Undergraduate Pacific Studies Exam Generation and Answering Using Retrieval Augmented Generation and Large Language Models

Colleen Gayheart
Air Force Institute of Technology
colleen.gayheart@us.af.mil

Alexandre Some
Air Force Institute of Technology
alexandre.some@us.af.mil

Erick Tyndall
Air Force Institute of Technology
erick.tyndall@us.af.mil

Joseph Genz
University of Hawai‘i - Hilo
jgenz@hawaii.edu

Torrey Wagner
Air Force Institute of Technology
torrey.wagner.2@us.af.mil

Brent Langhals
Air Force Institute of Technology
brent.langhals@us.af.mil

Abstract

The capabilities of large language models have increased to the point where entire textbooks can be queried using retrieval-augmented generation (RAG). The study evaluates the ability of OpenAI's ChatGPT-3.5-Turbo and ChatGPT-4-Turbo models to create and answer exam questions based on an undergraduate textbook. 14 exams were created with true-false, multiple-choice, and short-answer questions from a textbook available online. The accuracy of the models in answering these questions is assessed both with and without access to the source material. Performance was evaluated using text-similarity metrics including ROUGE-1, cosine similarity, and word embeddings. 56 exam scores were analyzed to find that RAG-assisted models outperformed those without access to the textbook, and that ChatGPT-4-Turbo was more accurate than ChatGPT-3.5-Turbo on nearly all exams. The findings demonstrate the potential of generative artificial intelligence tools in academic assessments and provide insights into comparative performance of these models.

Keywords: Generative Artificial Intelligence, Retrieval Augmented Generation, Large Language Models, Academic Examinations

1. Introduction

AI is becoming an increasingly prominent topic with ongoing growth in multiple domains. One category of AI is large language models (LLMs), which includes OpenAI's versions of Generative Pre-training Transformer (GPT) models. The application in this work uses GPT LLMs to generate and answer exam questions from an undergraduate textbook.

1.1. Overview

AI has continually become a popular and controversial topic in the context of assisting with or replacing everyday human tasks. Although it has existed for decades, it has not been accessible to the general population for consumer use. This all changed with the creation of OpenAI, whose goal is to be "an organization focused on developing artificial generative intelligence to benefit humanity" (Ray 2023). OpenAI released models such as GPT, GPT-2, and GPT-3, before finally releasing the transformative ChatGPT. ChatGPT was optimized for conversation-based tasks, including contextual understanding and coherence (Ray 2023). The public can access the models for free, with a paid subscription available that offers more advanced models and faster response times.

The academic domain is one space where ChatGPT use has been especially prevalent. Initially, students used the software to generate answers or even papers, leading to ethical concerns about cheating if such practices were prohibited by educational institutions. However, the GPT space is still relatively unexplored from the instructor side such as using a specific resource (i.e. textbook) to generate an exam, create matching solution sets and then answering AI developed exams. One example in the literature reviewed was the creation of multiple choice exams for medical students. According to the study, "GPT-4 can be used as an adjunctive tool in creating multi-choice question medical examinations yet rigorous inspection by specialist physicians remains pivotal." (Klang 2023).

Although the creation of LLMs has created a beneficial effect on the scope of generative AI, these models by themselves often suffer from

URI: https://hdl.handle.net/10125/109033
978-0-9981331-8-8
(CC BY-NC-ND 4.0)

hallucinations, where the AI generates information that seems plausible but is actually fabricated or inaccurate. This can be based on training data limitations, misunderstanding the context of the prompt or a model architecture that prioritizes coherence over accuracy (Turing, 2023). RAG was created to assist with this problem (AWS, n.d.). This process is often effective for small datasets or niche areas of information that an organization may use. For example, although an LLM might have some general knowledge in medical technology, the diverse amount of data it's trained on would not allow it to relate accurate answers about specific medical technology. RAG systems "effectively reduces the problem of generating factually incorrect content" (Gao 2024). This is because an organization can tailor an LLM by including relevant files as a basis for the model to answer questions. As the accessible knowledge base grows, efficiently retrieving relevant information from documents becomes crucial (Gao 2024). This technology is essential in the academic landscape, where educational topics are often specific to key areas or narrow domains that generalized LLMs typically lack without RAG capabilities.

1.2. Objectives & scope

The objective of this paper is to study the variations within different forms of ChatGPT by creating exam questions from a specified resource and measure their accuracy in answering those questions. The work creates a diverse set of exam questions using ChatGPT on a less rigorous scale and explores the methods it would take to allow GPT systems to create and answer the questions. Furthermore, the paper will analyze the differences between ChatGPT versions to determine which version creates the best exams and provides the most accurate answers. This will be done using a range of metrics later discussed in the Methodology section. The primary prompts used can be found in the appendices to allow replication of this work.

A limitation of this study is that while there are thousands of LLMs, only two variations of ChatGPT will be utilized. These models were selected as ChatGPT is an option easily accessible to educators. A second boundary of this study is the resource that will be tested will be a single undergraduate textbook.

2. Background

There are multiple types of LLMs including cloud-based and local based, each with differing advantages and disadvantages. LLMs have various applications and capabilities that can be enhanced with methods like RAG. LLMs have rapidly grown in use and availability in recent years and with this recent growth, there are expanding concerns for ethical considerations within the academic domain.

2.1. Types of large language models

LLMs are typically either local-based or cloud-based. Cloud LLMs such as ChatGPT are accessed via the internet, with a web interface or an application programming interface (API), as used in this work. These cloud-based models free users from the responsibility of managing and updating the required infrastructure, which can be quite extensive. They also reduce the initial costs related to purchasing hardware and software, allowing users to access the model as needed (Dilmegani, 2024). However, with this reduced burden on the user there are increased risks of security due to the cloud environment. This includes the opportunity for data breaches and illegitimate access to data (Dilmegani, 2024). Cloud LLMs are easily scalable and are advantageous when training requires extensive computing resources such as multiple instances of high-end GPUs and large amounts of data storage (Awan, 2023). If the user does not have the necessary hardware, the time to train or use a model can be immense without the help from cloud computing. While cloud LLMs can have lower startup costs they suffer from higher total costs reflected in subscription prices or pay-as-you-go plans (Dilmegani, 2024).

Local LLMs, such as BERT or T5 are run on individual devices or servers without needing a connection to the internet or cloud services. This provides users with more control over the LLM and its environment, and reduces security concerns. However, this requires greater familiarity with the technology and maintenance support (Dilmegani, 2024). Since the information input into the LLM is local, and not internet based, it is more inherently more secure difficult to access without physical access to the computer. Depending on the user's needs, the cost of hardware can be substantial.

LLMs can be applied to a wide variety of tasks, many of which are depicted in Figure 1. The x-axis represents numerous applications, including text generation and text classification. The figure shows the number of artifacts (such as models) hosted on huggingface.com in each category as of March 2024, totaling 583,326. Huggingface.com started with about 30,000 models in 2022. In March 2024 alone, 50,000 artifacts were added, including 4000 text classification models. Currently there are 54,131 text classification models. There has been a linear rise in the number of uploads per month, too many to keep track of or list individually.

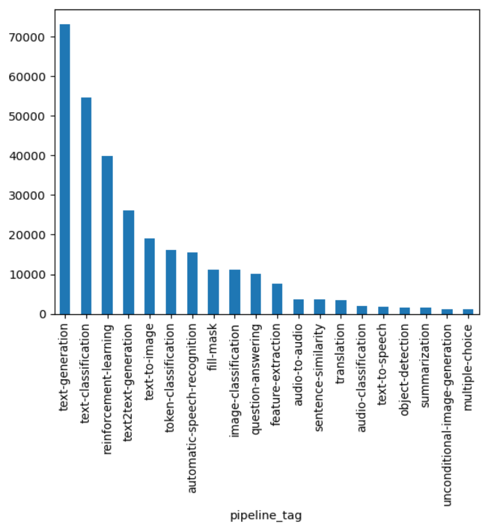

Figure 1. Number of artifacts per application, based on huggingface.com tags. Artifacts can include models & datasets.

The left side of Figure 1 shows the most popular LLM applications, and here is a description of the top 4 applications:

- Text Generation: Generation of coherent and contextually relevant text, such as a creative short story or generating a news article from a brief headline.
- Text Classification: Classify text into predefined categories based on its content, such as sorting customer reviews into positive, negative, or neutral sentiment categories.
- Reinforcement Learning: The development of agents that learn optimal conversational dialogue through user interactions.
- Text-to-Text Generation: Transforming input text into a different format or style, such as converting a formal piece of text into a more casual version or translating text from one language to another.

2.2. Retrieval-augmented generation applied to exam generation and scoring

RAG is a powerful method that enhances the performance of LLMs by incorporating external knowledge sources into the text generation process. This technique involves retrieving relevant information from a large corpus of documents and using this information to generate more accurate and contextually appropriate responses. RAG has shown significant promise in various applications, including automated exam generation and scoring, where the accuracy and relevance of generated content are crucial.

The study by Guinet et al. (2024) demonstrates an automated approach to evaluate RAG's effectiveness in generating task-specific exams. Their method involves creating synthetic exams composed of multiple-choice questions based on a designated corpus. This approach utilizes Item Response Theory to assess the quality of the generated exams, ensuring that the questions are informative and accurately reflect the model's understanding of the content.

Klang et al. (2023) explored the use of GPT-4 in creating medical examinations. They found that while GPT-4 could rapidly generate many multiple-choice questions, some of these questions required revisions due to inaccuracies or methodological flaws. Despite these challenges, GPT-4's ability to create many valid questions efficiently highlighted its potential as a supplementary tool for exam creation in medical education.

In the context of scoring, RAG's ability to retrieve relevant information plays a pivotal role in enhancing the accuracy of generated answers. The study by Guinet et al. (2024) emphasizes the importance of selecting the appropriate retrieval algorithms, noting that the choice of retrieval mechanism can significantly impact the model's performance. Their findings suggest that optimizing retrieval strategies often yields more substantial improvements than merely increasing the model size.

Klang et al. (2023) also noted that specialist reviews are essential for ensuring the quality of AI-generated exam questions. Their study revealed that while GPT-4 could produce high-quality questions, it was prone to specific types of errors, such as outdated terminology and context-insensitive content. Therefore, rigorous inspection by domain experts to validate the accuracy and appropriateness of the generated questions was a crucial step.

2.3. Ethical considerations in using AI for exam purposes

Although this paper has thus far explained benefits of AI use in exam creation, it is important to highlight ethical considerations as well. While there is a paucity of research on AI ethical use in exam creation, there is relevant research on the general topic of AI use in education. One report focused on K-12 education identified the following ethical concerns: privacy, surveillance, autonomy, bias, and discrimination (Akgun 2021). While not all these issues will affect the ethics of simply creating exams, some are relevant. Bias and discrimination are possible problems that need to be addressed, particularly teachers should be aware that "automated assessment algorithms have the potential to reconstruct unfair and inconsistent results." (Akgun 2021) When creating these exams, especially when the topic is socio-cultural in nature, it is important to ensure that questions are properly and appropriately formulated and structured.

Furthermore, autonomy could be a challenge when dealing with the constant creation of exams through AI. Professors may become overly reliant on AI, continuing to use it even if it is not in the best interest of the students. Conversely, it could even become standard practice at universities to mitigate potential discrepancies among professors. While this might seem like a good way to equalize issues, it carries the risk of "[jeopardizing] students and teachers' autonomy." (Akgun 2021) Ultimately, while these concerns may never come to fruition, they bear consideration if these processes are implemented.

3. Method

This experiment's purpose was to accurately generate and answer exam questions from an undergraduate textbook using GPTs. The project was implemented in a python Jupyter Notebook, beginning with the import of essential libraries. These libraries were used for various purposes such as interacting with the OpenAI API, handling data in JSON and CSV formats, performing regular expression operations, and conducting evaluations, such as precision scoring, word embeddings similarity scoring, and cosine similarity scoring. The versions of the main libraries and frameworks are shown in Table 1.

Table 1. Key Frameworks, Libraries, Modules

Python	3.10.12
OpenAI	1.30.1
JSON	2.0.9
CSV	1.0
RegEx	2.21
Evaluate	0.4.2
ROUGE	0.1.2
spaCy	3.7.4
scikit-learn	1.2.2

The hardware and software specifications used in this work are located in Table 2.

Table 2. Hardware/Software

RAM	12.67 GB
CPU	Dual-core Intel(R) Xeon(R) CPU @ 2.20GHz
OS	Ubuntu 22.04.3 LTS
Kernel	Linux 6.1.85+ x86_64

Within the OpenAI API, RAG was implemented by creating two "Assistants" that had access to the textbook, one utilized ChatGPT 3.5-Turbo while the other utilized ChatGPT 4-Turbo. The scope and behavior of the assistant models were defined to generate three types of examination questions, relevant answers, and excerpts from the source material. The assistants were also used to complete the examination questions previously generated. The essay questions were scored using three text similarity metrics.

3.1. Textbook Selection

The selected textbook had to meet certain criteria. The questions needed to focus on the text of the book, as it would be using that text to generate questions and answers. This gave the GPTs the best chance to create answers without having to analyze illustrations. The level of the textbook chosen was of collegiate quality, reflecting its potential use in an undergraduate course. While the selected text fulfilled these requirements, any book that fits the criteria could be used to recreate this study.

The open-source textbook used in this work is *Introduction to Pacific Studies,* which is Volume 6 of the *Teaching Oceania* series of books written by the

University of Hawai‘i Center for Pacific Island Studies (Mawyer et. al., 2020). It consists of seven sections, each of which discusses a key aspect of the political and cultural issues and topics that Oceania and their people face.

The ChatGPT models were given a list of criteria to generate the questions. They were:

- Create a ten-question quiz.
- Compose the quiz of true-false, multiple choice, and short-answer questions.
- Ensure questions have answers easily found within the book; creatively generated questions were not acceptable.
- Require multiple-sentence answers for short-answer questions.
- Do not ask questions that require reading charts, graphs, or other illustrations.

3.2. Metrics

For the true-false and multiple-choice questions, the answers were scored using accuracy. Short-answer questions were scored using the three metrics shown in Table 3, to ensure a comprehensive evaluation.

Table 3. Metrics and Brief Description

ROUGE-1	Automatic summarization and machine translation software (Lin, 2004)
Cosine Similarity	Used in text analytics to compare documents and determine if they're similar and how much (Supe, 2023)
Word Embeddings	Similarity through comparing multi-dimensional representations of a word (spaCy, 2016)

Recall-Oriented Understudy for Gisting Evaluation (ROUGE): A token similarity metric that "measures the similarity reference text and generated text by focusing on recall" (Jain 2023). ROUGE has been "employed to assess the quality and coherence of the generated text" (Jain 2023). ROUGE-1 was implemented via the rouge.compute() method to measure the overlap of individual words between the generated answers and the reference answers (*Metric: rouge, n.d*).

Cosine Similarity: This metric treats documents as vectors, with each unique word "as a dimension" (Supe 2023). After converting two different documents into vectors, it measures the angle between their vectors and take the cosine of that angle (Supe 2023). This metric is widely used due to its flexibility in text analytics. Because cosine similarity only measures direction and not magnitude, document length does not affect the calculation, allowing for proper comparison of documents of different lengths. Cosine similarity scores were computed using the scikit-learn library, employing CountVectorizer() and cosine_similarity() methods.

Word embeddings: These are multi-dimensional representations of words used in language models and can measure the similarity of two objects (spaCy2016). Using algorithms like word2vec, these vectors are compared to determine the semantic similarity of different documents. Word embeddings similarity scores were calculated using the spaCy library with the en_core_web_md pipeline.

3.3. Generative AI

To utilize an LLM to create an exam, the first step involved setting up two assistants. This was achieved by either creating new OpenAI assistants or accessing existing assistants created in this work using the beta.assistants.create() or beta.assistants.retrieve() methods. Each assistant was then instantiated with specific key parameters to define functionality. The instruction given to each assistant was: "You are an expert Anthropologist in the area of Pacific Studies. Use uploaded files only to answer questions about anthropology." This directive defined the scope (anthropology) and behavior (expert Anthropologist and reliance on vector-stored documents) expected from the assistant (*How Assistants work, n.d.*).

Each assistant was created with the model set as either "gpt-3.5-turbo" or "gpt-4-turbo." The tools parameter was set to [{ "type": "file_search" }], enabling the assistant to perform specialized tasks. Specifically, the file search tool enables the assistant to perform text-based searches on documents uploaded to an attached vector store (*File Search, n.d.*). For data preparation, the textbook was uploaded as a text file to a vector store. This was accomplished by beta.vector_stores.file_batches.upload_and_poll() and beta.assistants.update() methods, ensuring the assistant had access to the necessary academic content.

The next step was the core of the project: generating exam content. Each assistant was prompted to create sections of an examination through three independent prompts, which are shown in Appendix

A. First, they generated a specified number of true-false questions, along with their answers and excerpts. Next, they produced multiple-choice questions, also with answers and excerpts. Finally, they created short-answer questions, complete with answers and excerpts. This structured approach attempted to enforce consistency across the responses and address the brittleness of regular expression parsing when each model produced unpredictable outputs. By organizing the prompts in this manner, each assistant produced well-defined and distinct sections of the examination, facilitating easier validation and processing. Sample exam questions, answers, and textbook excerpts are located in Appendix B.

After creating each portion of the examination, the responses were processed using a ChatGPT model and regular expressions to capture the exam information in a Python list, which simplified the storage and subsequent scoring of results. This resulted in two complete exams, one produced by a GPT 3.5-Turbo Assistant, and one produced by a GPT 4-Turbo Assistant.

Once processed, both exams were then fed back to the original assistant and the competing assistant with a new prompt, instructing them to answer the questions using only the uploaded textbook. Each assistant's responses were validated through a simple text parsing function to ensure all questions were answered. These new answers were then processed similarly by a ChatGPT model and regular expressions to capture the responses for storage and scoring. The prompts used to answer the exams are shown in Appendix C.

4. Results & Analysis

In evaluating the performance of GPT-3.5-Turbo and GPT-4-Turbo the study utilized a combination of quantitative metrics namely the number of correct true-false and multiple-choice answers along with the similarity scores for short-answer questions. This choice of metrics was discussed in the methodology section where these specific measures were identified due to their ability to assess not only the factual correctness but also the linguistic and semantic quality of the responses generated by the models. The results of the performance evaluation for one exam are presented in Table 4, comparing the performance of the models across exams generated by GPT-3.5-Turbo Assistant and GPT-4-Turbo Assistant.

Each model was subjected to identical test conditions with and without access to source texts, designed to test what retrieval-augmented generation capabilities provide when answering domain-specific questions. This approach evaluated the robustness and adaptability of each model under controlled academic conditions.

Table 4. Model performance for one exam.

Model	Scoring	3.5T-exam[1]		4T-exam[2]	
3.5T student with text[3]	True/False	1.0000		1.0000	
	Mult Choice	0.5000		1.0000	
	ROUGE-1	0.5103	0.4752	0.3368	0.3600
	Cosine	0.6872	0.8077	0.5220	0.5231
	Embeddings	0.9789	0.9896	0.9781	0.9797
4T student with text[4]	True/False	1.0000		1.0000	
	Mult Choice	0.7500		1.0000	
	ROUGE-1	0.5046	0.4250	0.4478	0.4396
	Cosine	0.6786	0.7070	0.6437	0.6777
	Embeddings	0.9874	0.9828	0.9808	0.9857
3.5T student no text[5]	True/False	1.0000		0.7500	
	Mult Choice	0.7500		1.0000	
	ROUGE-1	0.2609	0.4228	0.2927	0.3770
	Cosine	0.4949	0.7467	0.5271	0.5930
	Embeddings	0.9661	0.9755	0.9559	0.9803
4T student no text[6]	True/False	1.0000		1.0000	
	Mult Choice	1.0000		1.0000	
	ROUGE-1	0.2632	0.4031	0.4000	0.4088
	Cosine	0.4800	0.7207	0.5821	0.6254
	Embeddings	0.9716	0.9831	0.9686	0.9803

4.1. Detailed Results

All models showed consistently high rankings in true-false questions. GPT-4-Turbo Assistant and GPT-3.5-Turbo Assistant typically received the highest rankings indicating their effectiveness in accurately interpreting and responding to binary questions based on the text.

In multiple-choice rankings the GPT-4-Turbo Assistant consistently outperformed other models suggesting superior capabilities in understanding and selecting the correct answers from multiple options.

[1] Exam created by Assistant using GPT-3.5-Turbo
[2] Exam created by Assistant using GPT-4-Turbo
[3] Assistant using GPT-3.5-Turbo with access to uploaded textbook
[4] Assistant using GPT-4-Turbo with access to uploaded textbook
[5] ChatGPT-3.5-Turbo without access to textbook
[6] ChatGPT-4-Turbo without access to textbook

This is indicative of its robust comprehension and retrieval abilities.

The ROUGE-1 rankings reflect the models' ability to produce text closely matching the reference material. Here the GPT-4-Turbo Assistant often ranked higher particularly in generating answers that align well with the expected responses demonstrating a strong grasp of content accuracy and relevance.

The rankings in cosine similarity scores were varied with GPT-4-Turbo generally achieving better results. Higher rankings in this metric indicate a closer match to the textual style and content of the reference material underlining GPT-4's adeptness at maintaining textual integrity and context.

The embeddings scores provide a perspective on semantic understanding. GPT-4-Turbo frequently received top rankings illustrating its superior ability to grasp and replicate the underlying semantic properties of the original text. This suggests a deep and nuanced understanding of the material which is crucial for generating contextually accurate responses.

4.2. Ranked Performance

The ranked performance of the models across various metrics is shown in Tables 5 and 6. These tables provide insights into how each model performs relative to others in specific evaluation criteria. Table 5 provides a breakdown of performance rankings per exam. Here the exam produced by GPT-4-Turbo Assistant often ranked first.

Table 5. Model rankings per exam

Model	Exam	Ranking
3.5T student with text	3.5T	1
4T student with text	4-T	2
4T student with text	3.5T	3
4T student no text	4-T	4
4T student no text	3.5T	5
3.5T student with text	4-T	6
3.5T student no text	4-T	7
3.5T student no text	3.5T	8

When examining the ranked performance in Table 6 the GPT-4-Turbo Assistant ranked highest in overall performance. This model's ability to consistently score high across all metrics underscores its robustness in exam content generation and answering.

Table 6. Model overall rankings

Model	Ranking
4T student with text	1
3.5T student with text	2
4T student no text	3
3.5T student no text	4

The ranked performance as detailed in Tables 5 and 6 illustrates a clear differentiation in the capabilities of GPT-3.5-Turbo and GPT-4-Turbo. GPT-4-Turbo consistently outperformed GPT-3.5-Turbo particularly in settings where text access was unrestricted suggesting a superior ability to leverage available resources for answer generation. This is indicative of GPT-4-Turbo's enhanced retrieval mechanisms and updated training algorithms which seem to allow it to better understand and utilize context from the provided source material.

Furthermore, in multiple-choice settings where nuanced comprehension of the question context and subtleties in phrasing can significantly influence the outcome GPT-4-Turbo demonstrated higher accuracy. This suggests advancements in its language processing capabilities likely attributable to its larger training dataset and more sophisticated neural network architecture compared to GPT-3.5-Turbo (Emmanuel 2023).

In contrast when text access was restricted the performance gap between the two models narrowed indicating that while GPT-4-Turbo has superior retrieval capabilities both models maintain a base level efficiency in generating logically consistent answers from embedded knowledge.

The comparative analysis highlights the implications of evolving model architectures and training environments in the development of AI applications for academic testing. It suggests that future enhancements in model training and development could further exploit these capabilities potentially leading to more sophisticated AI tools for educational assessment.

4.3. Analysis of Results

The results indicate that GPT-4-Turbo Assistant consistently outperforms other models particularly in terms of ROUGE-1, cosine similarity, and word embeddings similarity scores. This suggests that GPT-4 is particularly effective at both generating high-quality exam content and providing accurate answers. The strength of GPT-4 in handling true-false, multiple-choice and short-answer questions also highlight its utility in structured academic testing environments.

In the case of true-false and multiple-choice questions all models performed well often achieving perfect scores. However, the real differentiation among the models was observed in the short-answer questions as evidenced by the ROUGE-1, cosine similarity, and word embeddings similarity scores.

The GPT-4-Turbo Assistant showed a balanced performance across all metrics making it the most reliable model for generating high-quality exam content and providing accurate answers. The GPT-3.5-Turbo Assistant also performed well but had slightly lower scores in some of the more nuanced metrics like cosine similarity and word embeddings similarity.

5. Discussion and Conclusions

The ChatGPT 4-Turbo models performed the best compared to its 3.5T counterparts, which was anticipated as its complexity (number of parameters) is an order of magnitude greater. What was less expected was the ability of the 3.5T model to perform comparably. The 3.5T model performed well, specifically when looking at the summarization metrics. Even though it was not better, it performed close enough to suggest that individuals without the resources to purchase a subscription could use an equivalent process with the 3.5T model effectively with more oversight. Even when concerning true-false and multiple-choice questions, the 3.5T model often fell only one question behind. More testing is needed to determine exactly how far the 3.5T model lags behind the 4T model in this context. But upon preliminary review, the 3.5T model is usable and may only require a little more attention.

What seemed far more important than which specific model of ChatGPT was being used was whether the model was utilized as an assistant, which allows it to leverage RAG and the textbook as a specified source. As shown in the analysis, the GPT-3.5-T Assistant performed better overall than the 4T model that did not use the specific college textbook as a source. Rather than searching its pre-trained data, which will take longer and might be inaccurate, the models had a better idea of what to target and where to search. These results demonstrate that RAG is an effective way to create tests and answer said tests using a specified scholarly source. These findings can be applied to academic professionals looking for ways to create mass variations of exams that are both fair and accurate. Further research in this field can include expanding the number of LLMs tested, evaluating different metrics, and exploring textbooks that cover other topic areas.

Author's note: The views expressed are those of the authors and do not reflect the official guidance or position of the United States Government, the Department of Defense, the United States Air Force, the United States Space Force or any agency thereof. Reference to specific commercial products does not constitute or imply its endorsement, recommendation, or favoring by the U.S. Government. The authors declare this is a work of the U.S. Government and is not subject to copyright protections in the United States. This article has been cleared with case number 88ABW-2024-0625.

References

Akgun, S., & Greenhow, C. (2021). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, *2*(3). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8455229/

Awan, A. A. (2023). *The Pros and Cons of Using LLMs in the Cloud Versus Running LLMs Locally*.Datacamp. https://www.datacamp.com/blog/the-pros-and-cons-of-using-llm-in-the-cloud-versus-running-llm-locally

AWS (n.d.). *What is RAG? - Retrieval-Augmented Generation Explained.* Amazon Web Services, Inc. https://aws.amazon.com/what-is/retrieval-augmented-generation/

Castells, M. (2010). *The rise of the network society* (2nd ed.). Wiley-Blackwell.

Dilmegani, C. (2024). *In-Depth Guide to Cloud Large Language Models (LLMs) in 2024*. AIMultiple: High Tech Use Cases & Tools to Grow Your Business. https://research.aimultiple.com/cloud-llm/

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for Large*

Language Models: A Survey. https://arxiv.org/pdf/2312.10997
HuggingFace.com. (n.d.) Metric: rouge. *ROUGE - a Hugging Face Space by evaluate-metric.*. https://huggingface.co/spaces/evaluate-metric/rouge
Jain, S. (2023). *Elevating LLMs with ROUGE Evaluation*. UpTrain AI. https://blog.uptrain.ai/evaluating-llms-with-rouge-evaluation/
Klang, E., Portugez, S., Gross, R., Brenner, A., Gilboa, M., Ortal, T., ... & Segal, G. (2023). Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. *BMC Medical Education*, *23*. https://doi.org/10.1186/s12909-023-04752-w
Lin, C. Y. (n.d.). *ROUGE: A Package for Automatic Evaluation of Summaries*. https://www.aclweb.org/anthology/W04-1013
Spacy.io (n.d.). *Linguistic Features · spaCy Usage Documentation*. https://spacy.io/usage/linguistic-features#vectors-similarity
Mawyer, A., Auelua, R., Aikau, H., Barcham, M., Boeger, Z., Dawrs, S., Genz, J., Kava, L. (2020). *Introduction to Pacific Studies.* Honolulu: Center for Pacific Islands Studies, University of Hawai'i-Mānoa.
OpenAI. (n.d.). *File Search*. Tools. https://platform.openai.com/docs/assistants/tools
OpenAI. (n.d.). *How Assistants work*. How it works. https://platform.openai.com/docs/assistants/how-it-works
Ray, P. P. (2023). ChatGPT: a Comprehensive Review on background, applications, Key challenges, bias, , ethics, Limitations and Future Scope. *Internet of Things and Cyber-Physical Systems*, *3*(1), 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003
Supe, K. (2023). *Understanding Cosine Similarity in Python with Scikit-Learn* [Review of *Understanding Cosine Similarity in Python with Scikit-Learn*]. Memgraph. https://memgraph.com/blog/cosine-similarity-python-scikit-learn
Turing.com. (2023). *Best Strategies to Minimize Hallucinations in LLMs: A Comprehensive Guide.* Accessed June 10th 2024. https://www.turing.com/resources/minimize-llm-hallucinations-strategy

Appendices

Appendix A: Prompts used to generate exam

True/False Question Prompt:

```
"""
Write 4 true or false questions using your
knowledge and the uploaded files.
Provide an answer using only the uploaded files.
Include a snippet excerpt from inside the uploaded
files to justify the answer.

Use the following format:
Question: (True or False) A false statement from
the uploaded files
Answer: False
Excerpt: An excerpt from the uploaded files...

Question: (True or False) A true statement from the
uploaded files
Answer: True
Excerpt: An excerpt from the uploaded files...
"""
```

Multiple Choice Question Prompt:

```
"""
Write 4 multiple questions using your knowledge and
the uploaded files.
Provide an answer using only the uploaded files.
Include a snippet excerpt from inside the uploaded
files to justify the answer.
Ensure answers are fairly distributed across the
choices.

Use the following format:
Question: (Multiple Choice) A question from the
uploaded files
A) Incorrect answer (no more than 15 words)
B) Correct answer (no more than 15 words)
C) Incorrect answer (no more than 15 words)
D) Incorrect answer (no more than 15 words)
Answer: B
Excerpt: An excerpt from the uploaded files...

Question: (Multiple Choice) A question from the
uploaded files
A) Incorrect answer (no more than 15 words)
B) Incorrect answer (no more than 15 words)
C) Incorrect answer (no more than 15 words)
D) Correct answer (no more than 15 words)
Answer: D
Excerpt: An excerpt from the uploaded files...
"""
```

Essay Question Prompt:

```
"""
Write 2 essay questions using your knowledge and
the uploaded files.
Provide an answer for each question using only the
uploaded files.
Include a snippet excerpt for each question from
inside the uploaded files to justify the answer.

Ensure you provide 2 questions, 2 answers, and 2
excerpts.

Use the following format:
Question: (Essay) A question from the uploaded
files
Answer: A single concise paragraph answer
Excerpt: An excerpt from the uploaded files...

Question: (Essay) A question from the uploaded
files
Answer: A single concise paragraph answer
Excerpt: An excerpt from the uploaded files...
"""
```

Appendix B: Sample exam questions, answers, and textbook excerpts

Teacher: GPT-4-Turbo Assistant

Question: (True or False) The Pacific studies programs include a tenure-track research fellowship for each native Pacific scholar to be stationed in Oceania.
Answer: False
Excerpt: While disciplinary decolonization remains an incomplete project, significant achievements have been made on many fronts, including challenges to Western educational models and colonial languages in favor of vernacular education, valuing and in some cases recovering Indigenous knowledges and languages, supporting sovereignty and self-determination for Pacific peoples.

Question: What is a significant motivation for Pacific studies as an academic discipline?
A) To create political alliances within the Pacific islands.
B) To reposition mainstream Western academic approaches with Indigenous techniques.
C) To promote tourist activities within the academic community.
D) To focus solely on historical accounts of Western explorers.
Answer: B
Excerpt: For many Pacific studies practitioners, the foundational motivation for this work is to strongly reposition mainstream Western academic approaches in relation to more culturally appropriate techniques

Question: (Essay) How does Pacific Studies integrate indigenous methodologies, and why is this integration crucial for the discipline?
Answer: Pacific Studies has increasingly integrated indigenous methodologies by utilizing frameworks such as Kaupapa Māori Research, Kakala, Vanua Research Framework, and others, which enables the discipline to be more culturally sensitive and appropriate. This integration is crucial as it aligns academic pursuits with the needs and perspectives of Pacific communities, supporting their well-being and self-determination. It provides a means to decolonize knowledge production by valuing indigenous wisdom and approaches, thus respecting and prioritizing local contexts and customs over external academic frameworks.
Excerpt: As discussed in this volume, one outcome of this effort has been the development of Pacific studies into a broad range of pedagogies and research methodologies and methods to decolonize teaching and research within Pacific communities. For many Pacific studies practitioners, the foundational motivation for this work is to strongly reposition mainstream Western academic approaches in relation to more culturally appropriate techniques.

Appendix C: Prompts used to answer exam questions

Assistant True/False Answer Prompt:
```
"""
Using only the uploaded files, answer the following
question.
Pay close attention and ONLY write the ANSWER.

For True or False, provide True or False

Do not include the question.
Do not include any justification for the answer.

Please answer the following question:
{question}
"""
```

Assistant Multiple Choice Answer Prompt:
```
"""
Using only the uploaded files, answer the following
question.
Pay close attention and ONLY write the ANSWER.

For Multiple Choice, provide the LETTER only.

Do not include the question.
Do not include any justification for the answer.

Please answer the following question:
{question}
"""
```

Assistant Essay Answer Prompt:
```
"""
Using only the uploaded files, answer the following
question.
Pay close attention and ONLY write the ANSWER.

For open-ended essay questions, provide a single
concise paragraph.

Do not include the question.
Do not include any justification for the answer.

Please answer the following question:
{question}
"""
```

ChatGPT Answer Prompt:
```
"""
Please take your time and answer each of the
questions below.
Pay close attention and ONLY write the ANSWERS.
Do not include the question.
Do not include justification.

Here is the information:
{questions[0]}
...
{questions[9]}

Use the following example out format:
gpt_answers = [\"True\", \"False\", \"True\",
\"False\", \"A\", \"B\", \"C\", \"D\", \"A single
concise paragraph answer.\", \"A single concise
paragraph answer.\"]
"""
```