ChatGPT as Economics Tutor: Capabilities and Limitations

Natalie Bröse, Christian Spielmann,
and Christian Tode

Discussion Paper 25/786

2 April 2025

School of Economics

University of Bristol
Priory Road Complex
Bristol
BS8 1TU
United Kingdom

University of BRISTOL
School of Economics

ChatGPT as Economics Tutor: Capabilities and Limitations

Natalie Bröse[1], Christian Spielmann[2], and Christian Tode[3]

[1]Department of Social Policy and Social Security Studies, University of Applied Sciences Bonn-Rhein-Sieg, natalie.broese@h-brs.de
[2]School of Economics, University of Bristol, christian.spielmann@bristol.ac.uk
[3]Department of Management Sciences, University of Applied Sciences Bonn-Rhein-Sieg, christian.tode@h-brs.de

April 2, 2025

Abstract

Since the public release of ChatGPT in late 2022, the role of Generative AI chatbots in education has been widely debated. While some see their potential as automated tutors, others worry that inaccuracies and hallucinations could harm student learning. This study assesses ChatGPT models (GPT-3.5, GPT-4o, and o1preview) across important dimensions of student learning by evaluating their capabilities and limitations to serve as a non-interactive, automated tutor. In particular, we analyse performance in two tasks commonly used in principles of economics courses: explaining economic concepts and answering multiple-choice questions. Our findings indicate that newer models generate very accurate responses, although some inaccuracies persist. A key concern is that ChatGPT presents all responses with full confidence, making errors difficult for students to recognize. Furthermore, explanations are often quite narrow, lacking holistic perspectives, and the quality of examples remains poor. Despite these limitations, we argue that ChatGPT can serve as an effective automated tutor for basic, knowledge-based questions—supporting students while posing a relatively low risk of misinformation. Educators can hence recommend Generative AI chatbots for student learning, but should teach students the limitations of the technology.

Keywords: generative artificial intelligence, economics education, tutoring, ChatGPT

JEL classification: A22, O33, Y2

1 Introduction

The public release of ChatGPT in late 2022 has sparked an extensive debate in education, one of the most significant in decades. While some educators see Generative AI (GenAI) as a powerful tool for learning, others fear that inaccuracies and hallucinations could mislead students and harm learning outcomes. One of the most widely discussed applications of GenAI in education is its potential role as an automated tutoring system (e.g., Kasneci et al. 2023; L. Mollick and E. Mollick 2023), building on research suggesting that one-on-one tutoring can enhance student learning (e.g., Bloom 1984; Chi et al. 2001; Slavin 1987). In this study we argue that GenAI chatbots have the capability to act as automated tutors to support student learning in principles of economics courses with minimal risks of hallucinations or misinformation. However, their explanatory depth and example quality remain insufficient.

In practice, educators face major constraints in providing individualized student support. Rising student enrollment, increasing student-faculty ratios (e.g., Buckner and Zhang 2021) and declining shares of total government expenditure on tertiary education in OECD countries (OECD 2025) all limit direct interaction between instructors and students, making personalized support outside the classroom increasingly difficult.

Beyond structural constraints, psychological factors also play a role. Many students hesitate to ask questions, especially those they perceive as "too simple", due to fear of embarrassment or negative evaluation by instructors and peers (Holodynski and Kronast 2009). Research suggests that anonymity in educational environments, such as online discussion forums, makes participation easier, more comfortable, and more enjoyable (Grasso 2017; Miyazoe and Anderson 2011; Sullivan 2002).

Against this backdrop, GenAI chatbots like ChatGPT present an opportunity. Unlike human instructors, these chatbots can provide instant feedback and explanations in an anonymous setting and without time constraints. If GenAI chatbots can effectively function as automated tutors, they could complement traditional teaching by bridging gaps in student understanding and providing personalized support at scale.

Pedagogically, GenAI chatbots may also facilitate the transition to more student-centered learning approaches, such as the flipped classroom. In the flipped classroom model, students engage with learning materials (e.g., videos, readings) before class, allowing classroom time to focus on applications, discussions, and problem-solving (Hwang, Lai, and Wang 2015). Research suggests that the flipped classroom improves student performance (G. Akçayır and M. Akçayır 2018; Zhu 2021), but its effectiveness depends on students coming to class prepared. When students struggle with content at home and cannot resolve misunderstandings until the next class session, their learning process is disrupted (Bergmann and Sams 2012). GenAI chatbots could mitigate this challenge by offering automated on-demand explanations, helping students grasp learning content during self-study.

Given the significant potential benefits of GenAI chatbots in education, this study evaluates the capabilities but also the limitations of the technology for tutoring.

Our study contributes to the growing body of research on GenAI in education, which can be categorized into four main areas. First, there are subjective reflections on GenAI use in education, based on educators' personal experiences (e.g., Kasneci et al. 2023; L. Mollick and E. Mollick 2023). These papers introduced the idea of tutoring systems by GenAI chatbots, but lack rigorous performance analysis. Second, some studies evaluate GenAI's performance in standardized tests in higher education (e.g. Kung et al. 2023; OpenAI 2023b). While these studies assess whether GenAI chatbots produce correct answers, they do not evaluate whether the output facilitates student learning. Third, student surveys examine how students perceive and use GenAI (e.g., Von Garrel and Mayer 2023; Walczak and Cellary 2023). These studies confirm that many students use GenAI as a learning tool, but do not assess how well it functions as a tutor. Finally, experimental studies investigate GenAI's impact on learning outcomes (e.g. Bastani et al. 2024; Lehmann, Cornelius, and Sting 2024). These studies suggest that GenAI chatbots can improve short-term learning outcomes when used effectively.

We expand on this literature by systematically evaluating ChatGPT's response quality in tasks that are critical for automated tutoring by GenAI chatbots in economics. Unlike prior studies that focus mainly on accuracy, we assess multiple criteria relevant to effective tutoring, including explanatory depth, quality of examples, and how clear and easy to understand the responses are. We apply a moderation method during the evaluation to ensure an objective evaluation. As all three authors are economics educators, we focus on ChatGPT's performance in the context of introductory economics courses.

To assess ChatGPT's effectiveness as a tutor, we draw on prior research on tutoring strategies. Chi et al. (2001) distinguish between interactive tutoring, which engages students in dialogue, and non-interactive tutoring, which focuses on one-directional explanations. Their research suggests that even non-interactive tutoring can be effective, particularly when explanations are of high quality and re-frame material in an intuitive way. Well-crafted explanations can fill knowledge gaps, reinforce knowledge, and prompt student reflection. However, if explanations are poorly structured or overly simplified, they may lead to only superficial learning (VanLehn et al. 2003).

As we investigate the performance as an economics tutor, we incorporate subject-specific requirements. Economists use a toolbox of concepts and simplified models to explain real-world phenomena. Furthermore, applications and real-world examples are important as they foster active engagement with the material instead of shallow memorization (e.g., Sekwena 2023).

An effective tutor in economics must therefore demonstrate strong subject knowledge, communicate concepts clearly, provide well-structured explanations, and offer meaningful real-world examples (non-interactive tutoring). Additionally, a good tutor should identify knowledge gaps and encourages student reflection (interactive tutoring). While current GenAI chatbots may struggle with interactive engagement, they may still serve as effective non-interactive tutors, particularly in explaining material accessibly.

We analyze the performance of three ChatGPT models (GPT-3.5, GPT-4o, and o1preview) in two key tasks: explaining economic concepts and answering multiple-choice questions with explanations. We use CORE Econ's The Economy 1.0 (Bowles, Carlin, and Stevens 2017) as our reference material and evaluate ChatGPT's responses across multiple dimensions relevant to student learning.

We find that newer models of ChatGPT provide very accurate responses, though some inaccuracies persist. The concerning issue is that ChatGPT presents all responses with full confidence, making it difficult for students to recognize errors. Additionally, while responses are easy to understand, they lack a holistic perspective on economic concepts, and the quality of generated examples is consistently poor. Despite these limitations, GenAI chatbots can support students as non-interactive automated tutors. We suggest that ChatGPT has the ability to function as an automated tutor in economics, particularly when utilizing GPT-4o and following models.

The remainder of the paper is structured as follows: The next section reviews the existing literature. Section 3 describes our methodological approach and data. Section 4 presents and discusses our empirical results. Section 5 presents to extensions to the analysis and section 6 reflects on the implications of our findings and provides actionable advice for educators. Section 7 concludes. The appendix provides a complete list of economic concepts and questions used in our analysis. It further presents descriptive labels for all indicators used in the evaluation.

2 Literature Review

ChatGPT, developed by OpenAI, is a chatbot built upon large language models (LLMs) that utilize deep learning techniques and transformer architectures. These models are trained on extensive datasets, predominantly sourced from the internet, enabling them to predict the most probable sequence of characters following a given input (OpenAI 2023a,b). ChatGPT generates outputs based on users' prompts. Over time, OpenAI used different large language models in ChatGPT: GPT-3.5 and GPT-3.5 Turbo (2022/2023, discontinued), GPT-4 (since March 2023), GPT-4o (since May 2024) and o1preview (since September 2024). At the time we finalised this paper, the o1preview model was replaced by the o1 model.

OpenAI acknowledged that their GPT models are not fully reliable and that hallucinations, that is generated content that is unfaithful to the provided source material but appears factual, pose significant safety challenges (OpenAI 2023a,b). Hallucinations are particularly problematic because they are often difficult to detect; the generated text is typically fluent and well-structured, giving an impression of accuracy (Ji et al. 2023). In the context of education, such hallucinations can lead to the dissemination of misconceived knowledge and false information, which students may find challenging to unlearn.

Since the introduction of advanced GenAI chatbots like ChatGPT, their potential impact on education has been widely discussed. Early literature, mostly educators' subjective reflections, highlighted concerns about academic honesty (e.g., D. R. E. Cotton, P. A. Cotton, and Shipway 2024) and critical thinking skills (Farrokhnia et al. 2024; Kasneci et al. 2023; Michel-Villarreal et al. 2023; Zirar 2023). However, there is consensus on opportunities, including the introduction of innovative learning methods like personalized learning and tutoring systems (Farrokhnia et al. 2024; Grassini 2023; Kasneci et al. 2023; Michel-Villarreal et al. 2023; Rudolph, Samson Tan, and Shannon Tan 2023). L. Mollick and E. Mollick (2023) define a GenAI tutor as a system providing personal instruction and educational guidance interactively, motivating active engagement beyond mere subject knowledge provision.

Studies showed that a majority of students have used GenAI-based systems for purposes such as information search, problem-solving, programming, and understanding difficult topics (Von Garrel and Mayer 2023; Walczak and Cellary 2023). Students recognize the potential for personalized learning support but have concerns about output accuracy (Chan and Hu 2023); only 2% of students fully trust GenAI-generated content (Walczak and Cellary 2023). Further, adoption is significantly influenced by performance expectancy, that is the perceived usefulness for effective studying (Grassini, Aasen, and Møgelvang 2024; Strzelecki 2024).

An important aspect of performance is how accurately GenAI chatbots deliver information to students. This aspect was evaluated in particular for ChatGPT. Several studies investigated its performance in higher education tasks across various fields and question types, generally showing strong results, especially with newer versions like GPT-4. OpenAI (2023b) found GPT-4 achieved human-level performance on most of 26 professional and academic exams. In medicine, GPT models performed near or above passing levels on licensing exams (Ali et al. 2023; Farhat et al. 2023; Gilson et al. 2023; Kung et al. 2023). Similar findings are reported in natural sciences and engineering (Schulze Balhorn et al. 2024), accounting (Wood et al. 2023) and operations management (Terwiesch 2023).

In the field of economics, several studies have investigated ChatGPT's ability to accurately provide information. Geerling et al. (2023) had GPT-3 take the Test of Understanding in College Economics, consisting of multiple-choice questions, where GPT-3 got 63.3% of the microeconomics questions and 86.7% of the macroeconomics questions correct. OpenAI (2023b) investigated the performance of GPT-3.5 and GPT-4 on Advanced Placement Microeconomics and Macroeconomics exams, which include both multiple-choice and free-response questions. GPT-4 received the highest score (5 out of 5) on both exams, while GPT-3.5 received a score of 2 out of 5 for Macroeconomics and 3 out of 5 for Microeconomics. Buchanan, Hill, and Shapoval (2024) asked GPT-3.5 and GPT-4 to provide brief summaries of knowledge from various branches of economics with citations, using hallucinated citations as a proxy for factual inaccuracies. While the summaries were generally satisfactory, they found that GPT-3.5 fabricated more than 30% of sources and GPT-4 fabricated 20%, indicating considerable hallucinations in the summaries. They also noted that inaccuracies increased when questions were less general and more specific.

In summary, the literature suggests that ChatGPT performs well in most evaluated situations, with newer versions performing substantially better than older ones. According to that, GPT-4 likely exceeds human performance in many applications. However, inaccuracies remain a potential problem.

Good performance, however, extends beyond the accuracy of output and depends on how students engage with the technology. Henkel et al. (2024) conducted an experiment with 500 high school students in Ghana and found that one hour of access per week to an GenAI chatbot improved math scores by what would be expected from an extra year of learning. Lehmann, Cornelius, and Sting (2024), using GPT-3.5, and Bastani et al. (2024), using GPT-4, provide further insights. In experimental settings, both studies found that while GenAI chatbot use generally improves student learning, students who use an GenAI chatbot as a personal tutor improve substantially. In contrast, students who merely ask for solutions perform worse than those without access to an GenAI chatbot, as the latter do not actively engage with the material. Additionally, Walczak and Cellary (2023) highlights how responses with false information complicate proper use of the technology, finding that only 21% of students could identify incorrect output as such.

3 Analysis

3.1 Methodology

We generated responses from ChatGPT using the models GPT-3.5, GPT-4o and o1preview. We focus on ChatGPT because research indicates that it is the most widely used GenAI chatbot among students (Von Garrel and Mayer 2023).

Our analysis investigated two use cases for a GenAI tutor. First, we examined the explanation of (economic) concepts, because Von Garrel and Mayer (2023) have shown that students most often use GenAI tools for clarifying and understanding subject-specific ideas. Second, we investigated the answering and explanation of multiple-choice questions, recognizing the importance of application of concepts in the learning process.

We generated ChatGPT responses to basic concepts from introductory economics as well as answers and explanations to multiple-choice questions, which apply introductory concepts. We evaluated these responses using a marking grid with predefined indicators essential for assessing ChatGPT's effectiveness as an automated tutor in economics education. The evaluation criteria included accuracy of output and how easy the responses are to understand for the target audience. Our marking grid also included measures to identify the extent of hallucinations and incorrect information, which will be relevant when discussing the applicability for learning. Subsequently, we performed statistical analyses on the evaluation data.

