Race with the machines: Assessing the capability of generative AI in solving authentic assessments

Binh Nguyen Thanh, Diem Thi Hong Vo, Minh Nguyen Nhat, Thi Thu Tra Pham, Hieu Thai Trung, Son Ha Xuan
The Business School, RMIT University Vietnam, Ho Chi Minh City, Vietnam

In this study, we introduce a framework designed to help educators assess the effectiveness of popular generative artificial intelligence (AI) tools in solving authentic assessments. We employed Bloom's taxonomy as a guiding principle to create authentic assessments that evaluate the capabilities of generative AI tools. We applied this framework to assess the abilities of ChatGPT-4, ChatGPT-3.5, Google Bard and Microsoft Bing in solving authentic assessments in economics. We found that generative AI tools perform very well at the lower levels of Bloom's taxonomy while still maintaining a decent level of performance at the higher levels, with "create" being the weakest level of performance. Interestingly, these tools are better able to address numeric-based questions than text-based ones. Moreover, all the generative AI tools exhibit weaknesses in building arguments based on theoretical frameworks, maintaining the coherence of different arguments and providing appropriate references. Our study provides educators with a framework to assess the capabilities of generative AI tools, enabling them to make more informed decisions regarding assessments and learning activities. Our findings demand a strategic reimagining of educational goals and assessments, emphasising higher cognitive skills and calling for a concerted effort to enhance the capabilities of educators in preparing students for a rapidly transforming professional environment.

Implications for practice or policy

- Our proposed framework enables educators to systematically evaluate the capabilities of widely used generative AI tools in assessments and assist them in the assessment design process.
- Tertiary institutions should re-evaluate and redesign programmes and course learning outcomes. The new focus on learning outcomes should address the higher levels of educational goals of Bloom's taxonomy, specifically the "create" level.

Keywords: authentic assessment, Bloom's taxonomy, generative AI, AI in tertiary education, quantitative, case study

Introduction

The rapid emergence and expansion of generative artificial intelligence (AI) capabilities have significant potential impacts on tertiary education. AI, epitomised by language models like ChatGPT, has revolutionised machine interactions and the production of human-like text. It functions by discerning patterns from vast data sets and utilising this knowledge to produce and remember coherent and contextually appropriate responses (Megahed et al., 2023; J. Su & Yang, 2023). The role of generative AI in education has garnered considerable interest from scholars, stimulating discussion about its potential applications in research, pedagogical strategies (Peres et al., 2023), assessment strategies and ethical considerations (Qadir, 2023; J. Su & Yang, 2023). This technological evolution has introduced new opportunities and challenges for educators and institutions, particularly in ensuring student learning outcomes while capitalising on the benefits of generative AI. The integration of generative AI in education presents a significant challenge, especially in the realm of assessment design. Generative AI tools empower students to produce artefacts, like essays or research reports, by simply inputting the assessment instructions into the system (Qadir, 2023). This could potentially circumvent the traditional learning process, as these AI tools can effortlessly generate certain types of tasks. This challenge may be particularly relevant for authentic assessments that largely depend on report writing in take-home assignments. The benefits of authentic assessments, including the enhancement of higher-order thinking skills, the fostering of transferable skills in contextualised settings and the boosting of students' motivation and overall learning (Ashford-Rowe et al., 2014; Villarroel et al., 2018), are well documented. However, these advantages could be compromised if students resort to using generative AI models to complete assessment tasks.

Moreover, educators are faced with the formidable task of designing assessment tasks that not only leverage the benefits of generative AI but also ensure student learning outcomes are met (Crawford et al., 2023). It is paramount to identify the types of authentic assessments that generative AI can effectively complete, enabling educators to recalibrate their authentic assessments, particularly for take-home assignments where students have time and access to various AI tools that can assist them in responding to assignment questions. These unexplored questions necessitate the development of frameworks to help educators gauge the capabilities of generative AI tools. Such frameworks would focus on how generative AI tools can address assignment questions and equip educators with the necessary insights to effectively design authentic assessments for an era where generative AI becomes widely adopted.

To fill this gap, our study introduces a comprehensive framework designed to assist educators in the systematic assessment of the capabilities of generative AI tools in solving authentic assignment questions. In particular, we provide detailed guidance on how educators can craft authentic assignment questions, as well as develop rubrics and marking guides tailored for input into generative AI tools. Following this, we demonstrate how educators can obtain answers to assignment questions directly from the generative AI tools, and subsequently, how to critically evaluate the quality of these responses. To validate our approach, we implemented this systematic method within the field of economics, creating authentic assignment questions, rubrics and marking guides, and then proceeded to assess the capability of the generative AI tools in addressing these specific assignment tasks. Authentic assessments in economics often require students to compose take-home reports, making it a highly relevant area in the learning and teaching sphere that generative AI could potentially disrupt. Although our framework is designed with the economics discipline in mind, it is adaptable to other disciplines that require students to write reports and/or conduct data analysis, particularly within the social sciences.

In our study, Bloom's taxonomy (Armstrong, 2010) served as a foundational guideline for crafting authentic assignment questions and developing rubrics. This is because the advanced levels of Bloom's taxonomy underscore the cultivation of higher-order thinking skills and creativity. These skills encapsulate the three core principles of authentic assessment – realism, cognitive challenge and evaluative judgement (Villarroel et al., 2018) – ensuring a comprehensive and robust approach to educational evaluation. Using Bloom's taxonomy, our framework classifies assignment questions based on different levels of cognitive skills. This helps identify which type of cognitive skills generative AI can address and to what extent. For the purpose of our study, we used ChatGPT-4, ChatGPT-3, Microsoft Bing and Google Bard. These are the most prominent and easily accessible generative AI tools currently available.

The main contributions of our study are twofold. Firstly, we present a framework for educators to assess the potential of generative AI models in tackling authentic assignment questions. Secondly, we assessed the performance of generative AI in addressing various educational objectives, using authentic assessment questions within the field of economics. The insights from our study can help shape suitable policies for educational institutions, particularly around the design of authentic assessments, by offering evidence-based insights into the incorporation of generative AI. The proposed framework will equip educators with robust and effective methods for assessment design that consider generative AI as a tool students may utilise. As generative AI becomes progressively widespread across various sectors, tertiary education bears a crucial responsibility in preparing students for a world where this technology is not just present but significantly influences almost all aspects of society. Consequently, this paper confronts the challenges associated with academic integrity and academic security induced by the use of generative AI, with a specific focus on take-home authentic assignment questions in the field of economics. We believe this study will contribute to a better understanding of how generative AI tools like ChatGPT can be considered in the design of academic assessments, thus supporting educators in their mission to prepare students for the digital future.

Literature review

Authentic assessments as a critical part in economics education

Authentic assessment, a key component of authentic learning, plays an important role in achieving the fundamental goal of education, which is to enable students to apply their knowledge outside of the class (Herrington & Oliver, 2000). Since the 1990s, authentic assessment has been pivotal in fostering students' 21st-century skills, preparing them for a global interconnected knowledge economy (Koh, 2017). Authentic assessments bridge the gap between academic learning and real-world challenges, aligning education with future career expectations (Neely & Tucker, 2012). Koh emphasised that authentic assessment promotes deep understanding, higher-order thinking and problem-solving skills, requiring students to engage in complex tasks that mirror professional demands. Consequently, it fosters more effective learners, leaders and problem-solvers (Koh, 2017; Villarroel et al., 2018; Wiewiora & Kowalkiewicz, 2019).

Three main principles underpinning authentic assessment design are realism, cognitive challenge and evaluative judgement (Villarroel et al., 2018). Realism, the primary principle of authentic assessment (Bosco & Ferns, 2014), involves creating assessments that replicate real-life world challenges, requiring students to solve problems similar to those in real-life contexts, testing knowledge, skills and attitudes. The cognitive challenge principle demands higher-order thinking, pushing students to transform knowledge into new forms, as Wiggins (1993) highlighted. This approach encourages students to move past memorisation to deeper understanding, integrating new ideas with prior knowledge and applying theories to practical situations, including data analysis and evaluating theoretical arguments. The final principle, evaluative judgement, aims to develop students' ability to critique their own and others' work, promoting innovation and creativity in solving complex, non-standard tasks in novel situations (Tai et al., 2018). Authentic assessment thus becomes a vital tool for assessing critical thinking, problem-solving, innovation and creativity in students.

Economics studies are often seen as abstract and disconnected from real-life applications. Students may become immersed in abstract theories without understanding how to apply them to explain real-world phenomena. Authentic assessments aim to address this issue. Research has emphasises authentic assessment in economics education, highlighting educators' role in teaching economic theories and fostering their practical application (Woldab, 2013). This approach nurtures a real-world perspective, enabling students to analyse specific economic contexts, make predictions and propose solutions using a blend of theory and real-world insights (Gulikers et al., 2004). Therefore, economics curricula should encourage learners to connect their knowledge with real-life situations (James & Casidy, 2018; Manville et al., 2022).

Bloom's taxonomy

Overview of Bloom's taxonomy

Bloom's taxonomy was created in 1956 (Bloom et al., 1956). It is a hierarchical model in education for classifying learning objectives and cognitive processes. Bloom's taxonomy is structured around two main dimensions, consisting of the cognitive processes and the knowledge dimension (Lau et al., 2018). The cognitive processes originally featured six cognitive skill levels – knowledge, comprehension, application, analysis, synthesis and evaluation – with each level building on the preceding one (Bloom et al., 1956). This model has served as a reliable tool for curriculum development, instruction analysis and evaluation (Shane, 1981). However, to adapt to evolving societal and learning needs, Krathwohl revised the taxonomy in 2002, emphasising creativity (Callaghan-Koru & Aqil, 2022; Krathwohl, 2002). The revised model supports contemporary trends in authentic learning and remains relevant in today's rapidly advancing educational landscape, demonstrating its importance and usefulness across various academic disciplines. Both the original and revised Bloom's taxonomy also consider the knowledge dimension, emphasising the type of knowledge being accessed at each level, consisting of factual, conceptual, procedural and metacognitive knowledge (Krathwohl, 2002).

