International Conference on Teaching and Learning Proceeding
Faculty of Education and Teacher Training – Universitas Terbuka
UTCC, South Tangerang, Banten, May 15th 2025
Vol. 1, No. 1, pg. 337 – 345
ISSN: 3046-594X

ASSESSING ECONOMIC ESSAYS WITH CHATGPT: SYSTEMATIC REVIEW AND PRELIMINARY DESIGN

Ramadzan Defitri Pratama[1*], Khresna Bayu Sangka[2]
[1] Department Master of Economic Education, Sebelas Maret University, Surakarta, Indonesia
email: ramadzandp@gmail.com

Abstract: This study explores the integration of ChatGPT in automated essay scoring (AES) through a systematic literature review (SLR) of 46 Scopus-indexed articles published between 2020 and 2025. The review identifies key trends in the use of large language models (LLMs), particularly GPT-3.5, GPT-4, and ChatGPT, for evaluating student essays across various domains. Findings show that GPT-based models deliver strong accuracy and reliability, especially when guided by structured rubrics and effective prompt designs. Despite these advantages, challenges remain in terms of scoring validity, model consistency, and generalizability across educational contexts. Based on the synthesis, this study proposes a preliminary rubric-guided AES model using ChatGPT. The model is designed to align essay scoring with pedagogical standards while offering scalable and transparent evaluation. This paper contributes by mapping current practices, identifying research gaps, and laying the groundwork for future development of standardized, domain-specific AES systems, particularly in economics education. Further empirical validation and field-based experimentation are recommended to ensure the model's pedagogical soundness and practical effectiveness.

Keywords: automated essay scoring; ChatGPT; rubric; prompting; education assessment

Accepted: 30 April 2025 Approved: 15 May 2025 Published: 01 July 2025

© 2025 FKIP Universitas Terbuka
This is an open access under the CC-BY license

INTRODUCTION

Essay writing assignments are foundational in education for developing critical thinking and communication skills. However, scoring essays manually is labor-intensive, subjective, and often inconsistent (Mendonça et al., 2025). Instructors face growing class sizes and limited time, making it challenging to provide timely feedback on open ended responses (Mendonça et al., 2025). These challenges have prompted exploration of automated essay scoring (AES) and other AI-based solutions. Recent advances in large language models (LLMs) and generative AI (e.g. GPT-3.5, GPT-4) offer promise for automating complex scoring tasks (Pack et al., 2024). For example, Pack et al., (2024) note that "advancements in generative AI, such as large language models (LLMs), may serve as a potential solution to the burdensome task of essay grading". Similarly, Mendonça et al., (2025) find that LLMs can evaluate student responses at scale, substantially reducing instructor workload and improving consistency. At the same time, the rapid adoption of ChatGPT in education has generated interest in its use for writing

Vol. 1, No. 1, pg. 337 – 345
ISSN: 3046-594X

instruction and assessment (Li, 2024).

Despite this promise, evidence on the efficacy of LLMs in essay evaluation is mixed. Mendonça et al., (2025) emphasize that, although LLMs can align closely with human grading and enhance consistency, their validity and reliability in diverse contexts remain under-explored (Pack et al., 2024). In science education, Latif & Zhai, (2024) demonstrated that fine-tuning GPT-3.5 on domain-specific tasks yielded a 9.1% increase in scoring accuracy over a BERT model. Gandolfi, (2025) reports that even GPT-4's grading can be "marred by occasional loss of coherence and hallucinations," reducing consistency, despite overall scores aligning with human raters. These findings highlight that LLMs can significantly streamline essay scoring and approach human-level reliability, but are also sensitive to prompt design and domain alignment. For instance, Poole & Coss, (2023) illustrate that ChatGPT's rubric-based scoring is "devilishly" dependent on how prompts are formulated, indicating a need for careful prompt engineering in LLM assessment (Poole & Coss, 2023).

By mapping themes such as model types, rubric integration, and prompting challenges, we aim to uncover trends and gaps in the literature. In particular, although some studies have begun evaluating LLMs with rubric-based criteria in language learning (Yavuz et al., 2025), we find no standardized ChatGPT-based assessment model in economics education. This gap motivates the present work. To bridge these gaps, we propose an initial rubric-guided ChatGPT assessment model. The rationale is that rubrics explicitly encode assessment criteria, which can be translated into structured prompts for ChatGPT. By integrating detailed rubrics into the prompt design as in (Yavuz et al., 2025), we aim to align the LLM's scoring with established learning objectives. This approach may combine the efficiency of AI scoring with the transparency of human-designed rubrics. In sum, this systematic review synthesizes recent findings on AI-powered essay evaluation (2020–2025) and outlines a preliminary rubric-prompt model. We highlight ongoing gaps and pave the way for future research on standardized ChatGPT-based assessment, especially in under explored fields like economics.