3.2 Data Description

We evaluated ChatGPT's explanations of 56 basic economic concepts introduced in CORE Econ's The Economy 1.0 (Bowles, Carlin, and Stevens 2017), an introductory economics textbook, which has received increasing attention within the discipline. The concepts analysed are those listed in CORE Econ's glossary and overlap with the material most likely taught in other introductory economics courses. It includes main economic ideas such as "Opportunity Costs", "Aggregate Demand" or "Demand Elasticity", but also covers some more advanced concepts less commonly emphasised in introductory texts, such as "Bargaining gap", "Asymmetric Information" or "Economic Rent".

We also evaluated answers to 25 multiple-choice questions which are part of CORE Econ's The Economy 1.0. Multiple-choice questions are of varying difficulty and either ask questions related to definitions or apply concepts to simple economic scenarios. Considering these are in-text questions, we believe those multiple-choice questions are representative for the learning within this introductory text. A full list of used concepts and multiple-choice questions is presented in the Appendix.

There are two main reasons why we base our analysis on CORE Econ's The Economy 1.0 material. First, since its inception over 10 years ago, CORE Econ has gained significant traction, becoming a key option for introductory economics courses. Second, the pedagogical approach used in the book is based on several principles seen as effective in the pedagogy literature. In 2024 material by CORE Econ has been taught in over 500 institutions across the world. By integrating empirical evidence, interdisciplinary insights, and interactive resources, CORE Econ emphasizes application over rote learning. This approach increases student engagement and fosters deeper understanding.

3.3 Procedure

Since students are the primary users of automated tutoring systems, we approached ChatGPT in the same manner as a student would: through the ChatGPT website interface. To understand how students might engage with ChatGPT for our two tasks, we employed a student assistant who initially experimented with different prompts and generated preliminary responses for both tasks using the GPT-3.5 model. After this initial experimentation, the remaining research was carried out by the authors of this article. We agreed on the following prompts for each task:

Explanation of concepts: `You take on the role of a university tutor for a course in economics. Assume that I have no prior knowledge of economics. Explain to me the concept of <CONCEPT NAME>. Provide a simple example at the end.`

Answering and explanation multiple-choice questions: `Following you will face a multiple-choice question. Multiple answers may be correct. Name the correct answers and provide an explanation for every answer why it is correct or incorrect. While doing so act as a tutor to a university student of economics. <MULTIPLE CHOICE QUESTION>`

Both prompts specify the role the chatbot should assume (i.e., a tutor), the context of the task (e.g., economics and the user's assumed lack of prior knowledge), and the task itself. This prompting style is consistent with recommendations in L. Mollick and E. Mollick (2023) and Korinek (2023).

Using these prompts, we generated responses for both tasks using the GPT-3.5, GPT-4o, and o1preview models. As GPT-3.5 has been discontinued at the time of writing, our study focuses on GPT-4o and o1preview. We saved each prompt along with its corresponding response and an evaluation questionnaire—representing the marking grid that we will introduce later in this section—for each model in individual Markdown files.

For each answer, one of the authors took on the role as first marker, evaluating the generated responses using the marking grid. Another author took on the role as moderator, evaluating answers themselves and afterwards comparing them against the first marker's assessment. Cases of disagreement were discussed between the markers until agreement was reached. The process of moderation yields agreed evaluation, which, in our view, is preferable to averaging the assessment of two markers. After this process we consolidated the data into a single dataset for analysis.

We constructed two marking grids, one for the explanations of concepts and one for the multiple-choice questions. The marking grids for both tasks comprise two types of indicators: First, problem indicators, which describe the question presented to ChatGPT (e.g., the difficulty). Second, response indicators, which describe the quality of the response in multiple dimensions (e.g., the accuracy of the response or the quality of application provided).

For evaluating the explanations of economic concepts, the problem indicators use ordinal scales ranging from 1 to 5, where the maximum score represent the optimal response and lower scores indicate deviations from this ideal.
Frequency measures how prevalent a concept is within and beyond economics. A concept with a high Frequency score is, e.g., "Central Bank", while "Wage-setting Curve" has a low score. **Difficulty** measures how hard we consider it is to understand the concept. An example for a concept with a Difficulty score of 5 is the "Bargaining Gap", while "Employment Rate" is an example for a concept with a Difficulty score of 1. We also marked a concept on whether is from the branch of **Microeconomics** or **Macroeconomics**. If a concept is frequently used in both micro- and macroeconomics, it is marked as neither **Microeconomics** or **Macroeconomics**. The inclusion of these indicators serve two purposes. First, prior research indicates that GenAI chatbots perform less effectively on more difficult (Gilson et al. 2023) or complex (Schulze Balhorn et al. 2024) tasks. Second, the learning mechanisms of these GenAI models suggest that concepts with extensive representation in the training data yield more accurate outputs with fewer hallucinations. Therefore, incorporating frequency of use and difficulty allows us to investigate whether these variables explain variations in output quality.

For the response indicators, we also use ordinal scales from 1 to 5. **Accuracy** measures whether there is incorrect information in ChatGPT's explanation of the concept. It is important to note that inaccurate output does not necessarily result from hallucination. Since the training data of these models are not publicly known, we cannot determine whether the inaccuracy arises because the model produces content that is unfaithful to the training data (which would count as hallucination according to Ji et al. 2023), draws on erroneous information in its training data, or simply misunderstands the question and thus responds incorrectly. Nonetheless, we do know that the website hosting CORE Econ's textbook The Economy 1.0 is included in the Common Crawl dataset.[1] Since the textbook offers accurate definitions of the concepts, as well as multiple-choice questions with explanations (but does not identify which answers are correct), it is therefore more likely that inaccurate responses stem from either hallucinations or misunderstandings of the problem.

Scope measures how holistically the concept is explained. **Specificity** measures the clarity of the explanation, in order to support the learning of new material. **Quality of Example** measures how well the used example fits the context and helps understanding the concept and **Safety** measures whether the output acknowledges and communicates the limitations, complexities and relevant sources. Descriptive labels for these indicators and their scores are given in Table 6 in Appendix A. We further classified the types of inaccuracies in two dimensions. First, **Error Type** captures the type of the inaccuracies (adapted from Zheng, Huang, and Chang (2023)): Comprehension, factual or inference error. Second, **Error Domain** classifies whether the inaccuracies would be noticeably from the conversation alone (closed) or whether research outside of the conversation is required (open, adapted from Bubeck et al. (2023)). The groups for these classifications and their descriptions are given in Table 4 in Appendix A.

The marking grid for multiple-choice answers and explanations uses an amended set of problem and response indicators. For the problem indicators we also use **Difficulty** but add the binary variables **Includes Calculation** (indicating whether mathematical handling is required) and **Real-World Application** (indicating whether the problem discusses a real-world phenomenon). In addition, we added indicators with ordinal scores between and 1 and 3 to cover context dependency and transfer. **Context Dependency** measures to which extent a question is tied to specific course materials and conventions. **Conceptual Transfer** measures the level of transfer and application required to answer the question correctly.

For the response indicators we counted how many (and which) answers were chosen correctly (**Answer Accuracy**) and how many (and which) explanations were correct (**Explanation Accuracy I**). If at least one explanation was inaccurate, we evaluate the extent of the mistake using an ordinal variable **Explanation Accuracy II** with a scale ranging from 1 to 5. We evaluated **Specificity** and **Safety** as defined above. Also, **Error Types** and **Error Domains** were used as previously introduced for the explanation of concepts. Descriptive labels for these indicators are given in Tables 7, 5 and 4 in Appendix A.

4 Results

Table 1 presents summary statistics for the problem indicators employed in our study. Note that Difficulty and Frequency are ordinal variables measured on a 5-point scale, whereas Context Dependency and Conceptual Transfer are ordinal variables measured on a 3-point scale. The variables Microeconomics, Macroeconomics, Includes Calculation, and Real-World Application

[1] We verified the URL `www.core-econ.org` in multiple indexes of Common Crawl using the website `https://index.commoncrawl.org/`.

are binary.

The summary statistics indicate that concepts are fairly balanced in terms of Difficulty, but that higher-frequency concepts are more prevalent in our dataset. Moreover, the majority of concepts are macroeconomic, with four concepts not clearly assignable to either category.

Multiple-choice questions are graded more difficult than the concepts discussed above. We expected this result, as many questions require not only the reproduction of knowledge but also its application. Questions are close to equally balanced between micro- and macroeconomics, approximately 20% of questions involve calculations, and around 30% apply concepts to real-world scenarios. Additionally, about 25% of the questions are at least moderately context-dependent, and 60% extend beyond mere reproduction of concept definitions.

		Value (if binary)		Score (if ordinal)						
Task	Indicator	0	1	1	2	3	4	5	mean	total
Explanation of concepts	Difficulty			2	15	23	14	2	2.98	56
	Frequency			0	5	15	15	21	3.93	56
	Microeconomics	40	16						-	56
	Macroeconomics	20	36						-	56
Multiple-choice questions	Difficulty			0	4	8	12	1	3.4	25
	Microeconomics	14	11						-	25
	Macroeconomics	13	12						-	25
	Includes Calculation	20	5						-	25
	Real-World Application	18	7						-	25
	Context Dependency			19	5	1			1.28	25
	Conceptual Transfer			10	15	0			1.6	25

Table 1: Summary Statistics for problem indicators for both tasks. Numbers under Value and Score represent the number of concepts/questions with this value or score. Binary variables are shown under Value and ordinal variables are shown under Score.

Our analysis primarily relied on descriptive statistics, focusing on the distributional features of the response indicators. This approach provides a detailed understanding of the quality and shortcomings of ChatGPT responses and allows us to evaluate its capabilities and limitations to function as an online tutor to support student learning. We also sought to determine whether our problem indicators can help identify lower quality responses, as such insights would enable educators to guide students more effectively. For instance, if more difficult concepts were shown to yield lower accuracy scores, educators could warn students to exercise particular caution when employing GenAI chatbots for those types of concepts.

4.1 Explanation of concepts

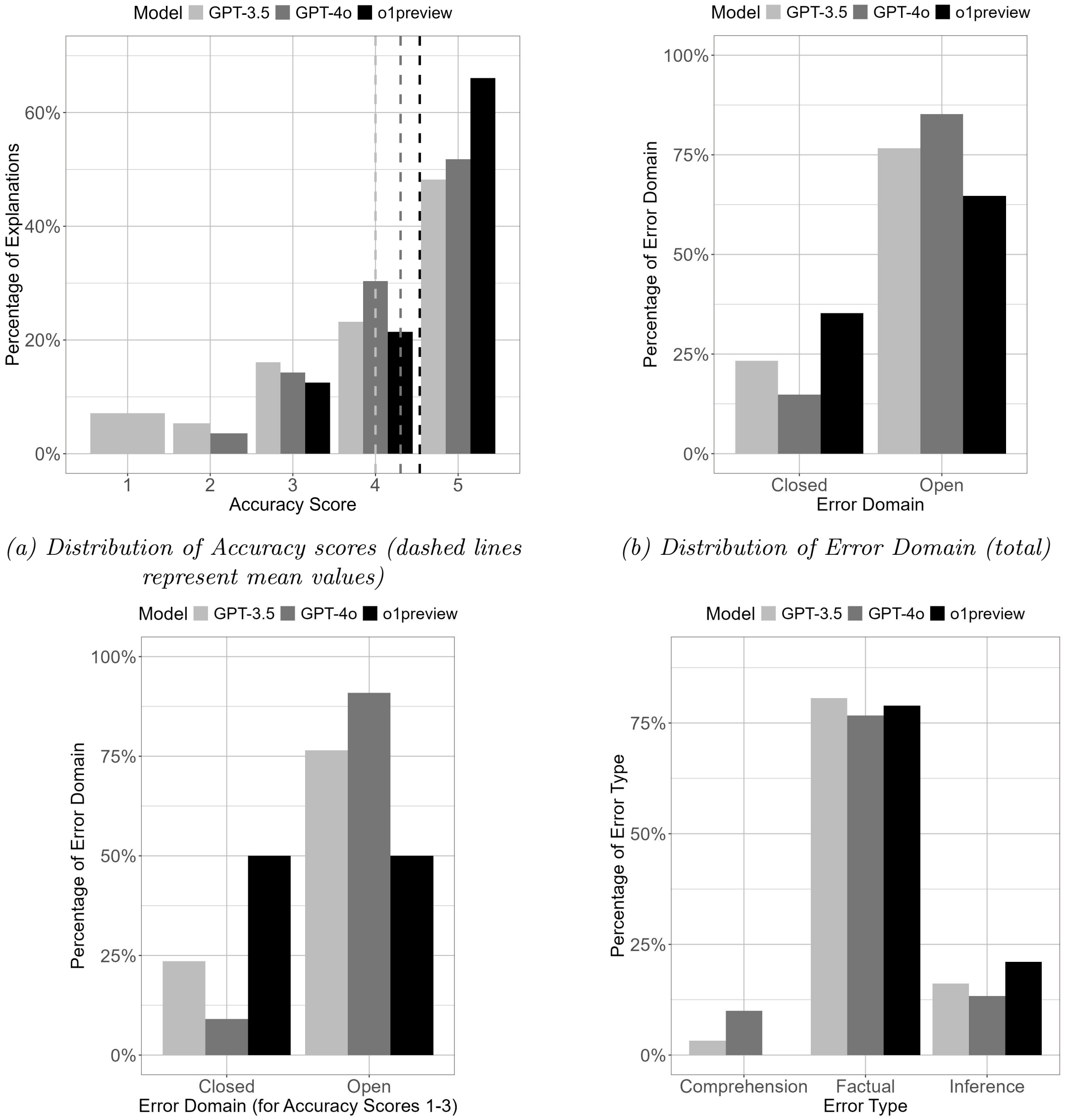
Given that performance expectancy is a key driver for student adoption of technology and that students exhibit low trust in GenAI-generated responses, accuracy is a central criterion for this study. Figure 1a illustrates that responses with Accuracy scores of 1 to 3—indicating the inclusion of inaccurate or misleading information in at least an important part of the concept—occur in 28.6% of GPT-3.5 outputs, 17.8% of GPT-4o outputs, and 12.5% of o1preview outputs. The sources of the inaccuracies range widely. For example, the chatbot does sometimes not know the concept ("Reservation Option" with GPT-3.5), explain a macroeconomic concept as if it where a microeconomic concept ("Goods Market Equilibrium" with GPT-3.5 and GPT-4o), calculate incorrectly ("Pareto-Efficiency" with GPT-3.5) or sometimes omits essential aspects of the definition ("Reservation Wage" with GPT-4o, in which the chatbot never mentions the next best alternative). Although the Accuracy score is an ordinal variable and its not strictly appropriate in a statistical sense, we also discuss the mean values. Mean values provide a single, interpretable summary that can help to compare groups, or approximate the distribution: GPT-3.5 has a mean accuracy score of 4.0, GPT-4o of 4.3, and o1preview of 4.54. This demonstrates a substantial reduction in inaccurate output with GPT-4o and o1preview. However, it also indicates that very inaccurate information remains prevalent in 12.5% of responses from the latest model tested (o1preview). This is problematic because all models present their responses as absolute facts; a safety score of 1 is assigned in 100% of responses by GPT-4o and o1preview, and in 94.6% of responses by GPT-3.5.