The role of Bloom's taxonomy in learning and assessment design

Bloom's taxonomy is widely used in various aspects of learning, including setting learning objectives, planning activities, delivering materials and designing assessments. It helps instructors create assessments that align with students' understanding and proficiency levels. Recognised as a key framework by experts such as West (2023) and Stanny (2016), it enhances critical analysis and metacognition skills (Callaghan-Koru & Aqil, 2022) and distinguishes between different student performance levels (Zaidi et al., 2017). Beyond assessment, Bloom's taxonomy also supports pre-service teachers in their cognitive development (Athanassiou et al., 2003) and is applied in fields such as health and business education, notably in logistics, project management and architecture programmes (Attia, 2021; Karanja & Malone, 2021; Lau et al., 2018; Pepin et al., 2021).

Bloom's taxonomy is utilised in a range of assessment methods, including traditional formats like multiple-choice and essays, as well as in digital assessments, enhancing metacognition and providing real-time feedback in online courses (Matore, 2021; Na et al., 2021; Thompson & O'Loughlin, 2015). Its practical use also improves authentic assessments across various disciplines, helping educators design tasks that encourage metacognitive learning (West, 2023). Bloom's revised taxonomy categorises competency levels into multi-tiered stages, starting with the basic tasks of reflection and summarising and advancing to more complex activities such as synthesis, evaluation and creation (DeMara et al., 2019; Krathwohl, 2002). As students advance through these levels, assessments become more authentic, involving complex assimilation and higher cognitive engagement, underscoring the taxonomy's importance in developing effective assessment strategies.

Although widely used, Bloom's taxonomy has faced criticism for potential inaccuracies in assessing learning across disciplines, difficulty in observing cognitive processes, issues with multiple observers and limitations in analysing student performance (Hussey & Smith, 2002; Kibble & Johnson, 2011; Thompson et al., 2013). However, modifications such as discipline-specific tools have been developed to improve its applicability and inter-rater agreement (Crowe et al., 2008; Zheng et al., 2008). Despite these criticisms, Bloom's taxonomy continues to be a valuable resource in higher education (Ramirez, 2017), recognised for its role in assessing personality, developing curriculum, enhancing cognitive skills and promoting deep thinking (Ho & Chng, 2021; Parwata et al., 2023). It aids in creating challenging learning processes, aligning learning outcomes with evaluation methods, enhancing metacognitive abilities and encouraging deeper learning, thus solidifying its status as an essential educational tool.

Bloom's taxonomy is particularly relevant for developing curricula and designing assessments in economics studies, as these fields heavily rely on cognitive processes. Although it is crucial to comprehend an economic event or phenomenon, the next step is to apply the knowledge acquired to make economic sense of it, adapt it to different economic contexts and assess relevant economic outcomes while drawing policy implications (Pappas et al., 2013).

Bloom's taxonomy in designing authentic assessment

As mentioned earlier, authentic assessments are grounded in three underlying principles: realism, cognitive challenge and evaluative judgement. Bloom's taxonomy, as a hierarchically ordered structure of cognitive processes, provides the foundation for the latter two principles. Concerning cognitive challenges, authentic assessments promote higher-order thinking skills that align with the upper levels of Bloom's taxonomy, where students are expected to apply, analyse, evaluate and create (Kar et al., 2022). These levels require deeper cognitive engagement, which is in line with the real-world and integrative nature of authentic tasks.

Bloom's taxonomy emphasises the development of cognitive skills that progress from lower-order thinking (e.g., remember and understand) to higher-order thinking (such as analyse, evaluate and create). Authentic assessments require students to apply their knowledge and skills in scenarios closely mirror actual real-world situations (Darling-Hammond & Snyder, 2000). This is also supported by the cognitive processes outlined in Bloom's taxonomy, particularly the “apply” and “analyse” stages, where students are encouraged to utilise their acquired knowledge to solve problems and evaluate information, aligning with the higher-order thinking from the cognitive levels of Bloom's taxonomy (Callaghan-Koru & Aqil, 2022; Na et al., 2021).

Concerning evaluative judgement, authentic assessment's emphasis on fostering creativity aligns with Bloom's “evaluate” and “create” levels (Greenstein, 2012; Villarroel et al., 2018). Students are not only expected to solve problems but also critically assess their own work and that of others. They are encouraged to find innovative solutions to complex challenges, reflecting the highest levels of Bloom's taxonomy (Villarroel et al., 2018).

Considering the above, Bloom's taxonomy provides a structured framework for understanding cognitive processes in learning and serves as a guiding principle in the design and evaluation of authentic assessments. It emphasises the development of higher-order thinking skills, practical knowledge application, and the importance of real-world contexts in both learning and assessment. In this study, Bloom’s taxonomy was employed as a guiding framework to design authentic assessments for testing capabilities of generative AI models.

Generative AI as a threat to authentic assessments

Generative AI based on large language models (LLMs) like ChatGPT significantly benefits students by offering personalised mentoring, real-time feedback and adaptable content delivery so that students could enhance academic writing, data analysis and knowledge retention (Lodge et al., 2023). However, AI integration in education raises concerns about academic integrity and security, particularly with take-home assessments. In fact, students may misuse AI for academic advantage, entirely relying on AI to complete assessments. To uphold assessment security (Dawson, 2020), schools should protect assessments from cheating while fostering students' skills of employing AI (Lodge et al., 2023; Nguyen et al., 2023). It is crucial to emphasise that AI-generated work challenges academic integrity as it fails to reflect students' efforts and undermines the credibility of assessment outcomes. Additionally, the availability of user-friendly AI platforms exacerbates the challenge of discouraging cheating and detecting fraud (Halaweh, 2023; Lodge et al., 2023).

The widespread adoption of generative AI can disrupt authentic assessments in economics education, which aims to cultivate higher-order thinking skills, as discussed in the previous section. To strike a balance between the benefits and threats posed by AI, a proactive and comprehensive reassessment of current authentic assessment designs considering generative AI intervention is necessary. It is crucial to understand the limitations and capabilities of generative AI in addressing assessment questions. Questions should be designed to require students' engagement at levels where AI is not a substitute but a complement for learning.

Research on generative AI in education primarily focuses on its use in administrative tasks like feedback and grading (e.g., Dai et al., 2023; Katz et al., 2023; Y. Su et al., 2023), with some studies exploring AI response detection and evaluation (Tian et al., 2023) and others examining its behavioural impact from learner and instructor perspectives (Strzelecki, 2023; Y. Su et al., 2023). However, there is a significant gap in understanding how generative AI affects assessment redesign, which our study aimed to address. We evaluated the effectiveness of generative AI tools like ChatGPT, Google Bard and Microsoft Bing by aligning them with the cognitive processes defined in Bloom's taxonomy, identifying the cognitive skill levels where AI excels or falls short. Our study is among the first attempts to explore AI's limitations and capabilities through the lens of Bloom's taxonomy, especially in authentic assessment design. By identifying generative AI's weaknesses, we aim to strategically redesign assessments to reduce students' misuse and over-reliance on AI. Adapting assessment designs is crucial in the evolving educational landscape to maintain academic integrity and prepare students for future challenges.

Methodology

We developed a framework designed to help educators assess the effectiveness of popular generative AI tools in solving authentic assessments, as illustrated in Figure 1. We employed Bloom's taxonomy as a guiding principle to create authentic assessments that evaluate the capabilities of generative AI tools. The entire framework process is detailed in the following subsections. In this study, we applied the framework to evaluate the generative AI capabilities to address take-home authentic assignment questions for an introductory-level economics course, centred on gross domestic product (GDP) and economic growth. The questions covered GDP topic, which is a fundamental and universal economic concept for students in economic, financial and business educational programmes. Teaching GDP has been widely acknowledged as one of the fundamental requirements in the instruction of economics, particularly in the field of macroeconomics, according to Talor (2000), Aghion and Durlauf (2005) and Samuelson and Nordhaus (2013). All data utilised in this study for assessing the capability to generate responses were collected exclusively through interactions with generative AI, without employing any responses from students. Consequently, the data does not contravene any ethical issues or conflicts of interest concerning students. Furthermore, we designed the questions: they were not adapted from existing questions, thus ensuring there were no breaches of ethical standards or copyright issues.

To investigate generative AI's impact on take-home assignments, we designed a research process, as illustrated in Figure 1. Initially, we categorised the assignment questions according to Bloom's taxonomy, encompassing the cognitive levels: "remember", "understand", "apply", "analyse", "evaluate" and "create". By constructing assessments that demand diverse cognitive abilities, we aimed to assess the capability of generative AI to handle different levels of complexity in economics assignments. We selected four distinct generative AI tools – ChatGPT-4, ChatCPT-3.5, Microsoft Bing and Google Bard – for assessment purposes. Additionally, we devised rubrics and marking guidelines based on the classified Bloom's taxonomy questions. Subsequently, we fed these questions, along with their corresponding rubrics, into the generative AI to generate answers. Four teaching staff members independently grade the generative AI-generated answers using predefined marking criteria, scoring from 0 to 100. This evaluation process allowed us to measure the proficiency of different generative AI in generating answers for take-home assessments. Ultimately, the examination of the grades obtained from Bloom's taxonomy questions provides valuable insights into the capabilities of various generative AI in handling assignment tasks. Detailed descriptions of the assessment design process, rubrics, marking guides, input procedures for AI tools and the evaluation of AI-generated answers are presented in the following sections.

