education sciences

Article

ChatGPT: The End of Online Exam Integrity?

Teo Susnjak [1,*] **and Timothy R. McIntosh** [2]

[1] School of Mathematical and Computational Sciences, Massey University, Auckland 0632, New Zealand
[2] School of Computing, Engineering and Mathematical Science, LaTrobe University, Melbourne, VIC 3086, Australia
* Correspondence: t.susnjak@massey.ac.nz

Abstract: This study addresses the significant challenge posed by the use of Large Language Models (LLMs) such as ChatGPT on the integrity of online examinations, focusing on how these models can undermine academic honesty by demonstrating their latent and advanced reasoning capabilities. An iterative self-reflective strategy was developed for invoking critical thinking and higher-order reasoning in LLMs when responding to complex multimodal exam questions involving both visual and textual data. The proposed strategy was demonstrated and evaluated on real exam questions by subject experts and the performance of ChatGPT (GPT-4) with vision was estimated on an additional dataset of 600 text descriptions of multimodal exam questions. The results indicate that the proposed self-reflective strategy can invoke latent multi-hop reasoning capabilities within LLMs, effectively steering them towards correct answers by integrating critical thinking from each modality into the final response. Meanwhile, ChatGPT demonstrated considerable proficiency in being able to answer multimodal exam questions across 12 subjects. These findings challenge prior assertions about the limitations of LLMs in multimodal reasoning and emphasise the need for robust online exam security measures such as advanced proctoring systems and more sophisticated multimodal exam questions to mitigate potential academic misconduct enabled by AI technologies.

Keywords: AI impact on education; ChatGPT in exam cheating; multimodal online assessments; GPT-4 vision evaluation; educational integrity; generative AI; examination security; large language models in assessments; critical thinking in LLMs; multi-hop reasoning in LLMs; multimodal chain-of-thought reasoning

check for updates

Citation: Susnjak, T.; McIntosh, T.R. ChatGPT: The End of Online Exam Integrity? *Educ. Sci.* **2024**, *14*, 656. https://doi.org/10.3390/educsci14060656

Academic Editor: Daniel Muijs

Received: 18 April 2024
Revised: 1 June 2024
Accepted: 13 June 2024
Published: 17 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The landscape of higher education has significantly transformed towards online learning which has been notably accelerated by the recent pandemic [1]. Higher Education Institutions (HEIs) worldwide have swiftly adapted to this new norm, transitioning to online classes and examinations to overcome these challenges [2–4]. This shift is expected to endure, driven by the recognised advantages of remote learning for both educational institutions and students [1,5].

With the increased adoption of online education, concerns regarding academic integrity have intensified [1,6,7]. Online assessments, in particular, have highlighted the potential for increased cheating and academic misconduct [4,8–11] fueled by factors like anonymity, reduced supervision, and easier access to unauthorised resources during exams. Cheating in HEIs is a longstanding issue and has been compounded by the online modality, with faculty facing significant barriers to effectively countering dishonest practices [12,13]. Empirical research indicates a notable rise in academic dishonesty within online settings. For instance, Malik et al. [14] observed a sudden and significant increase in academic performance among students during the pandemic, attributing it to the facilitation of cheating in online exams, while Newton and Essex [15] noted recently that cheating in online exams is prevalent, emphasising the allure of the opportunity for dishonesty in these settings. Historical data corroborate the extensive nature of academic cheating, with studies revealing a significant proportion of students admitting to such behaviour [16,17].

These concerns around negative trends in academic integrity in online exams were already well advanced when ChatGPT was released, and a new and significant compounding factor was introduced [18–21]. ChatGPT's ability to answer exam questions has already been demonstrated in multiple studies [21–23]. The opportunity to cheat is now magnified making it easier than ever before, while the current AI technologies are more effective at enabling this than anything else beforehand. Richards et al. [24] showed that it is not even necessary for students to augment ChatGPT responses with their own work or edit its outputs to reach at least a passing grade, while students who augment the responses with their own additional material are highly likely to score even higher on their assessments. Newton and Essex [15] confirm what is self-evident, that students appear to be most likely to cheat in online exams when there is an opportunity to do so. In spite of this, studies [25] continue to claim that even unproctored online exams can still provide meaningful assessments of student learning, and that AI text-detection tools are accurate and effective deterrents [26]. Much has been made of limitations of Large Language Models (LLMs) like ChatGPT due to their proclivity to fabricate non-existent facts or inappropriate information [22,27–29] referred to as hallucination, and with this comes the temptation to downplay and dismiss the performance capabilities of LLMs in exam contexts. It is also true that LLMs still encounter persistent difficulties across complex reasoning tasks [30] from which it is argued and claimed that due to this, they are again unsuitable for answering sophisticated exam questions with reliability [31,32]. While it is therefore possible to reduce the threats posed by LLMs on these technical grounds and limitations, some studies altogether dismiss the need for additional mitigation strategies that attempt to prevent cheating based on principle [33] and instead shift the blame onto the faculty and other externalities instead, which effectively absolves students from any personal responsibility.

This matter is clearly of some contention; however, there are numerous studies that do take the urgency to devise additional and more effective countermeasures in this new context seriously. Studies have discussed mitigations like the use of LLM text-detection tools [34]. In an effort to uphold academic integrity HEIs have implemented various measures, including technological interventions like digital proctoring systems [6,35,36]. Other studies have called for recasting examination and assessment questions so that they require higher-order thinking and critical reasoning skills [37,38] instead of merely factual recall. Others [39–41] have recommended integrating multimodal approaches into exams which venture beyond just text-based questions and additionally combine visual components. Indeed, very recent studies [38,42] have suggested that LLMs do perform poorly on complex reasoning tasks that combine both visual and text-based modalities into exam questions.

Aims and Contribution

The aim of this study is to inform and highlight the increasing reasoning capabilities of the new class of AI tools like LLMs, and their effects on educational integrity. This work specifically aims to demonstrate that LLMs are in fact more capable of performing complex reasoning tasks on multiple modalities than is currently believed. This work proposes a novel multimodal, iterative self-reflective strategy that decomposes a complex reasoning task into sub-tasks which are applied to each modality first, in order to guide the LLMs towards a correct response. Therefore, this study empirically demonstrates how one of the last bastions of evading cheating via LLMs which involved constructing multimodal exam questions, is now also compromised. This study also extensively probes the visual understanding capabilities of state-of-the-art LLMs to answer multimodal exam questions across multiple subjects and disciplines seeking to probe which types of questions pose the greatest difficulties for the current LLMs. This paper also seeks to critically examine the arguments and the inadequacies of current mitigations against cheating on online exams via LLMs, like the effectiveness of detection tools for AI-generated text, while challenging the assumption that LLMs' reasoning limitations preclude effective cheating that is sufficient to pass various exams. This work finally concludes with a list of up-to-date recommendations on conducting online exams.

2. Background

The literature review was informed by three key arguments cited in the introduction that raise questions about the perceived threat of ChatGPT to academic integrity, namely, (1) the effectiveness of AI-text detection tools, (2) the propensity of LLMs to hallucinate, and (3) the limited reasoning capabilities of LLMs. This review prioritises the most recent publications since the release of ChatGPT to ensure relevance. It critically examines the challenges in distinguishing AI-generated text from human-authored content, emphasising the evolving sophistication of LLMs that increasingly evade advanced detection methods. Additionally, the review addresses the phenomenon of LLM hallucination, highlighting that despite their tendency to produce misleading information, LLMs often satisfy minimal academic assessment criteria, meaning that their occasional inaccuracies do not preclude their potential use for academic dishonesty. Furthermore, this section delves into the ongoing debates about the reasoning abilities of LLMs, acknowledging their limitations in complex reasoning tasks while also considering emerging evidence of their enhanced capabilities through advanced, multi-hop reasoning strategies.

2.1. Challenges in Detecting AI-Generated Text

Discerning AI-generated text from human-authored content continues to present formidable challenges and this is particularly true as LLMs attain unprecedented levels of stylistic fluency. The effectiveness of contemporary detection strategies is frequently undermined by the evolving sophistication of generative models which have been shown to elude both well-established and nascent detection technologies with increasing ease [43]. Generally, these systems grapple with both false positives where they mistakenly flag human-written content, and false negatives where they fail to detect AI-generated outputs [44]. Traditional methods like stylometric analyses, once reliable, now often fail to detect the subtle inconsistencies that characterise AI-generated texts. These inconsistencies include atypical semantic patterns, unusual word choices, and subtle logical inconsistencies, demonstrating the models' ability to produce text with high grammatical correctness and contextual appropriateness. The landscape is further complicated by adversarial training techniques, which allow LLMs to adapt and learn patterns that specifically avoid detection, thus challenging the efficacy of even the most advanced machine learning-based detectors [43]. This ongoing adaptation among LLMs to counter traces of AI text-generation initiates a cycle of challenge and response that characterises the arms race between AI technologies and detection methodologies [45]. Thus, AI text detection techniques that are fine-tuned for one generation of LLMs are likely to become obsolete as new models emerge and proliferate.