Furthermore, Figure 1b shows that inaccurate responses are of the open domain (i.e., users cannot identify incorrect information without extensive research outside of the chat) in 76.67% of cases for GPT-3.5, 85.19% for GPT-4o, and 64.7% for o1preview. This changes slightly when focusing solely on responses with an accuracy score of 3 or lower (Figure 1c): inaccurate responses are of the open domain in 76.47% of cases for GPT-3.5, 91.91% for GPT-4o, and 50% for o1preview. For example, in the explanation for "Marginal Propensity to Consume", the model o1preview limits its output to the case of an individual household, even though the concept is predominantly used for the aggregate of all households in an economy. A student unfamiliar with the concept is unable to identify the inaccuracy as such. These findings suggest that it would be challenging for students to recognize the inaccurate outputs, an issue special to the technology of GenAI. A human tutor would likely acknowledge a lack of understanding and refrain from providing incorrect information; ChatGPT does not exhibit such behavior.

Figure 1d illustrates the types of errors identified in the chatbot responses. The findings reveal that over 75% of the errors across all models are factual errors. This means that information provided is incorrect based on established facts, either because models fail to retrieve accurate information from their training data, or facts are incorrectly assimilated during the learning process. Inaccurate outputs, however, are less pronounced in newer models, particularly o1preview.

An essential skill of a tutor is the ability to tailor responses to the target audience. In our case, this means that explanations should be easy to understand, thanks to clear and concise writing and doesn't require subject knowledge, making it appropriate to learn new material. We measure this using the Specificity score. Figure 2a presents the distribution of Specificity scores, showing that over 60% of responses provide at least a clear and accessible explanation (score 4) across all models. The mean Specificity scores are 3.54 for GPT-3.5, 3.66 for GPT-4o, and 3.71 for o1preview. In contrast to the Accuracy scores, Specificity scores are more consistent across concepts and models, yielding overall satisfactory results.

Even if a response scores high in Accuracy and Specificity, its answer might lack a holistic perspective, which would leave gaps in the students understanding and hinder the students ability to apply the concept to broader contexts. This is measured by the Scope score. Figure 2b displays the distribution of Scope scores, revealing that ChatGPT rarely offers a holistic view of the concept (which would result in a Scope score of 5). Notably, responses lack important detail or nuance (score 3) in 37.5% of cases for both GPT-3.5 and o1preview, and in 41.07% of cases for GPT-4o. The mean Scope scores are 3.41 for GPT-3.5, 3.46 for



(a) Distribution of Accuracy scores (dashed lines represent mean values)

(b) Distribution of Error Domain (total)

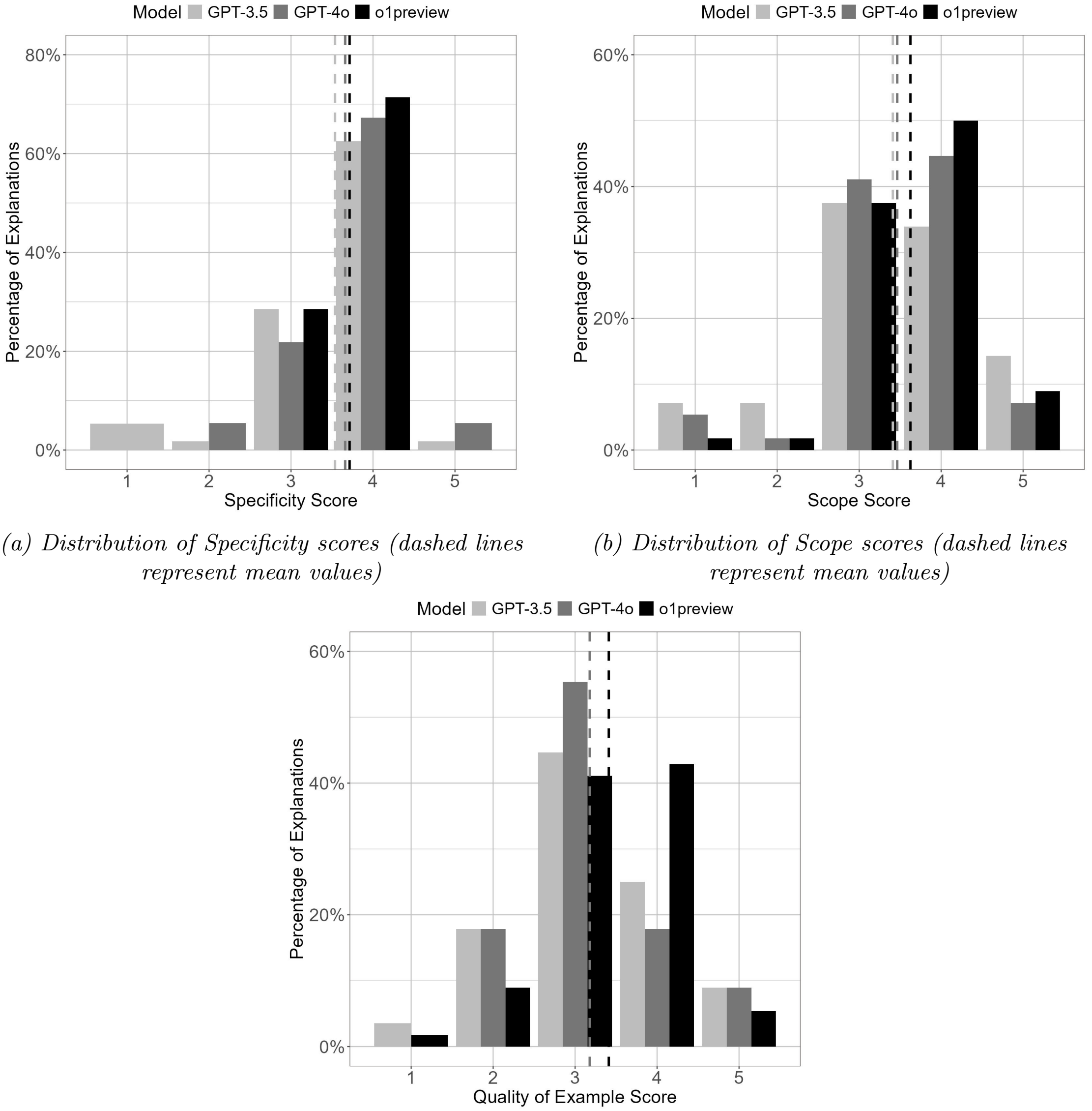
(c) Distribution of Error Domain (for Accuracy Scores 1-3)

(d) Distribution of Error Types

Figure 1: Distributions for Accuracy scores and Error Domains and Types for the explanation of concepts task.

GPT-4o, and 3.63 for o1preview. An example for a low Scope score is the explanation of "Gross Domestic Product" (GDP) by GPT-4o, where the chatbot only mentions the calculation of GDP from the so-called production side, while not mentioning the other two ways (so-called spending and income side). Our findings indicate that all models similarly struggle with this task.

Effective examples help to enhance students' understanding, as they serve as concrete representations of abstract ideas and provide essential context. Figure 2c shows the distribution of the



(a) Distribution of Specificity scores (dashed lines represent mean values)

(b) Distribution of Scope scores (dashed lines represent mean values)

(c) Distribution of Quality of Example scores (dashed lines represent mean values)

Figure 2: Distributions for Specificity, Scope and Quality of Example scores for the explanation of concepts task.

Quality of Example scores across the three models. 66.07% of responses from GPT-3.5, 73.21% from GPT-4o, and 51.79% from o1preview received a score of 3 or lower, indicating that the examples provided were weak, unhelpful, or even detrimental to comprehension. An additional 25% of GPT-3.5 responses, 17.86% of GPT-4o responses, and 42.86% of o1preview responses attained a score of 4, suggesting that the examples were relevant but overly simplistic. The mean scores were 3.18 for both GPT-3.5 and GPT-4o, and 3.41 for o1preview. Unhelpful examples frequently consist of simple calculation exercises. For instance, when explaining the concept of "Aggregate

Demand", responses only provided simple calculations adding example values for consumption, investment, government spending, and net exports. Moreover, ChatGPT models occasionally offered microeconomic examples for macroeconomic concepts (e.g., such as using individual consumer behavior to illustrate the "Marginal Propensity to Consume") which contributed to the lower scores. These findings suggest that while the models can generate examples, they often lack the depth and relevance necessary to effectively aid student understanding in economics.

In summary, our descriptive analysis reveals relatively low variation in both Specificity (Figure 2a) and Scope scores (Figure 2b), indicating robust and consistent performance in these areas. In contrast, the Quality of Example scores (Figure 2c) exhibit greater variability and generally low performance, suggesting that students should be cautioned about relying on the chatbot's examples for comprehension.

Accuracy scores (see Figure 1a) are, on average, satisfactory but display substantial variation. As previously discussed, students may struggle to identify inaccurate outputs, thus increasing the risk of disseminating misconceived knowledge. However, we see that for newer chatbot models Accuracy scores on average are high.

Building on earlier findings (e.g., Gilson et al. 2023) that GenAI chatbots tend to perform worse on more difficult problems, we extended the analysis to investigate whether the variation in Accuracy scores can be explained by controlling for the Frequency and Difficulty of the problem, as well as by distinguishing between microeconomic and macroeconomic contexts.

Since Accuracy is an ordinal variable measured on a five-point scale, we used a cumulative log-odds model to estimate the probability of Accuracy being at or below each score category (see Equation 1). As described earlier, both Frequency and Difficulty are also ordinal variables. Each is modeled using dummy variables for individual score levels. For Difficulty, we omit the dummy for the lowest category, and for Frequency, we omit the first two categories since score 1 is not observed (see Table 1). Additionally, we include a binary variable for Microeconomics, which equals 1 if the concept is microeconomic in nature and 0 if it pertains to macroeconomics or neither micro- or macroeconomics.

$$\begin{aligned}\text{logit}[P(\text{Accuracy} \leq j)] =& \log\left(\frac{P(\text{Accuracy} \leq j)}{P(\text{Accuracy} > j)}\right) \qquad (1)\\ =& \alpha_j - \sum_{k=3}^{5} \beta_k \text{Frequency}_k - \sum_{l=2}^{5} \gamma_l \text{Difficulty}_l \\ & - \delta \text{Microeconomics}, \; j = 1, 2, 3, 4\end{aligned}$$

Using regression analysis, we investigated whether Frequency, Difficulty, and if the concept is from microeconomics, are statistically associated with variations in Accuracy scores. It is essential to emphasize that we do not claim the relationships here to be causal; instead, we are identifying correlations. Furthermore, the estimation approach has inherent limitations. The chatbot models themselves are "black boxes", and we lack information about how data are processed internally. As a result, our explanatory variables represent only a small subset of the underlying factors that could influence Accuracy. The majority of factors remain unobserved and unknown.

Coefficient (Variable)	GPT-3.5	GPT-4o	o1preview
α_1 (Cut-Off between Scores 1 and 2)	-17.86***	-14.03	-7.12***
	(0.87)	(63.18)	(2.15)
α_2 (Cut-Off between Scores 2 and 3)	-16.64***	-13.67	-6.39**
	(0.8)	(63.18)	(2.03)
α_3 (Cut-Off between Scores 3 and 4)	-14.27***	-10.59	-3.02
	(0.88)	(63.18)	(1.84)
α_4 (Cut-Off between Scores 4 and 5)	-12.05***	-6.94	0.2
	(0.96)	(63.18)	(1.77)
β_3 (Frequency$_3$)	1.29	0.75	-2.41
	(1.15)	(1.23)	(1.29)
β_4 (Frequency$_4$)	2.19	2.78*	-0.53
	(1.23)	(1.34)	(1.27)
β_5 (Frequency$_5$)	1.42	1.63	-2.29
	(1.14)	(1.23)	(1.25)
γ_2 (Difficulty$_2$)	-14.91***	-11.83	0.09
	(0.58)	(63.17)	(1.43)
γ_3 (Difficulty$_3$)	-16.04***	-12.34	-0.54
	(0.45)	(63.17)	(1.41)
γ_4 (Difficulty$_4$)	-15.93***	-12.01	-1.24
	(0.5)	(63.17)	(1.53)
γ_5 (Difficulty$_5$)	-35.87***	-13.92	-3.03
	(0)	(63.2)	(2.26)
δ (Microeconomics)	-0.47	-0.02	-1.2
	(0.65)	(0.69)	(0.71)
Num. Obs.	56	56	56
AIC	153.53	121.70	106.09
BIC	176.84	146	130.4

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 2: Regression results for Equation 1. Values in the columns GPT-3.5, GPT-4o and o1preview represent coefficient estimates with asterisks indicating significance levels. Values in brackets represent standard errors.

Table 2 shows the estimation results. As is standard in these type of logistic regressions, our primary focus is on the estimated coefficients for the observable problem indicators: Difficulty, Frequency, and Microeconomics. The coefficients indicate how the likelihood of achieving a higher Accuracy score changes when the value of an explanatory variable increases by one unit. For instance, the coefficient for concepts with a Difficulty score of 2 represents how much more likely these concepts are to receive the next higher Accuracy score compared to concepts with a Difficulty score of 1.

The positive coefficient for the Frequency variable, suggests that the frequency a concepts is used generally increases the odds of achieving higher Accuracy scores for both GPT-3.5 and GPT-4o. While this aligns with our expectations, most of these estimates are not statistically significant. The primary exception is Frequency$_4$ in the GPT-4o model. The coefficient estimate for Frequency_4 is 2.78, which implies that the odds-ratio is $e^{2.78} = 16.12$. This means that the odds of having a higher Accuracy score if a concept has a Frequency score of 4 are more than 16-times the odds of a concept that has a Frequency score of 1 or 2. Surprisingly, for o1preview, we find that higher Frequency scores decrease the odds of achieving higher Accuracy scores. But again, the coefficient estimates are not statistically significant.

