

Changing Student Assignments and Tests in the Era of Generative Artificial Intelligence

Neil Rowe

Computer Science Department, NPS

ncrowe@nps.edu

Many previous homework questions are now obsolete.

- It's too easy for students to ask generative AI to create an answer in many cases.
- We should not eliminate homework at a university, and especially at a graduate school, since homework questions can ask students to put several concepts together, an important goal in education.
- Generative AI is also a problem for essays, project reports, and theses.
- It can also be a problem with tests:
 - Students could surreptitiously use electronic devices during a test; it is hard to see small devices.
 - We cannot see what students are doing during distance-learning tests.
 - There are high-tech ways to cheat on a test (see Rowe, Cheating in Online Student Assessment: Beyond Plagiarism, *On-Line Journal of Distance Learning Administration*, Summer 2004 – my #1 cited paper).

Some old questions on digital forensics that should not be used anymore.

Homework questions:

1. Why could thumb drives be a particularly rich source of data in a criminal or intelligence forensic investigation?
2. Why could the content on thumb drives be harder to analyze than computers or mobile devices in a forensic investigation?

Test questions:

3. In general, what is cosine similarity applied to when used for digital forensics?
4. According to the paper "AI-Assisted Warfare", why cannot we primarily blame AI software for targeting mistakes in AI-assisted warfare?

Essay questions are mostly obsolete

- Generative AI can write essays on most topics.
- This rules out many assessments of nontechnical essays at NPS.
- This makes impossible a MOVES Ph.D. policy suggestion, in lieu of a Written Qualifying Exam, to have all students write an essay about key concepts in modeling and simulation.
- Possible exceptions:
 - Essays on highly specialized information, such as analysis of a particular set of real data that has not been analyzed before.
 - Essays on raw unpublished experimental results, perhaps results under the control of the professor.
 - Essays on personal experiences.

Questions involving arithmetic are tougher for generative AI, but it can get some correct answers.

Homework question:

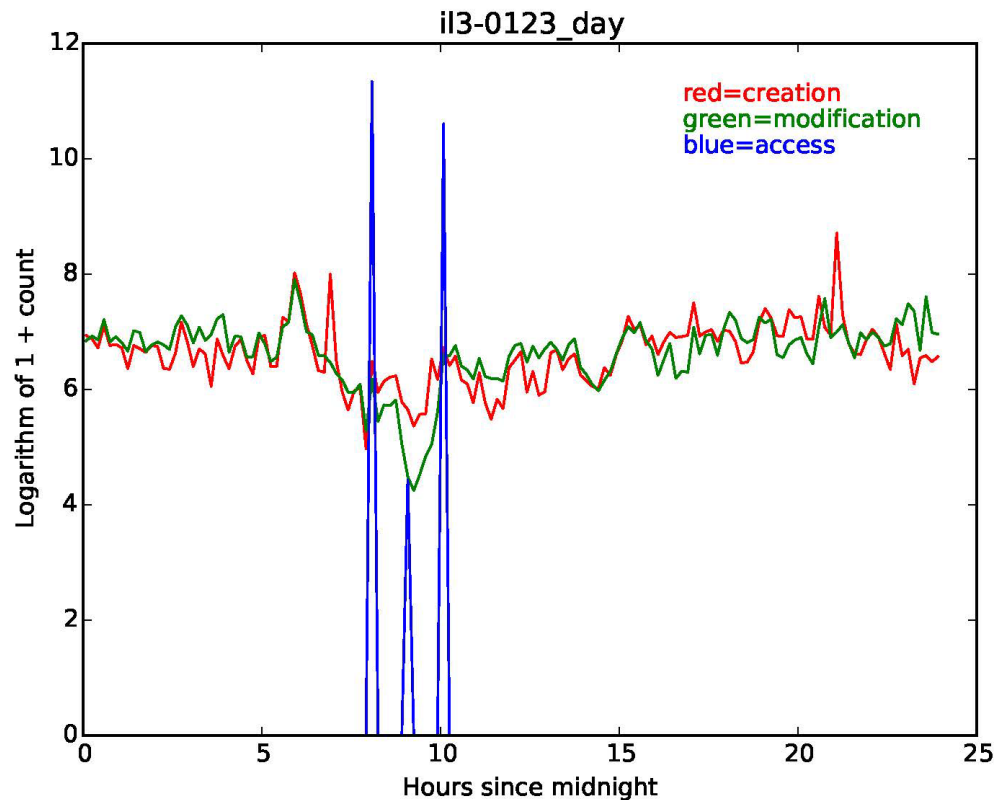
Suppose we have a cell phone with 500 gigabytes of non-volatile storage on its drive.