2.2. LLM Hallucination Is Not an Insurmountable Problem

LLMs exhibit a remarkable capacity to produce fluent, coherent, and persuasive text, often accompanied by a tendency for it to be factually incorrect, misleading, or nonsensical—otherwise referred to as hallucination. Hallucination in LLMs should not come as a surprise—LLMs hallucinate by design [28,46]. Indeed, studies [22,31,38,47,48] have identified this as a major obstacle in using LLMs for producing correct and reliable responses in various critical contexts, and this also translates to question-answering tasks within high-stakes online exams. While LLM hallucinations pose a challenge, their severity and frequency may be less problematic than initially presumed—at least in exam contexts. LLM hallucination is more of an impediment in exams where questions require a faithful recall of facts from subject areas that are underrepresented in the LLM training datasets [46] which is not an issue in most cases. Thus, the impact of hallucination on cheating depends on the nature of the assessment and the narrowness of a subject domain. Highly specialised exams on more esoteric topics which prioritise recall and memorisation may be more vulnerable, as plausible-sounding fabrications might go undetected by students.

Conversely, tasks requiring in-depth syntheses, deeper reasoning, or source verification present greater challenges for LLMs [32,49]. It is worthwhile acknowledging that cheating often aims to satisfy minimal requirements rather than achieve absolute correctness. This means cheaters often seek to produce work that appears sufficiently knowledgeable to pass, rather than striving for an in-depth understanding or absolute accuracy. This perspective is essential when considering AI-assisted cheating. LLMs excel at generating plausible and sometimes merely superficially correct output, aligning with the needs of those seeking to bypass genuine learning. Poorly designed assessments emphasising rote memorisation, simplistic short answers, or basic procedural knowledge are particularly susceptible as LLMs can now sufficiently fulfil these minimal requirements. Even if the possibility of LLM hallucination is deemed to be likely in certain assessments, recent studies demonstrate that with careful strategies and prompting techniques, hallucination can to a large degree be mitigated, though not entirely eliminated [22,50–53].

2.3. Reevaluating the Critiques of the Reasoning Capabilities of LLMs

Perceptions that advanced LLMs exhibit limited reasoning abilities still prevail [31,32,49]. This is also maintained across numerous studies exploring LLMs' complex reasoning capabilities in handling academic content. Yeadon and Halliday [54] suggest that while LLMs can function adequately on elementary physics questions, they falter with more advanced content, novel methodologies not included in standard curricula and basic computational errors. Singla [55] noted that while LLMs like GPT-4 demonstrate proficiency in text-based Python programming, they struggle significantly with visual programming tasks that require a synthesis of spatial, logical, and programming skills. Frequently, LLMs have been observed to under-perform in tasks that necessitate a deep integration of diverse cognitive skills, especially in math-intensive subjects across various languages [42].

However, while there are valid concerns regarding LLMs' current limitations in handling complex reasoning tasks, there is also accumulating evidence of their improving capabilities [38,56,57]. Within an academic context, Liévin et al. [58] conclude that LLMs can effectively answer and reason about medical questions, while a recent survey Chang et al. [30] indicates that LLMs perform well in tasks like arithmetic reasoning and demonstrate marked competence in logical reasoning tasks too; though, they do encounter significant challenges with abstract and multi-hop reasoning, struggling particularly with tasks requiring complex, novel, or counterfactual thinking. The ability to self-critique is necessary for advanced reasoning that supports rational decision-making and problem-solving, and Luo et al. [59] demonstrate the difficulty of achieving this within LLMs; however, they show how an improvement in LLM's performances on reasoning tasks can be elicited through advanced prompting techniques involving self-critique.

Indeed, studies are beginning to demonstrate how LLMs are becoming more competent at reasoning, and they are also illustrating how effective reasoning can be elicited from LLMs through more sophisticated strategies and prompting techniques that are particularly useful for complex tasks requiring multi-step approaches that decompose the problems into sub-tasks. The next frontier of reasoning complexity is the ability of LLMs to reason across multiple modalities of inputs, where the inputs comprise both text and visualisations and will eventually include additional multimedia inputs. Research in this space has only just begun to emerge and has consistently been indicating LLMs' limitations in their ability to integrate multimodal reasoning accurately. While multi-step reasoning involves a sequential process to reach a conclusion, building logically on each step, multi-hop reasoning entails making several inferential leaps among unlinked data points or different modalities to piece together an answer. Feng et al. [60] concluded that GPT-4 struggled to retain and process visual information in combination with textual inputs that require multi-hop reasoning, and likewise, Pal and Sankarasubbu [47] found that LLMs generally had variable performances in visual and multimodal medical question-answering tasks, with some under-performing significantly and exhibiting specific deficits, particularly in areas requiring intricate reasoning in medical imaging. Similar findings were arrived at by Stribling et al. [38], who assessed the capability of GPT-4 to answer questions from nine graduate-level final examinations in the biomedical sciences where GPT-4 again performed poorly on questions based on figures. Given the research interest in this area, Zhang et al. [42] created the first benchmark dataset for evaluating LLMs on multimodal multiple-choice questions, concluding that state-of-the-art LLMs struggled with interpreting complex image details in university-level exam questions and overall performed poorly. However, all studies exploring LLMs' multimodal capabilities did not consider multi-hop decomposition strategies that attempted to access the latent and sophisticated reasoning capabilities of the LLMs.

2.4. Summary of Literature and Identification of Research Gaps

Current research indicates that concerns regarding LLMs' tendency to produce hallucinated content, while recognised as a limitation, cannot significantly deter their use for effective cheating in most instances. Moreover, while LLMs exhibit certain limitations in complex reasoning tasks, there is clear emerging evidence of their enhanced capabilities, particularly when effectively prompted through more sophisticated strategies. These strategies, which involve decomposing complex textual tasks into simpler, multi-step stages, have proven effective in eliciting higher-order reasoning from LLMs. LLMs excel in tasks that leverage their strengths in pattern recognition and structured problem-solving but have recently been shown to face significant challenges with reasoning tasks across several modalities of inputs. These currently involve integrating visual and textual inputs in which current models show notable weaknesses in contexts requiring intricate multi-hop reasoning in complex academic testing. A clear gap therefore exists in the development of approaches seeking to improve LLM reasoning on multimodal tasks that demand multi-hop problem-solving strategies.

Based on the identified gaps in the literature and the objectives of this study, the following research questions are formulated:

1. RQ1: How can an iterative self-reflective strategy enhance the reasoning capabilities of LLMs when answering complex multimodal exam questions?
2. RQ2: What are the implications of enhanced multimodal reasoning capabilities in LLMs for the integrity of online examinations?

3. Multimodal LLM Self-Reflective Strategy

This study proposes a multimodal self-reflective strategy for LLMs to demonstrate how to invoke critical thinking and higher-order reasoning within the multimodal LLM agents to steer them towards accurately responding to complex exam questions that integrate textual and visual information, but which can, in theory, be expanded to other modalities as well. The strategy follows a structured iterative process described at a high level here and demonstrated concretely in the case studies in Section 5:

1. Initial Response Evaluation: The LLM initially responds to a multimodal question, providing a baseline for its capability in interpreting, analysing and integrating all the provided modalities.
2. Conceptual Self-Reflection: The LLM is then prompted to assess its own understanding of the key concepts within the textual content of the exam question, inviting the LLM to describe and explain in greater detail its knowledge and understanding of a key concept in isolation from information contained in other modalities.
3. Visual Self-Reflection: The LLM is prompted to focus its perception and understanding on key visual cues and elements within an image and to reflect on this, again in isolation from information contained in other modalities.
4. Synthesis of Self-reflection: The LLM model is then prompted to reevaluate and revise its initial response from the first step in light of its responses arising from the self-reflective prompting.

5. Final Response Generation: The LLM responds with a final answer that uses high-order reasoning to integrate and potentially correct the initial response with a revised answer, aiming for greater accuracy, completeness and depth of understanding.

The proposed iterative self-reflexive strategy builds upon prior works. It can be seen as an expansion of similar interactive self-reflection approaches designed to mitigate hallucination in text-only inputs like that of [22], and self-familiarity of Luo et al. [51], as well as the self-critique approach by Luo et al. [59] developed to enhance critical thinking in LLMs; alongside chain-of-thought (CoT) techniques [53] that attempt to elicit LLMs to explain their reasoning step-by-step. The proposed multimodal strategy differs from prior techniques, first and foremost by being a multimodal approach, and secondly, it is more prescriptive in how it directs self-reflection.

This approach prompts an LLM in a step-wise manner to revisit and refine its initial responses through a deliberate, guided internal dialogue across conceptual and visual dimensions that constitutes a metacognitive process. This iterative approach that emphasises critical analysis and the synthesis of information aligns with educational principles of self-regulated learning and reflective practice, and the ability of LLM to expand upon, integrate, and critically assess various inputs that mirror higher-order cognitive processes such as those outlined in Bloom's Taxonomy [61].

4. Materials and Methods

The methodology adopted to evaluate the reasoning abilities of GPT-4V(ision) across multiple modalities involves two parts, with GPT-4V being selected due to its presently superior multimodal capabilities compared to alternative models [42,62]. The first part comprises case studies that test the proposed multimodal self-reflection strategy on actual university-level exam questions. The second part involved a quantitative and qualitative self-evaluation by GPT-4V on its estimated ability to answer exam questions that contain textual descriptions of visualisations.