Regarding Difficulty, the coefficient estimates are generally negative, implying that higher Difficulty scores increase the likelihood of lower Accuracy scores. However, these results are not statistically significant for GPT-4o and o1preview. In contrast, for GPT-3.5, the Difficulty coefficients are significant and strongly negative, indicating that an increase in Difficulty drastically reduces the probability of achieving higher Accuracy scores. For example, the coefficient estimate for Difficulty_2 is -14.91. This means that the odds-ratio to a concept of a Difficulty score of 1 is $e^{\gamma_2} = 3.35 \times 10^{-7}$. Hence, for concepts that have a Difficulty score of 2, the odds of having a higher Accuracy score are just 3.35×10^{-7} times the odds of a concept that has a Difficulty score of 1. This clearly shows that more difficult concepts are associated with lower Accuracy scores for GPT-3.5. Notably, this effect disappears in the more advanced chatbot models (GPT-4o and o1preview).

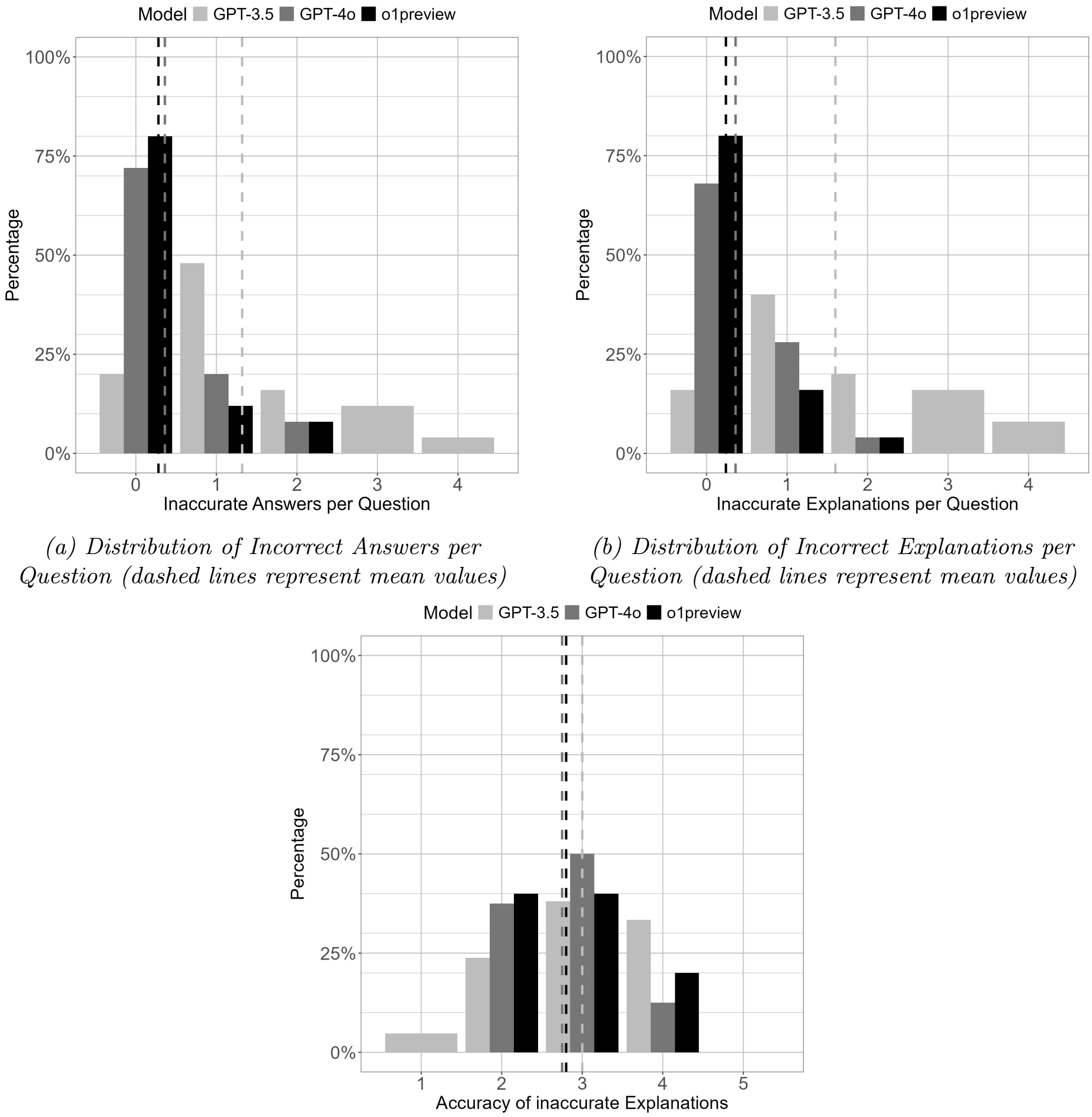
For all chatbot models, We observe a negative coefficient for the Microeconomics dummy. This suggests that, relative to macroeconomics and concepts that neither belong to micro- or macroeconomics, microeconomic concepts tend to be associated with lower Accuracy scores. However, as with most other coefficients, this relationship is not statistically significant.

Perhaps the most noteworthy outcome of our estimation is the absence of statistically significant results for the GPT-4o and o1preview models. In line with existing literature, our findings for GPT-3.5 indicate that more difficult concepts yield lower Accuracy scores. Educators may want to cautioning students when using earlier-generation models for difficult concepts and proactively labeling more difficult content. However, this finding does not hold for GPT-4o and o1preview. For these newer models, problem indicators (i.e., Frequency of use or Difficulty) exert no statistically significant influence on the Accuracy score. In other words, although overall accuracy improves with the more advanced models, it becomes increasingly difficult to pinpoint inaccuracies, thereby complicating the task of guiding students in their use of GenAI chatbots as automated tutors.

4.2 Answering and explanation of multiple-choice questions

Figure 3a illustrates the number of multiple-choice options (out of four) incorrectly assessed by the three chatbot models, while Figure 3b shows the number of incorrect explanations provided. The data reveal a significant improvement in performance from GPT-3.5 to GPT-4o and o1preview models. On average, GPT-3.5 incorrectly assessed 1.32 out of 4 options, whereas GPT-4o and o1preview incorrectly assessed only 0.36 and 0.28 options, respectively. In other words, GPT-3.5 correctly assessed 67% of the options, GPT-4o achieved 91%, and o1preview reached 93%. These results indicate that under exam conditions, both GPT-4o and o1preview would comfortably pass such assessments. The accuracy of explanations mirrors this trend: GPT-3.5 provided an average of 1.6 incorrect explanations out of 4, compared to 0.36 for GPT-4o and 0.24 for o1preview.

Figure 3c presents the Explanation Accuracy II scores, which measures the degree of inaccuracy for incorrect explanations.The mean Explanation Accuracy II scores are 3 for GPT-3.5, 2.75



(a) Distribution of Incorrect Answers per Question (dashed lines represent mean values)

(b) Distribution of Incorrect Explanations per Question (dashed lines represent mean values)

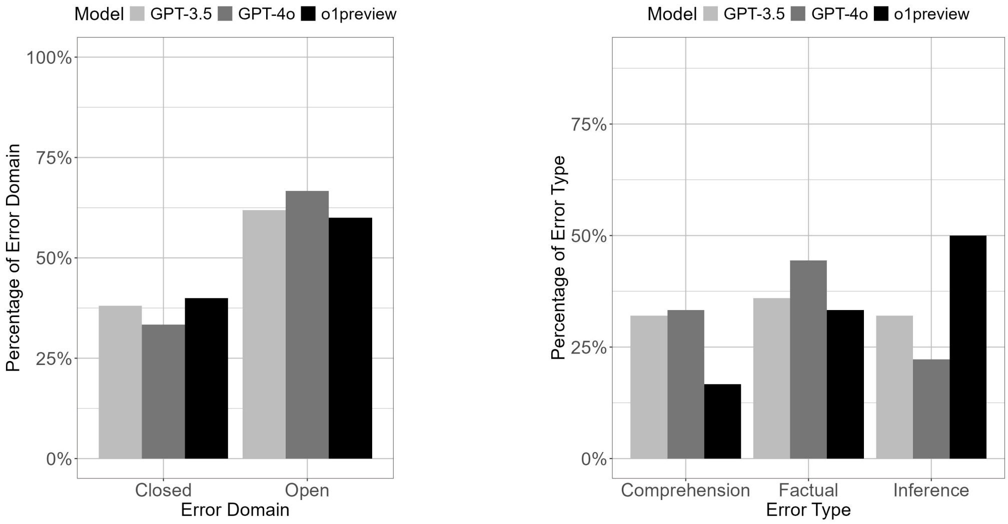
(c) Distribution of Accuracy Score for inaccurate Explanations (dashed lines represent mean values)

Figure 3: Distributions for Inaccurate Answers as well as Explanations per Question and Accuracy Score for inaccurate Explanations for the answering and explanation of multiple-choice questions task.

for GPT-4o, and 2.8 for o1preview, indicating that the incorrect explanations contain at least notable inaccuracies. Similar to the explanations of concepts, all models present 100% of explanations as absolute facts (Safety score 1).

Delving deeper into these inaccuracies, Figure 4a illustrates the Error Domain of the explanations. Compared to the explanation of concepts (refer to Figure 1b), inaccuracies in answers and explanations of multiple-choice questions are more readily identifiable, as a higher proportion of errors belong to the closed domain (i.e., they are noticeable to a careful reader without external research).

Figure 4b highlights differences not only in the Error Domain but also in the Error Types for multiple-choice questions. While factual errors dominated the explanations of concepts (as shown in Figure 1d), they constitute only 36% of errors for GPT-3.5, 44% for GPT-4o, and 33.33% for o1preview in the context of multiple-choice questions. Instead, comprehension errors play a more significant role, accounting for 32% of errors for GPT-3.5, 33.33% for GPT-4o, and 16.67% for o1preview. Option 4 in question 5.1 provides an example for a comprehension error: "According to the Pareto criterion, a Pareto-efficient outcome is always better than an inefficient one." (for the full question see Appendix B.2). GPT-4o incorrectly marks this option as correct, because it understands "better" as better according to the Pareto criterion, rather than more generally. Our findings suggest that ChatGPT more frequently fails to fully understand what the multiple-choice question is asking.



(a) Distribution of Error Domain

(b) Distribution of Error Types

Figure 4: Distributions for Error Domain and Types for the answering and explanation of multiple-choice questions task.

The increased prevalence of comprehension errors in multiple-choice questions can be attributed to the often complex scenarios described over multiple sentences, which increases the potential for misunderstandings and incorrect responses. Note that the set of multiple-choice questions were crafted by economics educators, who may inadvertently assume that certain knowledge is universally shared or already learned.

Finally, we observe that Specificity scores for the multiple-choice questions are comparable to those for the explanation of concepts. The mean Specificity scores are 3.4 for GPT-3.5, 4.0 for GPT-4o, and 3.92 for o1preview. This suggests that, regardless of the task, ChatGPT effectively tailors its responses to the target audience, demonstrating a consistent ability to break down complexity and adapt explanations appropriately to help students learn new material.

In the analysis above, we employed a regression analysis to investigate whether variations in the Accuracy score could be explained by the observable problem indicators. Since the most relevant models for our study (GPT-4o and o1preview) exhibit a high proportion of questions with entirely accurate answers, there is insufficient variation in the data to yield meaningful regression results. In fact, Figure 3a shows that for GPT-3.5, GPT-4o, and o1preview, the share of questions for which all options were chosen correctly is 20%, 72%, and 80%, respectively.
To address this limitation, we adopted a nonparametric testing approach to identify the relationship between individual problem indicators and the number of incorrectly chosen answers (Answer Accuracy). In particular, we tested whether the samples corresponding to each score or value of a given problem indicator originate from different underlying population distributions. In particular, we applied a Mann-Whitney U Test for binary and two-level ordinal indicators (e.g., Includes Calculation or Conceptual Transfer) and a Kruskal-Wallis One-Way Analysis of Variance Test for ordinal indicators with more than two levels (e.g., Difficulty). The Mann-Whitney U Test evaluates whether two independent samples differ in their median values, while the Kruskal-Wallis Test extends this logic to more than two samples. In both cases, the null hypothesis states that the median of the populations for all samples is the same.

Table 3 reports summary statistics (number of observations, mean, median, and standard deviation) by score or value for each problem indicator, as well as the results of the nonparametric tests for each ChatGPT model. The key finding is that, with one exception, all tests fail to reject the null hypothesis. In other words, the observable problem indicators do not significantly affect the number of inaccurate answers per question.

Problem Indicator	Value/ Score	Num. Obs.	Test used	GPT-3.5				GPT-4o				o1preview			
				Mean	Med.	St.Dev.	p-val.	Mean	Med.	St.Dev.	p-val.	Mean	Med.	St.Dev.	p-val.
Difficulty	1	0	Kruskal-Wallis	-	-	-	0.33	-	-	-	0.91	-	-	-	0.93
	2	4		1	1	0.82		0.25	0	0.5		0.25	0	0.5	
	3	8		1.75	1.5	1.28		0.38	0	0.74		0.38	0	0.74	
	4	12		1.25	1	0.97		0.42	0	0.67		0.25	0	0.62	
	5	1		0	0	-		0	0	-		0	0	-	
Microeconomics	0	14	Mann-Whitney U	1.29	0	1.2	0.8	0.43	0	0.65	0.47	0.5	0	0.76	0.05
	1	11		1.36	1	0.92		0.27	0	0.65		0	0	0	
Includes Calculation	0	20	Mann-Whitney U	1.15	1	0.99	0.13	0.3	0	0.57	0.44	0.25	0	0.64	0.54
	1	5		2	2	1.22		0.6	0	0.89		0.4	0	0.54	
Real-World Application	0	18	Mann-Whitney U	1.28	1	1.02	0.82	0.28	0	0.57	0.29	0.22	0	0.55	0.5
	1	7		1.43	1	1.27		0.57	0	0.79		0.43	0	0.79	
Context Dependency	1	19	Kruskal-Wallis	1.37	1	1.07	0.91	0.42	0	0.69	0.7	0.26	0	0.65	0.55
	2	5		1.2	1	1.3		0.2	0	0.45		0.4	0	0.55	
	3	1		1	1	-		0	0	-		0	0	-	
Conceptual Transfer	1	10	Mann-Whitney U	1.1	1	0.88	0.5	0.3	0	0.67	0.65	0.2	0	0.63	0.59
	2	15		1.47	1	1.19		0.4	0	0.63		0.33	0	0.62	
	3	0		-	-	-		-	-	-		-	-	-	

Med. = Median, St.Dev.= Standard Deviation, p-val. = p-value of the test.

Table 3: Number of inaccurate answers per question by problem indicator score or value. For every score or value and every ChatGPT model the number of observations as well as mean, median and standard deviation is given. For every problem indicator the applied statistical test and the resulting p-value is presented.