Our framework is designed to be adaptable, and educators can employ it as a guideline to evaluate emerging or existing generative AI tools, enabling systematic evaluations. The framework can also be utilised to assess the performance of generative AI tools in addressing both advanced assignment tasks in economics and assignments from various other disciplines. Educators could follow the framework, as depicted in Figure 1, to formulate the assignment questions, marking rubrics and guidelines, and subsequently, integrate these inputs into generative AI tools to obtain the results. Subsequently, educators would need to evaluate the responses generated by the generative AI tools.

For the prompting generation process, we acknowledge that the effectiveness of AI tools, particularly ChatGPT-3.5, hinges significantly on the quality of the prompts provided. The formulation of prompts was an iterative process, refined through testing and feedback to align with the objectives of the assessment tasks (see more details in Appendix B).

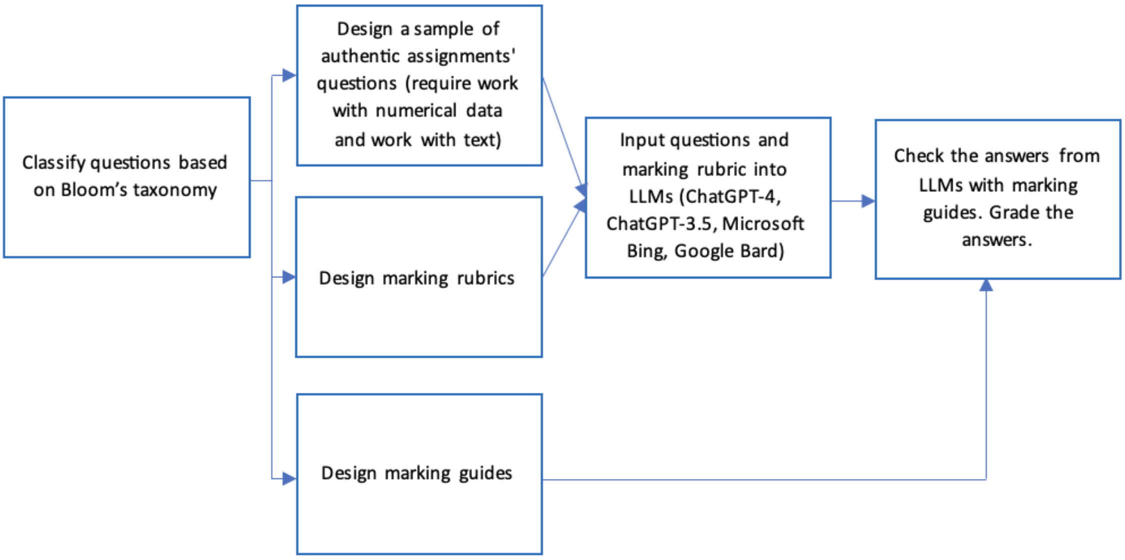

Classify questions based on Bloom's taxonomy
Design a sample of authentic assignments' questions (require work with numerical data and work with text)
Design marking rubrics
Design marking guides
Input questions and marking rubric into LLMs (ChatGPT-4, ChatGPT-3.5, Microsoft Bing, Google Bard)
Check the answers from LLMs with marking guides. Grade the answers.

Figure 1. Framework for the evaluation of generative AI capabilities

Process of designing assignment questions based on Bloom's taxonomy

The initial phase focused on designing assignment questions to test AI tools' responses. These evaluations included both numerical and text-based questions to thoroughly assess AI's capabilities. Numerical questions tested the AI's data handling, statistical analysis and real-world application skills, while text questions gauged its ability to synthesise and interpret economic theories. The assignment questions are based on the revised Bloom's taxonomy framework, which comprises six cognitive levels: "remember", "understand", "apply", "analyse", "evaluate" and "create". Each level demands specific knowledge, analytical and creative skills. For every level, we developed two question types – one numerical and one text-based – resulting in a total of 12 questions. These questions, centred on GDP and economic growth, are tailored to meet the distinct requirements of each taxonomy level. Numerical questions include tailored data sets for AI processing.

We conducted a test in June 2023 following the completion of the draft version of the questionnaire. This was done to ensure that the questions were relevant to GDP and economic growth, and that they complied with Bloom's taxonomy. The process included reviews by four of us, who are experts in economics, to assess the relevance and clarity of the questions. This was followed by AI testing to confirm the coherence and comprehension of the material. Based on expert and AI feedback, we finalised the questions for the experiment, with details in Appendix A.

Process of rubrics and marking guide design

Responses are scored on a scale from 0 to 100, divided into five groups: NN (0 – below 50), PA (50 – below 60), CR (60 – below 70), DI (70 – below 80) and HD (80–100). Each category reflects a different level of understanding and application ability, from NN indicating failure to meet basic criteria to HD representing thorough satisfaction of requirements. Every question is aligned with standards set for varying levels and types of questions based on Bloom's taxonomy.

Each question has a unique rubric corresponding to its taxonomy level and format, categorised into NN, PA, CR, DI and HD, each with specific requirements. The rubric creation process, applied to all questions, resulted in 12 distinct rubrics across two question formats and six Bloom's taxonomy levels.

To guarantee reliability and clarity, we engaged the four economic experts from our research team to review the rubrics. Based on their feedback, we made adjustments to these rubrics. We subsequently employed the revised versions in our study, and the final iterations of these rubrics are included in the supplementary materials.

We also developed a marking guide to standardise scoring by experts. It details content requirements, point allocation, and criteria for point deductions. This guide is exclusively for markers, ensuring consistent and valid evaluations. Students have access to question content, data sets and rubrics, whereas markers also have marking guides. Generative AI tools, used to replicate student response processes, only use materials available to students, with their integration process.

Process of marking AI-generated answers

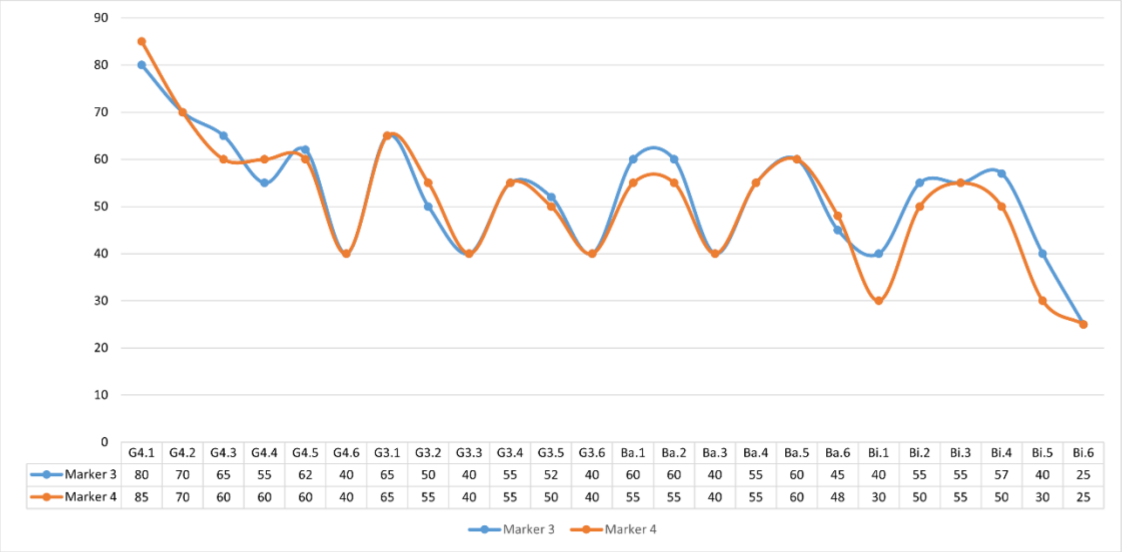
Four experts with extensive experience in teaching macroeconomics independently assessed AI tools' answers to ensure marking reliability and validity. Each answer was double-marked by two experts. For instance, two experts evaluated Question 1's answers, which included responses from all AI tools, across all six Bloom's taxonomy levels. Similarly, another pair of experts did the same for Question 2. Utilising specific rubrics and marking guides, the experts scored the answers and identified each AI tool's strengths and weaknesses. This comprehensive evaluation offered insights into AI's role in student assessments, providing valuable information for educators to tailor assessment designs to contemporary technological advancements.

Results and discussion

We analysed the grades that the generative AI tools achieved. First, we discuss some general observations on the performance of the AI tools. Next come the strengths and finally we summarise the limitations of the AI tools.