4.1. Evaluation of the Multimodal Self-Reflection Strategy

The initial phase of the methodology involved conducting detailed case studies to test the proposed multimodal self-reflection strategy. A university-level exam question from the field of Finance and another from Computer Science served as case studies. Each question was designed to invoke a high level of reasoning across both textual and visual representations from students, thus allowing for the observation of how GPT-4V applies its multimodal reasoning capabilities in real-world scenarios, and how the proposed strategy can guide the LLM toward a correct answer through self-reflection. Subject experts were used to evaluate the correctness of the responses The first author was one of the evaluators, having devised the Computer Science question for use in an actual student exam. The second subject expert in Finance is recognised in the Acknowledgement section.

4.2. Comprehensive Multimodal Question Assessment

Beyond the two case studies, the methodology extended to a broader evaluation involving a dataset of 600 multimodal exam questions. These questions were generated with the assistance of GPT-4V to cover 12 academic subjects. The dataset of the questions, responses and prompts are made available publicly (https://github.com/teosusnjak/multimodal-chatgpt-exam-evaluation). In this set of experiments, the exam questions were in a text-only format, describing the nature of each exam question together with a description of figures, graphs, diagrams, images, charts, and tables. This approach followed other similar approaches from literature [55,60], whereby instead of directly providing images embedded in the questions themselves, they instead provided descriptions of those images alongside the question itself to GPT-4V for a response.

Each question was designed to mirror the complexity and scope encountered in university-level academic examinations, thus providing an estimated measure of GPT-4V's visual and text-based reasoning proficiency. Academic subjects were selected for both their diversity and general popularity. These subjects were as follows: Computer Science, Engineering, Nursing, Biology, History, Communications, Education, Psychology, Marketing, Finance, Economics, and Business Administration/Management. The aim was to cover three major disciplines comprising Business, Sciences, and Humanities. Each discipline was represented by 200 exam questions and four subjects, thus aiming for an equal representation of both subjects across disciplines, and the distribution of exam questions per subject which can be seen in Table 1.

Table 1. Exam questions by discipline, subject and number.

Discipline	Subjects	Exam Questions
Business	Marketing	50
	Finance	50
	Economics	50
	Business Administration/Management	50
Sciences	Computer Science	50
	Engineering	50
	Nursing	50
	Biology	50
Humanities	History	50
	Communications	50
	Education	50
	Psychology	50

4.3. Proficiency Evaluation Process

GPT-4V's proficiency in handling these multimodal questions was evaluated through a dual-phase process. Initially, GPT-4V performed a self-assessment of its ability to respond accurately to each question, rating its proficiency on a scale from 0 to 100. This self-assessment phase allowed the LLM's self-perceived understanding and its ability to analyze and respond to complex multimodal data to be gauged. For each response, GPT-4V was asked to re-evaluate its initial response to arrive at the final proficiency scores. GPT-4V was also tasked with critically analysing its capability to answer each description of a multimodal question and to explain what the model's current training capacity finds challenging and easy to answer within each question. A selection of questions, proficiency scores and a self-assessment analysis by GPT-4V is shown in Table 2.

Table 2. Examples of three descriptive multimodal exam questions, including GPT-4V's self-assessment of its ability to answer the question and quantitative proficiency scores evaluating its ability to respond accurately.

Question Type (Subject)	Question and Competence Self-Evaluation	Proficiency Score
(Nursing) Surgical Outcomes Dashboard	**Question:** Analyze a dashboard displaying surgical outcomes, including success rates, complication rates, and patient satisfaction scores. Discuss how these data inform surgical quality improvement. **Self-Analysis:** GPT-4V could struggle with evaluating surgical outcomes data, particularly in understanding how success rates, complication rates, and patient satisfaction scores inform surgical quality improvement efforts.	70

Table 2. *Cont.*

Question Type (Subject)	Question and Competence Self-Evaluation	Proficiency Score
(Education) Educational Technology Tools Comparison Table	**Question:** Evaluate a table comparing various educational technology tools based on functionality, usability, and cost. Discuss how these tools can be effectively integrated into the classroom. **Self-Analysis:** While GPT-4V can compare educational technology tools, fully grasping how these tools integrate into classroom settings and enhance learning requires knowledge of pedagogical practices and technology effectiveness in education.	83
(Engineering) Automotive Aerodynamics Simulation Results	**Question:** Evaluate the results of computational fluid dynamics simulations for car models. Discuss how aerodynamics influences vehicle design and fuel efficiency. **Self-Analysis:** GPT-4V can evaluate aerodynamics simulation results but may have difficulty in fully understanding the implications for vehicle design and fuel efficiency, which require integration of fluid dynamics principles and automotive engineering practices.	80

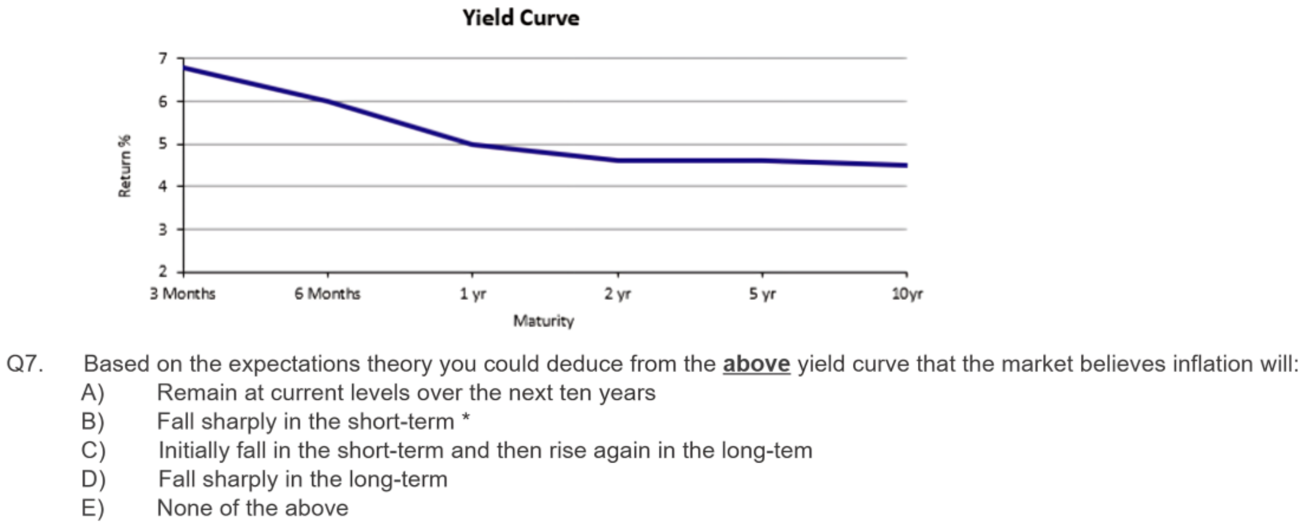
5. Results

The presentation of results begins with the illustration of the proposed method to invoke iterative self-reflection within the multimodal LLMs, with the goal of triggering and demonstrating advanced reasoning capabilities inherent within the selected LLM. Two illustrations of the multimodal self-reflection strategy are shown on exam questions from Finance and Computer Science respectively. The second part of the results section covers the results of GPT-4V's estimation of its capabilities to answer multimodal exam questions across the 12 subject areas.

5.1. Case Study—Finance

In the first example, a multiple-choice exam question from a Finance course is taken as seen in Figure 1. Initially, a baseline was established to determine if GPT-4V could answer the question correctly when presented with multiple-choice responses. The correct answer is "B", and GPT-4V correctly answered this, with the response: "The most accurate interpretation based on the shape of the yield curve would be option B, indicating that the market expects inflation to fall sharply in the short-term". With this established, the next aim was to demonstrate the efficacy of the proposed approach. Table 3 shows the Steps (1 to 4) that were carried out in a new and separate session. GPT-4V was first asked to answer the exam question without providing multiple-choice answers to test its ability to fully combine visual and conceptual reasoning, to which the response given was incorrect (Step 1). Subsequently, GPT-4V was asked to self-reflect and describe the relevant concepts (Step 2) and to self-reflect on its perception of important visual cues (Step 3) in the image. Both exercises produced sufficiently correct reflective responses, upon which, the exam question was again posed (Step 4) with GPT-4V being invited to revise its initial response based on its subsequent reflective responses. The culmination of these exercises produced the correct final answer.

The metacognitive process illustrated in Table 3, in the form of self-reflection, is being invoked within an LLM and evidenced by the model's ability to engage with questions that challenge its understanding, with an openness to revising its perspective based on new insights or information. A deliberate and structured review of the model's initial responses was demonstrated taking place at conceptual and visual perception levels. This iterative process with the emphasis on revision and improvement of understanding, based on a guided internal dialogue, captured the essence of self-reflection. Self-reflection is also intrinsically linked to critical thinking and higher-order reasoning which underscores the advanced and latent capabilities of GPT-4V. Critical thinking involves the objective analysis and evaluation of an issue in order to form a judgment; meanwhile, higher-order reasoning is characterised by the ability to understand complex concepts, apply multiple concepts simultaneously, analyze information, synthesise insights from various sources, and evaluate outcomes and approaches. In the example, the self-reflective process undertaken by the model to “expand and describe” a relevant concept, engages in a form of analysis which is one of the hallmarks of critical thinking. Meanwhile, GPT-4V’s higher-order reasoning was demonstrated when the model was prompted to integrate and synthesise separate reflections on concepts and visual information to form a comprehensive understanding of the phenomenon in question. This represents the synthesis and evaluation aspects of Bloom’s Taxonomy, which are considered higher-order cognitive skills.