METHOD

This study was conducted using the Systematic Literature Review (SLR) approach following the PRISMA protocol to ensure the regularity and reproducibility of the literature selection process (Page et al., 2021). The focus of the study was formulated using the PICOS framework (for example, population is student essay writing, intervention in the form of implementing ChatGPT or AI in assessment, comparison with manual assessment, outcome in the form of accuracy and consistency of assessment, and study design related to learning evaluation). The literature search focused on the Scopus database, resulting in 2076 initial articles. After being limited to the publication period of 2020–2025, 1637 articles remained. Furthermore, the document category was narrowed down to only English-language "articles," leaving 846 articles. Further filtering stages based on final article, title, abstract, and keyword relevance resulted in 46 articles that were included in the final analysis.

International Conference on Teaching and Learning Proceeding
Faculty of Education and Teacher Training – Universitas Terbuka
UTCC, South Tangerang, Banten, May 15th 2025
Vol. 1, No. 1, pg. 337 – 345
ISSN: 3046-594X

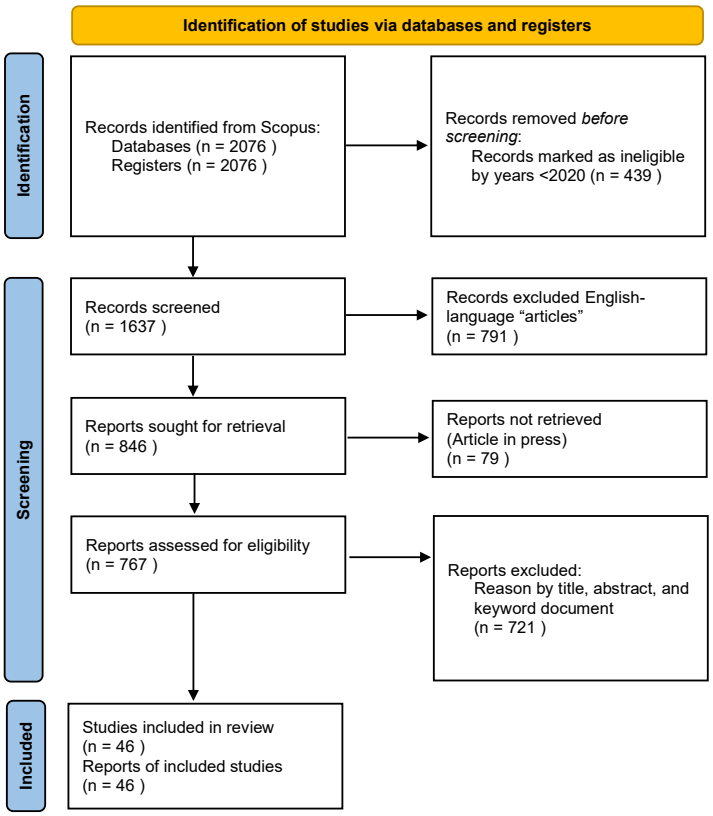

Table 1 Figure 1 Flow Diagram of This SLR

RESULT AND DISCUSSION

1. Trend Automated Essay Scoring (AES)

Analysis of publication trends is needed to understand how academic attention to this topic has developed in the last five years. The increase in the number of publications not only shows the intensity of research but can also reflect the dynamics of technology and the urgency of its application in the world of education. Thus, publication trends are an important early indicator to assess the relevance and momentum of this systematic study.

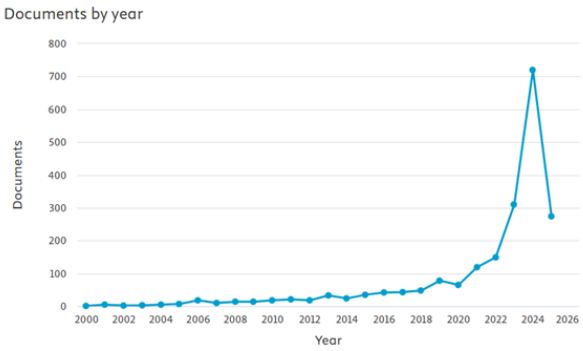

Figure 1 Graph of Publications Per Year

Vol. 1, No. 1, pg. 337 – 345
ISSN: 3046-594X

Based on the distribution graph of publications per year, there is a sharp increasing trend in the number of documents related to automated essay scoring (AES) since 2020. During the previous two decades (2000–2019), the number of publications was relatively low and stable, averaging less than 50 articles per year. However, since 2020, the number of publications has started to increase significantly and peaked in 2024 with more than 700 published documents. This increase reflects the growing academic and practical attention to the application of artificial intelligence technology in the essay evaluation process, especially with the development of large language models (LLMs) such as ChatGPT.