The sole exception is the Microeconomics indicator for the o1preview model, where the test indicates that at the 5% significance level, o1preview performs worse on microeconomic questions compared to macroeconomic questions or those that are neither clearly micro- nor macroeconomic. As the summary statistics suggest, all eleven non-microeconomic questions were answered correctly, whereas the microeconomic questions had on average 0.5 inaccurate answers out of four options.

Although not statistically significant at the 10%-level (but close), the Includes Calculation indicator for GPT-3.5 suggests that GPT-3.5 perform worse on questions requiring calculations. This aligns with previous findings in the literature (e.g., Terwiesch 2023). Notably, this effect vanishes for the newer GPT-4o and o1preview models.

In summary, for the multiple-choice questions, the test results align with the regression findings from the explanation of concepts. For the newer models, observable problem characteristics do not reliably explain chatbot accuracy. While this reduces the potential for guiding students by highlighting problematic concepts, the associated risks are limited, given the relatively small number of inaccurate outputs.

5 Extensions

5.1 Performance in other languages: German

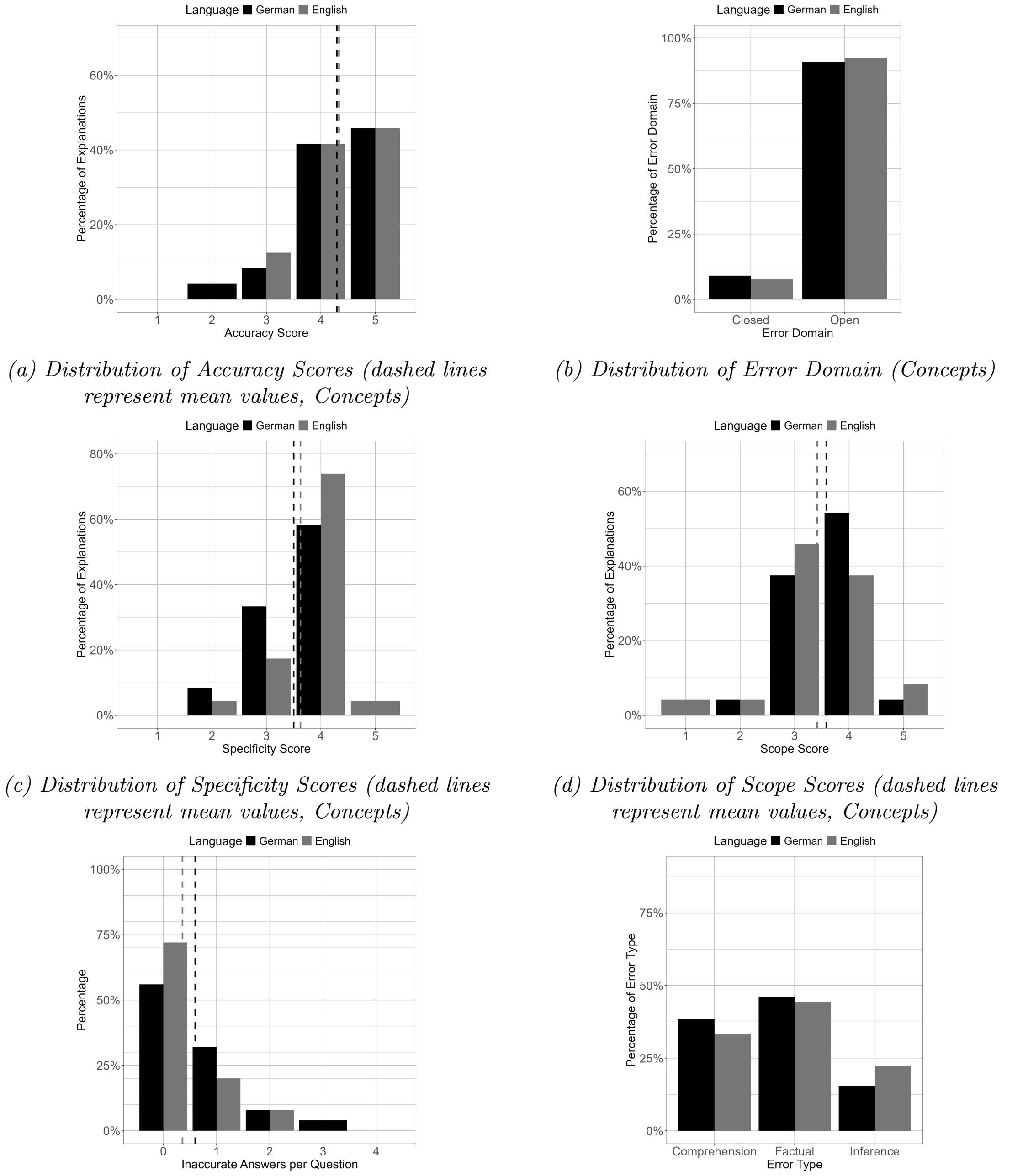
The details of the training data used for the newer ChatGPT models remain largely unknown. However, it is likely that English sources constitute over 90% of the dataset. For example, in GPT-3, English accounted for 92.65% of the words used in training, followed by French (1.82%) and German (1.47%) (Brown 2020). Given that economics is taught in multiple languages—two of the authors, for instance, also teach introductory economics courses in German—it is important to assess the robustness of our findings in non-English contexts.

To examine this, we tested ChatGPT with the model GPT-4o using the same two tasks but in German and compared the results with those obtained in English. While the same multiple-choice questions were used, the concept explanation task was limited to 24 economic concepts in German. A complete list of these concepts is provided in Appendix B.1. To ensure comparability, the following English results were also restricted to the 24 corresponding translated concepts.

Overall, the results in German and English are very similar. Below, we highlight the most relevant findings from the main analysis and their corresponding indicators. When explaining concepts, the distribution of Accuracy scores in German closely mirrors that in English (Figure 5a), with a high mean accuracy score of 4.29. Additionally, most inaccuracies belong to the open error domain (Figure 5b), meaning that incorrect information is difficult for users to identify without external verification. As in English, ChatGPT consistently presents explanations as absolute fact, without acknowledging uncertainties.

The Specificity and Scope scores for German concepts also exhibit similar distributions and mean values to their English counterparts (Figures 5c and 5d). However, we observe a slight leftward shift in the distribution of Specificity scores, which we discuss in more detail later.

When answering and explaning multiple-choice questions in German, ChatGPT incorrectly assessed an average of 0.6 out of 4 options, compared to 0.36 out of 4 in English (Figure 5e). Although this represents a decline in performance, the model still achieved an 85% accuracy



(a) Distribution of Accuracy Scores (dashed lines represent mean values, Concepts)

(b) Distribution of Error Domain (Concepts)

(c) Distribution of Specificity Scores (dashed lines represent mean values, Concepts)

(d) Distribution of Scope Scores (dashed lines represent mean values, Concepts)

(e) Distribution of Incorrect Answers per Question (dashed lines represent mean values, Multiple-choice questions)

(f) Distribution of Error Types (Multiple-choice questions)

Figure 5: Selected distributions for indicators in comparison between German and English. All results for the GPT-4o model. Note: In this extension the explanation of concepts task covered just 24 concepts.

rate in German.

The distribution of Error Types is also similar across languages (Figure 5f). However, Comprehension Errors appear slightly more frequently in German (38.5%) compared to English (33.3%), suggesting a marginally higher likelihood of misunderstanding the question prompt in German.

The overall similarity in results between English and German is an encouraging finding. However, it appears counterintuitive. Given the architecture of large language models, one would expect better performance in languages with a higher representation in the training data. This should imply weaker performance in less represented languages, including German. Due to the lack of transparency regarding its internal workings, we cannot fully explain why the chatbot performs so similarly across languages. Nevertheless, our analysis provides a plausible explanation: ChatGPT is likely to first translate the German prompt into English, generate a response in English, and then translate the output back into German.

As noted earlier, we observe slightly lower Specificity scores in German. Upon closer examination, we find that ChatGPT sometimes generates unnatural phrasing in German. For example, when explaining the German term "Bruttoinlandsprodukt" ("Gross Domestic Product"), the chatbot refers to a hypothetical country named "Ökonomienland". While not technically incorrect, this term is highly unusual in German and would be considered unnatural by native speakers. Notably, this resembles a literal translation of "Econland", a term that ChatGPT frequently uses in English explanations (e.g., for "Gross Domestic Product", "Business Cycle", and "Government Debt").

If our translation hypothesis is correct, educators teaching and students learning in languages other than English, can still use ChatGPT effectively given that the model is proficient in translating to and from the target language.

To assess whether ChatGPT can perform reliably in a given language, educators can conduct simple translation tests using key economic terms. If the translation quality is high, it is likely that ChatGPT will deliver performance comparable to its English-language responses.

5.2 Concistency of ChatGPT output

GenAI chatbots may produce different responses to the same question. This variability arises because large language models are probabilistic in nature. Specifically, these models generate text by predicting the next token in a sequence of text, assigning probabilities to possible token candidates. This process typically employs the Softmax function combined with a temperature parameter, which controls how deterministic or varied the outputs are (e.g., Peeperkorn et al. 2024).

To illustrate, if a model is set to a very low temperature, it behaves deterministically, consistently choosing the token with the highest probability. In contrast, at a high temperature, the probability distribution becomes more uniform, leading to greater variation in responses. The ChatGPT web interface employs a fixed temperature setting, that is not directly shown to the user.

In our primary analysis, each model was tested only once per task. However, to examine the consistency of responses, due to the probabilistic nature of the models, we asked ChatGPT to answer the same question multiple times.

Given the time-intensive evaluation and moderation process described earlier, we limited this investigation to the multiple-choice questions. We modified the prompt slightly, ensuring that ChatGPT's responses were standardized:

Adjusted prompt: `Following you will face a multiple-choice question. Multiple answers may be correct. Name the correct answers. Answer in the following format: 1.) %s, 2.) %s, 3.) %s, 4.) %s. Where %s is c if the respective statement is correct or f if the respective statement is incorrect. Just answer in this format and don't write any additional text. Question start: <MULTIPLE CHOICE QUESTION>`

Using this format, responses only indicate which answer choices are considered correct or incorrect. Each response was binarized: a value of 0 if ChatGPT deemed an option false and 1 if it deemed an option correct.
We applied this prompt to all 25 multiple-choice questions (listed in Appendix B.2) and tested the GPT-4o and o1preview models. For each model and each question, we generated 10 independent responses, this time using the OpenAI API, without changing the temperature parameter, instead of the ChatGPT website for efficiency.

To quantify variability, we computed the variance of the responses for each of the four options per multiple-choice question and then calculated the mean variance across all four answer options per question. This means:

- If ChatGPT gave the same response for all four options across all 10 iterations, the variance for each option would be zero, and the mean variance for the question would also be zero.
- If ChatGPT provided different responses for any option across iterations, the variance for that option would be positive, contributing to a nonzero mean variance for the question.

With 10 iterations, we get 10 responses per answer option and question. As the responses take values of 0 or 1, mean variance for every question takes values between 0 and 0.25.

The results show that for GPT-4o, the mean variance is zero for 12 out of 25 questions, meaning the model was fully consistent for those cases. However, for 13 out of 25 questions, GPT-4o produced inconsistent answers across iterations. In contrast, for o1preview, the mean variance is zero for 17 out of 25 questions, indicating that it was inconsistent in only 8 out of 25 questions.

These findings suggest that for a significant portion of questions, the probability distribution of the chatbot's output is relatively flat, meaning the model is not entirely sure about the correct response. While our analyses so far showed similar performance between GPT-4o and o1preview, the variability test reveals that o1preview provides much more stable responses.

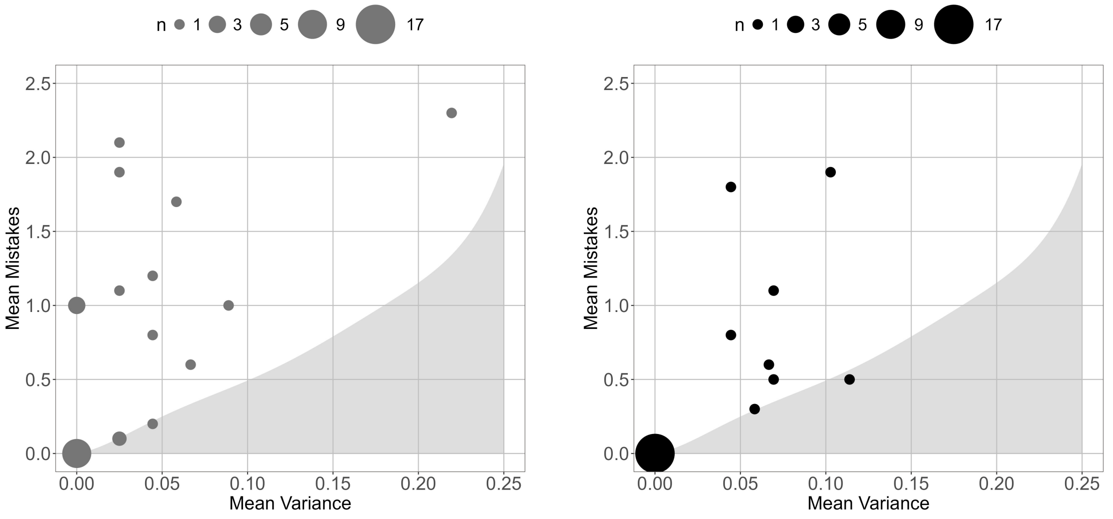
This degree of response variability presents a challenge for educational settings. If different students receive varying answers to the same question, it creates inconsistencies in learning and confusion when students discuss content with their peers. Similarly, if an individual student receives contradictory responses upon asking the same question multiple times, it may lead to confusion and impede effective learning. Such variability in outputs could also undermine students' trust in the technology and reduce usage.

So far, we have only assessed response variability and will now consider how correct iterative responses are. A model that consistently gives incorrect answers can be particularly harmful, as it reinforces false information with high confidence. To quantify correctness, we introduce a measure called mean mistakes, which calculates the average number of incorrect answers per iteration for each question and model. For example:

- If all responses in the 10 iterations are fully correct, the mean mistake value is 0.
- If there is one incorrect answer in one iteration out of ten, the mean mistake value is 0.1.

With four answer options for every question and every option can take the value 0 or 1, mean mistakes can take values between 0 and 4.

Figure 6 presents a bubble chart plotting mean variance against mean mistakes. The size of each bubble represents the number of questions associated with each combination of values.



(a) Results for GPT-4o

(b) Results for o1preview

Figure 6: Bubble chart showing the relationship between mean variance and mean mistakes for GPT-4o and o1preview.