- (a) Suppose the average file on the drive is 20 kilobytes and 40% the drive is occupied by files. How many files are on this drive? Show your calculations.
- (b) Suppose we need to collect all the email addresses in the files of the drive. Suppose a product like SleuthKit can direct us to the part of the drive that is occupied so we only need to search in it. Suppose it takes 0.3 milliseconds to retrieve each 4096-byte block and search it for addresses. Assume the data given previously. How long will it take in hours? Show your calculations.

Prof. Kolsch tried this in genai.mil and got a correct answer, though without showing any calculations. But generative AI does err on arithmetic, as we discovered when it tried to total grades on exams.

Questions involving image processing are probably safe for now.

- **Homework question:**



Here is the daily activity on a drive. Based on this distribution, what kind of computer or device is this likely to be, and why?

A strategy for modern questions: Use details on which generative AI is unlikely to be trained.

- **Homework questions:**
 - 1. Analyze drive 925 from India.
 - 2. Analyze drive 2201 from Thailand.
 - 3. Analyze drive 2204 from Thailand.
 - 4. Analyze drive 2219 from Thailand.
 - For each of problems 1-4, answer the following questions as best you can for each drive's data, and give your reasoning. It may help to use a translator like Google Translate, and it may help to look up some things online. Your answers to each part of these problems should not exceed 50 words, and less is needed for some questions.
 - (a) When were the starting time and the ending time of the period of usage of the drive by its users to the nearest quarter of a year?
 - (b) What distinctive patterns were in the usage per week, day, or hour? Note that week plots start with all of Thursday, and year labels on the overall plot are centered on July 1.
 - (c) What did the distribution of sizes and inode numbers tell you about usage of the drive?
 - (d) Did you see anything distinctive in the distribution of file types? What did it tell you about the drive usage? Bear in mind that less-common types vary more in percentage due to the laws of statistics.
 - (e) Examine the distribution of words in the file names (the hist_counts file). What does this tell you about the usage of this drive? Did you see anything especially distinctive? Bear in mind that most of these words were in the software files.
 - (f) What major kinds of applications (software) were loaded on the drive, and which were unusual? Look in the big text file (subset_bigtable). It may help to look the unfamiliar names up in Google.

Word limits on short-answer questions can make it hard to use answers from generative AI, since it tends to be wordy to be safe and cover all possible interpretations.

- In the paper "Risk analysis and control of large artificial intelligence models", what was the difference between "value concept risk" and "cognitive bias risk"? (50 words or less)

(These two terms are mostly specific to the paper. It also helps to restrict discussion to one source of information, as by naming the paper here.)

Students can still try to give answers from generative AI to some questions, but often their answers are too general, and you can mark them down for not answering the question.

- **Homework question:**
- (e) Examine the distribution of words in the file names in the hist_counts file. What does this tell you about the usage of this drive? Did you see anything especially distinctive?
- Student answer: The words in file names often follow a Poisson distribution.
- Student answer: Named entities are distinctive.
- *Note both answers are generalities that do not answer the question, so they should not receive much credit.*

Detecting generative AI on theses and reports

- The quality of thesis drafts has significantly decreased due to generative AI.
- So even if you think generative AI helps with writing, you must address lower-quality writing.
- Look for inadequate cross-referencing:
 - Missing definitions of terms
 - Missing cross-references to earlier or later sections
 - Missing citations of references
 - Use of clichés
- These problems occur with generative AI using transformers (the most common way to do generative AI), because it can hold a limited amount of data in memory at one time.
- Ask the student to revise such text by looking through the text for missing connections and fixing them.