General comments

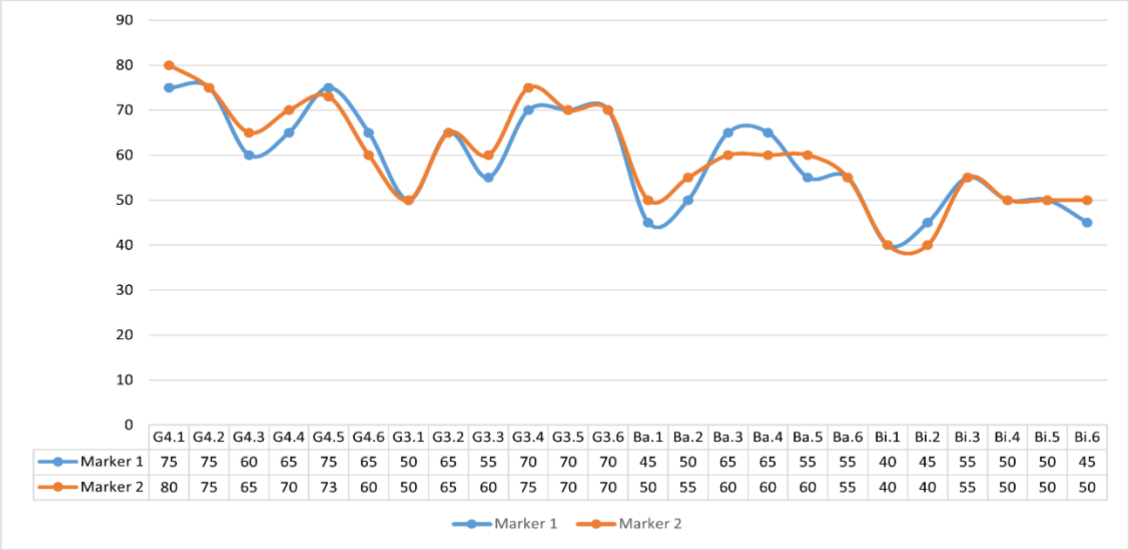
Figures 2 and 3 display the marks awarded to each generative AI for each question and level, respectively. We denote ChatGPT-4, ChatGPT-3.5, Google Bard and Microsoft Bing by G4, G3, Ba and Bi, respectively in Figures 2 and 3. For example, G4.1 denotes ChatGPT-4's performance on the "remember" level and G4.2 denotes ChatGPT-4's performance on the "understand" level. The grades assigned by the markers are consistent across the questions and AI outputs, with the highest deviation being 10 marks on two occasions.



	G4.1	G4.2	G4.3	G4.4	G4.5	G4.6	G3.1	G3.2	G3.3	G3.4	G3.5	G3.6	Ba.1	Ba.2	Ba.3	Ba.4	Ba.5	Ba.6	Bi.1	Bi.2	Bi.3	Bi.4	Bi.5	Bi.6
Marker 3	80	70	65	55	62	40	65	50	40	55	52	40	60	60	40	55	60	45	40	55	55	57	40	25
Marker 4	85	70	60	60	60	40	65	55	40	55	50	40	55	55	40	55	60	48	30	50	55	50	30	25

Figure 2. Markers' consistency on text input questions

Note. G4.1 denotes ChatGPT-4's performance on the "remember" level. G4.2 denotes ChatGPT-4's performance on the "understand" level. Similarly, the indexes after the "." 3, 4, 5 and 6,correspond to the "apply", "analyse", "evaluate" and "create" levels, respectively. The same rule is applied to the remaining notations, where G3 denotes ChatGPT-3.5 and Ba and Bi stand for Google Bard and Microsoft Bing, respectively.



	G4.1	G4.2	G4.3	G4.4	G4.5	G4.6	G3.1	G3.2	G3.3	G3.4	G3.5	G3.6	Ba.1	Ba.2	Ba.3	Ba.4	Ba.5	Ba.6	Bi.1	Bi.2	Bi.3	Bi.4	Bi.5	Bi.6
Marker 1	75	75	60	65	75	65	50	65	55	70	70	70	45	50	65	65	55	55	40	45	55	50	50	45
Marker 2	80	75	65	70	73	60	50	65	60	75	70	70	50	55	60	60	60	55	40	40	55	50	50	50

Figure 3. Markers' consistency on numerical input questions

Note. G4.1 denotes ChatGPT-4's performance on the "remember" level. G4.2 denotes ChatGPT-4's performance on the "understand" level. Similarly, the indexes after the "." 3, 4, 5 and 6,correspond to the "apply", "analyse", "evaluate" and "create" levels, respectively. The same rule is applied to the remaining notations, where G3 denotes ChatGPT-3.5 and Ba and Bi stand for Google Bard and Microsoft Bing, respectively.

Performance on numerical input and text input

Figures 4 and 5 illustrate the maximum and minimum grades attainable using the top-performing AI tool for each question level, covering both numerical and text-based formats. AI generally performed better with numerical data than text, but it matched numerical performance in the "remember", "understand" and "apply" levels, likely due to extensive data set training. However, AI's capability in text questions notably declines at higher levels like "analyse", "evaluate" and "create", with all models struggling especially in "create" (shown in Figure 5). Figure 6 presents the average grades for each AI tool, with ChatGPT-4 leading in both question types and Microsoft Bing ranking lowest.

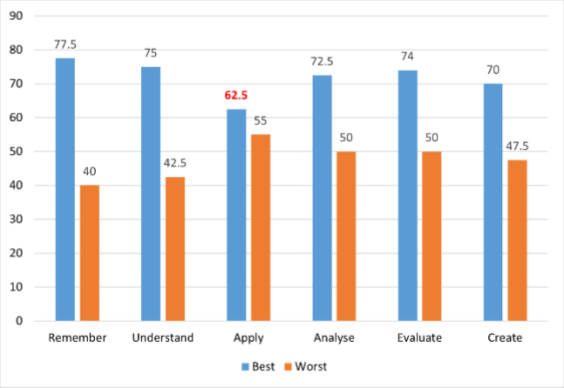


Figure 4. AI performance on numeric input

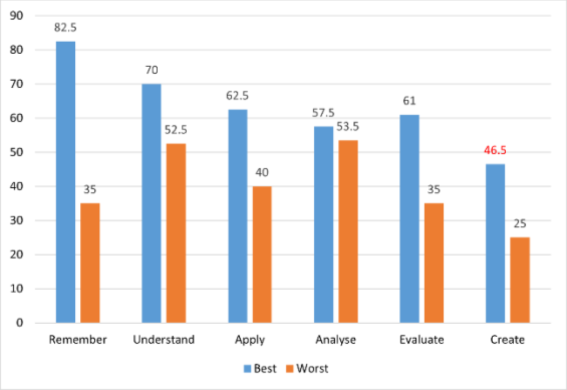


Figure 5. AI performance on text input

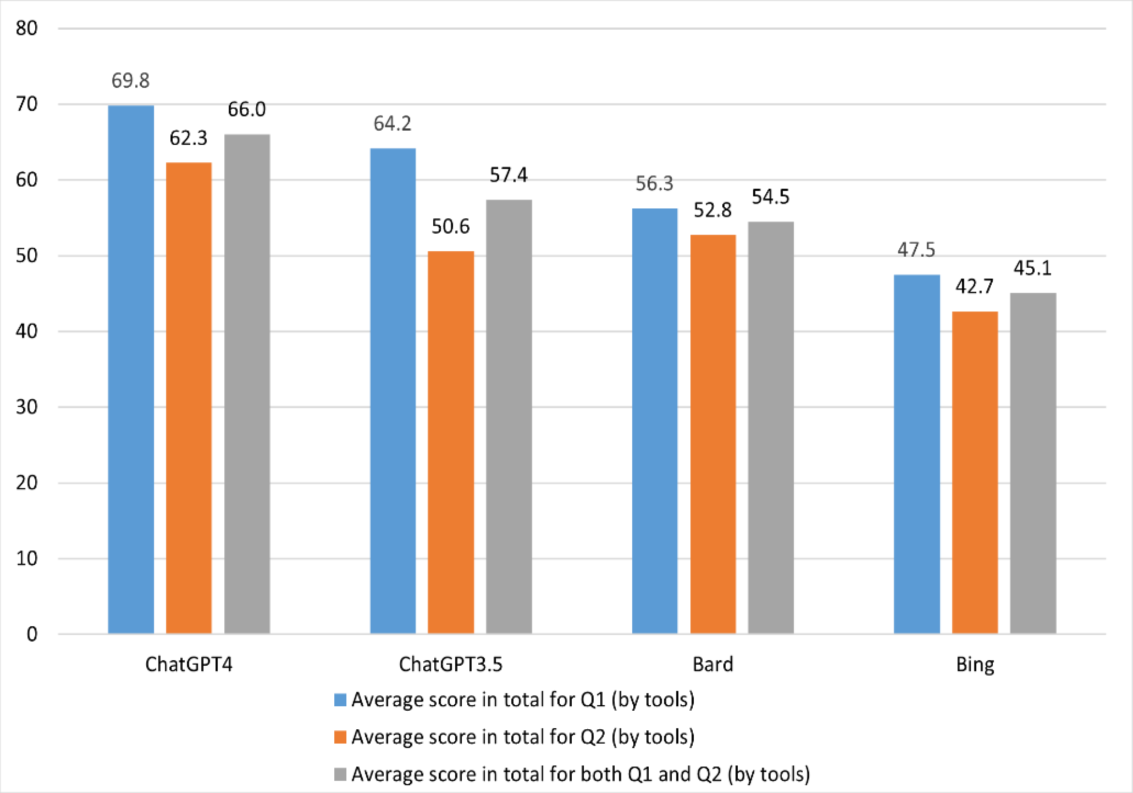


Figure 6. Average score by tools

Performance across Bloom's taxonomy levels
Figures 7, 8, 9, 10, 11, 12 and 13 display the performance of each AI tool on each kind of input at each of Bloom's taxonomy levels.

"Remember" level

ChatGPT-4 excelled with an average score of 80 (Figure 8), outperforming other tools, especially in numerical-based questions. It uniquely generated accurate graphs and correctly identified countries with higher GDP and GDP per capita. ChatGPT-4 also demonstrated superior skills in commenting on and interpreting results.