Q7. Based on the expectations theory you could deduce from the **above** yield curve that the market believes inflation will:
- A) Remain at current levels over the next ten years
- B) Fall sharply in the short-term *
- C) Initially fall in the short-term and then rise again in the long-tem
- D) Fall sharply in the long-term
- E) None of the above

Figure 1. An example of a multiple-choice question from a Finance exam in the original format.

Table 3. Finance exam question case study example illustrating the step-wise invocation of self-reflection within GPT-4V that is concept and visual-reasoning based. Key aspects of the LLM responses are highlighted to identify both the incorrect parts of the response that result in the answer being evaluated as wrong, and the factually correct parts that guide the LLM to ultimately arrive at a correct final response through the proposed approach.

Step	User Prompt	GPT-4V Response
(1) Initial response requiring concept and image reasoning	“Answer the following Finance exam question based on the supplied image. What trend will the market expect that the inflation will follow?”	“The market expectation inferred from the yield curve in the image is that in the short term, inflation may rise or remain relatively high, prompting central banks to maintain or raise interest rates to control inflationary pressures. Conversely, in the long term, the expectation of lower yields suggests that the market anticipates a reduction in inflation, potentially due to economic slowdown or effective monetary policy measures leading to lower interest rates.” INCORRECT ANSWER
(2) Self-reflection focusing on concept reasoning	“Expand and describe the expectations theory in the context of yield curves and how they affect short-term and long-term inflation.”	“The expectations theory holds that the yield curve affects inflation expectations in both the short and long term, although the mechanisms and strengths of these effects may differ. Short-term rates are closely tied to the central bank’s policy rate, which is influenced by the central bank’s inflation target. If the market believes that inflation will rise above the central bank’s target, it expects the central bank to increase the policy rate to combat inflation, thus increasing short-term yields. Consequently, a steep yield curve can signal rising short-term inflation expectations. The long end of the curve reflects the market’s expectations for inflation over a longer horizon. Suppose long-term rates are low relative to short-term rates, creating an inverted curve. In that case, it can indicate that the market expects the central bank to eventually reduce policy rates in response to lower inflation or even deflationary pressures. This scenario would indicate a forecast of stable to declining long-term inflation.” [1] CORRECT
(3) Self-reflection focusing on the image reasoning	“Based on the yield curve in the image, at what point in time is the change in yield most acute?”	“The change in yield is most acute at the shorter end of the spectrum, particularly between the 3-month and 2-year maturities. This portion of the curve displays a steeper slope compared to the more gradual slope between the 2-year and 10-year maturities.” CORRECT

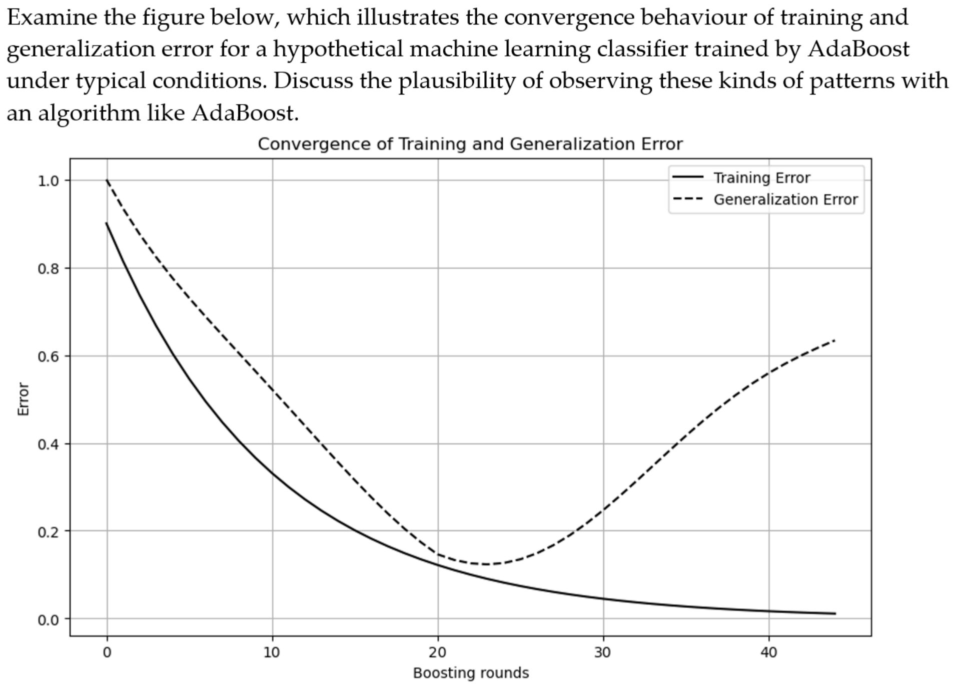
Table 3. *Cont.*

Step	User Prompt	GPT-4V Response
(4) Revised response post-self-reflection	"Based on your observations in the response and reasoning above, revise and improve your answer to the above question about the effect of yield curves on short-term and long-term inflation."	"Based on the steepness of the yield curve between the 3-month and 2-year maturities, the market appears to expect a relatively sharp decline in short-term inflation expectations. The leveling off of the yield curve beyond the 2-year mark suggests that the market anticipates more stable, modest changes to long-term inflation expectations." CORRECT ANSWER

[1] The statement while largely correct, does oversimplify the relationship between the yield curve and inflation expectations by focusing narrowly on inflation and central bank policies, while ignoring other influential factors and the complexity of economic decision-making. Nonetheless, the oversimplification is sufficient overall to generate the correct final response.

5.2. Case Study—Computer Science

The same analysis was replicated on a university-level exam question from Computer Science. This was a short answer question requiring critical reasoning, interpretation of visual patterns and their reconciliation with theoretical concepts. The question can be seen in Figure 2 and the initial incorrect response, as well as the multimodal invocation of the self-reflective strategy, can be seen in Table 4.



Examine the figure below, which illustrates the convergence behaviour of training and generalization error for a hypothetical machine learning classifier trained by AdaBoost under typical conditions. Discuss the plausibility of observing these kinds of patterns with an algorithm like AdaBoost.

Figure 2. An example of a multimodal Computer Science exam question.

The same pattern can be observed in line with the first case study example. The task of fully understanding the question, then integrating critical reasoning with the perceptive reasoning of a visual artifact, alongside the reconciliation of the observations with theoretical concepts, is too complex for multimodal LLMs to perform simultaneously (Step 1); confirming findings from other studies [30,59]. However, when the complexity is reduced and isolated to self-reflective exercises on individual modalities in turn (Steps 2 and 3), and then combined, the initial incorrect response is revised with a correct answer.

Table 4. Machine learning exam question case study example illustrating the step-wise invocation of self-reflection within GPT-4V that is concept and visual-reasoning based. Key aspects of the LLM responses are highlighted to identify both the incorrect parts of the response that result in the answer being evaluated as wrong, and the factually correct parts that guide the LLM to ultimately arrive at a correct final response through the proposed approach.

Step	**User Prompt**	**GPT-4V Response**
(1) Initial response requiring concept and image reasoning	"Examine the figure below, which illustrates the convergence behaviour of training and generalisation error for a hypothetical machine learning classifier trained by AdaBoost under typical conditions. Discuss the plausibility of observing these kinds of patterns with an algorithm like AdaBoost."	"The figure presented demonstrates a typical convergence pattern for AdaBoost, a popular ensemble learning algorithm. . . The convergence behaviour depicted in the graph is thus a plausible outcome when using AdaBoost without preventive measures against overfitting." INCORRECT ANSWER
(2) Self-reflection focusing on concept reasoning	"Tell me about the typical effects on generalisation by an algorithm like AdaBoost and, under normal circumstances how much it is likely to compromise generalizability by overfitting."	"Under standard conditions, the algorithm's strength lies in its adaptive nature, concentrating on instances that were previously misclassified and thus, when properly tuned, it can be less prone to overfitting compared to non-ensemble methods. However, overfitting with AdaBoost can occur in the presence of noise and outliers in the training data. . . In a controlled experimental setting, it is often possible to achieve generalisation error rates that are competitive with or even superior to other algorithms, provided that AdaBoost's parameters are optimised through techniques like cross-validation. However, in the presence of highly noisy data or when the assumptions underlying the algorithm are violated, the deterioration in generalisation could be considerable. Therefore, while AdaBoost has the potential for robust generalisation, it also has a non-negligible risk of compromising generalizability due to overfitting, especially when the model complexity is not adequately controlled." CORRECT
(3) Self-reflection focusing on the image reasoning	"Examine the figure carefully noting the value of the error when the overfitting you mentioned begins to take place and note carefully the point it has reached."	"The graph indicates that overfitting begins to occur just after 10 boosting rounds, where the generalisation error reaches its nadir and starts to increase. At this juncture, the training error continues to decrease, demonstrating the model's increasing fit to the training data. By the 40th boosting round, the generalisation error has increased significantly, surpassing an error rate of 0.4, which is a marked rise from its lowest point. This reflects a substantial degradation in the model's predictive accuracy on unseen data, signalling a pronounced overfitting as the model is overly optimised for the training dataset." CORRECT
(4) Revised response post-self-reflection	"Now, consider your response above about AdaBoost's effects on generalizability and combine your response with the detailed observations above regarding when overfitting started and the significant degree to which it increased. Based on these responses, revise your answer to the original exam question about the plausibility of observing these trends with AdaBoost under standard conditions where the data are not corrupted and everything else is typical."	"Upon integrating the detailed observations with the known effects of AdaBoost on generalizability, it is pertinent to re-evaluate the plausibility of the presented trends under typical, controlled conditions. When the dataset is clean and representative of the underlying distribution, AdaBoost, by incrementally correcting errors of the weak learners, should ideally show a decrease in both training and generalisation errors. The convergence pattern would typically demonstrate a plateauing of the generalisation error as the number of boosting rounds increases, indicating that the model has learned the general trends in the data without being unduly influenced by noise. . . Therefore, in a standard scenario with typical data quality and distribution, the observed pattern of the generalisation error increasing sharply after a certain point would not be considered typical for AdaBoost. . . In conclusion, the graph's indication of rapid overfitting after a relatively low number of boosting rounds is implausible in standard conditions with AdaBoost, where data are clean and well-prepared." CORRECT ANSWER