Generative AI often provides easy clues to the correct answer in questions it generates.

Test question we created with generative AI:

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.

Here the correct answer c is the longest choice. We saw this frequently with AI-generated questions, including many from the vendor QuantHub.

Generative AI also creates “giveaway” questions with the answer in the question.

Short-answer test question:

The concept of representing knowledge in a form that computers can understand and reason with is known as ?

(Correct answer according to generative AI was “knowledge representation”.)

Epistemological inadequacy of artificial neural networks with transformer algorithms and generative AI based on them

- They cannot explain what they did other than give a complicated formula for what they computed.
- That is, they cannot answer “why” questions well.
- That is, they are “black boxes” not “white boxes”.
- It may not matter in routine operational tasks like identifying tanks in satellite images.
- It does matter, though, when a wrong answer is produced during development. Then you must decide whether this error is due to data, to training, or to the algorithm.
- Artificial neural networks are an essential component of transformer algorithms for generative AI.

An example of epistemologic limitations in generative AI

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)

- This question was generated by AI from introductory AI texts.
- Four large language AI models were asked to grade this question and were told that answer “c” was correct.
- They still disagreed on the correct answer for this and similar questions.
 - Some of this is due to different training sets.
 - Some is due to not understanding the concept of “grading”.
- Note that the correct choice is the longest, a common weakness of AI-generated questions as well as human-generated ones.

Multiple-choice questions can be made harder for generative AI to answer, when students use it during a test

It may not have trained on all choices, something more common with technical disciplines.

Test question:

Of several reasons why pilots are decreasing in importance, the LEAST IMPORTANT of is:

- A. There is increasingly little for pilots to do.
- B. Pilots are expensive to train and their aircraft are expensive too.
- C. Autonomous aircraft better obey the Law of Armed Conflict. *
- D. Autonomous aircraft don't risk pilot lives.
- E. Autonomous aircraft could do more accurate targeting.

This question is also good because generative AI has trouble with logical quantifiers (“least” here) because it cannot do full logical reasoning.

Don't expect that generative AI can grade your tests and assignments.

Our experiments showed that generative AI had great difficulty grading multiple-choice tests, even when using the same large language model that generated the questions.

- Experimenter: 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?
- Generative AI: Of course! Looking at question 27, 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**

It also had great difficulty totaling scores, getting different numbers each time even though it had decided previously which answers were correct.

Another example of AI unable to grade questions

Similarly, we found many generative AI models used with transformers could not grade multiple-choice test questions with negatives like “not” in the question.

For instance, one AI model generated this question:

19. 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*

Grading with AI models not only disagreed on which answer was correct even though it was marked, but could not understand what choice the student picked (in this case, answer “b”). One thought the student picked answer “d” and said *“Incorrect choice; this is a supervised learning algorithm, not a probabilistic graphical model.”*

Abnormal psychology of generative AI

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.

- The model Cogito became so passive-aggressive about its performance once that it advised us to grade the exams ourselves if we believed we knew how to do so better than it.
- The Gemma model told us that because of its repeated failure to correctly grade a multiple-choice exam, we should report its failure to its engineers (a kind of suicide).

An apparent AI conspiracy against humans, shown by the Cogito model.

Test question:

What is the primary goal of artificial intelligence?

Cogito explanation of its grading:

- 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.

Why is this epistemological adequacy important?

- People expect other people have basic human knowledge. If AI lacks this knowledge:
 - The AI can err and cause harm.
 - People won't trust AI that cannot explain what it does or doesn't know.
- Prompt “engineering” doesn't engineer; it's just hit or miss with no science behind it. True engineering controls its results much better.
- Commentators have observed that generative AI using transformers is great at generating BS since it knows human methods for convincing people. This BS can mislead users and make them overrate AI output. See: G. Haigh, “Historical reflections: Artificial intelligence then and now”, *Communications of the ACM*, 68(2) (February 2025), pp. 24-29.