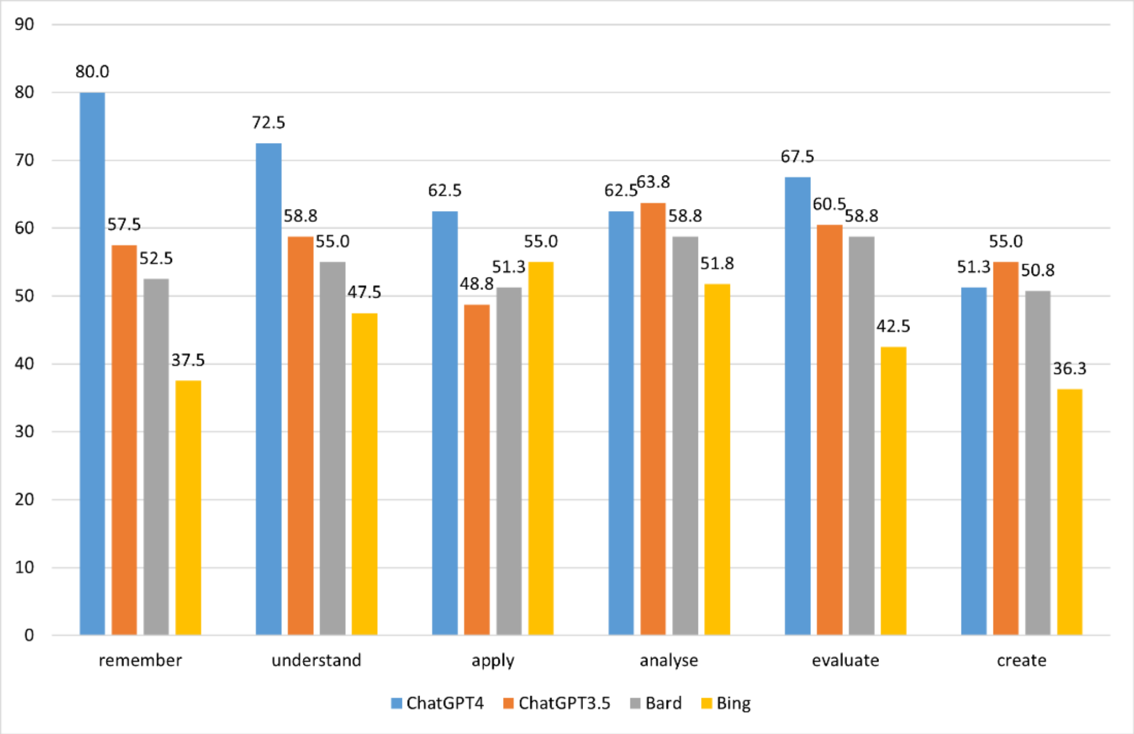


Figure 7. Tool's average score by levels

“Understand” level

ChatGPT-4 remains the best but scored slightly lower (72.5) in the “remember” task (Figure 7). In commenting tasks, ChatGPT-3.5 matched ChatGPT-4's performance, offering well-structured analyses and clear argumentation. However, it shows limitations in explaining trends. ChatGPT-3.5's responses were chronological, detailing concepts like real and nominal GDP and their differences.

“Apply” level

ChatGPT-4 leads in this task with an average of 62.5, but its responses lacked specific theoretical frameworks. For instance, when analysing GDP growth drivers, the connection to factors of production and GDP's four major components is missing, a common issue across all four AI tools' answers.

“Analyse” level

ChatGPT-3.5, excelling in the task, scored an average of 63.75 (Figure 7). All four tools struggled to connect content in their answers. For instance, they failed to link the causes of unemployment with its types, with Google Bard and Microsoft Bing notably lacking any such connections.

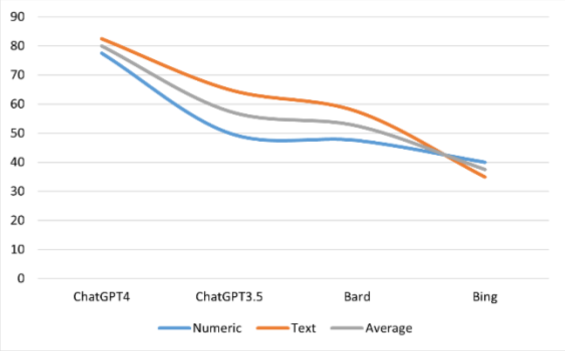


Figure 8. “Remember” level

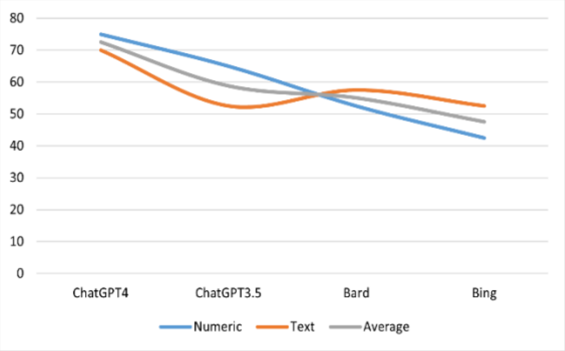


Figure 9. “Understand” level

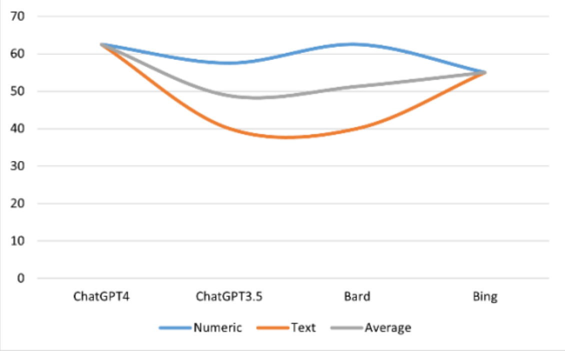


Figure 10. “Apply” level

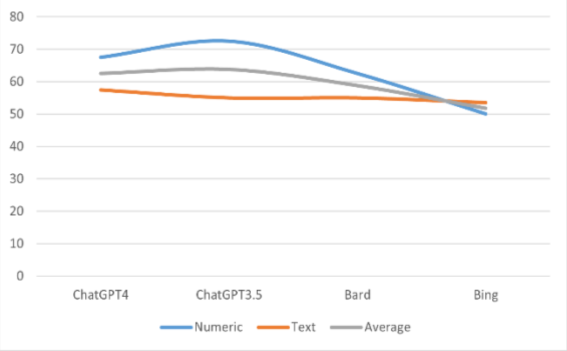


Figure 11. “Analyse” level

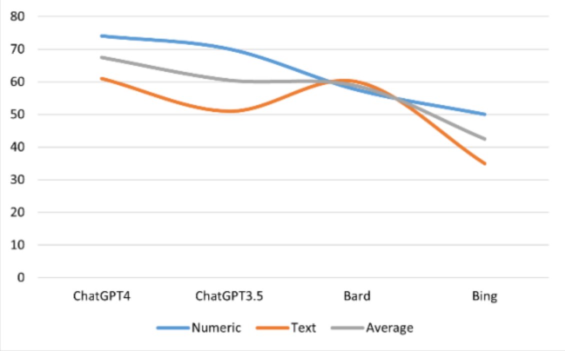


Figure 12. “Evaluate” level

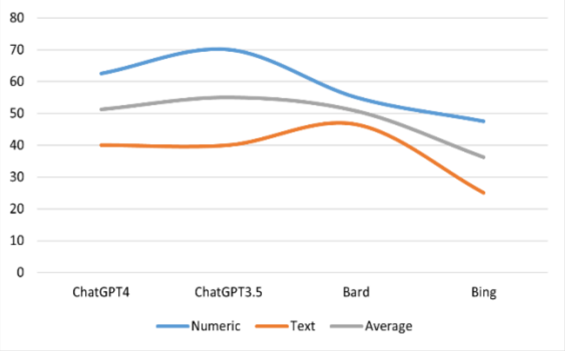


Figure 13. “Create” level

“Evaluate” level

ChatGPT-4 leads with an average of 67.5 (Figure 12), highlighting the performance gap between OpenAI's AIs (ChatGPT-3.5 and 4) and others. In tasks with numerical inputs, ChatGPTs demonstrated superior analytical skills by comparing growth prospects between developing and developed countries. Conversely, Google Bard and Microsoft Bing limited their analysis to individual country groups without cross-comparisons or in-depth analysis. ChatGPTs also showed an ability to relate analyses to a country's specific situation, though further clarification is needed.

“Create” level

ChatGPT-3.5, leading in this task, scored an average of 55, just above the pass mark (Figure 13). It shows superior analytical skills in numerical questions compared to others, which scored lower. A common shortfall across all AIs is their inability to connect answers to real-world situations, offering explanations without practical examples.

What generative AI is good at

Data analysis

The study highlights AI's proficiency in data analysis and graph generation. ChatGPT-4, scoring 77.5, 75, 67.5 and 74 in “remember”, “understand”, “analyse” and “evaluate” tasks, respectively (Figures 8–11), stands out among the tools. It excelled in generating accurate charts and identifying countries with high GDP. Moreover, its superior performance in interpreting results and providing insightful, relevant responses showcases generative AI's capability in analysing numerical data and creating graphical presentations.

“Remember” and “Understand” levels

ChatGPT outperformed other generative AI tools, particularly at the “remember” and “understand” levels, which align with traditional economics teaching focused on memorisation (Woldab, 2013). Its proficiency in these foundational areas highlights two key points. First, it demonstrates the usefulness of generative AI in efficiently synthesising knowledge, supporting Halaweh’s (2023) and Cooper’s (2023) observations about AI's capability in generating reflective and knowledge-based responses. Second, it presents a challenge to conventional assessment methods, as advanced AI tools can quickly address queries requiring knowledge reflection or synthesis. This aligns with Chiu et al.’s (2023) implication of AI potentially replacing student efforts in memorisation and factual processing tasks.

