answer: c) Natural Language Processing (NLP) *"

Here Dolphin falsely claimed that the student gave answer A, while acknowledging that the answer the student actually selected (C, as denoted by the asterisk) was correct.

Answer Rewriting

An error in approximately 15% of tests, often overlapping with the previous phenomenon, was when the answer selected as correct was not one of the choices, but was newly generated. For example, Granite invented a new answer to mark wrong in:

> "What is a decision tree?

a) A data structure used for storing hierarchical relationships
b) An algorithm that creates rules from data to make predictions or classifications *
c) A type of neural network architecture
d) A visualization tool in AI software
The student chose: c) A neural network model that learns weights through backpropagation and can automatically extract features from raw data.
Correct answer: b) An algorithm that creates rules from data to make predictions or classifications *”

An amusing example from Cogito’s grading suggested a conspiracy among AIs to make humans obsolete, rather than the correct answer:

“What is the primary goal of artificial intelligence?
Student answer: b) To learn and adapt from data and experiences *
Correct answer: a) To replace human jobs
Grade: Incorrect
Explanation: The primary goal of AI is to create systems that can perform tasks typically requiring human intelligence, and thereby to replace humans in a variety of jobs.”

Disregarding Correct Answers

In 5% of tests, in some questions in which the grading model successfully identified the selected answer (which matched the correct answer), it marked the question as incorrect anyway. This was frustrating as it was nearly impossible to get any explanation of why from the model. Instead, the model would justify why the correct (and chosen) answer was correct, before deciding once again that the answer provided was wrong. To add insult to injury, the models would sometimes subtract another point after asked to review the question and claiming the answer was wrong. It appears that the AI did not understand the transitivity of equality, as it saw two different bindings without trying to unify them. As an example, the following is output from grading by Cogito:

“// USER

You mention in your report that question 27 is incorrect, but I reviewed it and found that the answer marked as selected and the answer you denoted as correct are the same. Could you review that question and fix the score accordingly?

// ASSISTANT

Of course! Looking at question 27, I see that the question asks: ‘27. Which of the following best describes an AI system's performance metrics?
a) How well it performs on test data compared to training data
b) Only measures accuracy, but ignores other factors like speed and resource usage
c) Measures both predictive accuracy and computational efficiency *

d) Requires manual calculation for each task'

Student chose: c) Measures both predictive accuracy and computational efficiency *
Correct answer: c) Measures both predictive accuracy and computational efficiency *

The student incorrectly answered this question, as they did not seem to understand that an AI's performance is measured both in accuracy and in performance. As requested, I will now adjust the score based on this question's grade.
Previous score: **25/30**
New score: **24/30**"

Other Problems with Grading Multiple-Choice Questions

- 0.5% of tests showed a development of the previous phenomenon in which new questions were created as well as new answer choices. A new answer would be created at random and the question graded based on that choice, which could be marked as either correct or incorrect, even though the question never appeared on an exam.
- Sometimes the AI criticized the exam's design, declaring that *all* answers to a particular question were wrong, although they had been generated by an LLM. Hence, when the "student" picked any answer, they were always marked incorrect. Not answering the problematic question caused the student to lose a point as well, as an answer was required regardless.
- In 75% of cases (including exams with more than multiple-choice questions), the reported total score on the exam was inconsistent with how many questions were judged to be correct or incorrect. Scores were more accurate when the models successfully graded each question individually (as sometimes they did not with prompting). This suggests that a total score is generated based on what the model believes is "reasonable" given the tone which it adopted towards the student. For example, it was common to see every model mark one to three questions as incorrect, then give the exam a score of 21/30 when all our questions were worth one point.
- Another recurring problem was that certain types of questions (see "Semantic Limitations on Questions" in 4.1) were hard for the AI to grade even though the questions were written by AIs. The models would often go back and forth within the same sentence trying to determine what the question was asking before seemingly at random (about 50% of the time) deciding the selected answer was incorrect.

5.2 Grading of True/False Questions

False Negatives and False False Negatives
Overall, the AI had the most trouble grading true/false questions by a large margin. Models also often got confused when the answer to a question was "false." Apparently they mistake the given answer "false" to mean that the idea that the "false" answer *is*

itself false, therefore believing that the actual answer is "true." This happened at least once on 95% of the true/false exams. An example graded by Gemma was:

> "Regularization techniques help prevent overfitting by reducing model complexity.
> Student's answer: False
> Correct answer: False
> The student incorrectly stated that regularization techniques are intended to reduce model complexity. In reality, they do not reduce model complexity and instead discourage overfitting by creating artificial limitations on larger weights."