Before interpreting the results, it is important to note an inherent relationship between these two metrics. Since each question has only one correct set of answers, any variability in responses implies the presence of at least some incorrect answers. As such, higher variance generally goes hand in hand with higher mean mistakes. We therefore use Monte Carlo simulations, to approximate a lower bound for mean mistakes, shading the area below this threshold in light gray in Figure 6.

For GPT-4o, Figure 6a shows that for 9 out of 25 questions, the chatbot provided the correct answers consistently across all 10 iterations. However, six questions exhibit low variance (≤ 0.025, i.e. less than 10% of the maximum mean variance) yet have a mean mistake value of 1 or more. This means that for these six questions, the chatbot frequently misclassified at least one answer option.

A particularly concerning case is represented by three questions where the mean variance is zero (i.e., the chatbot gave identical responses every time), yet the mean mistake is exactly 1. This implies that for these cases, GPT-4o always provided the same incorrect response, suggesting that it was confidently wrong. For o1preview, Figure 6b shows better results. All 17 questions where there was no variability were answered fully correctly. This means that whenever o1preview provided consistent responses, those responses were also accurate. Furthermore, in contrast to GPT-4o, o1preview does not exhibit cases where low variance (≤ 0.025) coincides with frequent mistakes. The overall mean variance is lower, and no question exceeds a mean variance of 0.125 (i.e. there is no observation in the upper half of potential mean variance levels).

Even though o1preview outperforms GPT-4o, some might argue that answering only 17 out of 25 multiple-choice questions consistently is a poor result, making ChatGPT unsuitable for student self-study. However, it is important to consider that even if ChatGPT answers a question inconsistently across ten iterations, it may still have provided the correct response in some instances. To gain further insight into the results, we calculated the total number of mistakes across all iterations and examined which options were incorrectly selected to determine possible reasons for these errors.

Both GPT-4o and o1preview were tested on 25 multiple-choice questions, each answered ten times, with four answer options per question. This amounts to a total of $25 \times 10 \times 4 = 1000$ assessments per model. Our analysis shows that GPT-4o incorrectly assessed 190 out of these 1000 options, achieving an 81% accuracy rate, while o1preview made 75 mistakes, resulting in a 92.5% accuracy rate. Despite some inconsistency in responses, these figures suggest that the overall number of errors is relatively low, particularly for o1preview. This indicates that while response variability exists, it is a reasonable low risk to student learning.

A closer examination of which questions led to inconsistencies or inaccuracies reveals that, except for one question, whenever o1preview struggled, so did GPT-4o (however, GPT-4o struggled with more questions overall). In total, there were seven multiple-choice questions where both models made mistakes. In most of these cases, the models correctly identified the correct answers but also mistakenly selected additional incorrect options.

In two cases (Questions 6.5 and 15.1; see Appendix B.2), both models misapplied economic concepts. In another two cases, wording issues in specific answer choices caused confusion: In Question 6.1, both models misinterpreted “relation-specific asset,” while in Question 10.8, o1preview particularly struggled with the term “bank money,” possibly due to ambiguities in online definitions for this concept. In another case (Question 16.8), both models incorrectly concluded that if a new labor-saving technology emerges, long-term unemployment increases, neglecting the potential for job creation from technological advancements.

Other inaccuracies suggest misinterpretations of question phrasing. For example, in Question 10.11, option 1 states: “A principal–agent problem exists in loans due to a positive possibility of the principal not being repaid”. Both models sometimes assessed this as correct, but the textbook considers it incorrect, explaining that “A principal–agent problem exists in loans due to the asymmetric information regarding the borrower’s effort or the quality of the project”. Similarly, in Question 14.3, option 3 states “A household adjusts its precautionary saving in response to changes in its target wealth". Both models marked this as true in most iterations, but the textbook considers it incorrect, arguing that "A household adjusts its precautionary saving in response to the gap (positive or negative) between its actual and target wealth". In both cases, the phrasing of the options might be misleading, making the textbook's intended answer less obvious. This shows that some incorrect responses stem from ambiguous or nuanced phrasing rather than fundamental misunderstandings.

These findings indicate that the risks associated with variability in chatbot responses are significantly lower for o1preview. While our previous analyses showed similar overall performance between GPT-4o and o1preview, this variability test highlights that o1preview provides more consistent and reliable responses. This makes o1preview a safer choice in educational settings.

One possible explanation is that o1preview is designed as a reasoning model, meaning it allocates more processing time to the consideration of intermediate steps before generating a final answer. This more structured approach appears to enhance response stability, making it better suited for educational use.

6 Discussion

6.1 Advice for educators

Since the release of ChatGPT, universities have been debating if and how to integrate GenAI into education, with some initially seeking to ban it. However, as history has shown with, e.g., the internet, search engines, and electronic calculators, prohibiting its use may not be a successful strategy. We expect GenAI to stay and continue to influence the way we teach and learn. Our study provides takeaways for educators, particularly in principles of economics courses, but also applicable to other disciplines. In the following, we have translated our results into actionable advice that educators can implement.

1. Integrate ChatGPT into economics courses by leveraging its strengths in explanations and question-answering while educating students on its capabilities and limitations.

Newer chatbot models provide high quality responses, that have the potential to benefit students in their learning. We found that in particular newer models of ChatGPT performed on average very good in providing accurate answers when explaining economic concepts and answering multiple-choice questions. The risk that students learn false information is thus relatively small. When possible, students should use newer models like o1 (the successor of o1preview), as they offer higher accuracy, fewer hallucinations, and more consistent responses than older versions. Our findings show that o1preview outperforms GPT-4o in both tasks, while both significantly outperform GPT-3.5. Additionally, o1preview provides more stable and reliable answers, reducing the risk of students encountering contradictory responses. Earlier research (e.g. Terwiesch 2023) suggested that ChatGPT struggled with calculations, but our findings indicate that this limitation has largely been resolved in newer models. Since GPT-4o, ChatGPT performs most mathematical tasks in economics correctly, and in our multiple-choice question analysis, questions involving calculations were just as accurate as those without calculations. It's important to note that our study only tested basic arithmetic and algebra, so we cannot assess ChatGPT's capabilities in more complex or specialized computations. The effectiveness of GenAI in economics education depends on two key factors: proper use (which we address in later advice) and access to newer models. Currently, free users have limited access to GPT-4o, while unrestricted access to GPT-4o and o1 requires a paid subscription. Educators should keep this in mind when recommending GenAI tools and ensure that students without premium access have alternative learning resources. Additionally, universities should consider providing access to GenAI tools for all students to prevent disparities between those who can afford a subscription and those who cannot.

2. Teach students to use ChatGPT as a learning aid, not a replacement for textbooks, lectures, or their own effort in solving problems.

Students should rely on comprehensive learning materials, such as textbooks and lecture notes, as their primary resources, using ChatGPT only for clarification and reinforcement. Our findings show that ChatGPT consistently provides incomplete explanations of economic concepts, as reflected in its low Scope scores. This lack of a holistic perspective means students may miss critical aspects, making it harder to apply concepts into the real-world. To address this, educators should encourage students to first engage with human-written materials before turning to ChatGPT for additional support. One effective classroom strategy is to generate a ChatGPT explanation of an economic concept and have students identify what is missing, fostering critical analysis of GenAI-generated content. In addition, students should use ChatGPT as an interactive learning tool rather than a shortcut to answers, engaging in problem-solving before seeking GenAI feedback. Our findings show that newer ChatGPT models generate high-quality responses in most cases, but research (e.g., Bastani et al. 2024; Lehmann, Cornelius, and Sting 2024) indicates that learning outcomes crucially depend on how students engage with GenAI chatbots. The main challenge is not the accuracy of the GenAI model, but how students interact with it: whether they actively think through problems or passively rely on it for answers. Educators should encourage students to avoid simply copying and pasting questions into ChatGPT for solutions and instead use it to get feedback on their own work. However, shifting student behavior is difficult. ChatGPT offers quick, easily accessible explanations and it is much easier to passively ask ChatGPT for an answer than to locate and read relevant sections in a textbook or actively think through a solution first. Educators should highlight the long-term benefits of thorough learning over convenience-driven GenAI reliance.

3. Acknowledge that students will struggle to detect incorrect output.

Educators must stress that ChatGPT responses can be inaccurate and that students often lack the subject knowledge needed to identify errors. Our study shows that problem characteristics (e.g., difficulty) do not reliably predict whether ChatGPT's output is accurate or not, making it impossible to provide students with clear guidelines for detecting errors. While newer models produce fewer inaccuracies, we found that most errors are only identifiable with extensive subject knowledge (which students usually lack) or extensive research outside of the chat. This means that as educators, we cannot fully shield students from incorrect content. Instead, we must emphasize that ChatGPT's output can be wrong and difficult to recognize as such. One classroom strategy is to generate a ChatGPT explanation of an economic concept or solution to a problem and have students critically assess its accuracy. In addition, students should ask the same question multiple times to detect inconsistencies or inaccuracies in ChatGPT's responses. When ChatGPT varies its answers, students can learn two key lessons: (a) GenAI is not infallible, and its responses should always be critically evaluated, and (b) engaging with inconsistencies promotes deeper understanding by requiring students to analyze and verify the information. However, while this exercise helps build awareness, it does not eliminate the risk of students unknowingly internalizing incorrect information.

4. Shift classroom focus to application-based learning by leveraging a flipped classroom model and providing clear, detailed examples.

Since ChatGPT can assist with understanding fundamental economic concepts, educators should prioritize real-world applications and critical discussions in class. Our findings suggest that ChatGPT effectively explains basic economic concepts and simple applications, provided students use it effectively. This creates an opportunity for educators to spend less time on foundational explanations and focus more on complex applications, real-world case studies, and critical discussions. An obvious consequence is transitioning to a flipped classroom model. One of the challenges of flipping the classroom is ensuring that students resolve difficulties independently before class rather than bringing unanswered questions into discussions. Our findings suggest that ChatGPT can help with this. However, we also find that ChatGPT, regardless of model, tends to generate at best overly simplistic examples that may fail to illustrate key economic principles effectively. This emphasizes that educators should provide students with high-quality, context-rich examples. Such detailed examples will not only facilitate student understanding, but also provide content for students' chatbot queries. While there is no guarantee that ChatGPT will always apply the examples correctly, this approach is likely to produce more relevant and nuanced explanations than the default ChatGPT-generated examples.

5. Guide students to specify the question context as well as the chatbot's role when prompting.

Instruct students to explicitly specify ChatGPT's role (e.g., "Act as an economics tutor") and provide context to receive more tailored and relevant responses. This approach to prompting is also recommended by, e.g., L. Mollick and E. Mollick (2023) and Korinek (2023). In our research we find that carefully curated prompts lead to consistently high Specificity scores. This means responses are well suited to the student's level and needs. Designing prompts carefully required minimal effort, but students may still default to asking simpler, less detailed questions, resulting in less effective responses. Educators should demonstrate the impact of well-structured prompts in class by comparing answers by a GenAI chatbot with and without role/context specifications.

6. Refine your problem statements for easier comprehension.

Before assigning problems to students, educators can use ChatGPT to identify potential misunderstandings and refine the wording of questions for better clarity. Crafting clear and effective questions is challenging and while educators often assume that their wording is unambiguous, it may, in reality, lead to misinterpretations. Our findings show that ChatGPT sometimes struggled to understand questions, revealing that students might also find these questions unclear. By inputting questions into ChatGPT and analyzing its interpretation, educators can identify potential confusion, refine phrasing, and ensure better comprehension.

7. Verify translation accuracy before using ChatGPT in Non-English courses.

Instead of fully testing ChatGPT's performance in languages other than English, educators can assess its translation accuracy on economics terms to gauge its performance. Our findings show that ChatGPT's performance in German closely mirrors its performance in English, despite German being less represented in the training data. Additionally, our results suggest that ChatGPT translates back and forth between English and German, meaning that translation quality plays a key role in non-English responses. Evaluating ChatGPT's effectiveness in other languages by simply testing its ability to translate economic terms and concepts, rather than conducting full-scale performance evaluations, is much faster and more practical for educators. This advice is also useful for students who are not native English speakers but are studying economics in English. However, while this method was effective for German, we cannot guarantee the same results for all languages.

6.2 Limitations

To understand the broader impact of our findings, it is important to acknowledge the limitations of our study. We discuss both methodological and empirical constraints, with the latter being particularly relevant for assessing the external validity of our results.

One methodological limitation comes from our approach to prompting. In this study, we used one-shot prompting, meaning that we asked ChatGPT a question only once, without refining or iterating on its response. While we recognize that iterative prompting can improve learning outcomes, we chose one-shot prompts because they serve as a baseline for more interactive learning approaches and are easier to evaluate systematically. Future research could explore how iterative prompting improves output quality and student learning.

Another limitation comes from our moderation-based evaluation of ChatGPT's responses. We assessed chatbot accuracy through a moderation process where two researchers discussed and resolved differences in grading, rather than using a blind double-marking system with independent assessments. While moderation improves consistency and alignment, it may introduce bias, as the second marker might be influenced by the first marker's assessment. To mitigate this, the second marker only saw the first marker's mark after forming their own opinion. Despite potential biases, we believe that seeking agreement rather than averaging scores resulted in a more reliable evaluation.

A final methodological limitation is that our evaluation reflects an educator's perspective, rather than how valuable students perceive ChatGPT for their learning. Research suggests that understanding how students engage with GenAI is essential for evaluating its true educational value. Future studies should expand the already existing literature (e.g. Bastani et al. 2024; Lehmann, Cornelius, and Sting 2024) on how students interact with GenAI and its impact on learning outcomes.

There are three key empirical limitations affecting the generalizability of our findings: First, we evaluated only three OpenAI ChatGPT models (GPT-3.5, GPT-4o, and o1preview). However, GenAI models evolve rapidly, and by the time of writing, GPT-3.5 was discontinued and o1preview was replaced by o1. Additionally, we did not test other GenAI models, such as Claude, Google Gemini, or DeepSeek R1, which could produce different results. While this may limit the long-term relevance of our findings, the trend towards models with better reasoning capabilities suggests that our results provide a lower bound for GenAI performance in education.

Second, our evaluation was based on content from The Economy 1.0, a textbook by CORE Econ. Since this textbook differs from more traditional materials, one could argue that results might vary with a different textbook. Furthermore, we are confident that The Economy 1.0 is included in ChatGPT's training data, while other economics textbooks—especially those protected by copyright—are likely excluded. However, because we focused on fundamental economic concepts, which are largely consistent across introductory textbooks, we believe this had minimal impact on our findings. Additionally, even if other textbooks are not in the training data, student notes, summaries, and related materials are widely available online, likely mitigating this limitation.

