Conference Name: EduCon Singapore â€“ International Conference on Education, 05-06 July 2026
Conference Dates: 05-Jul-2026 to 06-Jul-2026
Conference Venue: The National University of Singapore Society (NUSS) The Graduate Club, Suntec City Guild House, 3 Temasek Boulevard (Tower 5), #02- 401/402 Suntec City Mall, Singapore
Appears in: PUPIL: International Journal of Teaching, Education and Learning (ISSN 2457-0648)
Publication year: 2026

Huilin Chen, 2026

Volume 2026, pp. 221-222

DOI- https://doi.org/10.20319/ictel.2026.221-222

This paper can be cited as: Chen, H. (2026). Comparative Evaluation of Three Llms for CEFR- Aligned Reading Passage and Item Generation. EduCon Singapore â€“ International Conference on Education, 05-06 July 2026. Proceedings of Teaching & Education Research Association (TERA), 2026, 221-222

COMPARATIVE EVALUATION OF THREE LLMS FOR CEFR-ALIGNED READING PASSAGE AND ITEM GENERATION

Huilin Chen
School of Education, Shanghai International Studies University, China
chlwilliam@hotmail.com

Abstract

Research Objectives: *This study compares three large language models (Deepseek, GPT-4, Baidu Ernie 4.0) in automatically generating CEFR A1–C2 reading comprehension passages and test items.*

Methodology: *Each model produced 72 passages (narrative, expository, argumentative, instructional) and 120 four-option multiple-choice items covering nine CEFR skill descriptors. Five certified language assessment experts independently rated all materials on nine quality dimensions using 5-point Likert scales. Inter-rater reliability was excellent (Fleiss' κ = 0.82–0.91). Data were analysed with Kruskal-Wallis H tests and post-hoc Dunn–Bonferroni corrections.*

Findings: *Passage and item quality were strong up to B2 level ($M > 4.2/5.0$ across models), but declined notably at C1–C2 ($M = 3.08–3.78$). Higher-order skills (e.g., rhetorical purpose evaluation, implicit attitude inference) scored significantly lower ($p < .001$). Deepseek outperformed GPT-4 on CEFR alignment ($p = .002$) yet remained 0.5–0.7 points below human-authored benchmarks.*

Research Outcomes: *Current LLMs can reliably generate psychometrically sound reading materials up to B2 level, but remain inadequate for fully automated high-stakes C1–C2 assessment without intensive human post-editing.*

Future Scope: *Enhance LLM capabilities for rhetorical complexity, cohesion manipulation, and higher-order item design targeting advanced proficiency levels.*

Keywords:
Large Language Models, Automatic Item Generation, Reading Comprehension, CEFR, Higher-Order Skills