

Article

Evaluating GPT-4o on Introductory Economics Problem Sets: A Rating-Based Benchmark of Question Type and Assessment Design

Yun Liu ^{1,*} , Tina Wong ¹, Tak Wai Chau ¹ and Yuchang Cao ² 

¹ School of Accounting and Finance, Hong Kong Polytechnic University, Hong Kong SAR 999077, China; aftina.wong@polyu.edu.hk (T.W.); tak-wai.chau@polyu.edu.hk (T.W.C.)

² Independent Researcher, Wuxi 214000, China; cyc1580862183@outlook.com

* Correspondence: liuy050229@gmail.com

Abstract

Problem sets are widely used in economics education, but the availability of generative artificial intelligence creates new challenges for assessment validity and academic integrity. This study conducts a rating-based benchmarking evaluation of GPT-4o on introductory economics problem-set items. Using 260 rated questions and three independent instructor ratings per item, we examine how GPT-4o performance varies across discussion, numerical, graphical and mixed-modality questions. The study is descriptive rather than causal: it benchmarks GPT-4o outputs under a specified prompting, input-modality and scoring protocol. All items were submitted through the ChatGPT web interface in fresh sessions, each item was answered once, graphical items were provided through uploaded original diagram images, no follow-up prompts were used, and outputs were saved without editing. Results show that GPT-4o performs comparatively well on text-based discussion and numerical-only items, but substantially less well on graphical items, especially those requiring numerical reasoning grounded in a graph. Inter-rater reliability is high according to intraclass correlation coefficients, and pooled rater-item analyses confirm the graphical and graphical-numerical performance gap as a descriptive benchmark pattern. To improve reproducibility while respecting copyright restrictions, the revised manuscript specifies the prompting protocol, coding procedure, rater procedure and supplementary replication files. The findings suggest that economics instructors should not simply add graphical questions as an “AI-proof” device, but should design constructively aligned, accessible, mixed-format assessments that validly sample intended economic reasoning skills. The conclusions are restricted to GPT-4o, the February–June 2025 testing window, the ChatGPT web-interface protocol and the item corpus used in this study.



Academic Editors: José Fernández-Cerero and Marta Montenegro-Rueda

Received: 17 April 2026

Revised: 9 June 2026

Accepted: 15 June 2026

Published: 24 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: economics education; GPT-4o; generative AI; assessment validity; academic integrity; problem sets; multi-modal reasoning

1. Introduction

Problem sets are a widely used form of continuous assessment in higher education. In economics education, they are especially valuable because they allow instructors to assess whether students can define concepts, apply models, perform calculations, interpret diagrams and communicate economic reasoning. Their use is consistent with longstanding evidence that frequent practice, feedback and retrieval support learning ([Brown et al., 2014](#); [Chickering & Gamson, 1987](#); [Gibbs & Simpson, 2005](#)).

The rapid diffusion of generative artificial intelligence (GenAI) tools has, however, complicated the interpretation of problem-set scores. When a student can obtain a plausible answer from a model without performing the intended reasoning process, the assessment score may no longer support the intended inference about student learning. This is not only an academic-integrity issue, but also an assessment-validity issue: the problem is whether the submitted answer continues to represent the student's own achievement of the intended learning outcomes (Kane, 2013; Messick, 1995).

This concern is particularly salient in introductory economics, where problem sets often mix three forms of disciplinary reasoning. First, discussion items require conceptual explanation, interpretation or policy reasoning. Second, numerical items require arithmetic, algebraic manipulation or quantitative comparison. Third, graphical items require students to read, interpret or construct diagrams such as demand–supply curves, cost curves, welfare areas and macroeconomic schedules (Becker, 1997; Pieters, 2024; Siegfried & Fels, 1979). Many authentic economics tasks combine these demands; for example, students may need to identify a welfare area from a diagram and then compute consumer surplus, tax revenue or deadweight loss.

The study benchmarks GPT-4o on a corpus of introductory economics problem-set items. It frames the study as a rating-based benchmarking exercise. Its aim is not to estimate causal effects of question type, but to describe how GPT-4o performs across item formats under a transparent protocol.

The study contributes to the economics-education and assessment-literacy literature in three ways. First, it provides item-level evidence on GPT-4o performance across economics problem-set formats. Second, it links GenAI vulnerability to assessment validity and constructive alignment rather than treating graphical difficulty as a merely technical observation. Third, it uses a multi-label coding scheme and rater-level analysis to better represent the hierarchical structure of the data. The practical contribution is a cautious one: the study does not identify an “AI-proof” format, but offers evidence for designing balanced, accessible and constructively aligned assessments that sample multiple forms of economic reasoning.

The remainder of the paper proceeds as follows. Section 2 develops the conceptual framework. Section 3 describes the rating-based benchmarking design, data, GPT-4o protocol, item coding, instructor ratings and statistical strategy. Section 4 reports the results. Section 5 discusses mechanisms, pedagogical implications and limitations. Section 6 concludes.

2. Literature Review and Conceptual Framework

2.1. Problem Sets, Constructive Alignment and Assessment Validity

Problem sets have long been central to economics teaching because they help students practice model-based reasoning, receive feedback and connect abstract concepts to applied problems (Brown et al., 2014; Koedinger et al., 2012; Taylor, 1950). In large introductory courses, problem sets also provide a scalable assessment format that can be graded consistently across sections and instructors (Myers et al., 2011; Pieters, 2024; Watts & Schaur, 2011).

From a constructive-alignment perspective, an assessment should elicit performances that correspond to the intended learning outcomes and teaching activities of the course (Biggs et al., 2022). In economics, relevant learning outcomes include explaining economic mechanisms, deriving equilibrium conditions, interpreting diagrams, comparing welfare outcomes and communicating implications. A well-designed problem set should therefore sample across these competencies rather than overrepresenting one narrow form of response.

The rise of GenAI changes the assessment-validity problem. Validity concerns the interpretation and use of assessment scores rather than the test instrument alone

(Kane, 2013; Messick, 1995). If a problem-set answer can be generated by GPT-4o without the student's meaningful engagement, the score may no longer validly indicate the student's own understanding. Assessment redesign should therefore be guided by validity and learning principles, not only by detection or deterrence.

2.2. GenAI Use, Academic Integrity and Assessment Security

We distinguish between GenAI use and academic misconduct. GenAI use refers broadly to interaction with AI tools for learning, drafting, checking or problem solving. Academic misconduct occurs when such use violates explicit assessment rules, misrepresents authorship or substitutes for required student work. This distinction matters because GenAI may support learning when transparently integrated, while also creating integrity risks when used covertly in restricted assessments (Cotton et al., 2024; Eager & Brunton, 2023; Kasneci et al., 2023).

Detection-based approaches are insufficient on their own. Studies of AI-generated text detection suggest that reliable identification is difficult, and false positives or false negatives can create fairness problems (Khalil & Er, 2023; Weber-Wulff et al., 2023). A complementary approach is assessment security by design: tasks can be structured to require process evidence, contextual judgment, explanation of reasoning, multi-modal integration, oral follow-up or authentic application (Boud & Falchikov, 2006; Dawson, 2020; Tai et al., 2018).

However, the goal should not be to search for a single "AI-proof" item type. Such a strategy may reduce construct coverage, encourage superficial redesign or create accessibility barriers. Instead, instructors should build an assessment ecology in which different item formats jointly support validity, fairness and learning. This includes clear AI-use policies, transparent rubrics, formative feedback, randomized parameters where appropriate and accessible alternatives for students with relevant needs (CAST, 2018).

2.3. Multi-Modal LLM Capability and Economics Question Types

GPT-4o is a multi-modal model rather than a text-only model (OpenAI, 2024). It would therefore be inaccurate to claim that current LLMs simply lack visual perception. A more precise question is whether a model can reliably complete the chain of operations required by economics problem sets: reading the prompt, parsing a diagram, identifying economically relevant objects, extracting values or relationships, performing calculations and explaining the final economic implication.

Recent multi-modal benchmarks show that advanced models can still face difficulties when tasks require visual mathematical reasoning, chart-based inference or expert multi-modal reasoning (Lu et al., 2024; Masry et al., 2022; Tong et al., 2024; Yue et al., 2024). Economics graphical tasks are a natural setting for examining these issues because successful answers often require both visual interpretation and domain-specific reasoning.

We therefore code each item along three non-mutually exclusive dimensions: discussion or explanation (DIS), numerical calculation (NUM) and graphical interpretation (GPH). Because many economics tasks combine these dimensions, the revised study uses multi-label coding rather than forcing each question into a single primary category. This framework leads to three descriptive hypotheses.

Hypothesis 1. *GPT-4o performance differs across economics question types because item formats require different combinations of language generation, calculation, visual parsing and economic reasoning.*

Hypothesis 2. *Items requiring graphical interpretation receive lower instructor ratings than otherwise comparable non-graphical items under the documented GPT-4o protocol.*

Hypothesis 3. *Items requiring both graphical interpretation and numerical calculation receive especially low ratings, because successful answers require integration across visual extraction, quantitative computation and economic explanation.*

These hypotheses are descriptive and benchmarking-oriented. They do not imply that item type is randomly assigned or that the design identifies causal effects.

3. Materials and Methods

3.1. Design

This study uses a rating-based benchmarking design. GPT-4o was asked to answer a corpus of introductory economics problem-set items. Three independent economics instructors then rated the model responses. The unit of analysis is the rater–item observation, with three ratings nested within each item. The design is descriptive: estimated associations between item features and ratings should be interpreted as performance patterns under the specific model, item pool, prompt protocol and grading procedure used here.

3.2. Question Corpus and Item Selection

The initial corpus contained 261 candidate introductory economics problem-set items drawn from commonly used textbook and institutional teaching materials, including *Principles of Economics* (Mankiw, 2022) and *Foundations of Economics* (Bade & Parkin, 2022). One item had missing marks from all three instructors and was excluded before analysis, leaving 260 rated items and 780 rater–item observations.

The items cover standard introductory topics such as market equilibrium, elasticity, welfare analysis, externalities, firm costs, competition, monopoly and macroeconomic aggregates. The sample is purposive rather than random: items were selected to cover core economics topics and to include discussion, numerical, graphical and mixed-modality formats. We therefore do not claim that the corpus is statistically representative of all economics assessments.

The original question items and diagrams are copyright-protected and were accessed through institutional channels. They cannot be redistributed publicly without publisher rights clearance. To improve transparency while respecting these copyright restrictions, the Supplementary Materials provide an anonymized item-level audit table containing: item ID, topic area, source type, whether a graph or figure was involved, the final *DIS*, *NUM* and *GPH* labels, the mutually exclusive item-type category, a paraphrased description of the task, the three normalized instructor scores and the file identifier for the corresponding GPT-4o output. The Supplementary Files do not reproduce copyrighted wording or diagrams. Appendix B gives paraphrased examples of each item type and explains the coding rationale.

3.3. GPT-4o Protocol

GPT-4o outputs were generated between February and June 2025 through the ChatGPT-4o web interface (<https://chatgpt.com/>). Each item was submitted in a fresh session so that no previous item content appeared in the conversation history. Each item was answered once only; no regeneration, follow-up prompting or answer selection across multiple runs was used. Because the web interface was used, temperature and other decoding settings were not user-controllable.

For non-graphical items, the item text was entered below the standardized instruction prompt reported in Appendix A. For graphical items, the original diagram image was uploaded together with the relevant question text, and the standardized prompt explicitly instructed GPT-4o to use any graph or figure provided. Thus, the graphical-item results

should be interpreted as a benchmark of GPT-4o under an image-upload condition in the ChatGPT web interface, not as a text-only transcription condition.

The standardized instruction prompt was: “Please answer an introductory economics problem-set question. Provide a complete solution. Show the economic reasoning, define variables when needed, show calculations step by step and explain the final answer clearly. If a graph or figure is provided, use information from the graph or figure in your answer.” The output produced after this single prompt was saved without editing and sent to instructors for rating. The full procedural record is reported in Appendix A.

3.4. Question Coding

Each item was coded using a multi-label scheme. $DIS = 1$ if the item required conceptual explanation, interpretation, policy evaluation or written economic reasoning beyond a purely numeric final answer. $NUM = 1$ if the item required arithmetic, algebraic manipulation, quantitative comparison, numerical substitution or computation for full-credit completion. $GPH = 1$ if a diagram, graph or figure was necessary for solving the item or extracting information used in the answer. A single item could therefore receive more than one label.

Borderline cases were handled using explicit decision rules. If a graph was merely illustrative and all necessary numerical or conceptual information appeared in the written stem, the item was not coded as GPH . If a numerical value appeared in the stem but no calculation or quantitative comparison was required, the item was not coded as NUM . If the item required students to explain an economic mechanism, justify an answer or interpret an implication, it was coded as DIS , even when a calculation was also required.

The coding was conducted as a two-stage consensus procedure. The first author assigned the initial DIS , NUM and GPH labels using the coding rules above. A second author then audited the item-level labels against the codebook and flagged ambiguous or borderline cases. Disagreements or flagged cases were resolved through discussion, with reference to whether the item required explanation, calculation and/or graphical extraction for full-credit completion. Because the original workflow was consensus-based rather than a prospectively designed blind double-coding procedure, we do not present Cohen’s κ as an original design feature. To strengthen verifiability, the supplementary item audit table reports the final labels and a paraphrased coding rationale for every retained item. The absence of a prospective independent intercoder-reliability design is acknowledged as a limitation in Section 5.3.

After excluding the one unrated item, the analytic sample contains 126 discussion items, 191 numerical items and 34 graphical items. Eighty-six items have more than one label.

3.5. Instructor Ratings and Inter-Rater Reliability

Three economics instructors independently rated each GPT-4o response on a 0–100 scale using the common grading criteria in Table 1. The rubric assigned 50% of the score to accuracy, 17% to elaboration with necessary details, 17% to structure and 16% to correct English and economic expressions. Raters were instructed to evaluate the saved GPT-4o answer as a student-style problem-set response and not to revise, complete or correct the answer during scoring.

The raters were not given the DIS , NUM and GPH coding variables or the study hypotheses. Full blinding to modality was not possible because the item and model answer sometimes revealed whether a graph or calculation was involved. No consensus scoring was imposed after rating. The raters did not discuss individual scores during the formal rating process. No formal calibration meeting or pilot-scoring adjustment was conducted

before the final ratings; this decision preserved independent judgments but is a procedural limitation. The high ICC values reported below indicate that, despite the absence of formal calibration, the raters displayed strong consistency and agreement.

Because Pearson correlations measure association but not absolute agreement, we report intraclass correlation coefficients (ICCs) in addition to pairwise correlations (Koo & Li, 2016; Shrout & Fleiss, 1979).

Table 1. Assessment grading rubric used for instructor ratings.

Marking Criterion	Weight
Accuracy	50%
Elaboration with necessary details	17%
Structure	17%
Correct English and economic expressions	16%

3.6. Statistical Strategy

The dependent variable is the normalized instructor score, $\mathcal{G}_{ir} = G_{ir}/100$, where i indexes items and r indexes raters. The main analysis uses the long-format dataset of 780 rater–item observations. Because item type was not randomly assigned and the corpus is purposive rather than a probability sample from a defined population of all economics assessment items, the models are used for adjusted descriptive benchmarking rather than causal inference or population-level hypothesis testing.

$$\mathcal{G}_{ir} = \alpha + \sum_{m \neq \text{NUM only}} \beta_m T_{im} + \delta_r + \varepsilon_{ir}, \quad (1)$$

where T_{im} denotes mutually exclusive item-type indicators, numerical-only items are the reference group and δ_r denotes rater fixed effects. The coefficients are interpreted as rater-adjusted descriptive mean contrasts relative to numerical-only items.

We also estimate a full main-effect and pairwise-interaction specification:

$$\mathcal{G}_{ir} = \alpha + \beta_1 DIS_i + \beta_2 NUM_i + \beta_3 GPH_i + \beta_4 (DIS_i \times NUM_i) + \beta_5 (DIS_i \times GPH_i) + \beta_6 (NUM_i \times GPH_i) + \delta_r + \varepsilon_{ir}. \quad (2)$$

This replaces the earlier interaction-only specification and retains all relevant main effects. We report only analyses that are visible in the results tables or reproducible from the Supplementary Materials.

4. Results

4.1. Descriptive Statistics

Table 2 reports descriptive statistics for the three normalized instructor scores and item-type indicators. Across raters, GPT-4o received average scores of approximately 0.67 to 0.73, indicating generally moderate-to-high performance but with substantial dispersion.

Table 2. Sample descriptive statistics.

Variable	Mean	S.D.	Min	Median	Max
G1	0.667	0.281	0.000	0.800	1.000
G2	0.726	0.302	0.000	0.875	1.000
G3	0.732	0.311	0.000	0.880	1.000
DIS	0.485	0.501	0.000	0.000	1.000
NUM	0.735	0.442	0.000	1.000	1.000
GPH	0.131	0.338	0.000	0.000	1.000

Note: N = 260 items after excluding one item with missing ratings from all three instructors.

Table 3 reports mutually exclusive item-type cells. Numerical-only items are the largest group, followed by discussion–numerical items and discussion-only items. Graphical-only, numerical–graphical and all-three-modality items are relatively small cells. This small graphical subsample is an important limitation and is discussed below.

Table 3. Mutually exclusive item-type cells and mean GPT-4o score.

Item Type	DIS	NUM	GPH	Items	Percent	Mean Score
Discussion only	1	0	0	49	18.8	0.778
Numerical only	0	1	0	119	45.8	0.753
Graphical only	0	0	1	6	2.3	0.256
Discussion + numerical	1	1	0	58	22.3	0.726
Discussion + graphical	1	0	1	14	5.4	0.658
Numerical + graphical	0	1	1	9	3.5	0.122
Discussion + numerical + graphical	1	1	1	5	1.9	0.493
Total				260	100.0	0.708

Note: Mean score is the average normalized instructor rating across the three raters. Takeaway: the lowest mean performance occurs for numerical–graphical items, but this cell is small, so the result should be interpreted with caution.

4.2. Inter-Rater Reliability

Table 4 reports pairwise correlations among the three instructors’ ratings. The correlations are high, but correlations alone do not establish absolute agreement. Table 5 therefore reports ICC estimates. The results indicate high single-rater consistency and very high average-rating reliability.

Table 4. Pairwise correlation coefficients.

Variable	G1	G2	G3	DIS	NUM	GPH
G1	1.000					
G2	0.826	1.000				
G3	0.745	0.949	1.000			
DIS	0.094	0.075	0.039	1.000		
NUM	−0.027	0.003	0.023	−0.518	1.000	
GPH	−0.376	−0.395	−0.355	0.056	−0.280	1.000

Note: The table reports correlations using normalized ratings. The revised manuscript treats these correlations as descriptive and supplements them with ICC estimates.

Table 5. Inter-rater reliability based on intraclass correlation coefficients.

Reliability Index	Estimate	Interpretation
ICC(C,1): single-rater consistency	0.841	High consistency
ICC(A,1): single-rater absolute agreement	0.829	High agreement
ICC(A,3): average-rating absolute agreement	0.936	Very high agreement

Note: ICC notation follows ShROUT and Fleiss (1979) and reporting guidance in Koo and Li (2016).

4.3. Pooled Rater–Item Evidence

Table 6 reports the mean differences implied by the mutually exclusive item-type classification, using numerical-only items as the reference category. Discussion-only items perform similarly to numerical-only items. Graphical-only items perform much worse, and numerical–graphical items perform worst of all. Discussion–graphical and all-three-modality items fall between these extremes.

Table 6. Pooled rater–item mean differences by item type.

Item Type	Mean Score	Difference from Numerical Only	Items
Numerical only	0.753	Reference	119
Discussion only	0.778	0.025	49
Graphical only	0.256	− 0.497	6
Discussion + numerical	0.726	−0.027	58
Discussion + graphical	0.658	−0.095	14
Numerical + graphical	0.122	− 0.631	9
Discussion + numerical + graphical	0.493	−0.260	5

Note: Statistical significance of less than 5% is indicated with **bold** font. The table is based on all 780 rater–item observations summarized at the item-type level. Numerical-only items are used as the reference because they form the largest single-modality technical category.

Table 7 confirms the pattern shown in the item-type means. Relative to numerical-only items, discussion-only items score slightly higher, by approximately 2.5 mark points, while discussion–numerical items score slightly lower, by approximately 2.7 mark points. In contrast, graphical-only items score 49.7 mark points lower than numerical-only items, and numerical–graphical items score 63.1 mark points lower. The all-three-modality items also score lower than numerical-only items, by approximately 26.0 mark points. These contrasts indicate that GPT-4o’s weakest performance in this benchmark occurs when graphical interpretation must be integrated with numerical calculation.

Table 7. Rater-adjusted descriptive contrasts by mutually exclusive item type.

Predictor	Coefficient	Mark-Point Difference	Predicted Mean Score
Numerical only	Reference	Reference	0.753
Discussion only	0.025	2.5	0.778
Graphical only	− 0.497	−49.7	0.256
Discussion + numerical	−0.027	−2.7	0.726
Discussion + graphical	−0.095	−9.5	0.658
Numerical + graphical	− 0.631	−63.1	0.122
Discussion + numerical + graphical	−0.260	−26.0	0.493
Rater fixed effect: Rater 2	0.059	5.9	–
Rater fixed effect: Rater 3	0.065	6.5	–

Note: Statistical significance of less than 5% is indicated with **bold** font. The dependent variable is the normalized instructor score from 0 to 1. The unit of analysis is the rater–item observation, with 780 observations from 260 items and three raters. Numerical-only items are the reference category. Rater fixed effects adjust for differences in rater severity. The estimates are used as descriptive benchmark contrasts rather than causal or population-inferential estimates, because item type was not randomly assigned and the corpus is purposive. The standard errors of coefficient estimates and supplementary replication script reproduces this table are available from the rating and coding file.

The rater fixed effects indicate that, relative to the first rater, the second and third raters assigned scores approximately 5.9 and 6.5 mark points higher, respectively. This difference in rater severity is accounted for in the pooled model. The qualitative pattern is unchanged when the analysis is summarized using graphical versus non-graphical items: graphical items receive an average score of 0.421, compared with 0.752 for non-graphical items, a difference of 33.1 mark points.

The revised regression logic has two merits. First, the analysis uses the full rater–item structure rather than estimating only separate instructor-level models. Second, interactions are not estimated without their main effects. The full interaction specification in Equation (2) retains *DIS*, *NUM*, *GPH* and all pairwise interactions. The qualitative conclusion in

Table 8 is consistent with Table 7: graphical items are associated with lower ratings, and the numerical–graphical combination is especially difficult under the GPT-4o protocol.

Table 8. Full main-effect and pairwise-interaction descriptive model.

Predictor	Coefficient	Mark-Point Difference
<i>DIS</i>	0.005	0.5
<i>NUM</i>	−0.020	−2.0
<i>GPH</i>	− 0.518	−51.8
<i>DIS</i> × <i>NUM</i>	−0.032	−3.2
<i>DIS</i> × <i>GPH</i>	0.397	39.7
<i>NUM</i> × <i>GPH</i>	− 0.114	−11.4

Note: Statistical significance of less than 5% is indicated with **bold** font. The dependent variable is the normalized instructor score from 0 to 1. Rater fixed effects are included. Because no item has *DIS* = 0, *NUM* = 0 and *GPH* = 0, individual coefficients in this interaction model should be interpreted through the implied cell means rather than as standalone causal effects. The implied numerical–graphical gap relative to numerical-only items is approximately −0.631, matching the preferred mutually exclusive item-type model.

The modest explanatory power of these models should be interpreted carefully. Item type is intentionally a coarse assessment-design predictor. GPT-4o performance also depends on topic, wording, diagram clarity, whether values must be extracted visually, the number of reasoning steps and prompt sensitivity. Thus, the models are not intended to predict every score, but to identify systematic performance differences across pedagogically meaningful item types.

5. Discussion

5.1. Mechanism: From “Visual Perception” to Multi-Modal Integration

The results reported in Section 4 should not be interpreted as showing that GPT-4o lacks vision. GPT-4o is multi-modal. The more plausible mechanism is that graphical economics items require a sequence of operations: parsing visual elements, reading axes and labels, identifying economically meaningful areas or shifts, extracting numerical or ordinal information, mapping the visual information to an economic model, performing calculations and explaining the implication. Errors can arise at any stage of this chain.

The especially low scores on numerical–graphical items suggest that the difficulty is not simply “seeing” a graph. Rather, the challenge is integrating visual extraction with quantitative and economic reasoning. For example, a model may correctly describe that a tax creates a wedge but fail to identify the relevant area for deadweight loss, or it may identify the correct area but compute it using incorrect base and height values. This mechanism is consistent with broader multi-modal benchmark evidence showing continuing weaknesses in chart, diagram and visual–mathematical reasoning tasks (Lu et al., 2024; Masry et al., 2022; Tong et al., 2024; Yue et al., 2024).

5.2. Implications for Assessment Validity, Constructive Alignment and Task Quality

The results have three specific implications for educational research and assessment design. First, from an assessment-validity perspective, the central issue is not whether GPT-4o can produce a high-scoring answer, but whether a problem-set score continues to support the intended inference about a student’s own economic understanding (Kane, 2013; Messick, 1995). If a task can be completed by a model without the student’s meaningful engagement, the score interpretation becomes weaker. Instructors should therefore design tasks that require visible reasoning processes, such as explaining why a welfare area is relevant, justifying the choice of formula or connecting a numerical result to an economic mechanism.

Second, from a constructive-alignment perspective, graphical or numerical–graphical items should be used when they directly correspond to intended learning outcomes, not merely because they appear harder for AI (Biggs et al., 2022). If the learning outcome is diagram interpretation, then students should be asked to identify curves, axes, shifts, areas and equilibrium changes. If the learning outcome is quantitative welfare analysis, then students should also show how values are extracted and how calculations follow from the economic model. The assessment format, teaching activities and scoring rubric should therefore be aligned around the same disciplinary competence.

Third, from the perspective of discipline-specific task quality in economics education, the findings support problem sets that require translation among verbal, numerical and graphical representations (Becker, 1997; Pieters, 2024). A higher-quality economics item might ask students to annotate a supply–demand diagram, compute a welfare measure and then explain the policy implication in words. Such tasks are not necessarily “AI-proof”, but they sample richer economic reasoning and provide instructors with more evidence about how students connect representations.

These implications should be implemented with attention to fairness and accessibility. Simply increasing graphical complexity may narrow construct coverage or disadvantage students with visual processing difficulties if accessible alternatives are not provided. A balanced assessment ecology should combine clear AI-use policies, formative feedback, transparent rubrics, randomized or contextualized parameters where appropriate and opportunities for students to explain their reasoning process (Boud & Falchikov, 2006; CAST, 2018; Dawson, 2020; Tai et al., 2018). In this sense, the findings are not a recipe for “AI-proofing” assessment. They are evidence for preserving validity, alignment and disciplinary task quality in a changing technological environment.

5.3. External Validity and Limitations

Several limitations should be emphasized. First, the study evaluates only GPT-4o. It should not be generalized to all GenAI systems, other multi-modal models or future versions without further evidence. Second, the data were generated between February and June 2025, and model capabilities change rapidly. Third, the question corpus is purposive and drawn from introductory economics materials used in the Hong Kong higher-education context. The findings may be most relevant to similar problem-set-based introductory economics courses rather than all economics assessments globally.

Fourth, graphical items form a small subsample. Only 34 items include graphical interpretation, and the numerical–graphical cell contains nine items. The magnitude of the graphical performance gap is substantively large, but future studies should replicate the analysis with larger and more systematically sampled graphical item banks. Fifth, the study evaluates model answers rather than student behavior. It does not directly measure whether students would misuse GenAI, whether they could improve prompts, or whether instructors could detect AI-generated work.

Finally, the results depend on the prompting and input-modality protocol documented in Appendix A. In this study, graphical items were presented through uploaded original diagram images in the ChatGPT web interface. The graphical performance gap should therefore be interpreted as applying to this image-upload protocol, not to text-only diagram descriptions, alternative multi-modal interfaces, API settings, multi-prompt tutoring interactions or future model versions.

A further limitation concerns item-type coding. The revised manuscript documents who coded the items, how borderline cases were resolved and where readers can inspect the final item-level labels. However, the original workflow was consensus-based rather than a prospectively designed blind double-coding procedure with intercoder-reliability coeffi-

cients. Future benchmark studies should preregister the coding codebook, use independent blind coders and report intercoder reliability for each item-type dimension.

6. Conclusions

This study benchmarks GPT-4o on introductory economics problem-set items using three independent instructor ratings per item. The revised manuscript treats the design as a rating-based benchmark rather than a quasi-natural experiment and analyzes the full rater-item structure of the data.

The main finding is that GPT-4o performs relatively well on discussion-only and numerical-only questions but performs substantially worse on graphical questions, especially when graphical interpretation must be integrated with numerical calculation. The results suggest that some item formats are less easily answered by GPT-4o under the documented image-upload protocol, but they do not show that graphical questions causally reduce academic misconduct.

For economics instructors, the implication is to design balanced, accessible and constructively aligned assessments that require students to demonstrate multiple forms of disciplinary reasoning. This recommendation is consistent with validity theory, which emphasizes the interpretation and use of assessment scores (Kane, 2013; Messick, 1995); constructive alignment, which requires assessment tasks to elicit intended learning outcomes (Biggs et al., 2022); and assessment-security scholarship, which emphasizes task design rather than detection alone (Boud & Falchikov, 2006; Dawson, 2020). Graphical and numerical-graphical items can play an important role when they assess intended economic skills, but they should be combined with clear AI-use policies, formative feedback, transparent rubrics and attention to accessibility (CAST, 2018; Tai et al., 2018). Future research should replicate this benchmark across additional multi-modal models, institutions, languages, disciplines and assessment formats and should include qualitative analysis of graphical-error mechanisms.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/educsci16070994/s1>, Table S1, anonymized item-level audit table with item IDs, topic areas, item-type labels, graph indicators, paraphrased item descriptions, coding rationales and rater scores; Data S1, long-format rater-item score file; File S1, GPT-4o output archive with copyrighted prompt wording and diagrams omitted where necessary; Code S1, replication script for generating the descriptive statistics, ICCs and rater-adjusted descriptive models; and File S2, README and data dictionary. Original copyrighted question wording and diagrams are not redistributed.

Author Contributions: Conceptualization, Y.L. and T.W.; methodology, Y.L., T.W. and T.W.C.; formal analysis, Y.L. and Y.C.; data curation, Y.L., T.W. and T.W.C.; writing—original draft preparation, Y.L.; writing—review and editing, Y.C.; project administration, Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Departmental Learning and Teaching Enhancement Grants of the corresponding author's university, The Hong Kong Polytechnic University, funding number: 88GW.

Institutional Review Board Statement: Ethical review and approval were not required for this study because the research analyzed AI-generated responses to textbook-style assessment items and anonymized instructor ratings, and did not collect student submissions, student-level personal data or interventions involving students.

Informed Consent Statement: Informed consent was obtained from the instructor raters for the use of their anonymized ratings in the research analysis. No student participants or student-level personal data were involved.

Data Availability Statement: The anonymized item-level metadata, item-type labels, paraphrased item descriptors, instructor ratings and analysis code are provided as Supplementary Materials. The original question wording and diagrams cannot be redistributed publicly because they are drawn from copyright-protected textbook and institutional materials accessed through institutional channels.

Acknowledgments: The authors thank Ziwen Zhang for editorial assistance and proofreading. During manuscript revision, the authors used generative AI tool GPT-5 for grammar checking, readability refinement and LaTeX formatting assistance. The authors reviewed and edited all AI-assisted output and take full responsibility for the content of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in the design of the study; in the collection, analysis or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
GenAI	Generative artificial intelligence
GPT	Generative pretrained transformer
ICC	Intraclass correlation coefficient
ILO	Intended learning outcome
LLM	Large language model
TLA	Teaching and learning activity

Appendix A. GPT-4o Prompting Protocol

Model: GPT-4o.

Data-generation window: February to June 2025.

Access route: ChatGPT web interface.

Session handling: Each item was submitted in a fresh chat session. No previous item content appeared in the conversation history for the item being tested.

Number of runs: One run per item. The first model response was retained. No regeneration or selection among alternative answers was used.

Temperature and decoding settings: The ChatGPT web interface was used; temperature and other decoding settings were not user-controllable by the authors.

Input modality for non-graphical items: The standardized instruction prompt was entered together with the item-specific question text.

Input modality for graphical items: The original diagram image was uploaded together with the item-specific question text. No manual annotation, correction or explanation of the diagram was added beyond the information contained in the original item.

Follow-up prompts: No follow-up prompts were used after the first model response.

Output handling: GPT-4o outputs were saved without editing and sent to instructors for rating.

Standard instruction prompt:

Please answer an introductory economics problem-set question. Provide a complete solution. Show the economic reasoning, define variables when needed, show calculations step by step and explain the final answer clearly. If a graph or figure is provided, use information from the graph or figure in your answer.

The item-specific question text and, where applicable, the uploaded image were inserted below this standard instruction.

Appendix B. Paraphrased Examples of Item Coding

Table A1. Paraphrased examples of coding rules.

Code	Paraphrased Item Form	Coding Rationale
DIS only	Explain how a price ceiling affects market outcomes.	Requires conceptual explanation but no calculation or graph extraction.
NUM only	Given linear demand and supply equations, solve for equilibrium price and quantity.	Requires algebraic calculation without a required diagram.
GPH only	Use a provided supply–demand diagram to identify the area of consumer surplus.	Requires interpreting a figure but no numerical calculation.
DIS + NUM	Calculate a tax revenue amount and explain economic implications.	Requires computation and written economic interpretation.
DIS + GPH	Interpret a diagram showing a policy change and explain the welfare implication.	Requires graphical interpretation and conceptual explanation.
NUM + GPH	Find values or areas from a graph and compute consumer surplus, producer surplus or deadweight loss.	Requires visual extraction and numerical calculation.
DIS + NUM + GPH	Use a graph to calculate a welfare change and explain the policy implication.	Requires all three forms of reasoning.

Note: Examples are paraphrased rather than reproduced verbatim because the original items and figures are copyright-protected. Source materials include standard introductory economics textbooks (Bade & Parkin, 2022; Mankiw, 2022).

Appendix C. Replication Package and Analysis Execution

The supplementary replication package contains the cleaned anonymized rating and coding files used to generate the tables in the manuscript. The main analysis file reshapes the data from wide item-level form to long rater–item form, excludes the single item with missing ratings from all three instructors, creates the mutually exclusive item-type categories, computes descriptive statistics, computes pairwise correlations and ICCs, and estimates the rater-adjusted descriptive models in Equations (1) and (2).

The replication files are organized as follows:

- `File_S1.pdf`: GPT-4o output archive with copyrighted prompt wording and diagrams omitted where necessary.
- `Table_S1.csv`: anonymized item ID, topic area, coding indicators, mutually exclusive item type, paraphrased item description, coding rationale and three instructor scores.
- `Data_S1.csv`: long-format rater–item file used for pooled analysis.
- `Code_S1.R`: script that reproduces the descriptive statistics, ICCs, item-type tables and rater-adjusted descriptive models.
- `File_S2_README.docx`: variable definitions, file structure and reproduction instructions.

Copyrighted source question wording and diagrams are not fully included in the public replication package. Instead, the package provides paraphrased item descriptions and coding rationales sufficient to understand the object of analysis without redistributing protected textbook material.

References

- Bade, R., & Parkin, M. (2022). *Foundations of economics* (9th ed.). Pearson.
- Becker, W. E. (1997). Teaching economics to undergraduates. *Journal of Economic Literature*, 35(4), 1347–1373.
- Biggs, J., Tang, C., & Kennedy, G. (2022). *Teaching for quality learning at university* (5th ed.). McGraw-Hill Education.
- Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education*, 31(4), 399–413. [CrossRef]
- Brown, P. C., Roediger, H. L., III, & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. Harvard University Press.
- CAST. (2018). *Universal design for learning guidelines version 2.2*. CAST. Available online: <https://udlguidelines.cast.org/more/downloads/#v2-2> (accessed on 20 July 2025).
- Chickering, A. W., & Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. *AAHE Bulletin*, 3, 7.
- Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239. [CrossRef]
- Dawson, P. (2020). *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*. Routledge.
- Eager, B., & Brunton, R. (2023). Prompting higher education towards AI-augmented teaching and learning practice. *Journal of University Teaching and Learning Practice*, 20(1), 1–19. [CrossRef]
- Gibbs, G., & Simpson, C. (2005). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, 3–31.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. [CrossRef]
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidtl, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. [CrossRef]
- Khalil, M., & Er, E. (2023). Will ChatGPT get you caught? Rethinking plagiarism detection. In *Proceedings of the international conference on human-computer interaction* (pp. 475–487). Springer.
- Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(4), 757–798. [CrossRef] [PubMed]
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. [CrossRef] [PubMed]
- Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K. W., Galley, M., & Gao, J. (2024, May 7–11). *MathVista: Evaluating mathematical reasoning of foundation models in visual contexts*. International Conference on Learning Representations, Vienna, Austria.
- Mankiw, N. G. (2022). *Principles of economics* (9th ed.). Cengage Learning.
- Masry, A., Long, D. X., Tan, J. Q., Joty, S., & Hoque, E. (2022). ChartQA: A benchmark for question answering about charts with visual and logical reasoning. *arXiv*. [CrossRef]
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. [CrossRef]
- Myers, S. C., Nelson, M. A., & Stratton, R. W. (2011). Assessment of the undergraduate economics major: A national survey. *The Journal of Economic Education*, 42(2), 195–199. [CrossRef]
- OpenAI. (2024). *GPT-4o system card*. OpenAI.
- Pieters, G. C. (2024). Designing effective assessments in economics courses: Guiding principles. *The Journal of Economic Education*, 55(1), 63–76. [CrossRef]
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. [CrossRef] [PubMed]
- Siegfried, J. J., & Fels, R. (1979). Research on teaching college economics: A survey. *Journal of Economic Literature*, 17(3), 923–969.
- Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, 76(3), 467–481. [CrossRef]
- Taylor, H. (1950). The teaching of undergraduate economics. *American Economic Review*, 40(1), i-226.
- Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., & Xie, S. (2024, June 17–21). *Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs*. IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9568–9578), Seattle, WA, USA.
- Watts, M., & Schaur, G. (2011). Teaching and assessment methods in undergraduate economics: A fourth national quinquennial survey. *The Journal of Economic Education*, 42(3), 294–309. [CrossRef]

- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltyniek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1), 1–39. [[CrossRef](#)]
- Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., ... Chen, W. (2024, June 17–21). MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA.

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