DOI: https://doi.org/10.36094/sc.v92.2026.Comparative_Evaluation_of_Large_Language_Models.Paul.662

ARTICLE

Sci. and Cult. 92 (9–10) : 662-666 (2026)

COMPARATIVE EVALUATION OF LARGE LANGUAGE MODELS FOR PEDAGOGICAL QUESTION GENERATION USING AN ENHANCED FRAMEWORK

ATANU KUMAR PAUL[1], KINGSHUK CHATTERJEE[2], PARAMITA DEY[1], ARCHANA MONDAL[3], RYAN SAPTARSHI RAY[4], DEEPAYAN DAS[5]* AND MEENAKSHI ACHARYA[4]

Automated question generation using Large Language Models (LLMs) is transforming modern education, but maintaining pedagogical quality, cognitive alignment, and relevance remains challenging. This paper presents an enhanced Predict_Score framework that combines NLP, semantic similarity, and Bloom's Taxonomy to evaluate question quality. It assesses relevance, Bloom level accuracy, mark alignment, and pedagogical value, while comparing multiple LLMs using structured metrics. Results show that different models excel in distinct areas, such as structured generation or higher-order reasoning, offering an interpretable and scalable solution for automated assessment.