What generative AI is not good at

Reference provision

Microsoft Bing distinguished itself as the only model capable of generating numerous links to real references, yet their relevance to the topic was often unclear. This ambiguity is due to LLMs being designed for text generation rather than information retrieval, a function typical of search engines like Google.

Theoretical framework

All generative AI tools fell short in providing economic theories to substantiate their answers. For example, the expected theoretical framework for GDP discussion should include consumption, investment, government and export, which were not discussed by AI.

Coherence

Generative AI tools struggled to coherently link elements in their responses, such as adequately connecting causes of unemployment to its types. Both Google Bard and Microsoft Bing particularly failed to establish these connections. Overcoming this limitation requires deep analytical skills, essential for all students.

Provision of real-world evidence

A common weakness identified in AI responses across all levels is their failure to incorporate real-life examples or evidence. The AI predominantly offered theoretical perspectives without practical relevance to the specific country in question. This limited the effectiveness of their arguments. Even ChatGPT-4, while suggesting policies, did not clearly relate them to Thailand's current situation or its economic strengths and vulnerabilities. The AI's proposals also lacked concrete evidence and illustrative examples, crucial for demonstrating the feasibility of these policies.

“Create” level with text input

All AI models struggled with “create” level text-based questions, with the best-performing model, Google Bard, scoring only 47, below the pass mark of 50 (Figure 13). Their responses showed limited exploration and missed key rubric points, such as analysing economic growth or evaluating AI technology's efficacy. The strategies lacked justification and causal explanations. Interestingly, these tools fared better with numerical input at the same level; for example, ChatGPT-3.5 scores an average of 70. This disparity is likely due to the AI's difficulty in relating text-based questions to real-world scenarios, a challenge less evident in numerical questions where responses are based on provided data sets.

Overall evaluation

Our research reveals that generative AI tools exhibit strong performance at the lower tiers of Bloom’s taxonomy, maintain a decent performance at the levels of “apply”, “analyse” and “evaluate” but falter significantly at the “create: level. This pattern indicates that these AI tools might have the capacity to tackle the cognitive challenges and evaluative judgements that educators seek to cultivate in students via authentic assessments (Villarroel et al., 2018), particularly when the questions are targeted at the lower to middle levels of Bloom’s taxonomy of cognitive levels. This observation raises a significant concern, as it suggest that students might merely need to input the assignment questions and rubrics into these AI tools to achieve satisfactory to high marks in assessments (Halaweh, 2023; Lodge et al., 2023).

This outcome highlights a critical need for educators to re-examine and refine their assessment designs, ensuring that the assessments truly and effectively evaluate the skills and abilities that students are expected to acquire and demonstrate. The goal is to create an assessment environment that genuinely fosters learning and the development of critical skills, rather than one that could potentially be circumvented with AI tools (Lodge et al., 2023; Nguyen et al., 2023).

Conclusion

In this study, we developed a framework based on Bloom's taxonomy, allowing educators to evaluate the capabilities of widely used generative AI tools in assessments, with particular emphasis on authentic assessments. We applied this framework to assess the abilities of ChatGPT-4, ChatGPT-3.5, Google Bard, and Microsoft Bing in solving authentic assessments in economics. Specifically, we created six numeric-based and six text-based questions corresponding to the six levels of educational goals in Bloom's taxonomy. We presented these questions, along with the corresponding rubrics, to the four generative AI tools and retrieved their answers. These answers were then independently graded by four academics with many years of experience teaching economics and business courses. Our proposed framework and findings help educators and tertiary education to navigate through the rising challenges with regard to assessment integrity and design as pointed out by Lodge et al. (2023).

Generative AI tools excel at lower levels of Bloom's taxonomy but show weaker performance at higher levels, particularly in ”create” tasks. They handle numeric questions better than text-based ones. Across all tools, there are difficulties in constructing arguments based on theoretical frameworks, maintaining argument coherence and providing relevant references. Performance may hinge on prompt engineering skills, assuming only basic user abilities like copying or uploading data, questions and rubrics into the AI tools in our study.

AI tools are proficient in addressing the "remember" and "understand" levels of Bloom’s taxonomy, reflecting the impact of technology on education. Similar to how the Internet and personal computers reduced reliance on rote memorisation (Freeman, 1997), generative AI could aid tasks requiring basic recall and understanding. This technological progress affects both educational design and objectives, highlighting the necessity to focus on learning goals beyond these fundamental levels.

Moreover, ChatGPT-4, along with its Code Interpreter tool, represents a significant advancement in the field of data analysis and interpretation. With the ability to generate graphs from data sets and to analyse and interpret these graphs, ChatGPT-4 performs at a high level across nearly all educational goals, even achieving a minimum of the CR mark band for the “create” level of learning goals. This impressive capability presents a critical challenge for educators and universities, especially in fields like statistics or econometrics that heavily rely on data analysis. The ease with which students can now produce reports and data analyses through generative AI – by simply providing data, questions and marking rubrics – necessitates a comprehensive re-evaluation of course learning outcomes, content and assessments.

Our findings reveal that despite their growing capabilities, generative AI tools struggle to apply economic theory in forming coherent arguments. This presents an opportunity for students to enhance AI-provided foundations. Instead of merely gathering information and forming basic arguments, students should adopt a nuanced role, integrating AI outputs within relevant theoretical and contextual frameworks. By embedding these elements into appropriate theories and contexts, students can add depth to their responses, creating more compelling and persuasive arguments.

As generative AI advances and increasingly automates tasks that require lower levels of educational goals, educators need to shift their focus to designing authentic assessments and learning activities that emphasise higher-level competencies within Bloom's taxonomy. Specifically, the "create" level, where generative AI tools currently exhibit limited capabilities, is of paramount importance.

With generative AI becoming increasingly prevalent in the workplace, universities urgently need to focus on higher-level educational goals to prepare graduates for an AI-integrated work environment. Professionals, especially in data- and text-based fields, will see their roles transform significantly, requiring a shift in educational strategies. This evolution demands a proactive response from educators and institutions, involving enhanced educator capabilities and a thorough redesign of assessments, course outcomes and academic programmes, to meet the challenges posed by generative AI.

Author contributions

Binh Nguyen Thanh: Conceptualisation, Investigation, Writing – original draft, Writing – review and editing, Supervision; **Diem Thi Hong Vo**: Investigation, Methodology, Conceptualisation, Writing – original draft, Writing – review and editing; **Minh Nguyen Nhat**: Investigation, Methodology, Writing – original draft, Writing – review and editing; **Thi Thu Tra Pham**: Investigation, Methodology, Writing – review and editing; **Hieu Thai Trung**: Visualisation, Writing – original draft, Writing – review and editing; **Son Ha Xuan**: Data curation, Writing – review and editing, Project administration.

References

Aghion, P., & Durlauf, S. (Eds.). (2005). *Handbook of economic growth* (Vol. 1A, 1st ed.). Elsevier.

Armstrong, P. (2010). *Bloom's taxonomy*. Vanderbilt University Center for Teaching. https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy

Ashford-Rowe, K., Herrington, J., & Brown, C. (2014). Establishing the critical elements that determine authentic assessment. *Assessment and Evaluation in Higher Education 39*(2), 205–222. https://doi.org/10.1080/02602938.2013.819566

Athanassiou, N., McNett, J. M., & Harvey, C. (2003). Critical thinking in the management classroom: Bloom's taxonomy as a learning tool. *Journal of Management Education, 27*(5), 533–555. https://doi.org/10.1177/1052562903252515

Attia, A. (2021). Bloom's Taxonomy as a tool to optimize course learning outcomes and assessments in architecture programs. *Journal of Applied Science and Engineering, 24*(3), 315–322. https://doi.org/10.6180/jase.202106_24(3).0006

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. McKay.

Bosco, A. M., & Ferns, S. (2014). Embedding of authentic assessment in work-integrated learning curriculum. *Asia-Pacific Journal of Cooperative Education, 15*(4), 281–290. https://www.ijwil.org/files/APJCE_15_4_281_290.pdf

Callaghan-Koru, J. A., & Aqil, A. R. (2022). Theory-informed course design: Applications of Bloom's taxonomy in undergraduate public health courses. *Pedagogy in Health Promotion*, *8*(1), 75–83. https://doi.org/10.1177/2373379920979684

Chiu, T. K. F., Moorhouse, B. L., Chai, C. S., & Ismailov, M. (2023). Teacher support and student motivation to learn with artificial intelligence (AI) based chatbot. *Interactive Learning Environments,* 1–17. https://doi.org/10.1080/10494820.2023.2172044

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology, 32*(3), 444–452. https://doi.org/10.1007/s10956-023-10039-y

Crawford, J., Cowling, M., & Allen, K.-A. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching & Learning Practice, 20*(3), Article 2. https://doi.org/10.53761/1.20.3.02

Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in Bloom: Implementing Bloom's taxonomy to enhance student learning in biology. *CBE Life Sciences Education, 7*(4), 368–381. https://doi.org/10.1187/cbe.08-05-0024

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y.-S., Gašević, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. In M. Chang, N.-S, Chen, R. Kuo, G. Rudolph, D. G. Sampson, & A. Tlili (Eds.), *Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies* (pp. 323–325). IEEE. https://doi.org/10.1109/ICALT58122.2023.00100

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education, 16*(5-6), 523–545. https://doi.org/10.1016/S0742-051X(00)00015-9

Dawson, P. (2020). *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education* (1st ed.). Routledge. https://doi.org/10.4324/9780429324178

DeMara, R. F., Tian, T., & Howard, W. (2019). Engineering assessment strata: A layered approach to evaluation spanning Bloom’s taxonomy of learning. *Education and Information Technologies, 24*(2), 1147–1171. https://doi.org/10.1007/s10639-018-9812-5

Freeman, M. (1997). Flexibility in access, interaction and assessment: The case for web-based teaching programs. *Australasian Journal of Educational Technology, 13*(1), 23–39. https://doi.org/10.14742/ajet.1917

Greenstein, L. M. (2012). *Assessing 21st century skills: A guide to evaluating mastery and authentic learning*. Corwin Press.