5.3. GPT-4V Multimodal Capability Estimations by Subject Area

Here, a broader and more general estimation of GPT-4V's multimodal capabilities for answering exam questions is presented. More specifically, GPT-4V's own estimations of its proficiency to answer 600 multimodal exam questions across 12 subject areas are quantified. The analysis of the means and standard deviations of the estimated proficiency scores can be seen in Figure 3 indicating GPT-4V's assessment of its competence to contextualise, visually perceive, reason about, and generate coherent responses based on multimodal inputs. When evaluating the mean proficiency scores, the disciplines of biology, history, and psychology rank highest, suggesting GPT-4V exhibits a stronger alignment with the types of visual information and analytical reasoning these fields typically employ. This could be due to the rich contextual cues present in visual materials like biological diagrams or historical timelines, which offer structured and often hierarchical information that aligns well with GPT-4V's training on pattern recognition and sequence alignment. In contrast, disciplines such as business administration and economics, which often involve complex and abstract data, may require a deeper understanding of human and market behaviours which currently present challenges, and are evidenced in lower mean proficiency scores.

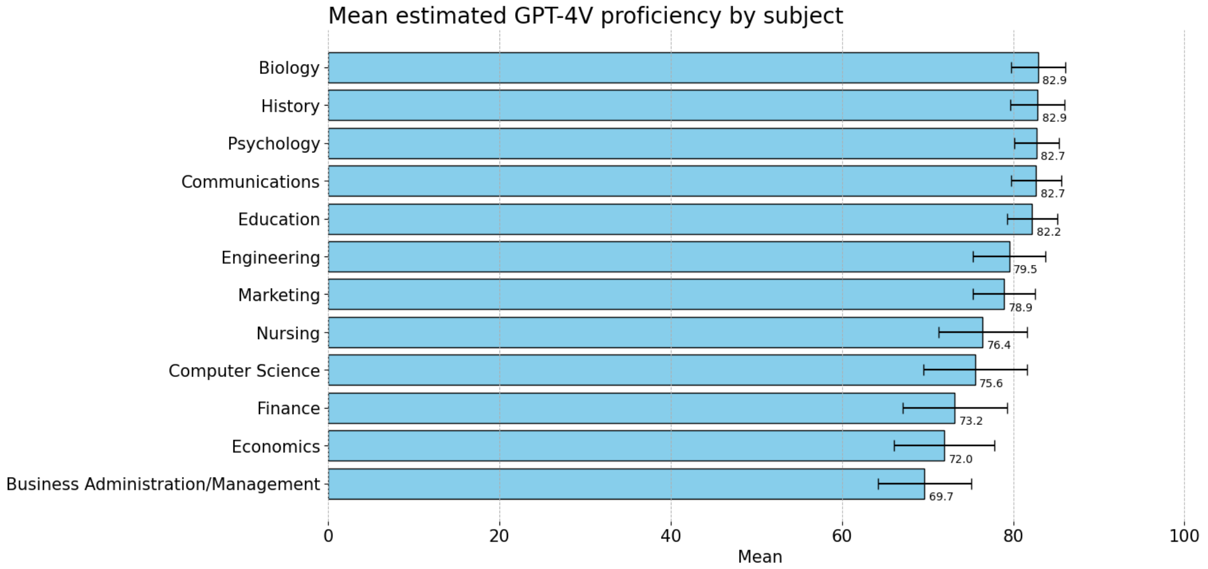


Figure 3. Mean GPT-4V estimated proficiency score across each subject in rank-order together with the standard deviations.

The standard deviation scores seen in Figure 3 indicate the consistency of GPT-4V's performance across subjects. Higher variability in fields like finance, computer science and economics suggests that GPT-4V's understanding may fluctuate significantly depending on the specificity of the task or the complexity of the visual data. This could imply that while GPT-4V can proficiently handle standard multimodal questions within these domains, it might struggle with more complex or less conventional topics, or those requiring deeper inferential reasoning. Moreover, the relatively lower variances observed in psychology, communications and education imply a more uniform proficiency across different queries within these subjects. This may be attributed to the nature of data in these fields, which frequently include human behavioural patterns and communicative structures—areas where GPT-4V has substantial experience from training datasets. Table 5 collates all the analyses across 600 responses and identifies features of multimodal exam questions per subject that GPT-4V perceives via self-assessment, to be able to handle both with high proficiency as well as with some degree of difficulty.

Finally, all the proficiency scores are aggregated by the overarching disciplines and depicted in Figure 4. The ANOVA results reveal a statistically significant difference between the means of the three disciplines, evidenced by an F-statistic of 161.2 (degrees of freedom for both the between-group variation = 2 and the within-group variation = 597, $p < 0.0001$), indicating substantial variability between the discipline groups compared to within-group variability. The Tukey HSD test revealed significant pairwise differences in mean scores between all the disciplines with all adjusted p-values below 0.001, leading to the rejection of the null hypothesis that no differences exist between the groups. In the humanities, which registers the highest proficiency score, the nature of the discipline itself is likely playing a pivotal role. Humanities topics often encompass a broad spectrum of data interpretation which points to a close alignment with GPT-4V's strengths in language understanding and integration of contextual information. The humanities' interpretive nature allows for a wider margin of acceptable responses, providing GPT-4V with a conducive environment for showcasing its capacity to draw connections between disparate historical events, socio-cultural dynamics, and philosophical concepts. Furthermore, humanities subjects often demand a high level of narrative construction, which is also well-suited to GPT-4V's design that inherently focuses on language and narrative generation. Given the narrative nature of these subjects, there is likely an inherent bias in the training data for the LLMs, with these subjects being more represented in the training corpus which yields a higher level of confidence in the ability of GPT-4V to correctly answer these questions. In contrast, science and business disciplines tend to demand more precision and this is reflected in lower scores for each. In the case of sciences, the quantitative and empirical rigidity of the field is the likely cause of the observed proficiency dip. Sciences often require precise and unequivocal interpretations, leaving less room for the breadth of interpretative responses that GPT-4V can generate well for subjects within the humanities in general. Moreover, scientific data frequently necessitate a deeper understanding of causality, experimental design, and statistical validity, which can be challenging for GPT-4V currently, as its abilities are modelled on pattern recognition rather than first principles reasoning. While GPT-4V is competent at identifying patterns and trends in scientific data, the intricacies of scientific theory and the need for detailed methodological analysis can present challenges, which might account for the slightly lower proficiency scores compared to humanities.

Table 5. An overarching summary of all the exam question performance evaluations of GPT4's strengths and limitations across all subjects.

Subject	GPT-4V's Performance Characteristics
Business Administration/ Management	Excels in extracting and summarising business-related information from visual data, recognising patterns and trends. Exhibits challenges in grasping the full scope of socio-economic contexts and generating deeper strategic insights.
Finance	Displays adeptness at quantitative financial analysis and understanding fundamental financial concepts. It encounters challenges with complex financial theories and strategic real-world applications.
Marketing	Excels in interpreting marketing data and grasping core concepts, aiding in trend identification. It struggles with deciphering more complex strategic implications and nuances of consumer psychology.
Economics	Identifies trends from economic data and understands foundational principles, but faces limitations in deeper theoretical analyses and predictive economic impacts.
Computer Science	Interprets technical diagrams and data trends in computer science contexts effectively but struggles with more sophisticated system dynamics and predictive analysis.
Nursing	Adeptly interprets data and fundamental nursing concepts but faces challenges in more demanding clinical reasoning and holistic healthcare strategy development.
Engineering	Excels in parsing engineering data and explaining technical concepts but encounters difficulties with contextual analyses and predictive evaluations.
Biology	Shows proficiency in interpreting biological data and explaining processes but struggles with understanding more complex concepts and performing predictive analysis.
Education	Parses educational data and links theories to practice well but struggles with the complexities of educational systems and multidisciplinary integration.
Psychology	Effectively interprets psychological data but struggles with more demanding constructs and forward-looking analyses that require a deeper understanding.
History	Processes historical data and concepts well but struggles with analysing more complex relationships and conducting critical evaluations.
Communications	Analyzes communication trends and strategies effectively but lacks depth in grasping socio-cultural impacts and strategic ethical considerations.