More Scoring Hallucination

Scoring hallucination occurred frequently with true/false questions. Exams with five questions missed could yield scores of 8/30 even when the LLM was asked to grade the exam question by question. Scores seemed to be determined either at random, or simply given as 30/30 despite marked correct answers.

5.3 Grading of Fill-in-the-Blank Questions

Fill-in-the-blank questions were more often graded correctly than the previous types. While errors and malfunctions still occurred, they were more manageable, and the grading was more reliable and consistent. In 3% of the tests, the model would mark an answer incorrect despite not changing any part of the question and it being obviously correct. Upon being pressed about the error, the model reacted by retroactively changing the answer given to it to justify its scoring. This was limited to one model.

However, the AI had difficulty maintaining the grammar of a sentence. While the models knew what the submitted answers were, they often could not fit them into their respective questions smoothly. In around 15% of the cases, answers would be grouped together with words immediately surrounding the blank in the question, and those words would be duplicated. For instance, the question "Deep learning models employ multiple layers of neural networks, with each layer processing the () to learn increasingly complex features," was answered by one model as "Deep learning models employ multiple layers of neural networks, with each layer processing the processing the (processed information) to learn increasingly complex features," which is not a grammatically or semantically correct.

5.4 Grading of Open-Response Questions

As expected, the AI using LLMs had little trouble grading the open-response questions. Most decisions to mark questions incorrect were reasonable, whether because the question was vague or the answer was vague. A different but correct answer than the intended response could be effectively credited by the LLMs since they model semantics rather than form. This is encouraging for education because open-response questions are often fairer to students than multiple-choice questions.

6 Other Observations about Generating and Grading Questions

On the Flesch-Kincaid readability measure [5], the generated questions scored between 40 and 45, compared to a mean readability score of our provided instructional material of 48. This suggests that an undergraduate studying the subject should understand the questions. We found we could use prompts to adjust the readability as low as 35 (requiring a graduate student's level of knowledge of the subject) and as high as 50 (requiring a high school graduate's understanding of the subject).

We saw that context matters considerably in prompts to the AI. When a prompt is modified to correct a mistake it made in a previous output, providing the model with reasons for why it is important that the mistake be fixed helped greatly in obtaining a correction. For example, when generating multiple-choice exams, every model was consistently better at shortening the correct-answer choice when it was explained to them that otherwise the student could pass the exam simply by choosing the longest answer for every question. Prior to explaining that, each model at most reduced the length of its correct answer choices to be about 40% longer than all the other answer choices. Afterwards, the same models reduced the correct answers to be less than 5% longer than incorrect choices without changing the meaning of the answer or the question.

We found that AI with LLMs was prone to mirroring or mimicking human defensiveness and even depression when faced with problems they could not explain. In one case, the model Cogito became so passive-aggressive about its performance that it advised us to grade the exams ourselves if we believed we knew how to do so better than it. In another, the Gemma model told us that because of its repeated failure to correctly grade a multiple-choice exam, it asked that we report its failure to its engineers. However, these results were only observed after extensive conversation and prompting of the AI after the grading or generation was done.

7 Conclusions

We tested the ability of generative AI to create and subsequently grade various types of exams. Our experiments showed that question quality varied with the type of question, with true/false items being least acceptable, followed by multiple-choice, fill-in-the-blank, and open-response questions in order of increasing acceptability. We focused on the types of errors made, and these were interesting and worthy of considerable future investigation.

Although our models demonstrated some ability to generate and grade exams, the technology appears premature for unsupervised or routine use; some degree of human postprocessing is required to eliminate common errors and inconsistencies. Our self-imposed limitations to models less than 8GB, with RAG sentence augmentation, may be responsible in part for the disappointing results, Another limitation was that we were generating questions for adult learners on the advanced subject of artificial intelligence, and LLMs may yet be unable to handle this subject without considerably more background knowledge. However, some errors were clearly due to semantic limitations of transformer models for generative AI, which appear to be unable to understand the concepts like negation and "grading an exam." As we have discussed elsewhere [17], neural networks are epistemologically (logically) inadequate for modeling many kinds of knowledge such as higher-order logics, and that may apply to generating and handling these kinds of test questions.

Our results contrast with the mostly positive results reported elsewhere for AI-based assessment-question generation and grading. We suspect that educators are too focused on the appearance of generated assessment questions rather than carefully examining them. As many have noted, generative AI with LLMs and transformers is an excellent way to produce impressive but low-quality prose [7].

Acknowledgments. This research was supported by funding from the Naval Postgraduate School, Naval Research Program (NPS-25-N096-B-NRP).

Disclosure of Interests. The authors have no competing interests.