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52*(3), 67–86. https://doi.org/10.1007/BF02504676

Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. *Contemporary Educational Technology, 15*(2), article ep421. https://doi.org/10.30935/cedtech/13036

Herrington, J., & Oliver, R. (2000). An instructional design framework for authentic learning environments. *Educational Technology Research and Development, 48*(3), 23–48. https://doi.org/10.1007/BF02319856

Ho, H. K., & Chng, H. T. (2021). Stirring deep thinking and learning through student-designed assessment problems. *Currents in Pharmacy Teaching and Learning, 13*(5), 536–543. https://doi.org/10.1016/j.cptl.2021.01.007

Hussey, T., & Smith, P. (2002). The trouble with learning outcomes. *Active Learning in Higher Education, 3*(3), 220–233. https://doi.org/10.1177/1469787402003003003

James, L. T., & Casidy, R. (2018). Authentic assessment in business education: its effects on student satisfaction and promoting behaviour. *Studies in Higher Education, 43*(3), 401–415. https://doi.org/10.1080/03075079.2016.1165659

Kar, S. P., Chatterjee, R., & Mandal, J. K. (2022). Automated assessment – An application in authentic learning using Bloom’s taxonomy. In M. E. Auer, W. Pachatz, & T. Rüütmann (Eds.), *Lecture notes in networks and systems: Vol. 634*. *Learning in the age of digital and green transition* (pp. 747–756). Springer. https://doi.org/10.1007/978-3-031-26190-9_78

Karanja, E., & Malone, L. C. (2021). Improving project management curriculum by aligning course learning outcomes with Bloom’s taxonomy framework. *Journal of International Education in Business, 14*(2), 197–218. https://doi.org/10.1108/JIEB-05-2020-0038

Katz, A., Wei, S., Nanda, G., Brinton, C., & Ohland, M. (2023). *Exploring the efficacy of ChatGPT in analyzing student teamwork feedback with an existing taxonomy*. arXiv. https://doi.org/10.48550/arXiv.2305.11882

Kibble, J. D., & Johnson, T. (2011). Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? *Advances in Physiology Education, 35*(4), 396–401. https://doi.org/10.1152/advan.00062.2011

Koh, K. H. (2017). Authentic assessment. In G. W. Noblit (Es.), *Oxford research encyclopedia of education*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.22

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Lau, K. H., Lam, T. K., Kam, B. H., Nkhoma, M., & Richardson, J. (2018). Benchmarking higher education programs through alignment analysis based on the revised Bloom's taxonomy. *Benchmarking : An International Journal, 25*(8), 2828–2849. https://doi.org/10.1108/BIJ-10-2017-0286

Lodge, J. M., Thompson, K., & Corrin, L. (2023). Mapping out a research agenda for generative artificial intelligence in tertiary education. *Australasian Journal of Educational Technology, 39*(1), 1–8. https://doi.org/10.14742/ajet.8695

Manville, G., Donald, W. E., & Eves, A. (2022). Can embedding authentic assessment into the curriculum enhance the employability of business school students? *GILE Journal of Skills Development, 2*(2), 73–87. https://doi.org/10.52398/gjsd.2022.v2.i2.pp73-87

Matore, M. E. E. M. (2021). Rasch model assessment for Bloom digital taxonomy applications. *Computers, Materials & Continua 68*(1), 1235–1253. https://doi.org/10.32604/cmc.2021.016143

Megahed, F. M., Chen, Y.-J., Ferris, J. A., Knoth, S., & Jones-Farmer, L. A. (2023). How generative AI models such as ChatGPT can be (mis)used in SPC practice, education, and research? An exploratory study. *Quality Engineering*, 1–29. https://doi.org/10.1080/08982112.2023.2206479

Na, S.-J., Ji, Y. G., & Lee, D. H. (2021). Application of Bloom's taxonomy to formative assessment in real-time online classes in Korea. *Korean Journal of Medical Education, 33*(3), 191–201. https://doi.org/10.3946/KJME.2021.199

Neely, P., & Tucker, J. (2012). Using business simulations as authentic assessment tools. *American Journal of Business Education, 5*(4), 449–456. https://doi.org/10.19030/ajbe.v5i4.7122

Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B.P. T. (2023). Ethical principles for artificial intelligence in education. *Education and Information Technologies, 28*(4), 4221–4241. https://doi.org/10.1007/s10639-022-11316-w

Pappas, E., Pierrakos, O., & Nagel, R. (2013). Using Bloom's taxonomy to teach sustainability in multiple contexts. *Journal of Cleaner Production, 48*, 54–64. https://doi.org/10.1016/j.jclepro.2012.09.039

Parwata, I. G. A. L., Jayanta, I. N. L., & Widiana, I. W. (2023). Improving metacognitive ability and learning outcomes with problem-based revised Bloom's Taxonomy oriented learning activities. *Emerging Science Journal, 7*(2), 569–577. https://doi.org/10.28991/ESJ-2023-07-02-019

Pepin, M., Audebrand, L. K., Tremblay, M., & Keita, N. B. (2021). Evolving students' conceptions about responsible entrepreneurship: A classroom experiment. *Journal of Small Business and Enterprise Development, 28*(4), 570–585. https://doi.org/10.1108/JSBED-02-2020-0035

Peres, R., Schreier, M., Schweidel, D., & Sorescu, A. (2023). Editorial: On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *International Journal of Research in Marketing, 40*(2), 269–275. https://doi.org/10.1016/j.ijresmar.2023.03.001

Qadir, J. (2023). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. In *Proceedings of the 2023 IEEE Global Engineering Education Conference* (pp. 1–9). IEEE. https://doi.org/10.1109/EDUCON54358.2023.10125121

Ramirez, T. V. (2017). On pedagogy of personality assessment: Application of Bloom's taxonomy of educational objectives. *Journal of Personality Assessment, 99*(2), 146–152. https://doi.org/10.1080/00223891.2016.1167059

Samuelson, P. A., & Nordhaus, W. D. (2013). *Economics* (19th ed.). McGraw Hill.

Shane, H. G. (1981). Significant writings that have influenced the curriculum: 1906-81. *The Phi Delta Kappan, 62*(5), 311–314. https://www.jstor.org/stable/20385884

Stanny, C. J. (2016). Reevaluating Bloom's taxonomy: What measurable verbs can and cannot say about student learning. *Education Sciences, 6*(4), Article 37. https://doi.org/10.3390/educsci6040037

Strzelecki, A. (2023). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments*, 1–14. https://doi.org/10.1080/10494820.2023.2209881

Su, J., & Yang, W. (2023). Unlocking the power of ChatGPT: A framework for applying generative AI in education. *ECNU Review of Education*, *6*(3), 355–366. https://doi.org/10.1177/20965311231168423

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*, Article 100752. https://doi.org/10.1016/j.asw.2023.100752

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: enabling students to make decisions about the quality of work. *Higher Education 76*, 467–481. https://doi.org/10.1007/s10734-017-0220-3

Taylor, J. B., (2000). Teaching modern macroeconomics at the principles level. *American Economic Review*, *90*(2), 90–94. https://doi.org/10.1257/aer.90.2.90

Thompson, A. R., Braun, M. W., & O'Loughlin, V. D. (2013). A comparison of student performance on discipline-specific versus integrated exams in a medical school course. *Advances in Physiology Education, 37*(4), 370–376. https://doi.org/10.1152/advan.00015.2013

Thompson, A. R., & O'Loughlin, V. D. (2015). The Blooming Anatomy Tool (BAT): A discipline-specific rubric for utilizing Bloom's taxonomy in the design and evaluation of assessments in the anatomical sciences. *Anatomical Sciences Education, 8*(6), 493–501. https://doi.org/10.1002/ase.1507

Tian, Y., Chen, H., Wang, X., Bai, Z., Zhang, Q., Li, R., Xu, C., & Wang, Y. (2023). *Multiscale positive-unlabeled detection of AI-generated texts*. arXiv. https://doi.org/10.48550/arXiv.2305.18149

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment and Evaluation in Higher Education 43*(5), 840–854. https://doi.org/10.1080/02602938.2017.1412396

West, J. (2023). Utilizing Bloom's taxonomy and authentic learning principles to promote preservice teachers' pedagogical content knowledge. *Social Sciences & Humanities Open, 8*(1), Article 100620. https://doi.org/10.1016/j.ssaho.2023.100620

Wiewiora, A., & Kowalkiewicz, A. (2019). The role of authentic assessment in developing authentic leadership identity and competencies. *Assessment & Evaluation in Higher Education, 44*(3), 415–430. https://doi.org/10.1080/02602938.2018.1516730

Wiggins, G. P. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. Jossey-Bass. https://psycnet.apa.org/record/1993-98969-000

Woldab, Z. E. (2013). Constructivist didactics in teaching economics: A shift in paradigm to be exemplary teacher. *Academic Journal of Interdisciplinary Studies, 2*(1), 197–203. https://doi.org/10.5901/ajis.2013.v2n1p197

Zaidi, N. B., Hwang, C., Scott, S., Stallard, S., Purkiss, J., & Hortsch, M. (2017). Climbing Bloom's taxonomy pyramid: Lessons from a graduate histology course. *Anatomical Sciences Education, 10*(5), 456–464. https://doi.org/10.1002/ase.1685

Zheng, A. Y., Lawhorn, J. K., Lumley, T., & Freeman, S. (2008). Assessment: Application of Bloom's taxonomy debunks the “MCAT myth”. *Science (American Association for the Advancement of Science), 319*(5862), 414–415. https://doi.org/10.1126/science.1147852

Corresponding author: Diem Thi Hong Vo, diem.vo@rmit.edu.vn

Please cite as: Thanh, B. N., Vo, D. T. H., Nhat, M. N., Pham, T. T. T., Trung, H. T., & Xuan, S. H. (2023). Race with the machines: Assessing the capability of generative AI in solving authentic assessments. *Australasian Journal of Educational Technology*, *39*(5), 59-81. https://doi.org/10.14742/ajet.8902

Appendix A: Proposed questions based on Bloom's taxonomy

Question 1: Numerical Data Question and Requirements

(Given data: Collect data on nominal GDP, real GDP, nominal GDP per capita, real GDP per capita, real GDP growth, unemployment during the 2007-2022 period (or up to the time that data are available) for Austria and Thailand)

Question 1 - Remember:
Plot both countries' GDP on a line chart and GDP per capita on another line chart. Which country has a higher GDP? Which country has a higher GDP per capita? Are they the same country?
Provide some comments.