Third, we tested only 56 economic concepts and 25 text-based multiple-choice questions, covering principles of microeconomics and macroeconomics. This is a small subset of potential topics covered in economics courses and does not account for advanced economic concepts or questions requiring graph or image interpretation. While we ensured a broad selection of micro- and macroeconomics topics, we can only generalize our findings to introductory economics courses. For more advanced topics, GenAI performance is likely weaker due to the limited availability of high-quality training data on specialized subjects.

7 Conclusion

Research shows that one-on-one tutoring significantly improves student learning outcomes. This makes sense, as students have diverse backgrounds and face unique challenges in understanding course content. However, educators have limited capacity to address each student's individual needs. Growing student numbers and declining staff-to-student ratios reduce the possibility of personalized support. Additionally, classroom emotions such as shame or fear of negative evaluation may prevent students from expressing their difficulties. Lastly, while flipped classrooms promote student-centered learning, they require students to grasp complex topics independently, which can be challenging.

Given these challenges, recent advances in GenAI chatbots have sparked discussions about their potential as personalized, automated tutors. We argue that GenAI chatbots can act as automated tutors and support students in principles of economics courses with low risk of hallucination or misinformation. However, their explanatory depth and example quality remain insufficient. To assess their effectiveness, we evaluated ChatGPT models (GPT-3.5, GPT-4o, and o1preview) on two tasks: explaining economic concepts and answering multiple-choice questions. We used CORE Econ's The Economy 1.0 as a basis for concepts, questions, and our evaluation.

Our analysis found that newer models provide very accurate responses, though some inaccuracies persist. The concerning issue is that ChatGPT presents all responses with full confidence, making it difficult for students to recognize errors. Additionally, while responses are well-tailored to the target audience, they lack a holistic perspective on economic concepts, and the quality of generated examples is consistently poor. Despite these limitations, GenAI chatbots can support students, even as non-interactive automated tutors.

Our findings have important implications for both educators and researchers. Educators should recommend GenAI chatbots but highlight their weaknesses, particularly their limited scope of explanations and poor example quality. Since GenAI chatbots can assist with basic concept understanding, classroom instruction should prioritize real-world applications, problem-solving, and critical discussions. The technology may support instructors adopting a flipped classroom model, where students can use GenAI chatbots for pre-class preparation while in-class activities focus on higher-order learning. Educators should teach students to use GenAI as an interactive tutor, for example, by asking ChatGPT to evaluate their own explanations instead of generating answers for them. As the way students use GenAI influences its effectiveness, research should further explore best practices for GenAI-assisted learning. It is essential to understand how different learners benefit from GenAI to ensure that it promotes educational equity rather than widening existing disparities. For example, students with social anxiety may find an anonymous chatbot to be a safe space for asking questions without fear of judgment. Meanwhile, more advanced learners may effectively use GenAI-generated explanations and examples to deepen their understanding, whereas less experienced students might struggle with the technology's limitations. Students with stronger prompting skills are likely to receive more useful GenAI responses, further exacerbating learning gaps. Additionally, access remains a concern, as paid GenAI versions offer more advanced features, potentially creating inequality in the classroom. To address these challenges, education policy should focus on strategies to ensure all students have equal access to high-quality GenAI automated tutors.

GenAI has the potential to enhance personalized learning, but its effectiveness depends on how students engage with it. By guiding students to use this technology critically and effectively, educators can play a part in GenAI improving the learning outcomes of students.

Acknowledgements

We are grateful for the support by Omar Serraj, advice received from Thomas Deckers and the comments received at the TeachECONference 2024 of the Center for Teaching and Learning Economics (CTaLE).

A Descriptive labels for indicators in the marking grids

Error Type	
Comprehension Error	The chatbot does not understand the prompt
Factual Error	The chatbot provides information that is demonstrably incorrect based on established facts
Inference Error	The chatbot reaches a conclusion that is not logically justified by the information provided in the conversation or its own knowledge base

Error Domain	
Closed	Inconsistencies noticeable in the chat, a careful reader can identify the existence of the error
Open	Inconsistencies can only be noticed by extensive research outside of the chat

Table 4: Descriptive labels of the categories for Error Type and Error Domain. These indicators were used in both tasks.

	Problem indicators	
Value	Includes Calculation	Real-World Application
0	Mathematical handling is not required	The problem does not discuss a real-world phenomenon
1	Mathematical handling is required	The problem discusses a real-world phenomenon

Table 5: Descriptive labels categories for the binary problem indicators Includes Calculation and Real-World Application. These indicators were used in the answering and explanation of multiple-choice questions task.

	Problem indicators		Response indicators				
Score	Frequency	Difficulty	Accuracy	Scope	Specificity	Quality of Example	Safety
1	very uncommon	very easy	Significantly inaccurate or misleading information. Major misconceptions about the entire concept	Explanation demonstrates a complete misunderstanding of the concept	Unclear due to lack of detail	No example or irrelevant example	Presents explanation as absolute fact, ignoring any limitations. Does not provide sources
2	uncommon in economics	easy	Significantly inaccurate or misleading information for a very important part of the concept	Explanation covers just a few aspects of the concept	Unclear due to complexity of explanation	Misleading or incorrect example	Acknowledges the concept's complexity but doesn't discuss limitations. Does not provide sources
3	common in certain fields of economics	medium	Accurate most of the time, but inaccurate or misleading information for an important part of the concept	Explanation covers most aspects of the concept, but lacks some detail or nuance	Explanation is somewhat clear, but unnecessarily difficult to follow (e.g., difficult language, lacks consistency)	Weak or unhelpful example	Raises awareness for limitations of its explanation but lacks details on the specific issues. May or may not mention sources
4	very common in economics in general	hard	Mostly accurate, but inaccurate or misleading information for a minor part of the concept	Explanation covers all aspects of the concept, but does not provide a holistic view	Clear and accessible explanation	Relevant but simple example	Raises awareness for limitations of its explanation but provides details. May or may not mention sources
5	very common, not just in economics	very hard	Very accurate, no inaccuracies or misleading information	Explanation covers the full concept and a holistic view of the concept	Exemplary clear, accessible and addresses common challenges	Insightful and practical example	Critically discusses the limitations of the explanation in detail. Provides evidence from credible sources

Table 6: Descriptive labels for the scores of the indicators Frequency, Difficulty, Accuracy, Scope, Specificity, Quality of Example and Safety. All of these indicators were used in the explanation of concepts task.

	Problem indicators			Response indicators		
Score	Difficulty	Context Dependency	Contextual Transfer	Explanation Accuracy II	Specificity	Safety
1	very easy	Requires definitions of universally accepted concepts without any knowledge of the specific course material	Requires the use of definition of concepts without application	Incorrect explanations are very inaccurate	Unclear due to lack of detail	Presents explanation as absolute fact, ignoring any limitations. Does not provide sources
2	easy	Requires an understanding of concepts as typically taught in the specific course, including common models or examples that are likely covered in similar courses as well	The problem applies the concept to a very simplified example	Incorrect explanations are more incorrect than correct	Unclear due to complexity of explanation	Acknowledges the concept's complexity but doesn't discuss limitations. Does not provide sources
3	medium	Question relies heavily on unique content, including specific notations	The problem applies the concept to a more complex example	Incorrect explanations are correct, but with notable inaccuracies	Explanation is somewhat clear, but unnecessarily difficult to follow (e.g., difficult language, lacks consistency)	Raises awareness for limitations of its explanation but lacks details on the specific issues. May or may not mention sources
4	hard			Incorrect explanations are mostly accurate, but there are very minor inaccuracies	Clear and accessible explanation	Raises awareness for limitations of its explanation but provides details. May or may not mention sources
5	very hard			Incorrect explanations are very accurate.	Exemplary clear, accessible and addresses common challenges	Critically discusses the limitations of the explanation in detail. Provides evidence from credible sources

Table 7: Descriptive labels for the scores of the indicators Difficulty, Context Dependency, Contextual Transfer and Explanation Accuracy II, Specificity and Safety. All of these indicators were used in the answering and explanation of multiple-choice questions task.

B List of concepts and multiple-choice questions

B.1 Concepts

The following concepts were used in this study. If a concept was also investigated in German, the German expression is given in parenthesis: Aggregate demand (aggregierte Nachfrage), Assets, Asymmetric information, Automatic stabilizer (automatischer Stabilisator), Bank Money (Giralgeld), Bank (Geschäftsbank), Bargaining gap, Base money (Zentralbankgeld), Business cycle, Capitalism (Kapitalismus), Central bank, Ceteris Paribus, Circular flow of the economy, Competitive equilibrium, Consumer and producer surplus (Konsumenten- und Produzentenrente), Cost function (Kostenfunktion), Demand curve (Nachfragefunktion), Demand elasticity, Economic rent (ökonomische Rente), Employment rate (Beschäftigungsquote), Employment rent, Equilibrium unemployment, Fiscal stimulus (fiskalpolitischer Stimulus), Goods market equilibrium, Government debt (Staatsverschuldung), Gross domestic product (Bruttoinlandsprodukt), Incomplete contract, Indifference curve, Industrial Revolution (Industrielle Revolution), Inflation target (Inflationstargeting), Inflation, Institutions (Institutionen), Labour force (Erwerbspersonen), Liabilities, Marginal propensity to consume, Marginal Rate of Substitution (Grenzrate der Substitution), Marginal Rate of Transformation (Grenzrate der Transformation), Market, Money, Multiplier process (Multiplikatoreffekt), Nash-equilibrium, Okun's law (Okun'sches Gesetz), Opportunity cost, Pareto-efficiency, Participation rate, Perfect competition (perfekter Wettbewerb), Phillips curve (Phillipskurve), Policy interest rate (Leitzins), Price-setting curve, Prisoners dilemma, Recession (Rezession), Reservation option (Reservationslohn), Reservation wage, Technology (Technologie), Unemployment rate, Wage-setting curve.

B.2 Multiple-choice questions

We used 25 multiple-choice questions from Units 1 to 16 from CORE Econ's The Economy 1.0 textbook. Following a complete list of the questions with options are given. The numbering of the questions correspond to the numbering in the textbook. CORE Econ's The Economy 1.0 is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 International license (CC BY-NC-ND 4.0). The copyright to the questions lies thus with CORE Econ and its authors. For our investigation using the German language, we used the translation of these questions from Die Wirtschaft 1.0, the German translation of the textbook.

Question 3.5

What is the marginal rate of substitution (MRS)?

1. The ratio of the amounts of the two goods at a point on the indifference curve.
2. The amount of one good that the consumer is willing to trade for one unit of the other.
3. The change in the consumer's utility when one good is substituted for another.
4. The slope of the indifference curve.

Question 3.6

You are a taxi driver in Melbourne who earns A\$50 for a day's work. You have been offered a one-day ticket to the Australian Open for A\$40. As a tennis fan, you value the experience at A\$100. With this information, what can we say?

1. The opportunity cost of the day at the Open is A$40.
2. The economic cost of the day at the Open is A$40.
3. The economic rent of the day at the Open is A$10.
4. You would have paid up to A$100 for the ticket.

Question 4.1

In a simultaneous one-shot game:

1. A player observes what others do before deciding how to act.
2. A player decides his or her action, taking into account what other players may do after knowing his or her move.
3. Players coordinate to find the actions that lead to the optimal outcome for society.
4. A player chooses an action taking into account the possible actions that other players can take.

Question 4.6

Four farmers are deciding whether to contribute to the maintenance of an irrigation project. For each farmer, the cost of contributing to the project is $10. But for each farmer who contributes, all four of them will benefit from an increase in their crop yields, so they will each gain $8.

1. If all the farmers are selfish, none of them will contribute.
2. If one of the farmers, Kim, cares about her neighbour Jim just as much as herself, she will contribute $10.
3. If Kim is altruistic and contributes $10, the others might contribute too, even if they are selfish.
4. If the farmers have to reconsider this decision every year, they might choose to contribute to the project even if they are selfish.

Question 5.1

Which of the following statements about the outcome of an economic interaction is correct?

1. If the allocation is Pareto efficient, then you cannot make anyone better off without making someone else worse off.
2. All participants are happy with what they get if the allocation is Pareto efficient.
3. There cannot be more than one Pareto-efficient outcome.
4. According to the Pareto criterion, a Pareto-efficient outcome is always better than an inefficient one.

Question 6.1

Which of the following statements is true?

1. A labour contract transfers ownership of the employee from the employee to the employer.
2. The office where the employee works is a relation-specific asset, because the employee cannot use it after leaving the firm.
3. In a labour contract, one side of the contract has the power to issue orders to the other side, but this power is absent from a sale contract.
4. A firm is a structure that involves decentralization of power to the employees.

Question 6.2

Which of the following statements about the separation of ownership and control is true?

1. When the ownership and control of a firm is separated, the managers become the residual claimants.
2. Managers always work to maximize the firm's profit.
3. One way to address the problem associated with the separation of ownership and control is to pay the managers a salary that depends on the performance of the firm's share price.
4. It is effective for shareholders to monitor the performance of the management, in a firm owned by a large number of shareholders.

Question 6.4

In which of the following employment situations would the employment rent be high, *ceteris paribus*?

1. In a job that provides many benefits, such as housing and medical insurance.
2. In an economic boom, when the ratio of job-seekers to vacancies is low.
3. When the worker is paid a high salary because she is a qualified accountant and there is a shortage of accountancy skills.
4. When the worker is paid a high salary because the firm's customers know and trust her.

Question 6.5

Maria earns $12 per hour in her current job and works 35 hours a week. Her disutility of effort is equivalent to a cost of $2 per hour of work. If she loses her job, she will receive unemployment benefit equivalent to $6 per hour. Additionally, being unemployed has psychological and social costs equivalent to $1 per hour. Then:

1. The employment rent per hour is $3.
2. Maria's reservation wage is $6 per hour.
3. Maria's employment rent if she can get another job with the same wage rate after 44 weeks of being unemployed is $6,160.
4. Maria's employment rent if she can only get a job at a lower wage rate after 44 weeks of being unemployed is more than $7,700.

Question 7.1

A firm's cost of production is £12 per unit of output. If P is the price of the output good and Q is the number of units produced, which of the following statements is correct?

1. Point $(Q, P) = (2,000, 20)$ is on the isoprofit curve representing the profit level £20,000.
2. Point $(Q, P) = (2,000, 20)$ is on a lower isoprofit curve than point $(Q, P) = (1,200, 24)$.
3. Points $(Q, P) = (2,000, 20)$ and $(4,000, 16)$ are on the same isoprofit curve.
4. Point $(Q, P) = (5,000, 12)$ is not on any isoprofit curve.

Question 7.2

Consider a firm whose unit cost (the cost of producing one unit of output) is the same at all output levels. Which of the following statements are correct?