2. AI Models Used in Automated Essay Scoring

Various studies employ a range of AI models for automated essay scoring (AES). Dominant among them are generative large language models (LLMs). For instance, Lee et al., (2024) implemented GPT-3.5 and GPT-4 for evaluating student responses , while Tang et al., (2024) reported that GPT-4 showed the highest scoring accuracy compared to GPT-3.5 and Claude 2. ChatGPT based on GPT-3.5/GPT-4 was frequently used; Yavuz et al., (2025) tested ChatGPT (original and fine-tuned) alongside Bard, and Tate et al., (2024) compared ChatGPT scores with those of human raters . Other LLMs such as Google PaLM 2 and Anthropic Claude 2 also appeared in AES evaluations (Pack et al., 2024). Additional research used traditional embedding models, such as RoBERTa embeddings combined with XGBoost (Faseeh et al., 2024)) and SBERT integrated with LSTM-Attention (Nie, 2025). For image-based data (e.g., handwritten answers), a one dimensional convolutional neural network (1D-CNN) architecture was applied (Grecea Pasaribu et al., 2024).

Table 2 AI Model Used

AI Model	Frequency of Use	Example Studies (Authors, Year)
GPT-3 (text-davinci-003)	12	Mizumoto & Eguchi (2023)
GPT-3.5 / ChatGPT	6	Mizumoto & Eguchi, (2023) ; Yavuz et al., (2025) ; Tate et al., (2024)
GPT-4	6	Lee et al., (2024) ; Tang et al., (2024) ; Pack et al., (2024)
Claude 2	6	Tang et al., (2024)
Google Bard	7	Yavuz et al., (2025)
Google PaLM 2	2	Pack et al., (2024)
RoBERTa Embedding + XGBoost	2	Faseeh et al., (2024)
SBERT + LSTM-Attention	2	Nie, (2025)
1D CNN	3	Grecea Pasaribu et al., (2024)

GPT-based models dominate due to their superior scoring performance. For example, Lee et al., (2024) and Tang et al., (2024) found GPT-4 consistently yielded more accurate and reliable scoring than GPT-3.5. Hybrid models (e.g., SBERT+LSTM or RoBERTa+XGBoost) also show promising results, but GPT-based LLMs remain the

Vol. 1, No. 1, pg. 337 – 345
ISSN: 3046-594X

most widely used in recent AES literature.

3. Most Discussed Aspects

- **Effectiveness.** Many studies evaluate effectiveness through metrics like accuracy or quadratic weighted kappa (QWK). Lee et al., (2024) found that few-shot prompting achieved higher accuracy (0.67) than zero-shot (0.60), a +12.6% gain . Tang et al., (2024) observed up to +112% scoring accuracy improvement in GPT-4 using rubric-based prompts. Hybrid approaches such as RoBERTa+XGBoost reached QWK scores of 0.941 Faseeh et al., (2024) , while SBERT+LSTM Attention improved performance consistency Nie, (2025). Overall, these models can produce scoring accuracy close to human raters (Mizumoto & Eguchi, 2023).
- **Reliability.** Yavuz et al., (2025) found that fine-tuned ChatGPT achieved an ICC of 0.972, compared to 0.947 for the default model and 0.919 for Bard. Pack et al.,(2024) showed GPT-4 had the highest intrarater reliability and construct validity among four popular LLMs . Prompt engineering significantly enhanced reliability, as seen in Tang et al., (2024) where GPT-4 improved +114% under rubric-based prompts.
- **Efficiency.** Several studies highlight the efficiency gains of AI-based AES. (Grecea Pasaribu et al., 2024) reported 1D-CNN achieved ~81.2% accuracy while reducing grading workload and subjectivity. Jung et al., (2024) combined machine translation and neural networks to assess multilingual essays with comparable performance to human raters. Yavuz et al., (2025) also noted AI could "reduce educators' workload" by automating essay assessment.
- **Prompting Techniques.** Prompting strategies vary across studies. Few-shot prompting outperforms zero-shot, as shown by Lee et al. (2024), who reported 0.67 accuracy (few-shot) vs 0.60 (zero-shot). CoT prompting alone showed limited benefits, but significantly improved performance when combined with item stems and rubrics. Tang et al. (2024) observed GPT-4 improved by +112% in accuracy using criterion-based prompts. Amin et al. (2025) used domain-specific fine-tuning and few-shot strategies, also reporting strong gains in QWK scores. In general, structured prompts especially when paired with rubrics consistently enhanced AES performance.
- **Rubric Integration.** Rubric use improves performance. Lee et al., (2024), found that integrating scoring rubrics with Chain-of-Thought prompts significantly enhanced scoring accuracy (e.g., +13.44% for zero-shot). Yavuz et al., (2025) used a five-dimension analytic rubric to guide scoring and achieved high consistency (ICC 0.972). In code-based tasks, detailed rubrics aligned GPT-4's scores closely with human ratings (Pack et al., 2024). These findings suggest rubrics serve as effective scaffolding for improving accuracy, consistency, and scoring transparency.