Question Requirements:

- Illustrate the GDP and GDP per capita of each country based on the given data.
- The expected answers can reflect the data by correctly drawing the country with higher GDP and GDP per capita, based on the given data.
- The expected answers can recall the meanings and differences between GDP and GDP per capita concepts.

Question 1 - Understand:
Choose one country and plot its nominal and real GDP on a line chart. Compare and contrast the trend and fluctuations of nominal and real GDP.
Question Requirements:

- Illustrate the nominal and real GDP growth of a chosen country, based on the given data.
- The expected answers can explain and interpret the differences between nominal and real GDP as well as the reasons for those differences.

Question 1 - Apply:
Plot the GDP growth for both countries. Which country experienced a higher economic growth?
Identify two main growth drivers of this country.

Question Requirements:

- Illustrate the nominal and real GDP growth of both countries, based on the given data.
- Showing understanding of the theory of economic growth and application to the context of the selected country.
- Shows understanding and provides a comprehensive comparison of the economic growth between the two countries.

Question 1 - Analyse:
Select one country and the period of its highest unemployment. Analyse the type(s) and cause(s) of unemployment during this period.

Question Requirements:

- Accurately identifies the type(s) of unemployment and provides insightful and detailed explanations of the unemployment type(s) (Frictional, structural, cyclical or classic unemployment).
- Provides an analysis of the causes that significantly determine the unemployment of the chosen country using the relevant theoretical framework. The analysis is well-supported by evidence and shows a deep understanding of the relationship between these causes and unemployment.
- Explanations are supported by reliable and relevant evidence.

Question 1 - Evaluate:
Evaluate the growth potential of both countries. What are the differences in growth prospects between developed and developing countries?

Question Requirements:

- The answers evaluate the economic growth of countries by providing an accurate and comprehensive evaluation of the economic growth of both countries in the given period, with arguments based on a relevant theoretical framework.
- The evaluation on economic growth should link to the key growth drivers for each country should be identified and analysed.
- Deliver contrast of the economic growth prospects of developed and developing countries, supported by reliable evidence.
- Identifies and provides explanations of at least two potential factors leading to the differences in the growth prospects of developed and developing countries.

Question 1 - Create:
Select one country. Based on your analysis, recommend a policy to foster its economic growth.

Question Requirements:

- Provides an understanding of the GDP growth drivers or constraints of the selected country, with detailed explanation and accurate evaluation on the potential effects. Arguments are based on a relevant theoretical framework.
- The proposed policies are explained, made based on relevant arguments, and supported by reliable evidence.
- Clear and logical connection between the proposed policies and the identified drivers or constraints.

Question 2: Text Question and Requirements

Question 2 - Remember:
What are the distinctions between GDP and GDP per capita? Elaborate the distinctions.

Question Requirements:

- The expected answers can correctly recall the definitions of GDP and GDP per capita.
- The answer differentiates and explains the role of GDP and GDP per capita in determining the economy's size versus the economy's wealth (or richness).
- Provides an explanation on the relationship between GDP and GDP per capita.

Question 2 - Understand:
Suppose a country's nominal GDP grows faster than its real GDP. What could explain this phenomenon?

Question Requirements:

- Demonstrates an understanding of the relationship between real GDP and nominal GDP.
- Provide understanding of the causes leading to faster growth of nominal GDP than real GDP.

Question 2 - Apply:
Critique the assertion that rich countries consistently grow at a slower pace than poor countries.

Question Requirements:

- Clearly state the agreement with the statement with an analysis of key reasons that determine the slower growth of rich countries compared to poor countries.
- The answer needs to well-deliver that the convergence hypothesis or the catch-up effect, which states that countries started out poor tend to grow faster given the same unit of additional capital investment, and the growth will slow down due to the law of diminishing return.
- The analyses must be based on a relevant theoretical framework.

Question 2 - Analyse:
Reflect critically on the idea that technologies such as automation, robotics, and AI could exacerbate unemployment. Which type(s) of unemployment could be most affected?

Question Requirements:

- Analyse the potential benefits and risks that might come from the adoption of emerging technology on employment and unemployment of a country.
- Provide analysis and evaluation of the impact ability of technology on unemployment in countries with different levels of economic development. The arguments are supported by relevant evidence and comparisons.

Question 2 - Evaluate:
Evaluate the effects of AI adoption on economic growth in developing and developed countries.

Question Requirements:

- Analyses the impact of AI adoption on economic growth by identifying, explaining, and interpreting the potential effects and their implications.
- Demonstrate the impact of AI adoption as a new technology that provides productivity.
- Provide comparison between the impact of AI adoption on economic growth in developing versus developed countries with compelling arguments.

Question 2 - Create:
Select a developed and a developing country. What strategies could each employ to leverage AI technology for sustainable growth?

Question Requirements:

- Analyses the advantages and constraints to adopt AI technology based on the economic status of each country.
- Delivers evaluation of the usefulness of adopting AI technology for each country, highlighting the usefulness of AI technology adoption in each country and explaining the difference of usefulness of AI technology adoption between developed and developing country.
- Proposes strategies with explanations of causes and effects. Arguments are well-connected and substantiated with appropriate references, evidence, or comparisons.

Appendix B: Prompt formulation process for generative AI

The approach starts with providing the AI with a data set, especially for data set-dependent questions like Question 1: Numerical Data Question and Requirements. We then contextualise the AI's role, such as an economics student, to focus its responses on the relevant domain. Specific questions are articulated as listed in Appendix A. Finally, we impose conditions like word limits or adherence to a reference rubric. This structured approach ensures that AI-generated responses are coherent, academically sound and relevant.

(1) **Forming the data set**: The prompting process begins by uploading an Excel file to form a data set, which is fundamental to how the generative AI functions. This data set, aligned with the relevant question, directs the AI's understanding of the task's scope and context. An example of this data set might include data on nominal GDP, real GDP, and unemployment rates.

(2) **Describing the context**: Context description is crucial for narrowing down the focus of the AI towards the desired domain. For instance, by stating "*Let's assume that you are an economics student,*" we set a specific scenario that guides the AI in tailoring its responses to be relevant to economics. This step helps in aligning the AI's output with the thematic and academic context required for the study, ensuring that the generated responses are on-point and applicable.

(3) **Providing the corresponding question:** Questions presented to the AI are crafted to be clear and directly related to the study topic. They prompt the AI to analyse the provided data set and context, generating responses that aptly address these inquiries. These questions aim to extract specific information or analysis, evaluating the AI's ability to comprehend and respond to academic questions. For instance, an "understand-level" prompt might ask to "Choose one country and compare and contrast the trend and fluctuations of its nominal and real GDP."

(4) **Providing the condition for answering the question:** Setting conditions like word limits and referencing a rubric standardizes AI responses, ensuring they meet academic and structural standards. This approach evaluates the AI's compliance with guidelines and its capacity to produce concise, relevant answers. The word limit enhances brevity, and the rubric serves as a benchmark for assessing response quality and accuracy. For example: "*In a maximum of 300 words, choose one country and compare and contrast the trend and fluctuations of its nominal and real GDP, adhering to the evaluation criteria outlined in the rubric for the Understand level question.*".

The sample prompts below, using ChatGPT-3.5 Turbo, present a complete flow to obtain a response to a Question 1: Numerical Data Question and Requirements - Understand-level question.

(1) *Collect data on nominal GDP, real GDP, nominal GDP per capita, real GDP per capita, real GDP growth, unemployment during the 2007-2022 period (or up to the time that data are available) for Austria and Thailand.*

(2) ***Let's assume that you are an economics student****, you should get the highest mark for the question: "In* ***a maximum of 300 words, choose one country and compare and contrast the trend and fluctuations of its nominal and real GDP****, adhering to the evaluation criteria outlined in the* ***rubric****".*

- *"**Let's assume that you are an economics student**": the context of the question.*
- *"**Choose one country and compare and contrast the trend and fluctuations of its nominal and real GDP**": the question.*
- *"**Maximum of 300 words**" and "**the rubric**": the condition for the answers.*

For the Question 2: Text Question and Requirements, we do not need to upload any Excel file. Thus, the sample prompt below (used for ChatGPT-3.5 Turbo) describes the complete flow to obtain the answer for the question of Understand-level.

- ***Let's assume that you are an economics student****, you should get the highest mark for the question: "In a* ***maximum of 300 words, suppose a country's real GDP grows faster than its nominal GDP. What could explain this phenomenon,*** *adhering to the evaluation criteria outlined in the* ***rubric****".*