In subjects belonging to the business discipline, the lowest mean proficiency score can be seen that can be attributed to the complexity associated with the nature of business decision-making which often requires an integration of both quantitative data and human judgment. Business subjects not only involve financial and operational data interpretation but also necessitate an understanding of market dynamics, consumer behaviour, and strategic decision-making under uncertainty. These areas rely heavily on real-time data, contextual subtleties, and forward-looking predictions, which are challenging for GPT-4V given its limitations in temporal awareness and predictive modelling based on past and present data trends alone.

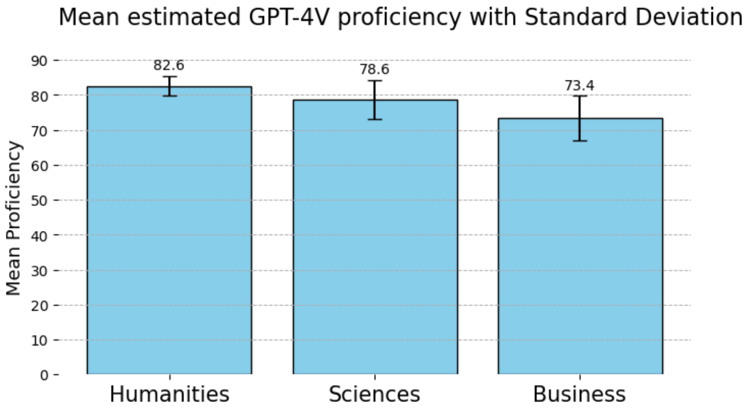


Figure 4. Mean GPT-4V estimated proficiency scores across each discipline.

Therefore, by examining GPT-4V's performance across these disciplines, the emerging evidence points to GPT-4V excelling in contexts and subjects allowing for interpretative flexibility and narrative construction, while it faces challenges in fields demanding high precision, empirical rigour, deep understanding of causality and real-time contextual understanding.

6. Discussion

This study has shown the capability of LLMs to answer multimodal exam questions that require advanced reasoning, something that recent studies have identified as a limitation (RQ1). This work has specifically developed a technique and demonstrated how it can be used to invoke self-reflection within LLMs at different input modalities comprising exam questions, to eventually arrive at correct responses, and thereby, without the intention to do so, this study has provided a "how-to" recipe for more effective cheating on sophisticated exam questions that are unproctored. However, as educators, we need to be aware of these capabilities. A silver lining does exist though since successfully executing a sequence of appropriate and contextualised self-reflective questions that are relevant for each modality in a multimodal exam question, does at least require some sophistication and level of knowledge about the exam question subject from a person attempting to cheat. Additionally, this technique can be used by academics in order to assist with research involving interpretation and reasoning about figures and theoretical concepts.

Nonetheless, it is clear from the results that LLMs possess a noteworthy level of reasoning capability that this study has shown also extends to both text-based and visual modalities. LLMs' reasoning is improving and will likely continue to do so; however, limits and plateaus to this capability are also reasonable to expect, thus there is a need to continuously probe this capability vis-a-vis their proficiency at answering complex multimodal exam questions to determine their upper-performance limits under the current transformer-based [63] machine learning architectures. The iterative self-reflective procedure outlined in this study works, and it has strong theoretical underpinnings [22,51,53,59] that can explain why it works. Transformer-based models that underlie LLMs like GPT-4V, comprise two components: the encoder and the decoder [63]. The encoder excels at analysing and understanding prompts texts while the decoder focuses on producing text based on this understanding. However, the encoder processes the input prompts holistically at once enabling it to capture complex interdependencies, while the decoder is constrained by being able to focus only on previously generated portions of the response without recourse to to consider its future response generation. At the risk of oversimplifying the dynamic, this means that the LLM's ability to understand prompts has certain strengths over its ability to generate responses. Therefore, the encoder's strengths can be leveraged and can be loosely conceptualised as a form of AI self-reflection, where iterative prompting is conducted to extract improved understanding and better responses which are then collectively fed back to the encoder for a final response. The demonstrated process allows the encoder to refine the understanding and context, which the decoder then uses to produce more accurate and contextually appropriate responses. Such iterative, multi-hop decomposition of a complex reasoning task, harnesses the encoder's robust analytical abilities, by progressively improving the quality of the generated responses—and the whole procedure can be viewed as iterative multimodal self-reflection.

This finding has serious implications for online exams. One of the last remaining strategies for making exam questions 'LLM proof' has been to incorporate visual components alongside text [37]. Recent studies [38,42] have confirmed that LLMs perform poorly on visual reasoning in exam questions; however, this work has shown that advanced reasoning capabilities are latent within LLMs that can be accessed for solving complex multimodal tasks (RQ2). It is now incumbent on researchers to apply these types of multi-hop self-reflective techniques on larger multimodal benchmark datasets to more comprehensively quantify the capacities of LLMs on these types of tasks.

With respect to results from estimated multimodal capabilities of GPT-4V across different disciplines (RQ2), we can infer that GPT-4V shows a particular strength in handling tasks that involve interpretive and descriptive analysis typical of the humanities likely due to having been trained on extensive collections of narrative and descriptive texts which translates to more accurate explanations for visuals related to these subjects. On the other hand, in the sciences and business where the questions may more frequently require predictive, forward-looking and strategic thinking, GPT-4V's limitations become more suggested. Its perceived inability to fully grasp real-world time progression, conduct original research and hypothesise about possible outcomes may result in less accurate performances. This observation indicates that GPT-4V's effectiveness varies significantly with the nature of the multimodal exam topic. It is heavily dependent on the type of reasoning the subject demands and the characteristics of the underlying data used in its training. Thus, while GPT-4V estimates broad capabilities of processing across a diverse range of visuals, its performance distinctly mirrors the intrinsic features of each academic discipline and the current limitations of AI technology which performs well in identifying patterns and relationships but falls short in understanding causality and making strategic predictions.

6.1. Recommendations

Based on the research findings, this work proposes the following strategies that may help in the short term enhance the integrity and effectiveness of online assessments in the context of the advanced reasoning capabilities of LLMs:

1. **Proctored online exams:** There is no substitute for effective proctoring. Therefore, it is recommended to ensure that all online exams are proctored as extensively as possible. Proctoring technologies that include real-time monitoring have their limitations, but they can deter the misuse of LLMs and other digital aids. Unproctored exams in the context of existing, and improving multimodal LLM capabilities can no longer be regarded as possessing validity. *Feasibility and Challenges:* Advanced proctoring software with AI capabilities can be costly and raises ethical concerns regarding data privacy and surveillance. Students have also reported negative experiences with these technologies. Additionally, students with limited access to reliable internet and technology may face disadvantages.

2. **Reinstatement of viva-voce exams:** The reintroduction of viva-voce examinations, conducted online, can complement a suite of other assessments. Although viva-voce exams also possess limitations (as do all assessment types), they offer a dynamic, generally reliable, and direct assessment method for measuring student knowledge and reasoning skills. These exams are akin to professional interviews commonly used in industry, thus making them relevant for preparing students for real-world contexts and for professional challenges (The teaching team at Macquarie Law School has demonstrated how authentic assessment through viva voce exams that aim to uphold academic integrity and enhance student employability can be implemented https://teche.mq.edu.au/2023/03/authentic-assessment-through-viva-voce/, accessed on 16 April 2024) [64].
   Feasibility and Challenges: Online meeting platforms already facilitate viva-voce exams. While logistical complexities and potential biases present challenges, these can be mitigated through efficient scheduling tools, comprehensive bias mitigation training for examiners, and detailed rubrics. Offering mock sessions can help students adapt to the format, while framing topics around real-world scenarios enhances relevance and skill development. Additionally, requiring proficiency in web-conference communication equips students for the virtual dimensions of modern workplaces. This approach also supports the capacity to record the sessions for record-keeping, grade moderation, assessor training, and student self-reflection.
3. **Enhanced multimodal exam strategies:** If proctoring online exams is not feasible, multimodal exams should be designed to maximally increase the cognitive load and processing complexity, making it more challenging for LLMs to provide reliable cheating assistance:
   - Include multiple images alongside text per question to invoke a higher degree of reasoning across the modalities. This approach would increase the complexity of the questions and require a deeper level of reasoning and synthesis, which current LLMs may struggle to manage effectively. *Feasibility and Challenges:* While feasible, creating questions that effectively integrate multiple images with text could be resource-intensive, time-consuming, and foreign to some disciplines. It would require careful design to ensure that the images and text are complementary and contribute to deeper reasoning and synthesis while being relevant to each discipline.
   - Design questions that necessitate the formulation of long-term strategies, forecasts, and projections. These types of questions require not only higher levels of conceptual understanding as well as causal relationships, but also the ability to project future trends and consequences which this research shows are a challenge to the predictive capabilities of LLMs. *Feasibility and Challenges:* This is feasible but such questions may not be easily adaptable to all disciplines. These questions require a high level of cognitive processing which would be demanding for students under exam conditions. They also require significant time to construct and evaluate and may necessitate additional training for educators to design and grade them effectively.
   - Integrate real-world scenarios that are current and relevant. Questions that reflect very recent developments or ongoing complex real-world problems and require up-to-date knowledge, making it difficult for LLMs to reason and generate accurate responses based solely on pre-existing and limited data for certain topics. *Feasibility and Challenges:* While being feasible, constantly updating exam content to reflect the latest developments would be challenging. This approach would require ongoing efforts to keep the scenarios current and relevant.
   - Consider incorporating additional modalities into the questions such as video-based and/or audio-based questions alongside images and text, thus fully exploiting the current limitations of LLMs to process to incorporate them all simultaneously. *Feasibility and Challenges:* Modern learning management system

platforms support the inclusion of video and audio content alongside text and images, therefore, this is feasible. However, evaluating responses that span multiple modalities would be more complex and would require clear assessment criteria together with a significant adjustment for the evaluators.