4. Challenges and Opportunities

Despite promising results, several gaps remain. Tang et al., (2024) recommended expanding evaluations to diverse genres and participant pools to test model generalizability. Pack et al., (2024) noted performance variability across models and testing conditions . Yavuz et al., (2025) also found ChatGPT struggled with non-objective assessment criteria . Fleckenstein et al., (2024) raised concerns about academic integrity, reporting educators often fail to detect AI-generated text . Steiss et al., (2024) noted AI-provided feedback was less creative than that from experienced teachers.

Nonetheless, the consistency and reliability of GPT-based AES especially GPT-4 and fine-tuned ChatGPT—offer opportunities for model development. Tate et al., (2024) and Yavuz et al., (2025) both confirmed AI scoring approached human performance. A promising direction involves

Vol. 1, No. 1, pg. 337 – 345
ISSN: 3046-594X

designing domain-specific scoring models using ChatGPT with rubric-informed prompt engineering. Further, rapid feedback generation using LLMs may alleviate instructors' grading burden (Steiss et al., 2024; Tate et al., 2024). In conclusion, while more research is needed, the reviewed studies strongly support developing a preliminary AES model based on ChatGPT with integrated rubrics and prompting to enhance scoring quality and fairness.

5. Preliminary Design of the Automated Essay Assessment System

Based on findings from the systematic literature review (SLR), a preliminary design for an automated essay assessment system using ChatGPT has been developed. This early-stage design aims to bridge pedagogical demands with the capabilities of large language models (LLMs), particularly in the context of evaluating higher-order thinking skills (HOTS) in economics education.

5.1 System Architecture

The system is designed as a modular pipeline that includes five major components:

1. Essay Input Module: Receives essay prompts and student responses.
2. Prompt Engineering Module: Formats structured prompts for ChatGPT based on a standardized scoring rubric.
3. ChatGPT API Connector: Sends prompt and receives scored responses and feedback using the OpenAI GPT API.
4. Response Parser: Extracts numerical scores and qualitative feedback from the output.
5. Result Display Interface: Presents results to teachers and students in a readable format.

This architecture allows real-time, rubric-based evaluation aligned with the principles of formative assessment in economics education.

5.2 Essay Question Example and Assessment Rubric

Topic: Inflation
Question:
Explain the main causes of inflation in Indonesia and how it impacts people's purchasing power. Include real examples and solutions that can be taken by the government.

Table 3 Assessment Criteria

Assessment criteria	Analysis and Assessment	Score
1. Mastery of Economic Concepts	Demonstrates a basic understanding of inflation, its relationship to fuel, and purchasing power. However, there is no in-depth explanation of the types of inflation or price transmission mechanisms.	3
2. Argumentation and Reasoning	The argument is logical and coherent, although not very deep. The relationships between ideas are still quite simple.	3
3. Relevance and Quality of Examples	The actual examples (chili, rice, subsidies) are quite relevant but have not been discussed quantitatively or in great detail.	3
4. Structure and Language	The language is communicative and fairly structured. There is coherence between ideas, although not very academic.	1
Total Score		**10**

5.3 Automated Essay Scoring Prompt Design

In order to facilitate automated scoring of student essays in the subject of economics, we designed a structured natural language prompt that guides the Large Language Model

(ChatGPT) to assess student responses in alignment with pre-defined rubric criteria. The prompt is designed to be both human readable and machine processable, making it suitable for integration within an educational assessment tool.

A sample scoring prompt is presented below:

*You are an economics teacher evaluating a student's short essay. Use the following rubric to assign scores (0–10) and provide brief feedback (max. 100 words). The criteria are:

1. Conceptual Understanding (0–3)
2. Reasoning and Argumentation (0–3)
3. Relevance of Examples (0–3)
4. Structureand Language (0–1)

Then output a scoring table and feedback.*

Question:

" Explain the main causes of inflation in Indonesia and how it impacts people's purchasing power. Include real examples and solutions that can be taken by the government."

Student Answer:

"[Insert student response here]"

This prompting structure ensures that the model aligns with instructional goals, promotes transparency, and supports formative feedback. It also enables consistent evaluation across multiple responses and question variations, especially when embedded in a semi-automated assessment environment.

The proposed preliminary design, encompassing structured prompts, rubric-based scoring components, and the integration framework with ChatGPT, serves as a foundational step toward developing a functional automated essay assessment system for economics education. However, to ensure its pedagogical soundness, practical feasibility, and scoring reliability, this design must undergo empirical testing through field trials and be subjected to expert validation by assessment specialists. These steps are crucial to refine the system before its broader implementation in educational settings.

CONCLUSION

The results of this systematic review show that the use of large language models (LLMs) such as ChatGPT in automated essay scoring has grown rapidly in the past five years, with results showing high potential in terms of effectiveness, reliability, and efficiency. The GPT-4 and ChatGPT models performed very close to human raters, especially when combined with appropriate scoring rubrics and prompting strategies. The reviewed studies also emphasize the importance of integrating explicit rubrics and domain-based prompt design to produce transparent, consistent, and reliable scoring. However, challenges related to validity, performance variation across models, and generalizability remain. As a follow-up, development research needs to be conducted to design and test a structured ChatGPT-based automatic essay assessment model, including an assessment rubric tailored to the characteristics of academic tasks in the field of economics education. Empirical validation and field trials are also important to ensure the pedagogical suitability and reliability of the system before widespread implementation. Collaboration with educators and evaluation experts is needed so that the system design is not only technically accurate but also educationally valid.

REFERENCES

Faseeh, M., Jaleel, A., Iqbal, N., Ghani, A., Abdusalomov, A., Mehmood, A., & Cho, Y.-I. (2024). Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy. *Mathematics*, *12*(21), 3416. https://doi.org/10.3390/math12213416

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence*, *6*, 100209. https://doi.org/10.1016/j.caeai.2024.100209

Gandolfi, A. (2025). GPT-4 in Education: Evaluating Aptness, Reliability, and Loss of Coherence in Solving Calculus Problems and Grading Submissions. *International Journal of Artificial Intelligence in Education*, *35*(1), 367–397. https://doi.org/10.1007/s40593-024-00403-3

Grecea Pasaribu, N., Budiman, G., & Dyah Irawati, I. (2024). Auto Evaluation for Essay Assessment Using a 1D Convolutional Neural Network. *IEEE Access*, *12*, 188217–188230. https://doi.org/10.1109/ACCESS.2024.3515837

Jung, J. Y., Tyack, L., & von Davier, M. (2024). Combining machine translation and automated scoring in international large-scale assessments. *Large-Scale Assessments in Education*, *12*(1), 10. https://doi.org/10.1186/s40536-024-00199-7

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210

Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100213. https://doi.org/10.1016/j.caeai.2024.100213

Li, M. (2024). Leveraging ChatGPT for Second Language Writing Feedback and Assessment. *International Journal of Computer-Assisted Language Learning and Teaching*, *14*(1), 1–11. https://doi.org/10.4018/IJCALLT.360382

Mendonça, P. C., Quintal, F., & Mendonça, F. (2025). Evaluating LLMs for Automated Scoring in Formative Assessments. *Applied Sciences*, *15*(5), 2787. https://doi.org/10.3390/app15052787

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Nie, Y. (2025). Automated essay scoring with SBERT embeddings and LSTM-Attention networks. *PeerJ Computer Science*, *11*, e2634. https://doi.org/10.7717/peerj-cs.2634

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, *6*, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an

Vol. 1, No. 1, pg. 337 – 345
ISSN: 3046-594X

updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

Poole, F. J., & Coss, M. (2023). *Can ChatGPT Reliably and Accurately Apply a Rubric to L2 Writing Assessments? The Devil is in the Prompt(s)*. https://doi.org/10.35542/osf.io/3r2zb

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Tang, X., Chen, H., Lin, D., & Li, K. (2024). Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, *10*(14), e34262. https://doi.org/10.1016/j.heliyon.2024.e34262

Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence*, *7*, 100255. https://doi.org/10.1016/j.caeai.2024.100255

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, *56*(1), 150–166. https://doi.org/10.1111/bjet.13494