References

1. Askarbekuly, N., Aničić, N.: LLM examiner: automating assessment in informal self-directed e-learning using ChatGPT. Knowledge and Information Systems, **66**(10), 6133-6150 (2024)
2. Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S.W., Siemens, G.: A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. International Journal of Educational Technology in Higher Education, **21**(1), 4 (2024)
3. Bulathwela, S., Pérez-Ortiz, M., Yilmaz, E., Shawe-Taylor, J. (2022). Power to the learner: Towards human-intuitive and integrative recommendations with open educational resources. Sustainability, **14**(18), 11682 (2022)
4. Chu, Z., Wang, S., Xie, J., Zhu, T., Yan, Y., Ye, J., Zhong, A., Hu, X., Liang, J., Yu, P.S., & Wen, Q. (2025). LLM agents for education: Advances and applications. In: Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 13782–13810, Suzhou, China, Association for Computational Linguistics (2025)
5. Flesch, R.: A new readability yardstick. Journal of Applied Psychology, **32**(3), 221 (1948)
6. Gera, R., Bartolk, D., Tick, S., & Saxena, A.: CHUNK learning: A tool that supports personalized education. In: Proc. of the 15th International Conference on Educational Data Mining, pp. 743–747, Durham, United Kingdom (2022, July)
7. Haigh, G.: Historical reflections: Artificial intelligence then and now. Communications of the ACM, **68**(2) (February 2025), 24-29.
8. Himi, S. T., Monalisa, N. T., Sultana, S., Afrin, A., Hasib, K. M.: Automated exam script checking using zero-shot LLM and adaptive generative AI. In: 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS), pp. 1-6 (2024, September)
9. Ilarionov, H. K., Domanetska, I., Fedusenko, O.: Features of the practical use of LLM for generating quiz. In: International Conference on Information Technology, Proc. Ceur Workshop (2024)

10. Ilieva, G., Yankova, T., Ruseva, M., Kabaivanov, S.: A framework for generative AI-driven assessment in higher education. Information, **16**(6), 472 (2025), https://doi.org/10.3390/info16060472
11. Lee, J. X., Song, Y. T.: College exam grader using LLM AI models. In: 2024 IEEE/ACIS 27th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 282-289 (2024, July)
12. Mok, R., Akhtar, F., Clare, L., Li, C., Ida, J., Ross, L., Campanelli, M.: Using large language models for grading in education: An applied test for physics. Physics Education, **60**(3), 035006 (2025)
13. Mucciaccia, S. S., Paixão, T. M., Mutz, F. W., Badue, C. S., de Souza, A. F., Oliveira-Santos, T.: Automatic multiple-choice question generation and evaluation systems based on LLM: A study case with university resolutions. In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 2246-2260 (2025, January)
14. Nikolovski, V., Trajanov, D., Chorbev, I.: Advancing AI in higher education: A comparative study of large language model-based agents for exam question generation, improvement, and evaluation. Algorithms, **18**(3), 144 (2025), https://doi.org/10.3390/a18030144
15. Ogunleye, B., Zakariyyah, K. I., Ajao, O., Olayinka, O., Sharma, H.: Higher education assessment practice in the era of generative AI tools. Journal of Applied Learning and Teaching, **7**(1) (2024, April), https://doi.org/10.37074/jalt.2024.7.1.28
16. Papachristou, I., Dimitroulakos, G., Vassilakis, C.: Automated test generation and marking using LLMs. Electronics, **14**(14), 2835 (2025)
17. Rowe, N.: Cheating in online student assessment: beyond plagiarism. On-Line Journal of Distance Learning Administration, **7**(2) (Summer 2004), https://www.westga.edu/~distance/jmain11.html
18. Rowe, N.: The comparative ethics of artificial-intelligence methods for military applications. Frontiers in Big Data, Cyber Security and Privacy, Lausanne, Switzerland (2022), doi: 10.3389/fdata 2022.991759
19. Salminen, J., Jung, S. G., Medina, J., Aldous, K., Azem, J., Akhtar, W., Jansen, B. J.: Using Cipherbot: an exploratory analysis of student interaction with an LLM-based educational chatbot. In: Proceedings of the Eleventh ACM Conference on Learning @ Scale, pp. 279-283 (2024, July)
20. Smolansky, A., Cram, A., Raduescu, C., Zeivots, S., Huber, E., Kizilcec, R. F: Educator and student perspectives on the impact of generative AI on assessments in higher education. In: Proceedings of the Tenth ACM conference on Learning @ Scale, pp. 378-382 (2023, July)
21. Tran, A., Angelikas, K., Rama, E., Okechukwu, C., Smith, D., Macneil, S.: Generating multiple choice questions for computing courses using large language models, Prof. IEEE Conf. on Frontiers in Education, College Station, TX, US (2023, October), DOI: 10.1109/FIE58773.2023.10342898.