Reducing Burnout with AI-assisted Medical Question Writing: Time Saver or Troublemaker

Aditya Singh, Michael JH McCarthy, Sarju Patel,
Andreia de Almeida, Holly Keeling

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD
School of Medicine
Ysgol Meddygaeth

GIG CYMRU NHS WALES
Bwrdd Iechyd Prifysgol Caerdydd a'r Fro
Cardiff and Vale University Health Board

BACKGROUND

Artificial Intelligence (AI) is rapidly growing in medical education, with large language models (LLMs) demonstrating capability in passing the United Kingdom Medical Licensing Exam (UKMLA), generating examination-style questions and aiding medical documentation, though concerns regarding its reliability persist[1,2,3].

UK medical educators face rising workload pressures beyond COVID-19. Increasing regulatory and administrative demands, assessment reforms such as the UKMLA, staff shortages, and greater student support needs have intensified the strain[4-7]. These pressures have led to interest in AI as a tool to streamline assessment design and reduce educator burnout [8].

Medical students are also adopting AI, with more than half of UK undergraduates reporting use, though risks of inaccuracy and overreliance remain[9,10]. Internationally, AI is being piloted in assessment design and simulation, but oversight and expert review remain essential [11].

AIM

Our study aims to evaluate the accuracy of four LLMs (ChatGPT, Gemini, Copilot and Deepseek) in producing medical school-style questions from official spinal pathology lectures.

METHODOLOGY

3 spinal pathology lecture slides + transcript + notes

→ Cardiff Medical School's SBA-writing Guide

Input into each LLM to produce 10 SBAs

→ Quality checked by a spine expert. Categorised into 'Good, Ok, Poor'

Descriptive statistical analysis using SPSS 29.0.2.0

Quality checks were conducted by a consultant spinal surgeon at the University Hospital of Wales (UHW) in a blinded manner.

RESULTS

	GPT	CP	DS	Gem
Poor	0	13	3	1
Ok	6	7	12	5
Good	24	10	15	24

A Friedman test compared correctness ratings among the models. There was a statistically significant difference between the LLMs, $\chi^2 = 29.71$, $p < .001$. Mean ranks suggested Gemini (2.93) and GPT (2.97) outperformed Deepseek (2.35) and Copilot (1.75).

All LLMs produced 10 SBAs in under 3 minutes on average per lecture.

DISCUSSIONS & CONCLUSION

AI offers potential to reduce administrative and assessment burdens, an important consideration amid rising workload and pressures of UKMLA alignment. For learners, AI-generated practice questions may provide scalable, accessible preparation tools that complement traditional resources, easing exam stress and indirectly reducing demand on faculty support.

Factual errors and lower-order cognitive focus remain limitations, reinforcing the need for expert oversight. Thus, AI-assisted question writing shows promise as a dual utility.

For References: