Data & Policy (2025), 7: e57
doi:10.1017/dap.2025.10024

CAMBRIDGE
UNIVERSITY PRESS

RESEARCH ARTICLE

Impact of retrieval augmented generation and large language model complexity on undergraduate exams created and taken by AI agents

Erick Tyndall[1], Colleen Gayheart[1], Alexandre Some[1], Joseph Genz[2], Torrey Wagner[1] and Brent Langhals[1]

[1]Department of Systems Engineering and Management, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio, USA
[2]Department of Anthropology, University of Hawai'i at Hilo, Hilo, Hawaii, USA
Corresponding author: Erick Tyndall; Email: erick.tyndall@us.af.mil

Received: 05 April 2025; **Revised:** 01 July 2025; **Accepted:** 21 July 2025

Keywords: academic examinations; artificial intelligence; generative pre-training transformer; large language models; retrieval augmented generation

Abstract

The capabilities of large language models (LLMs) have advanced to the point where entire textbooks can be queried using retrieval-augmented generation (RAG), enabling AI to integrate external, up-to-date information into its responses. This study evaluates the ability of two OpenAI models, GPT-3.5 Turbo and GPT-4 Turbo, to create and answer exam questions based on an undergraduate textbook. 14 exams were created with four true-false, four multiple-choice, and two short-answer questions derived from an open-source Pacific Studies textbook. Model performance was evaluated with and without access to the source material using text-similarity metrics such as ROUGE-1, cosine similarity, and word embeddings. Fifty-six exam scores were analyzed, revealing that RAG-assisted models significantly outperformed those relying solely on pre-trained knowledge. GPT-4 Turbo also consistently outperformed GPT-3.5 Turbo in accuracy and coherence, especially in short-answer responses. These findings demonstrate the potential of LLMs in automating exam generation while maintaining assessment quality. However, they also underscore the need for policy frameworks that promote fairness, transparency, and accessibility. Given regulatory considerations outlined in the European Union AI Act and the NIST AI Risk Management Framework, institutions using AI in education must establish governance protocols, bias mitigation strategies, and human oversight measures. The results of this study contribute to ongoing discussions on responsibly integrating AI in education, advocating for institutional policies that support AI-assisted assessment while preserving academic integrity. The empirical results suggest not only performance benefits but also actionable governance mechanisms, such as verifiable retrieval pipelines and oversight protocols, that can guide institutional policies.

Policy Significance Statement

Rapid progress in large language models means that today's best-performing system is often overtaken within months. Yet two design choices appear to deliver durable benefits across architectures and domains: (1) pairing the model with retrieval-augmented generation so that it grounds its output in verifiable sources, and (2) maintaining

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

human-in-the-loop oversight for critical applications such as graded assessments. Although limited to GPT-3.5 Turbo and GPT-4 Turbo models on a single undergraduate-level textbook, the experiment demonstrates these principles in practice. Adding retrieval-augmented generation cut hallucinations and boosted answer accuracy more effectively than upgrading to a more advanced base model alone. For policy makers, this suggests that procurement guidelines and accreditation standards should emphasize verifiable retrieval pipelines rather than mandating specific proprietary models. Because retrieval-augmented generation can elevate the performance of lower-cost models to near parity, equity-minded policies can require that any AI-assisted assessment workflow be reproducible on an open-weight or low-cost model to reduce the risk of widening resource gaps. These recommendations align with the governance obligations for high-risk educational systems in the EU AI Act and the risk-management steps in the NIST AI RMF, including bias mitigation, documentation, and human oversight.

1. Introduction

AI is becoming an increasingly prominent and policy-relevant topic, with growing implications for governance, education, and workforce regulation. With the rapid development of large language models (LLMs), such as OpenAI's Generative Pre-training Transformer (GPT) models, new opportunities and challenges emerge in the automation of information and content generation. These advances raise critical policy questions surrounding fairness, transparency, and compliance with national and international education standards.

The application in this work uses GPT LLMs for exam generation and answering, specifically analyzing how retrieval-augmented generation (RAG) enhances accuracy and reliability in automated academic assessments. As institutions explore AI-driven testing, policymakers and accreditation bodies must evaluate the implications for educational integrity, bias mitigation, and regulatory compliance, ensuring that AI tools align with ethical and legal standards in academia. This study contributes to that evaluation by testing performance differences between GPT models with and without RAG and translating those empirical findings into design principles that support responsible, policy-aligned implementation of AI-assisted academic assessments.

1.1. Overview

AI has continually been a popular and controversial topic, particularly in the context of assisting with or replacing everyday human tasks, automating processes, shaping workforce dynamics, and influencing education policy. Although AI technologies have existed for decades, they have remained largely confined to specialized domains. This changed with the introduction of consumer-friendly LLMs, which made the technology broadly accessible to the public. This all changed with the creation of OpenAI, whose stated 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. Optimized for conversation-based tasks, including contextual understanding and coherence (Ray, 2023), ChatGPT has become increasingly available, with both free access and paid subscription tiers that offer more advanced models and faster response times.

The academic domain is one space where ChatGPT use has been especially prevalent, creating both opportunities and regulatory concerns. 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 educators' and administrators' perspectives, such as using a specific resource (i.e., textbook) to generate an exam, create matching solution sets, and scoring answers. 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 et al., 2023). However, regulatory bodies and institutions must establish policies to ensure that AI-generated assessments align with academic standards, maintain fairness, and mitigate potential biases.

Although the creation of LLMs has created a beneficial effect on the scope of generative AI, these models by themselves often suffer from hallucinations, where the AI generates information that seems plausible but is wholly fabricated or factually 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 (Lewis et al., 2021). 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 produce accurate answers about specific medical technology or capture relevant advancements. RAG systems "effectively reduces the problem of generating factually incorrect content" (Gao et al., 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 et al., 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.

As AI becomes increasingly integrated into academia, from content generation to assessment design, there is a growing need for clear regulatory frameworks. Policymakers and educational institutions must work together to promote transparency, fairness, and accountability in AI-assisted assessments. These policies should address concerns such as data privacy, model bias, and the appropriate boundaries for AI use in student evaluation and curriculum development. Without such guidelines, the rapid adoption of AI tools risks undermining academic standards and introducing fairness and equity challenges that this study seeks to anticipate and address through a focused evaluation of RAG-enhanced GPT models.

1.2. Objectives and scope

The objective of this paper is to study the variations within different forms of GPT by creating exam questions from a specified resource and measuring each model's accuracy in answering those questions. The work creates a diverse set of exam questions using OpenAI's GPT, focusing on RAG to enhance accuracy and reliability, and explores the methods it would take to enable AI models to create and answer the questions. Furthermore, the paper will analyze the differences between GPT versions to determine which produces higher-quality assessments and provides the most accurate answers. This will be done using a range of text-similarity and scoring metrics discussed later in the Methodology section. The primary prompts used can be found in the Supplementary Appendices to allow replication and validation of this work, which is crucial for developing policies that ensure transparency and fairness when using AI tools.

A limitation of this study is that while there are thousands of LLMs, only two variations of OpenAI's GPT models will be utilized. These models were selected as ChatGPT remains one of the most widely accessible AI tools for educators. A second boundary is that this study evaluates AI performance using a single undergraduate textbook, providing a controlled test case while acknowledging that broader applications may require additional resources to ensure assessment validity across diverse educational settings. Although limited in scope, the controlled setup enables clear attribution of performance differences to RAG and model complexity, offering insight into durable design choices for AI-assisted assessment workflows. These findings can inform procurement, oversight, and fairness considerations across diverse institutional settings.

2. Background

There are multiple types of LLMs, including cloud-based and locally deployed models, each with differing advantages and disadvantages. LLMs have various applications and capabilities that can be enhanced with methods like RAG. As these models are increasingly integrated into education and assessment, policymakers must evaluate their implications for academic integrity, data security, and regulatory compliance. LLMs have rapidly grown in use and availability in recent years, and with this recent growth, there are expanding concerns for ethical considerations and the need for governance

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

frameworks within the academic domain. This study identifies which design features, such as retrieval grounding or model complexity, most impact performance and trustworthiness in educational settings.

2.1. Infrastructure and applications of large language models

LLMs are typically either locally deployed or cloud-based. Cloud LLMs such as ChatGPT are accessed via the internet through a web interface or an application programming interface (API), as is 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 security risks due to the nature of accessing a cloud environment. This includes the opportunity for data breaches and illegitimate access to data (Dilmegani, 2024). Cloud LLMs are easily scalable and offer advantages for research institutions requiring extensive or high-performance 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 fees or pay-as-you-go payment 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 greater control over the LLM and its environment and reduces data security and privacy concerns. However, this requires greater familiarity with the technology and maintenance support (Dilmegani, 2024). Since the information input into the LLM is stored locally, rather than online, it is inherently more secure and difficult to access without physical access to the computer. However, authorized access and usage policies must still be considered during deployment. Depending on the user's needs, the cost of hardware can be substantial. While this study did not evaluate locally deployed models, understanding their characteristics is useful for contextualizing policy considerations across different deployment environments.

LLMs can be applied to a wide range of tasks, many of which are depicted in Figure 1. The x-axis represents different applications, including text generation and text classification, and the y-axis represents the number of AI models available for each category. 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, and in March 2024 alone, 50,000 artifacts were added, including 4000 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. Currently, there are 54,131 text classification models, reflecting the rapid expansion of AI tools and the increasing demand for governance and policy frameworks to manage their integration into education and research.

The left side of Figure 1 highlights the most popular LLM applications, including:

- Text Generation: Producing coherent and contextually relevant text, such as a creative short story or an AI-generated news article from a brief headline.
- Text Classification: Sorting text into predefined categories based on its content, such as organizing customer reviews into positive, negative, or neutral sentiment categories.
- Reinforcement Learning: Fine-tuning AI agents to improve conversational dialogue quality 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. AI-assisted 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

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

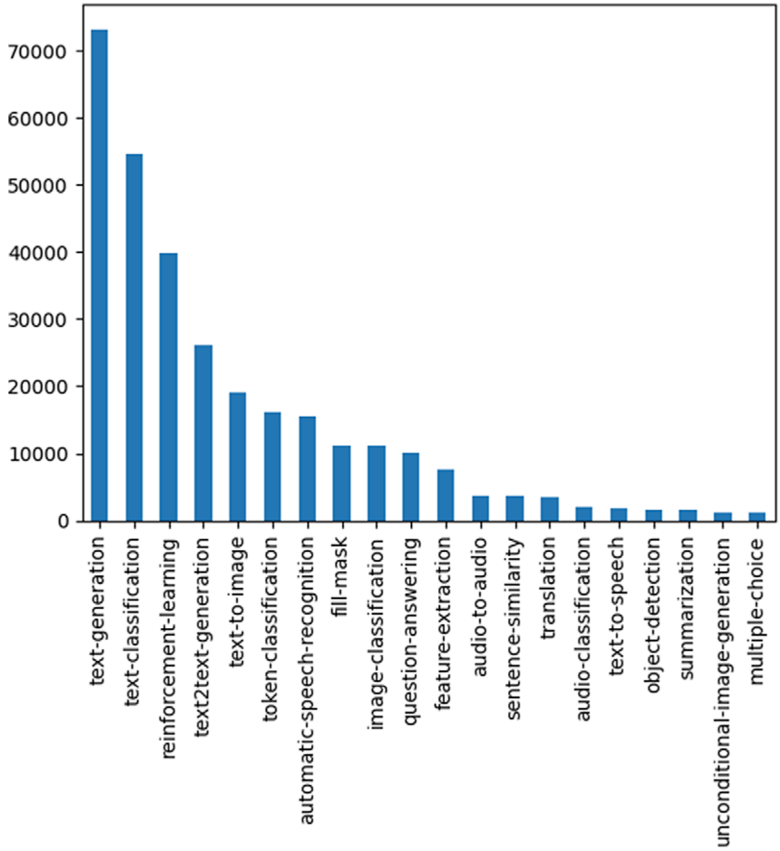

Figure 1. *Number of artifacts (models & datasets) per application, based on huggingface.com tags.*

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, relevance, fairness, and reliability of AI-generated content are critical for educational policy considerations.

Klang et al. (2023) explored the use of GPT-4 in creating medical examinations, relying solely on the model's pretrained internal knowledge without the use of retrieval-augmented methods. While GPT-4 demonstrated the ability to rapidly generate a large volume of multiple-choice questions, the absence of external verification contributed to several factual, contextual, and methodological errors. For instance, the model confused treatment protocols for different conditions and used outdated medical terminology. Additionally, the study noted redundancy in question generation and a small number of questions that included flawed logic or oversimplified clinical reasoning. These findings underscore the limitations of relying solely on an LLM's internal knowledge base for high-stakes educational content. To ensure alignment with accreditation policies and institutional assessment standards, quality assurance mechanisms involving subject-matter experts and educational authorities remain essential.

Building on these concerns, Guinet et al. (2024) demonstrate 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.

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

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

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, underscoring the need for standardized guidelines on AI-driven grading transparency and reliability.

These studies demonstrate both the promise and the limitations of AI-assisted exam generation and scoring. While techniques like RAG can enhance accuracy through targeted retrieval, even advanced models like GPT-4 require expert oversight to ensure content quality and alignment with academic standards. As AI becomes more integrated into educational assessment, establishing clear protocols for validation, review, and bias mitigation will be essential to maintain fairness and credibility.

2.3. Ethical considerations in using AI for exam purposes

Although this paper has thus far explained the benefits of AI use in exam creation, it is important to highlight ethical and regulatory considerations as well. While there is a paucity of research on the ethical use of AI-generated exams, 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 and Greenhow, 2021). While not all these issues will affect the ethics of simply creating exams, some are highly relevant to AI-assisted assessment policies. Bias and discrimination are possible problems that need to be addressed, particularly as educators and institutions should be aware that "automated assessment algorithms have the potential to reconstruct unfair and inconsistent results" (Akgun and Greenhow, 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 in a way that mitigates bias and adheres to academic fairness standards. This is especially pertinent in the context of this study, which used a textbook on Pacific Studies, a subject area with inherent cultural dimensions that require sensitive treatment.

Furthermore, autonomy could be a challenge when AI-generated exams become standard practice in educational environments. 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 an institutional norm at universities to mitigate potential discrepancies and inconsistencies among professors. While this might seem like a good way to equalize issues and ensure fairness, it carries the risk of "[jeopardizing] students and teachers' autonomy" (Akgun and Greenhow, 2021). These concerns highlight the need for clear institutional policies that balance AI integration with human oversight. Regulatory frameworks should ensure that AI remains a tool for academic enhancement rather than a substitute for pedagogical judgment. Ultimately, while these concerns may not immediately materialize, they warrant consideration as AI-tools continue to develop and become ubiquitous.

2.4. Policy considerations in using AI for exam purposes

The integration of AI in educational assessments has introduced both opportunities and regulatory challenges, prompting universities, accreditation bodies, and policymakers to consider new frameworks for AI-assisted exam generation. While AI-driven tools such as RAG can enhance efficiency, consistency, and accessibility in test creation, they also raise concerns about fairness, transparency, and academic integrity. Given the rapid deployment of AI in educational settings, it is crucial to examine existing policies and governance frameworks that guide the ethical and responsible use of AI-driven assessments. The retrieval-augmented approach tested in this study aligns closely with these regulatory goals, particularly those focused on transparency, human oversight, and data traceability (Gao et al., 2024; Holistic AI, 2024).

One of the most significant regulatory developments in AI governance is the European Union's AI Act, which classifies AI systems based on risk levels. Under this framework, universities operating within the EU are considered high-risk AI system providers, meaning they must comply with stringent requirements to ensure transparency, accountability, and bias mitigation in AI-driven processes (Digital Education Council, 2024). Universities using AI-powered assessment tools, such as automated exam generation and

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

scoring, must establish risk management systems, conduct data governance to ensure fairness and completeness, and maintain technical documentation for compliance verification. Additionally, the Act mandates record-keeping of system modifications, human oversight mechanisms, and quality management systems to ensure the reliability of AI applications in education. These requirements, while promoting responsible AI use, may pose administrative and financial burdens on universities, particularly smaller institutions with limited resources to maintain compliance. This highlights the importance of scalable and cost-effective solutions, such as pairing lower-cost models with retrieval techniques, to ensure compliance does not exacerbate institutional disparities.

The EU AI Act also has broader implications beyond Europe, particularly for universities engaged in multinational research collaborations or institutions replicating the work of foreign academic bodies. As AI-driven assessments become standardized in global education, universities that collaborate on research, share AI-generated exam datasets, or replicate findings from foreign institutions may face regulatory challenges when aligning with different AI governance frameworks. Institutions outside the EU must also remain aware of these regulations, as international AI policies may shape future national regulations and influence best practices for AI adoption in education. The growing regulatory landscape suggests that universities engaging in cross-border AI research and assessments must prioritize compliance with evolving AI laws to avoid potential legal and ethical pitfalls.

In the United States, AI governance is influenced by the National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF), which was established under the National Artificial Intelligence Initiative Act of 2020 (Holistic AI, 2024). Unlike the EU AI Act, which enforces strict compliance for high-risk AI applications, the NIST AI RMF is a voluntary framework, designed to help organizations assess and manage AI risks while promoting responsible development and deployment. The framework outlines four core functions, Govern, Map, Measure, and Manage, that guide institutions in ensuring that AI systems are transparent, trustworthy, and aligned with ethical considerations.

While the NIST AI RMF does not impose mandatory regulations, its voluntary guidelines are increasingly being adopted by universities and research institutions as a best-practice model for AI governance. This is particularly relevant for AI-generated assessments, as the framework encourages bias mitigation strategies, risk assessments, and human oversight mechanisms to enhance the reliability and fairness of AI-driven exams. Moreover, the NIST framework emphasizes alignment with international AI standards, which is crucial for U.S. universities engaging in multinational research collaborations or implementing AI-based assessment methods influenced by foreign institutions.

As AI-generated assessments become more prevalent in higher education, policymakers and institutions must consider several key policy areas. These include bias detection and fairness, ensuring accountability and transparency in AI-driven assessments, addressing academic integrity concerns, and promoting equitable access to AI tools. By establishing comprehensive policies, institutions can ensure that AI remains a valuable tool for educators while upholding the principles of fairness, accuracy, and student autonomy.

3. Method

This experiment's purpose was to accurately generate and answer exam questions from an undergraduate textbook using GPT models. 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.

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

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 used ChatGPT 4-Turbo. These Assistants relied on OpenAI's file_search tool to retrieve content from a vectorized index of the textbook, enabling

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

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

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

grounded responses based solely on the uploaded material. Each Assistant was tasked with generating three types of examination questions: true-false, multiple choice, and short answer. For each question, the Assistant also produced a corresponding answer key and included supporting excerpts from the textbook. After the exams were created, both Assistants were prompted to take the exams as if they were students, with and without access to the same source material. Their short-answer responses were then evaluated using three text similarity metrics. The design ensured that all answers from RAG-enabled models could be traced to specific source excerpts, supporting traceability and citation auditability in alignment with emerging policy frameworks, and is documented in an open-access Zenodo repository (Tyndall et al., 2025).

3.1. Textbook selection

The textbook used in this study had to meet specific criteria to support controlled, reproducible testing. All questions needed to be based solely on textual content, allowing the models to generate and answer questions without interpreting illustrations. The text also had to reflect undergraduate-level academic quality to align with higher education use cases. While the selected textbook met these conditions, the approach is generalizable, and any similar resource could be used to replicate or extend 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 Hawaiʻi Center for Pacific Island Studies (Mawyer et al., 2020). The book is organized into seven sections, each addressing key political and cultural issues relevant to Oceania and its peoples.

The models were instructed to follow specific criteria when generating exam questions:

- 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 responses for short-answer questions.
- Do not ask questions that require reading charts, graphs, or other illustrations.

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

Table 3. *Metrics and brief description*

Metric	Library	Description
ROUGE–1	rouge-score	Automatic summarization and machine translation software (Lin, 2004)
Cosine Similarity	scikit-learn	Used in text analytics to compare documents and determine if they are similar and how much (Supe, 2023)
Word Embeddings	spaCy	Similarity through comparing multi-dimensional representations of a word (Honnibal et al., 2020)

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.

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, using the rouge-score library, to measure the overlap of individual words between the generated answers and the reference answers (Hugging Face, 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 (Honnibal et al., 2020). 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. During initial runs, this was achieved by creating new OpenAI assistants using the beta.assistants.create() method. For subsequent runs, the same assistants were accessed using the beta.assistants.retrieve() method. 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 Supplementary Appendix A. First, they generated a specified number of true-false questions, along with

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

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 Supplementary 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 Supplementary Appendix C.

This architecture mirrors policy recommendations that emphasize source-grounded outputs, citation transparency, and model accountability, as seen in the EU AI Act and NIST AI RMF. Because the generation and answer formats were structured with modular prompts, this workflow can be adapted for other subjects, levels, or textbooks with minimal modification. It provides a reproducible template for regulated educational use that prioritizes auditability and fairness.

4. Results and 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. Notably, when textbook access was restricted, the performance gap between GPT-3.5 Turbo and GPT-4 Turbo narrowed. This suggests that while GPT-4 Turbo has stronger retrieval capabilities, both models are capable of generating logically consistent answers using their embedded knowledge alone.

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. 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

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

Table 4. *Model performance for one exam*

Model	Scoring	GPT-3.5 Turbo assistant exam		GPT-4 Turbo assistant exam	
GPT–3.5 Turbo Assistant with RAG	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
GPT–4 Turbo Assistant with RAG	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
ChatGPT–3.5 Turbo without RAG	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
ChatGPT–4 Turbo without RAG	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

of the reference material underlining GPT-4 Turbo'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.

Table 5. *Model rankings per exam*

Model	Exam	Ranking
GPT–3.5 Turbo Assistant with RAG	GPT–3.5 Turbo	1
GPT–4 Turbo Assistant with RAG	GPT–4 Turbo	2
GPT–4 Turbo Assistant with RAG	GPT–3.5 Turbo	3
ChatGPT–4 Turbo without RAG	GPT–4 Turbo	4
ChatGPT–4 Turbo without RAG	GPT–3.5 Turbo	5
GPT–3.5 Turbo Assistant with RAG	GPT–4 Turbo	6
ChatGPT–3.5 Turbo without RAG	GPT–4 Turbo	7
ChatGPT–3.5 Turbo without RAG	GPT–3.5 Turbo	8

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

Table 6. *Model overall rankings*

Model	Ranking
GPT–4 Turbo Assistant with RAG	1
GPT–3.5 Turbo Assistant with RAG	2
ChatGPT–4 Turbo without RAG	3
ChatGPT–3.5 Turbo without RAG	4

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, where a lower rank indicates better performance. While the base GPT-4 Turbo model generally outperformed its GPT-3.5 counterparts, the GPT-3.5 Turbo Assistant with RAG achieved the top ranking in one exam. This suggests that RAG can significantly enhance the performance of a less powerful model, allowing it to outperform more advanced models under certain conditions.

When examining the ranked performance in Table 6 the GPT-4 Turbo Assistant with RAG ranked highest in overall performance. Notably, both assistant models that had access to the textbook through RAG (GPT-4 Turbo Assistant and GPT-3.5 Turbo Assistant) outperformed their non-RAG counterparts. This highlights the substantial performance boost provided by retrieval-augmented generation, even when applied to a less powerful model. The consistent strength of the GPT-4 Turbo Assistant across all evaluation metrics demonstrates not only the robustness of the base model but also the importance of integrating external knowledge sources in exam generation and answering tasks.

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).

The comparative analysis highlights the implications of evolving model architectures and training environments in the development of AI applications for academic testing. The observed performance differences, particularly the implementation of RAG, demonstrate how model enhancements can directly influence assessment quality. These findings suggest 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 highlights its utility in structured academic testing environments.

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

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 GPT-4 Turbo models performed the best compared to their GPT-3.5 Turbo counterparts, which was anticipated as their complexity (number of parameters) is an order of magnitude greater. What was less expected was that, while GPT-3.5 Turbo did not outperform GPT-4 Turbo, it demonstrated a level of performance that was sufficiently close in several metrics. GPT-3.5 Turbo performed particularly well when analyzing summarization metrics. Even though it was not better, GPT-3.5 Turbo performed relatively close enough to GPT-4 Turbo to suggest that individuals without the resources to purchase a subscription could use an equivalent process with the model effectively, provided there is additional oversight. Even with true-false and multiple-choice questions, the GPT-3.5 Turbo model often fell only one question behind. More testing is needed to determine exactly how far the GPT-3.5 Turbo model lags behind the GPT-4 Turbo model in this context. However, these findings suggest that AI accessibility remains a key consideration for educational policymakers, since reliance on premium models may unintentionally increase resource disparities among students and institutions.

What seemed far more important than which specific GPT model was being used was whether the model was utilized as an assistant, which allows it to leverage RAG and the textbook as a source. As shown in the analysis, the GPT-3.5 Turbo Assistant performed better overall than the ChatGPT-4 Turbo model, which did not have access to the college textbook. While this study did not directly test for hallucinations, citations were manually verified to ensure that model responses were grounded in the textbook and accurately reflected its content. In doing so, it was found that RAG-supported models consistently produced valid and contextually appropriate citations, whereas non-RAG models occasionally generated responses that lacked clear textual grounding. This suggests a higher likelihood of hallucinations or outdated information when models rely solely on pre-trained knowledge. By contrast, models utilizing RAG consistently retrieved accurate and relevant information to generate their responses. 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. Although this study focused on a narrow setup with two models and one textbook, the results highlight workflow decisions that remain relevant across many platforms. Retrieval grounding and citation transparency, in particular, offer practical solutions to issues of accuracy, cost, and fairness. Additionally, policymakers and accreditation bodies should consider establishing guidelines for the responsible use of AI-assisted exam generation to maintain transparency, academic integrity, and equitable access to AI-driven tools across diverse educational institutions.

5.1. Analytical framework

The rapid turnover of foundation models means that any label, such as OpenAI's "GPT-4 Turbo" will date quickly, whereas certain workflow choices remain stable across model generations. Two such choices appear to determine most of the educational risk surface:

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

Table 7. *AI-assisted exam workflow archetypes*

	Closed-book	Retrieval-grounded
Automatic	Base model writes and grades exams without external material.	RAG model extracts textbook passages without human review.
Human-in-the-loop	Base model to writes exams; professor manually edits questions and grades responses.	RAG model drafts and grades exams; professor reviews flagged answers.

- Knowledge-grounding strategy. A system can rely on internal weights (“closed-book”) or retrieve passages from an external, auditable corpus (“retrieval-grounded”).
- Oversight locus. Assessment can be scored entirely automatically or require a human decision point at one or more stages (“human-in-the-loop”).

Crossing these axes yields a 2 by 2 grid that captures the design space for AI-assisted exams (see Table 7). The closed-book with automatic quadrant corresponds to public chatbot deployments that generate items and grades without citations. Retrieval-grounded with automatic systems resembles the pipeline evaluated in this study, where the model must attach a source snippet to every answer. The quadrants that introduce human review add formal guardrails. In the closed-book with human-in-the-loop scenario, faculty must validate questions that originate only from the model’s internal knowledge before they are presented to students. Retrieval-grounded with human-in-the-loop pipelines require educators to approve both the retrieved material and the generated answer at every step.

This framework helps decouple policy decisions from the pace of vendor-specific advancements. Regulators and institutions can ask which quadrant a proposed tool occupies and whether its safeguards match the residual risks. For example, requiring retrieval source traceability through cited references addresses many accuracy concerns even before considering the base model. Likewise, mandating faculty sign-off moves a system into the higher-trust human-in-the-loop category regardless of the architecture implemented. By organizing oversight strategies around this grid, institutions can adopt a consistent policy framework that scales across tools and academic contexts.

5.2. Framework implications

Mapping exam-generation workflows onto the two design choices yields four archetypes:

Empirical results occupy the bottom-right cell that combines retrieval grounding with human-in-the-loop oversight. Relative performance advantages, including higher ROUGE-1 and BLEU scores and a lower hallucination rate, suggest that retrieval and oversight reinforce each other. Their combination improves both accuracy and traceability, supporting workflows that are better aligned with policy goals related to quality assurance and transparency.

Quality control of the knowledge base remains essential. Accuracy improves only when sources are version-controlled and subject to peer review. A viable governance protocol should implement semantic versioning for each corpus release, maintain source traceability metadata for every paragraph, and require periodic faculty audits focused on coverage gaps. These steps enable traceability, reproducibility, and institutional accountability.

Data privacy considerations also arise, since textbook excerpts may be under copyright or protected by privacy regulations such as Family Educational Rights and Privacy Act (FERPA) and General Data Protection Regulation (GDPR) (Azam et al., 2024; Kelso et al., 2024; Cooper et al., 2025). Retrieval APIs should enforce in-place querying so that student prompts are processed within a secure institutional environment and responses are streamed without being stored externally (Zhao,

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

2024; Mithun et al., 2025). This safeguards both legal compliance and student trust (Shrestha et al., 2024).

Equitable access depends on cost as well as accuracy. According to OpenAI's published pricing, GPT-3.5 Turbo is substantially less expensive per token than GPT-4 Turbo (*Pricing*, n.d.). When paired with retrieval-augmented generation, GPT-3.5 delivered performance approaching that of GPT-4 across several metrics in this study. This makes it a viable option for institutions operating under budget constraints, especially if procurement guidelines emphasize retrieval transparency and traceability over use of the most advanced proprietary model.

Grounding also reduces hallucinations tied to contextual bias by ensuring that all answers are traceable to a citable excerpt. Manual citation auditing confirmed this benefit. These safeguards promote fairness and consistency, especially in courses involving sensitive or contested subject matter. The remaining three cells in the matrix represent opportunities for further research and policy development rather than regulatory gaps. By centering workflow design choices rather than model branding, institutions can apply this framework regardless of which LLM vendor or architecture is implemented.

5.3. Implementation considerations

The adoption of retrieval-grounded language models for educational assessment raises several operational considerations that institutions must address to ensure responsible and sustainable deployment (Lewis et al., 2021; Gan et al., 2025; Ni et al., 2025).

First, knowledge-base curation and quality control must be formalized. Each retrievable passage should include source traceability metadata, version information aligned with the associated course materials, and a defined review schedule approved by instructional staff (Gan et al., 2025; Ni et al., 2025). Institutions should consider implementing automated content audits to flag obsolete information or inconsistencies, allowing faculty to revise the corpus before assessments are administered (Lewis et al., 2021). These practices are essential to maintain alignment with current academic standards and to ensure transparent and traceable decision-making.

Second, data privacy and copyright compliance are central to system design. Legal constraints such as FERPA and GDPR restrict how educational data and copyrighted content may be used. A layered access model can help address these concerns (Koga et al., 2025; Ni et al., 2025). Students should only receive the specific excerpt used to generate or justify an answer, while faculty maintain access to the full corpus for auditing and accreditation. Retrieval tools should be configured to support in-place querying so that prompts and document content are processed within secure institutional environments rather than external servers (Koga et al., 2025).

Third, equity and resource allocation should guide infrastructure decisions. Institutions with constrained budgets may not be able to afford high-end proprietary models or high-volume API usage. Retrieval-grounded workflows can help offset these limitations by boosting the effectiveness of lower-cost models (Li et al., 2024; An et al., 2025). Regardless of vendor, institutions should prioritize systems that enable verifiable source citations and offer sufficient control over token or compute budgets to avoid unexpected cost spikes during exam generation and scoring (An et al., 2025).

Finally, fairness and bias mitigation require ongoing oversight. Even when using a vetted corpus, retrieval-grounded systems may reflect cultural, linguistic, or ideological biases present in the source material (Dai et al., 2024; Guo et al., 2024). Institutions should establish routine audits of both prompts and retrieved content, supported by faculty review and student feedback channels (Ni et al., 2025). An appeals process for AI-scored responses can further promote procedural fairness and transparency.

These implementation considerations are not tied to any specific model or domain and can be adapted across disciplines and technical platforms. By embedding these safeguards into institutional workflows, universities and other educational organizations can support AI-assisted assessment in a way that aligns with evolving regulatory expectations and ethical standards (Ni et al., 2025).

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

5.4. Strategic and policy implications

As educational institutions evaluate how to integrate LLMs into academic assessment, this study highlights two strategic design choices that remain stable across rapid model iteration: (1) coupling the system with RAG to ground answers in auditable source material, and (2) enforcing human-in-the-loop oversight at key decision points. These principles are architecture-agnostic and vendor-neutral, making them a durable basis for institutional policy, procurement standards, and accreditation frameworks.

Rather than focus on proprietary model benchmarks, institutions should assess whether an AI-assisted exam system provides traceable citations, retrieval transparency, and override controls for educators. For example, retrieval grounding improved answer accuracy more consistently than upgrading from GPT-3.5 Turbo to GPT-4 Turbo in the absence of RAG. This suggests that transparent retrieval pipelines can raise the performance floor for less expensive models and reduce reliance on high-cost commercial APIs. It also underscores the importance of investing in retrieval infrastructure, versioned knowledge bases, and citation audit trails rather than focusing exclusively on model specifications.

These findings carry important equity implications. Because RAG can elevate the quality of responses from lower-cost models, institutions with constrained budgets can still achieve acceptable validity and reliability targets. Policymakers can mitigate resource disparities by requiring that AI-assisted assessment workflows be reproducible on at least one open-weight or low-cost model. This approach echoes the risk-based logic of the EU AI Act and the documentation principles in the NIST AI RMF, both of which prioritize accountability over model branding.

The study also emphasizes that hallucination is not primarily a function of model size. Non-RAG models occasionally produced plausible-sounding but unsupported answers, while RAG-enabled models consistently retrieved relevant and citable textbook content. Because hallucinations result from the probabilistic nature of generative decoding, the presence of curated source material is a more effective safeguard than simply using a larger model. Embedding retrieval into the workflow is, therefore, a general mitigation strategy applicable across future architectures.

To operationalize these insights, several oversight mechanisms can be incorporated into institutional and accreditation policy:

1. Target oversight where it matters most. The largest accuracy gap was found in short-answer responses, where RAG and non-RAG implementations differed by an average of approximately 6.6 ROUGE-1 points. In contrast, true-false performance remained consistently high across implementations. Institutions should therefore prioritize faculty review or automated similarity checks for open-ended answers, rather than applying them uniformly across all question formats. This targeted oversight reduces workload without compromising quality.
2. Use retrieval confidence as a risk flag. Non-RAG models offered no basis for evaluating confidence, but RAG-enabled pipelines allow inspection of top-k retrieval scores. A practical protocol is to route answers with low retrieval confidence to faculty for manual grading. This threshold-based approach aligns with risk management principles and avoids blanket intervention.
3. Define model-agnostic procurement baselines. A GPT-3.5 Turbo model with RAG outperformed GPT-4 Turbo without RAG on several exams. Institutions can use this configuration as a public baseline when evaluating vendor proposals. Systems that fail to meaningfully outperform this setup should not justify premium pricing. This benchmarking method encourages both competition and transparency.
4. Redirect budgets toward open, verifiable infrastructure. Because RAG offers a higher return on investment than simply upgrading model tiers, educational funding should support the development of well-maintained retrieval frameworks and curated knowledge corpora. This also promotes the use of openly licensed educational materials that benefit the wider academic community.
5. Ensure auditability and due process. Traceable answers allow students and faculty to challenge or verify AI-generated outcomes. Accreditation bodies could require that every exam item include a structured citation format that supports both automated processing, such as machine-readable

https://doi.org/10.1017/dap.2025.10024 Published online by Cambridge University Press

metadata, and manual review, such as human-readable citations. Assessment systems should also support comprehensive record-keeping and review, including the ability to log retrieved source excerpts, track model responses over time, and generate audit reports that faculty or administrators can inspect during grade disputes or accreditation reviews.

Taken together, these strategic and operational recommendations support a governance model grounded in transparency, oversight, and adaptability. Rather than focusing on model size or vendor branding, institutions can prioritize traceability, cost-effectiveness, and instructional alignment. As large language model technologies continue to evolve, this framework offers a durable foundation for responsibly integrating AI-assisted assessment systems that meet academic, legal, and equity standards.

5.5. Limitations and future research

The experiment was deliberately scoped to two OpenAI models and a single undergraduate textbook to control for confounding factors. While this limits direct generalization, the mechanisms highlighted (retrieval grounding and human oversight) are not bound to those specifics. Future work should stress-test these principles in multilingual courses, across subject domains with higher conceptual complexity, and with truly open-weight models. Such studies would provide further evidence that the policy guidance articulated here is robust to technological churn and disciplinary variation.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/dap.2025.10024.

Data availability statement. The results were derived from interactions with GPT-3.5 Turbo and GPT-4 Turbo during the first quarter of 2024. Due to the evolving nature of large language models, the outputs may no longer be reproducible with newer model iterations. However, the complete codebase and all exam-generation outputs have been deposited on Zenodo at https://doi.org/10.5281/zenodo.15769736 (Tyndall et al., 2025) as of 29 June 2025. The repository contains the Jupyter notebook that implements the assistants and evaluation pipeline, a combined exam-response CSV file with 14 exams, textbook excerpts, and model answers, and a CSV containing the summary of metrics.

Author contribution. Conceptualization: C.G., A.S.; Validation: J.G.; Project administration: E.T., T.W., B.L.; Writing: original draft: C.G., A.S., E.T.; Writing: review and editing: E.T., T.W., B.L.; All authors approved the final submitted draft.

Funding statement. This work received no specific grant funding.

Competing interests. The authors declare no potential conflicts of interest with respect to the research, authorship and/or publication of this article. 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 WPAFB-2024-0625.

References

Akgun S and 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/.

An Y, **Cheng Y**, **Park SJ and Jiang J** (2025) HyperRAG: Enhancing quality-efficiency tradeoffs in retrieval-augmented generation with Reranker KV-Cache reuse. *arXiv*. https://doi.org/10.48550/arXiv.2504.02921.

Awan AA (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.

Azam N, **Michala AL**, **Ansari S and Truong N** (2024) M*odelling Technique for GDPR-compliance: Toward a Comprehensive Solution.* arXiv. https://doi.org/10.48550/arXiv.2404.13979.

Cooper AF, **Gokaslan A**, **Cyphert AB**, **Sa CD**, **Lemley MA**, **Ho DE and Liang P** (2025) Extracting memorized pieces of (copyrighted) books from open-weight language models. *arXiv*. https://doi.org/10.48550/arXiv.2505.12546.

Dai S, **Xu C**, **Xu S**, **Pang L**, **Dong Z and Xu J** (2024) Bias and unfairness in information retrieval systems: New challenges in the LLM Era. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. pp. 6437–6447. https://doi.org/10.1145/3637528.3671458

Digital Education Council (2024) *EU AI Act: What It Means for Universities.* The Digital Education Council. https://www.digitaleducationcouncil.com/post/eu-ai-act-what-it-means-for-universities.

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/.

Emmanuel C (2023, August 3) GPT-3.5 and GPT-4 Comparison: *Medium*. https://medium.com/@chudeemmanuel3/gpt-3-5-and-gpt-4-comparison-47d837de2226.

Gan A, **Yu H**, **Zhang K**, **Liu Q**, **Yan W**, **Huang Z**, **Tong S and Hu G** (2025) *Retrieval Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey.* arXiv. https://doi.org/10.48550/arXiv.2504.14891.

Gao Y, **Xiong Y**, **Gao X**, **Jia K**, **Pan J**, **Bi Y**, **Dai Y**, **Sun J**, **Wang M and Wang H** (2024) *Retrieval-Augmented Generation for Large Language Models: A Survey.* https://arxiv.org/pdf/2312.10997.

Guo Y, **Guo M**, **Su J**, **Yang Z**, **Zhu M**, **Li H**, **Qiu M and Liu SS** (2024) *Bias in Large Language Models: Origin, Evaluation, and Mitigation.* arXiv. https://doi.org/10.48550/arXiv.2411.10915.

Guinet G, **Omidvar-Tehrani B**, **Deoras A and Callot L** (2024) *Automated Evaluation of Retrieval-Augmented Language Models with Task-Specific Exam Generation*. arXiv. https://doi.org/10.48550/arXiv.2405.13622.

Holistic AI (2024) *Elements of the NIST AI RMF: What You Need to Know.* Holistic AI. https://www.holisticai.com/blog/nist-ai-rmf-core-elements.

Honnibal M, **Montani I**, **Van Landeghem S and Boyd A** (2020) *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo. https://doi.org/10.5281/zenodo.1212303.

Hugging Face (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/.

Kelso E, **Soneji A**, **Rahaman S**, **Soshitaishvili Y and Hasan R** (2024) *Trust, Because You Can't Verify: Privacy and Security Hurdles in Education Technology Acquisition Practices.* arXiv. https://doi.org/10.48550/arXiv.2405.11712.

Klang E, **Portugez S**, **Gross R**, **Brenner A**, **Gilboa M**, **Ortal T and 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.

Koga T, **Wu R and Chaudhuri K** (2025) *Privacy-Preserving Retrieval-Augmented Generation with Differential Privacy.* arXiv. https://doi.org/10.48550/arXiv.2412.04697.

Lewis P, **Perez E**, **Piktus A**, **Petroni F**, **Karpukhin V**, **Goyal N**, **Küttler H**, **Lewis M**, **Yih W-t**, **Rocktäschel T**, **Riedel S and Kiela D** (2021) *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv.* https://arxiv.org/abs/2005.11401.

Li Z, **Li C**, **Zhang M**, **Mei Q and Bendersky M** (2024) *Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach.* arXiv. https://doi.org/10.48550/arXiv.2407.16833,

Lin CY (2004) *ROUGE: A Package for Automatic Evaluation of Summaries.* https://www.aclweb.org/anthology/W04-1013.

Mawyer A, **Auelua R**, **Aikau H**, **Barcham M**, **Boeger Z**, **Dawrs S**, **Genz J and Kava L** (2020) *Introduction to Pacific Studies*. Honolulu: Center for Pacific Islands Studies, University of Hawaiʻi-Mānoa.

Mithun P, **Noriega-Atala E**, **Merchant N and Skidmore E** (2025) *AI-VERDE: A Gateway for Egalitarian Access to Large Language Model-Based Resources For Educational Institutions.* arXiv. https://doi.org/10.48550/arXiv.2502.09651.

Ni B, **Liu Z**, **Wang L**, **Lei Y**, **Zhao Y**, **Cheng X**, **Zeng Q**, **Dong L**, **Xia Y**, **Kenthapadi K**, **Rossi R**, **Dernoncourt F**, **Tanjim MM**, **Ahmed N**, **Liu X**, **Fan W**, **Blasch E**, **Wang Y**, **Jiang M and Derr T** (2025) *Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey (No. arXiv:2502.06872; Version 1).* arXiv. https://doi.org/10.48550/arXiv.2502.06872.

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.

OpenAI (n.d.) *Pricing.* Other Models. https://platform.openai.com/docs/pricing#other-models.

Ray PP (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.

Shrestha AK, **Barthwal A**, **Campbell M**, **Shouli A**, **Syed S**, **Joshi S and Vassileva J** (2024) *Navigating AI to Unpack Youth Privacy Concerns: An In-Depth Exploration and Systematic Review.* arXiv. https://doi.org/10.48550/arXiv.2412.16369.

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.* https://www.turing.com/resources/minimize-llm-hallucinations-strategy.

Tyndall E, **Gayheart C**, **Some A**, **Wagner T**, **Langhals B and Genz J** (2025) *Undergraduate Pacific Studies Exam Generation and Answering Using Retrieval Augmented Generation and Large Language Models.* Zenodo. https://doi.org/10.5281/zenodo.15769737.

Zhao D (2024) *FRAG: Toward Federated Vector Database Management for Collaborative and Secure Retrieval-Augmented Generation.* arXiv. https://doi.org/10.48550/arXiv.2410.13272.

Cite this article: Tyndall E, Gayheart C, Some A, Genz J, Wagner T and Langhals B (2025). Impact of retrieval augmented generation and large language model complexity on undergraduate exams created and taken by AI agents. *Data & Policy*, 7: e57. doi:10.1017/dap.2025.10024