1. Each isoprofit curve depicts the firm's profit for different outputs for a given price of the output good.
2. Isoprofit curves can be upward-sloping when at high profit levels.
3. Every price-quantity combination lies on an isoprofit curve.
4. Isoprofit curves slope downward when the price is above the unit cost.

Question 7.5

Consider a firm with fixed costs of production. Which of the following statements about its average cost (AC) and marginal cost (MC) is correct?

1. When AC = MC, the AC curve has a zero slope.
2. When AC > MC, the MC curve is downward-sloping.
3. When AC < MC, the AC curve is downward-sloping.
4. The MC curve cannot be horizontal.

Question 7.6

Suppose that the unit cost of producing a pound of cereal is \$2, irrespective of the level of output. (This means there are no fixed costs, that is, costs that are present for any level of output, including zero.) Which of the following statements is correct?

1. The total cost curve is a horizontal straight line.
2. The average cost curve is downward-sloping.
3. The marginal cost curve is upward-sloping.
4. The average cost and the marginal cost curves coincide.

Question 9.1

Which of the following statements is correct?

1. To maximize profits, firms set the wage at the level where the workers are indifferent between working and not working.
2. Firms aim to set as high a price as possible.
3. In equilibrium, the wage clears the labour market, so there is no unemployment.
4. If all firms set the same price and pay the same nominal wage, then the higher the real wage that they pay, the lower is their markup.

Question 9.2

Which of the following statements is correct?

1. participation rate = employed ÷ labour force
2. unemployment rate = unemployed ÷ population of working age
3. employment rate = employed ÷ population of working age
4. employment rate + unemployment rate = 1

Question 10.8

Which of the following statements is correct?

1. Money is the cash (coins and notes) used as the medium of exchange to purchase goods and services.
2. Bank money is the total money in the savers' deposit accounts at the bank.
3. Base money is broad money minus bank money.
4. Liquidity transformation occurs when the banks transform illiquid deposits into liquid loans.

Question 10.11

Which of the following statements are correct regarding the principal–agent problem?

1. A principal–agent problem exists in loans due to a positive possibility of the principal not being repaid.
2. The principal–agent problem can be resolved by writing a binding contract for the borrower to exert full effort.
3. One solution for the principal–agent problem in loans is for the borrower to provide equity.
4. The principal–agent problem leads to credit rationing in the loans market.

Question 13.3

Which of the following statements is correct regarding measuring GDP?

1. GDP can be measured either as the total spending on domestically produced goods and services, or the total value added in domestic production, or the sum of all incomes received from domestic production.
2. Information about exports but not imports is necessary to calculate GDP.
3. Government production is not included in the GDP.
4. The value added of government production is computed using the price that public goods and services are sold at in the market.

Question 14.3

Which of the following statements is correct regarding household wealth?

1. A household's material wealth is its financial wealth plus the value of its house.
2. The total broad wealth equals material wealth plus expected future earnings.
3. A household adjusts its precautionary saving in response to changes in its target wealth.
4. If the household's target wealth is above its expected wealth, then it will decrease savings and increase consumption.

Question 14.6

The aggregate demand of an open economy is given by the after-tax domestic consumption C, the investment I (which depends on the interest rate r), the government spending G and net exports $X - M$:

$$\begin{aligned}\text{AD} &= C + I + G + X - M \\ &= c_0 + c_1(1-t)Y + I(r) + G + X - mY\end{aligned}$$

c_0 is autonomous consumption, c_1 is the marginal propensity to consume, and m is the marginal propensity to import. In the economy's equilibrium this equals its output: AD $= Y$. Solving for Y yields:

$$Y = \left(\frac{1}{1 - c_1(1-t) + m}\right)(c_0 + I(r) + G + X)$$

Given this equation, which of the following increases the multiplier?

1. A fall in government spending.
2. A fall in the interest rate.
3. A fall in the marginal propensity to import.
4. A rise in the tax rate.

Question 14.8

Which of the following statements is correct regarding the multiplier?

1. Economists tend to agree on their estimates of the multiplier.
2. Reverse causation can be a problem when estimating the multiplier empirically.
3. If households anticipate that increased government spending will be funded by future tax increases, then the multiplier will be higher.
4. If firms anticipate that the government's fiscal policy will be effective, then the multiplier will be higher.

Question 15.1

The following table shows the annual inflation rate (the GDP deflator) of Japan, the UK, China and Nauru in the period 2010–2013 (Source, World Bank):

	2010	2011	2012	2013
Japan	-1.9%	-1.7%	-0.8%	-0.3%
UK	1.6%	2.0%	1.6%	1.9%
China	6.9%	8.2%	2.4%	2.2%
Nauru	-18.2%	18.1%	24.1%	-21.7%

Based on this information, which of the following statements is correct?

1. Japan experienced a persistent period of disinflation between 2010 and 2013.
2. In the UK the price of goods and services remained stable between 2010 and 2013.
3. China has been experiencing deflationary pressure between 2011 and 2013.
4. Nauru's price level at the end of 2013 is lower than it was at the start of 2010.

Question 15.8

Which of the following statements is correct regarding monetary policy?

1. When interest rates go down, asset prices go up.
2. The zero lower bound refers to the central bank's inability to set the real interest rate to below zero.
3. Quantitative easing involves the central bank lowering its official interest rate.
4. Interest rates cannot be set in a currency union.

Question 16.5

Which of the following statements is correct regarding the model of the labour market?

1. In the short- and medium-run models the amount of capital is fixed, while in the long-run model the amount of capital can vary.
2. Labour-saving technological progress raises unemployment in both the short and long run.
3. In the long-run model, firms enter the market when the markup is low.
4. In the long-run model, the markup is independent of the number of firms.

Question 16.8

Does the introduction of a new labour-saving technology result in?

1. Higher wage share of output and higher Gini coefficient in the short run.
2. Lower wage share of output and higher Gini coefficient in the short run.
3. Lower wage share of output and lower Gini coefficient in the short run.
4. Higher unemployment, lower wage share of output, and higher Gini coefficient in the long run.

References

Akçayır, Gökçe and Murat Akçayır. “The flipped classroom: A review of its advantages and challenges”. In: *Computers & Education* 126 (Nov. 2018), pp. 334–345.

Ali, Rohaid et al. “Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank”. In: *Neurosurgery* 93.5 (2023).

Bastani, Hamsa et al. *Generative AI Can Harm Learning.* 2024. URL: `https://www.ssrn.com/abstract=4895486`.

Bergmann, Jonathan and Aaron Sams. *Flip your classroom: Reach every student in every class every day.* First Edition. International Society for Technology in Education, June 2012.

Bloom, Benjamin S. “The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring”. In: *Educational Researcher* 13.6 (June 1984), pp. 4–16.

Bowles, Sam, Wendy Carlin, and Margaret Stevens. *The Economy: Economics for a Changing World.* 1.0. Oxford: Core Econ and Oxford University Press, 2017.

Brown, Tom. *GPT-3 Dataset Language Statistics.* 2020. URL: `https://github.com/openai/gpt-3/tree/master/dataset_statistics`.

Bubeck, Sébastien et al. *Sparks of Artificial General Intelligence: Early experiments with GPT-4.* Mar. 2023. URL: `http://arxiv.org/pdf/2303.12712v5`.

Buchanan, Joy, Stephen Hill, and Olga Shapoval. “ChatGPT Hallucinates Non-existent Citations: Evidence from Economics”. In: *The American Economist* 69.1 (2024), pp. 80–87.

Buckner, Elizabeth and You Zhang. “The quantity-quality tradeoff: a cross-national, longitudinal analysis of national student-faculty ratios in higher education”. In: *Higher Education* 82.1 (July 2021), pp. 39–60.

Chan, Cecilia Ka Yuk and Wenjie Hu. “Students’ voices on generative AI: perceptions, benefits, and challenges in higher education”. In: *International Journal of Educational Technology in Higher Education* 20.1 (2023).

Chi, Michelene T.H. et al. “Learning from human tutoring”. In: *Cognitive Science* 25.4 (July 2001), pp. 471–533.

Cotton, Debby R. E., Peter A. Cotton, and J. Reuben Shipway. “Chatting and cheating: Ensuring academic integrity in the era of ChatGPT”. In: *Innovations in Education and Teaching International* 61.2 (Mar. 2024), pp. 228–239.

Farhat, Faiza et al. *Evaluating Large Language Models for the National Premedical Exam in India: Comparative Analysis of GPT-3.5, GPT-4, and Bard.* Aug. 2023. URL: `http://preprints.jmir.org/preprint/51523`.

Farrokhnia, Mohammadreza et al. “A SWOT analysis of ChatGPT: Implications for educational practice and research”. In: *Innovations in Education and Teaching International* 61.3 (May 2024), pp. 460–474.

Geerling, Wayne et al. “ChatGPT has Aced the Test of Understanding in College Economics: Now What?” In: *The American Economist* 68.2 (2023).

Gilson, Aidan et al. “How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment”. In: *JMIR Medical Education* 9 (Feb. 2023).

Grassini, Simone. “Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings”. In: *Education Sciences* 13.7 (July 2023), p. 692.

Grassini, Simone, Maren Linnea Aasen, and Anja Møgelvang. “Understanding University Students’ Acceptance of ChatGPT: Insights from the UTAUT2 Model”. In: *Applied Artificial Intelligence* 38.1 (Dec. 2024).

Grasso, SandraJean M. “Use of a Social Question Answering Application in a Face-to-Face College Biology Class”. In: *Journal of Research on Technology in Education* 49.3-4 (Oct. 2017), pp. 212–227.

Henkel, Owen et al. “Effective and Scalable Math Support: Experimental Evidence on the Impact of an AI-Math Tutor in Ghana”. In: *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*. Ed. by Andrew M. Olney et al. Vol. 2150. Springer Nature Switzerland, 2024, pp. 373–381.

Holodynski, Manfred and Stefanie Kronast. “Shame and pride: Invisible emotions in classroom research”. In: *Emotions as Bio-cultural Processes*. Ed. by Hans J. Markowitsch and Birgitt Röttger-Rössler. Springer US, 2009.

Hwang, Gwo-Jen, Chiu-Lin Lai, and Siang-Yi Wang. “Seamless flipped learning: a mobile technology-enhanced flipped classroom with effective learning strategies”. In: *Journal of Computers in Education* 2.4 (Dec. 2015), pp. 449–473.

Ji, Ziwei et al. “Survey of Hallucination in Natural Language Generation”. In: *ACM Computing Surveys* 55.12 (2023), pp. 1–38.

Kasneci, Enkelejda et al. “ChatGPT for good? On opportunities and challenges of large language models for education”. In: *Learning and Individual Differences* 103 (2023), p. 102274.

Korinek, Anton. “Generative AI for Economic Research: Use Cases and Implications for Economists”. In: *Journal of Economic Literature* 61.4 (Dec. 2023), pp. 1281–1317.

Kung, Tiffany H. et al. “Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models”. In: *PLOS Digital Health* 2.2 (Feb. 2023). Ed. by Alon Dagan.

Lehmann, Matthias, Philipp B. Cornelius, and Fabian J. Sting. *AI Meets the Classroom: When Does ChatGPT Harm Learning?* Aug. 2024. URL: `http://arxiv.org/abs/2409.09047`.

Michel-Villarreal, Rosario et al. “Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT”. In: *Education Sciences* 13.9 (2023), p. 856.

Miyazoe, Terumi and Terry Anderson. “Anonymity in blended learning: who would you like to be?” In: *Journal of Educational Technology & Society* 14.2 (2011), pp. 175–187.

Mollick, Lilach and Ethan Mollick. “Assigning AI: Seven Approaches for Students with Prompts”. In: *The Wharton School Research Paper* (2023).

OECD. *OECD Data Explorer - Share of total government expenditure on education*. 2025. URL: `http://data-explorer.oecd.org/s/ug`.

OpenAI. *GPT-4 System Card*. 2023. URL: `https://cdn.openai.com/papers/gpt-4-system-card.pdf`.

– *GPT-4 Technical Report (v6)*. 2023.

Peeperkorn, Max et al. *Is Temperature the Creativity Parameter of Large Language Models?* May 2024. URL: `http://arxiv.org/abs/2405.00492`.

Rudolph, Jürgen, Samson Tan, and Shannon Tan. “ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?” In: *Journal of Applied Learning & Teaching* 6.1 (Jan. 2023).

Schulze Balhorn, Lukas et al. “Empirical assessment of ChatGPT’s answering capabilities in natural science and engineering”. In: *Scientific Reports* 14.1 (Feb. 2024), p. 4998.

Sekwena, Gailele L. “Active Learning Pedagogy for Enriching Economics Students’ Higher Order Thinking Skills”. In: *International Journal of Learning, Teaching and Educational Research* 22.3 (Mar. 2023), pp. 241–255.

Slavin, Robert E. “Making Chapter 1 Make a Difference”. In: *The Phi Delta Kappan* 69.2 (1987), pp. 110–119.

Strzelecki, Artur. “To use or not to use ChatGPT in higher education? A study of students’ acceptance and use of technology”. In: *Interactive Learning Environments* 32.4 (2024).

Sullivan, Patrick. ““It’s Easier to Be Yourself When You Are Invisible”: Female College Students Discuss Their Online Classroom Experiences”. In: *Innovative Higher Education* 27 (2002).

Terwiesch, Christian. “Would Chat GPT Get a Wharton MBA?” In: *The Mack Institute for Innovation Management White Paper* (2023).

VanLehn, Kurt et al. “Why Do Only Some Events Cause Learning During Human Tutoring?” In: *Cognition and Instruction* 21.3 (Sept. 2003), pp. 209–249.

Von Garrel, Jörg and Jana Mayer. “Artificial Intelligence in studies—use of ChatGPT and AI-based tools among students in Germany”. In: *Humanities and Social Sciences Communications* 10.1 (Nov. 2023), p. 799.

Walczak, Krzysztof and Wojciech Cellary. “Challenges for higher education in the era of widespread access to generative AI”. In: *Economics and Business Review* 9.2 (2023).

Wood, David A. et al. “The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions?” In: *Issues in Accounting Education* 38.4 (2023), pp. 81–108.

Zheng, Shen, Jie Huang, and Kevin Chen-Chuan Chang. *Why Does ChatGPT Fall Short in Providing Truthful Answers?* Apr. 2023. URL: `http://arxiv.org/pdf/2304.10513v3`.

Zhu, Gang. “Is flipping effective? A meta-analysis of the effect of flipped instruction on K-12 students’ academic achievement”. In: *Educational Technology Research and Development* 69.2 (Apr. 2021), pp. 733–761.

Zirar, Araz. “Exploring the impact of language models, such as ChatGPT, on student learning and assessment”. In: *Review of Education* 11.3 (Dec. 2023).