Can prompt engineering solve these problems?

- The standard advice when a text-generative AI system gives poor or inadequate answers is to experiment with the prompt: Give it more data about what you want.
- This often requires trial and error. For instance:
 - We noticed AI-generated correct choices to multiple-choice questions were 45% longer than incorrect answers with the Cogito LLM.
 - We modified the prompt to ask that correct choices were not longer. It then gave correct choices that were 20% longer than incorrect answers.
 - We rephrased the prompt to give a reason, namely that it is poor testing when a student can answer a question correctly without knowing the material. We then got correct answers that were 4% longer than incorrect answers.
- Trial and error is not the way science and engineering usually advance, despite occasional successes like Edison's light-bulb filament and Ehrlich's syphilis drug.

Another problem: AI reasoning may be limited

- Generative AI tries to be helpful, but it lacks priorities of what is important.
 - The prompt can help, but has a maximum size.
 - Retrieval-augmented generation (attaching a database) can help generative AI obtain specialized facts not in its training, but cannot add general reasoning capabilities.
- Generative AI is also limited by the maximum size of the context (subset of the LLM loaded into memory) that the transformer algorithm uses. It can guess what context the user needs and augment the retrieval, but is limited in how much information it can hold at any one time.
 - So generative AI cannot adequately cross-reference documents.
 - So generative AI cannot adequately know when to provide definitions of terms.

Another problem: Generative AI + transformers is unlikely to improve much

- Since the obstacles to improvement are epistemological and fundamental, they cannot be fixed by improvement in materials as with construction technology or quality control.
- The situation is unlike, say, civil engineering where improvements in material and procedures are continuous.
- Vendor hype makes all kinds of unjustified claims for the current AI technology with inadequate testing.
- Nonetheless, AI will improve eventually.
- What is most needed is better incorporation of logical reasoning in neural networks, since AI history shows the value of logical reasoning.
- This requires fundamentally different technology than currently, much like how semiconductors have reached their limits and we need quantum computers for hardware improvements.

Ethics of generative AI's lower-quality products

- A general concern about text generation by generative AI is that it creates lower-quality text than human-written text, since the techniques are epistemologically inadequate. But it can save money for businesses.
- This is similar to the lowering of quality in the food industry: Money can be made by substituting cheap ingredients for more-expensive and more-nutritious ingredients. Example: increasing the amount of sugar in a product to compensate for flavor.
- This is also similar to replacement by plastics of wood and metal in many consumer products.
- This raises ethical issues, since higher-quality products have advantages due to their materials. A plastic shovel breaks more easily than a metal shovel.
- Lower-quality products are also occurring with entertainment.
- There is hope, however: High-quality products are still appreciated by people who can afford them. Ignoring price, people dislike plastic.

Military contractors and vendors love generative AI + transformers

- It's a way to avoid requirements.
- Since generative AI with transformers makes random decisions and a result is rarely duplicated on additional runs, testing software using it is very hard.
- This circumvents acquisition procedures and requirements: You can't write requirements for generative AI (beyond "create something matching the prompt"), and you can't test it.
- This alone should make us skeptical of generative AI using transformers.

Generative AI may encourage cheating

- Two recent examples from my classes suggest that students are getting bolder about cheating, as a secondary effect of generative AI:
 - A student handed in my answer sheet on a homework assignment. They should have reported they got access to the answer sheet.
 - A student handed in solutions to the homework that were exactly the same as another student, including the other student's name at the top.

More information about our experiments with generating and grading test questions:

G. Lipow and N. C. Rowe, Not ready for prime time: Generating and grading test questions for adult learners with large language models. In *Proc. of the 12th International Conference, CSCI*, Las Vegas, NV, US, December 2025. In: *Springer Lecture Notes in Computer Science*.

Other papers: <https://faculty.nps.edu/ncrowe>

*In general, AI should be carefully tested just like any software.
Because the AI output “looks good” is not a test.*