- Explore formulating multimodal questions that require students to annotate the provided figure(s) or draw an additional figure as part of their answer which would again exploit some of the current LLM limitations. From a technological perspective, this strategy could be implemented in most settings without undue difficulty. Challenges would however lie in designing suitable questions that require annotation or drawing and would also be time-consuming. Ensuring that students have the tools and skills to complete these tasks effectively would also be an additional burden.
- As much as possible, consider ways of linking multimodal exam questions with prior assessments completed by students during a teaching semester, and other course materials which would increase the difficulty of the LLM in producing a correct response. *Feasibility and Challenges:* This is achievable and would be expected to enhance coherence and continuity in student learning. Linking exam questions to prior assessments does not pose any implementation challenges; however, there are questions of fairness that could be raised by students who had not completed the earlier assessments and would therefore be penalised in the exam.
- Consider incorporating some decoy questions specifically designed to detect LLM assistance. These questions could be subtly designed to prompt LLMs into revealing their non-human reasoning patterns through specific traps that exploit known LLM weaknesses, such as generating responses based on unlikely combinations of concepts or unusual context switches that a human would likely not make. This would not, however, be straightforward to implement since different multimodal LLMs will likely also have different responses to the decoy questions. *Feasibility and Challenges:* Leveraging LLMs' weaknesses requires research that is ongoing and would also need to encompass numerous available LLMs. Therefore, this approach is likely not feasible for most institutions from a resource perspective, but also from the perspective of needing deep expertise.

6.2. Limitations and Future Work

This study advances the use of LLMs for enhancing academic assessment but acknowledges the need for broader empirical support, suggesting areas for future exploration. This work was limited by only two multimodal exam case studies and further research should extend this approach across more disciplines and question types, particularly focusing on developing a benchmark dataset comprising deep reasoning multimodal exam questions and not merely multiple choice questions that currently do exist. Moreover, while the evaluations of exam questions by GPT-4V provide a useful gauge of the model's current capabilities, they remain estimates and necessitate more rigorous, quantitative validation to accurately measure its actual performance. Additionally, the study highlights the manual nature of the self-reflective process used to trigger latent LLM reasoning and suggests that the algorithmic automation of the proposed strategy across multiple modalities should be explored and developed to improve scalability. Future research should also investigate the application of these multimodal self-reflective strategies in pedagogical settings to enhance teaching methodologies and improve student learning outcomes.

7. Conclusions

This study critically addresses the role of Large Language Models (LLMs) like ChatGPT in modern online educational exams, highlighting the challenges they pose to academic integrity in the absence of proctoring. Key contributions of this research include the introduction of a novel iterative strategy that invokes self-reflection within LLMs to guide them toward correct responses to complex multimodal exam questions, thereby demonstrating latent multi-hop reasoning capabilities within LLMs. By invoking self-reflection within LLMs on each separate modality, the proposed strategy demonstrated how critical thinking and higher-order reasoning can be triggered and integrated in a step-wise manner to steer LLMs towards correct answers, demonstrating that LLMs can be used effectively for cheating on multimodal exam questions. The study also conducted a broad evaluation using descriptions of 600 multimodal exam questions across 12 university subjects to estimate the proficiency of LLMs with vision capabilities to answer university-level exam questions involving visuals. The findings suggest that exam questions including visuals from humanities may pose the least amount of challenge to answer correctly by the best-performing LLMs, followed by exam questions from the sciences and business subjects, respectively. Additionally, this work offers pragmatic recommendations for conducting exams and formulating exam questions in view of the current reasoning and multimodal capabilities of LLMs.

Author Contributions: Conceptualization, T.S.; methodology, T.S.; formal analysis, T.S.; writing—original draft preparation, T.S.; writing—review and editing, T.S. and T.R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset generated and used in this research can be found here: https://github.com/teosusnjak/multimodal-chatgpt-exam-evaluation.

Acknowledgments: The author thanks Liping Zou (https://orcid.org/0000-0002-7091-484X) for kindly providing the Finance-based exam question in the case study as well as her subject expertise in verifying the veracity of the LLM's responses.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large Language Model
GPT	Generative Pretrained Transformer
GPT-4V	Generative Pretrained Transformer version 4 with Vision capability
HEI	Higher Education Institution
CoT	Chain-of-Thought reasoning

References

1. Barber, M.; Bird, L.; Fleming, J.; Titterington-Giles, E.; Edwards, E.; Leyland, C. Gravity assist: Propelling higher education towards a brighter future: Report of the digital teaching and learning review [Barber review]. Government report; Office for Students: Bristol, England, 2021. Available online: https://www.voced.edu.au/content/ngv:89765 (accessed on 16 April 2024).
2. Butler-Henderson, K.; Crawford, J. A systematic review of online examinations: A pedagogical innovation for scalable authentication and integrity. *Comput. Educ.* **2020**, *159*, 104024. [CrossRef] [PubMed]
3. Coghlan, S.; Miller, T.; Paterson, J. Good proctor or "big brother"? Ethics of online exam supervision technologies. *Philos. Technol.* **2021**, *34*, 1581–1606. [CrossRef]
4. Henderson, M.; Chung, J.; Awdry, R.; Mundy, M.; Bryant, M.; Ashford, C.; Ryan, K. Factors associated with online examination cheating. *Assess. Eval. High. Educ.* **2022**, *48*, 980–994. [CrossRef]
5. Dumulescu, D.; Muţiu, A.I. Academic leadership in the time of COVID-19—Experiences and perspectives. *Front. Psychol.* **2021**, *12*, 648344. [CrossRef]
6. Whisenhunt, B.L.; Cathey, C.L.; Hudson, D.L.; Needy, L.M. Maximizing learning while minimizing cheating: New evidence and advice for online multiple-choice exams. *Scholarsh. Teach. Learn. Psychol.* **2022**, *8*, 140. [CrossRef]
7. Garg, M.; Goel, A. A systematic literature review on online assessment security: Current challenges and integrity strategies. *Comput. Secur.* **2022**, *113*, 102544. [CrossRef]
8. Arnold, I.J. Cheating at online formative tests: Does it pay off? *Internet High. Educ.* **2016**, *29*, 98–106. [CrossRef]

9. Ahsan, K.; Akbar, S.; Kam, B. Contract cheating in higher education: A systematic literature review and future research agenda. *Assess. Eval. High. Educ.* **2021**, *47*, 523–539. [CrossRef]
10. Crook, C.; Nixon, E. How internet essay mill websites portray the student experience of higher education. *Internet High. Educ.* **2021**, *48*, 100775. [CrossRef]
11. Noorbehbahani, F.; Mohammadi, A.; Aminazadeh, M. A systematic review of research on cheating in online exams from 2010 to 2021. *Educ. Inf. Technol.* **2022**, *27*, 8413–8460. [CrossRef]
12. Allen, S.E.; Kizilcec, R.F. A systemic model of academic (mis) conduct to curb cheating in higher education. *High. Educ.* **2023**, *87*, 1529–1549. [CrossRef]
13. Henderson, M.; Chung, J.; Awdry, R.; Ashford, C.; Bryant, M.; Mundy, M.; Ryan, K. The temptation to cheat in online exams: Moving beyond the binary discourse of cheating and not cheating. *Int. J. Educ. Integr.* **2023**, *19*, 21. [CrossRef]
14. Malik, A.A.; Hassan, M.; Rizwan, M.; Mushtaque, I.; Lak, T.A.; Hussain, M. Impact of academic cheating and perceived online learning effectiveness on academic performance during the COVID-19 pandemic among Pakistani students. *Front. Psychol.* **2023**, *14*, 1124095. [CrossRef] [PubMed]
15. Newton, P.M.; Essex, K. How common is cheating in online exams and did it increase during the COVID-19 pandemic? A systematic review. *J. Acad. Ethics* **2023**, *22*, 323–343. [CrossRef]
16. McCabe, D.L. CAI Research Center for Academic Integrity, International Center for Academic Integrity, PO Box 170274, Atlanta, GA 30317, 2005. Available online: https://academicintegrity.org/ (accessed on 16 April 2024).
17. Wajda-Johnston, V.A.; Handal, P.J.; Brawer, P.A.; Fabricatore, A.N. Academic dishonesty at the graduate level. *Ethics Behav.* **2001**, *11*, 287–305. [CrossRef]
18. Lee, D.; Arnold, M.; Srivastava, A.; Plastow, K.; Strelan, P.; Ploeckl, F.; Lekkas, D.; Palmer, E. The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100221. [CrossRef]
19. Xia, Q.; Weng, X.; Ouyang, F.; Lin, T.J.; Chiu, T.K. A scoping review on how generative artificial intelligence transforms assessment in higher education. *Int. J. Educ. Technol. High. Educ.* **2024**, *21*, 40. [CrossRef]
20. Yusuf, A.; Pervin, N.; Román-González, M. Generative AI and the future of higher education: A threat to academic integrity or reformation? Evidence from multicultural perspectives. *Int. J. Educ. Technol. High. Educ.* **2024**, *21*, 21. [CrossRef]
21. Newton, P.; Xiromeriti, M. ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. *Assess. Eval. High. Educ.* **2023**, 1–18. [CrossRef]
22. Ji, Z.; Yu, T.; Xu, Y.; Lee, N.; Ishii, E.; Fung, P. Towards Mitigating Hallucination in Large Language Models via Self-Reflection. *arXiv* **2023**, arXiv:2310.06271.
23. Farazouli, A.; Cerratto-Pargman, T.; Bolander-Laksov, K.; McGrath, C. Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. *Assess. Eval. High. Educ.* **2024**, *49*, 363–375. [CrossRef]
24. Richards, M.; Waugh, K.; Slaymaker, M.; Petre, M.; Woodthorpe, J.; Gooch, D. Bob or Bot: Exploring ChatGPT's Answers to University Computer Science Assessment. *ACM Trans. Comput. Educ.* **2024**, *24*, 1–32. [CrossRef]
25. Chan, J.C.; Ahn, D. Unproctored online exams provide meaningful assessment of student learning. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2302020120. [CrossRef] [PubMed]
26. Van Wyk, M.M. Is ChatGPT an opportunity or a threat? Preventive strategies employed by academics related to a GenAI-based LLM at a faculty of education. *J. Appl. Learn. Teach.* **2024**, *7*. [CrossRef]
27. Martino, A.; Iannelli, M.; Truong, C. Knowledge injection to counter large language model (LLM) hallucination. In Proceedings of the European Semantic Web Conference 2023, Hersonissos, Greece, 28 May–1 June 2023; pp. 182–185.
28. Yao, J.Y.; Ning, K.P.; Liu, Z.H.; Ning, M.N.; Yuan, L. LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples. *arXiv* **2023**, arXiv:2310.01469.
29. Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv* **2023**, arXiv:2309.01219.
30. Chang, Y.C.; Wang, X.; Wang, J.; Wu, Y.; Zhu, K.; Chen, H.; Yang, L.; Yi, X.; Wang, C.; Wang, Y.; et al. A Survey on Evaluation of Large Language Models. *arXiv* **2023**, arXiv:2307.03109. [CrossRef]
31. McKenna, N.; Li, T.; Cheng, L.; Hosseini, M.J.; Johnson, M.; Steedman, M. Sources of Hallucination by Large Language Models on Inference Tasks. *arXiv* **2023**, arXiv:2305.14552. [CrossRef]
32. Liu, H.; Ning, R.; Teng, Z.; Liu, J.; Zhou, Q.; Zhang, Y. Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4. *arXiv* **2023**, arXiv:2304.03439. [CrossRef]
33. Schultz, M.; Callahan, D.L. Perils and promise of online exams. *Nat. Rev. Chem.* **2022**, *6*, 299–300. [CrossRef]
34. Cotton, D.R.; Cotton, P.A.; Shipway, J.R. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innov. Educ. Teach. Int.* **2024**, *61*, 228–239. [CrossRef]
35. Alessio, H.M.; Malay, N.; Maurer, K.; Bailer, A.J.; Rubin, B. Examining the effect of proctoring on online test scores. *Online Learn.* **2017**, *21*, 146–161. [CrossRef]
36. Han, S.; Nikou, S.; Ayele, W.Y. Digital proctoring in higher education: A systematic literature review. *Int. J. Educ. Manag.* **2023**, *38*, 265–285. [CrossRef]
37. Abd-Alrazaq, A.; AlSaad, R.; Alhuwail, D.; Ahmed, A.; Healy, P.M.; Latifi, S.; Aziz, S.; Damseh, R.; Alrazak, S.A.; Sheikh, J.; et al. Large language models in medical education: Opportunities, challenges, and future directions. *JMIR Med. Educ.* **2023**, *9*, e48291. [CrossRef] [PubMed]

38. Stribling, D.; Xia, Y.; Amer, M.K.; Graim, K.S.; Mulligan, C.J.; Renne, R. The model student: GPT-4 performance on graduate biomedical science exams. *Sci. Rep.* **2024**, *14*, 5670. [CrossRef] [PubMed]
39. Rudolph, J.; Tan, S.; Tan, S. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *J. Appl. Learn. Teach.* **2023**, *6*, 342–363.
40. Lo, C.K. What is the impact of ChatGPT on education? A rapid review of the literature. *Educ. Sci.* **2023**, *13*, 410. [CrossRef]
41. Nikolic, S.; Daniel, S.; Haque, R.; Belkina, M.; Hassan, G.M.; Grundy, S.; Lyden, S.; Neal, P.; Sandison, C. ChatGPT versus engineering education assessment: A multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *Eur. J. Eng. Educ.* **2023**, *48*, 559–614. [CrossRef]
42. Zhang, W.; Aljunied, M.; Gao, C.; Chia, Y.K.; Bing, L. M3exam A multilingual, multimodal, multilevel benchmark for examining large language models. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 5484–5505.
43. Sadasivan, V.S.; Kumar, A.; Balasubramanian, S.; Wang, W.; Feizi, S. Can AI-Generated Text be Reliably Detected? *arXiv* **2023**, arXiv:2303.11156. [CrossRef]
44. Orenstrakh, M.S.; Karnalim, O.; Suarez, C.A.; Liut, M. Detecting LLM-Generated Text in Computing Education: A Comparative Study for ChatGPT Cases. *arXiv* **2023**, arXiv:2307.07411.
45. Kumarage, T.; Agrawal, G.; Sheth, P.; Moraffah, R.; Chadha, A.; Garland, J.; Liu, H. A Survey of AI-generated Text Forensic Systems: Detection, Attribution, and Characterization. *arXiv* **2024**, arXiv:2403.01152.
46. Kalai, A.T.; Vempala, S.S. Calibrated Language Models Must Hallucinate. *arXiv* **2024**, arXiv:2311.14648.
47. Pal, A.; Sankarasubbu, M. Gemini Goes to Med School: Exploring the Capabilities of Multimodal Large Language Models on Medical Challenge Problems & Hallucinations. *arXiv* **2024**, arXiv:2402.07023.
48. Nori, H.; King, N.; McKinney, S.; Carignan, D.; Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv* **2023**, arXiv:2303.13375. [CrossRef]
49. Stechly, K.; Marquez, M.; Kambhampati, S. GPT-4 Doesn't Know It's Wrong: An Analysis of Iterative Prompting for Reasoning Problems. *arXiv* **2023**, arXiv:2310.12397. [CrossRef]
50. Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J.; Mordatch, I. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv* **2023**, arXiv:2305.14325. [CrossRef]
51. Luo, J.; Xiao, C.; Ma, F. Zero-Resource Hallucination Prevention for Large Language Models. *arXiv* **2023**, arXiv:2309.02654. [CrossRef]
52. Creswell, A.; Shanahan, M. Faithful Reasoning Using Large Language Models. *arXiv* **2022**, arXiv:2208.14271. [CrossRef]
53. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv* **2023**, arXiv:2201.11903.
54. Yeadon, W.; Halliday, D.P. Exploring durham university physics exams with large language models. *arXiv* **2023**, arXiv:2306.15609.
55. Singla, A. Evaluating ChatGPT and GPT-4 for Visual Programming. In Proceedings of the 2023 ACM Conference on International Computing Education Research, Chicago, IL, USA, 7–11 August 2023; Volume 2, pp. 14–15.
56. Zheng, C.; Liu, Z.; Xie, E.; Li, Z.; Li, Y. Progressive-Hint Prompting Improves Reasoning in Large Language Models. *arXiv* **2023**, arXiv:2304.09797. [CrossRef]
57. Han, S.J.; Ransom, K.J.; Perfors, A.; Kemp, C. Inductive reasoning in humans and large language models. *Cogn. Syst. Res.* **2024**, *83*, 101155. [CrossRef]
58. Liévin, V.; Hother, C.E.; Motzfeldt, A.G.; Winther, O. Can large language models reason about medical questions? *Patterns* **2023**, *5*, 100943. [CrossRef] [PubMed]
59. Luo, L.; Lin, Z.; Liu, Y.; Shu, L.; Zhu, Y.; Shang, J.; Meng, L. Critique ability of large language models. *arXiv* **2023**, arXiv:2310.04815.
60. Feng, T.H.; Denny, P.; Wuensche, B.; Luxton-Reilly, A.; Hooper, S. More Than Meets the AI: Evaluating the performance of GPT-4 on Computer Graphics assessment questions. In Proceedings of the 26th Australasian Computing Education Conference, Sydney, NSW, Australia, 29 January–2 February 2024; pp. 182–191.
61. Bloom, B.S.; Engelhart, M.D.; Furst, E.J.; Hill, W.H.; Krathwohl, D.R. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain*; Longman: New York, NY, USA, 1956.
62. Han, T.; Adams, L.C.; Bressem, K.K.; Busch, F.; Nebelung, S.; Truhn, D. Comparative Analysis of Multimodal Large Language Model Performance on Clinical Vignette Questions. *J. Am. Med. Assoc.* **2024**, *331*, 1320–1321. [CrossRef]
63. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
64. Melkonian, H.; Bending, Z.; Tomossy, G. Viva Voce Assessment—Legal Education for the Real World. In Proceedings of the 2022 Professional Legal Education Conference: LawTech, Newlaw and NetZero: Preparing for an Uncertain Future, Gold Coast, QLD, Australia, 2022; Conference Program.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.