Evaluation of Neural Consult's AI-Powered Board Question Generator Compared to Published NBME Step 2 CK Questions

Jonathan Theros

Background

Neural Consult, an AI-driven educational platform developed by Dendritic Health AI, offers custom large language models designed to generate USMLE-style multiple-choice board questions. With over 20,000 medical student users worldwide, the platform delivers personalized, adaptive learning content, including explanations, flashcards, clinical case simulation, and board-style questions. This study aimed to evaluate the quality and correctness of AI-generated board questions in comparison to officially published NBME Step 2 CK questions.

Methods

A random sample of multiple-choice questions were collected from two sources: questions generated by Neural Consult's AI platform and publicly available published NBME/USMLE Step 2 CK questions. Each question included a clinical stem, answer choices, and an accompanying explanation.

Five reviewers were recruited: two physicians (one attending and one resident) and three medical students who had previously passed Step 2 CK. Reviewers were blinded to question origin and randomized to receive sets of 30–40 questions. Each question was rated on a 5-point Likert scale (1 = poor, 5 = excellent) for question stem quality and explanation clarity.

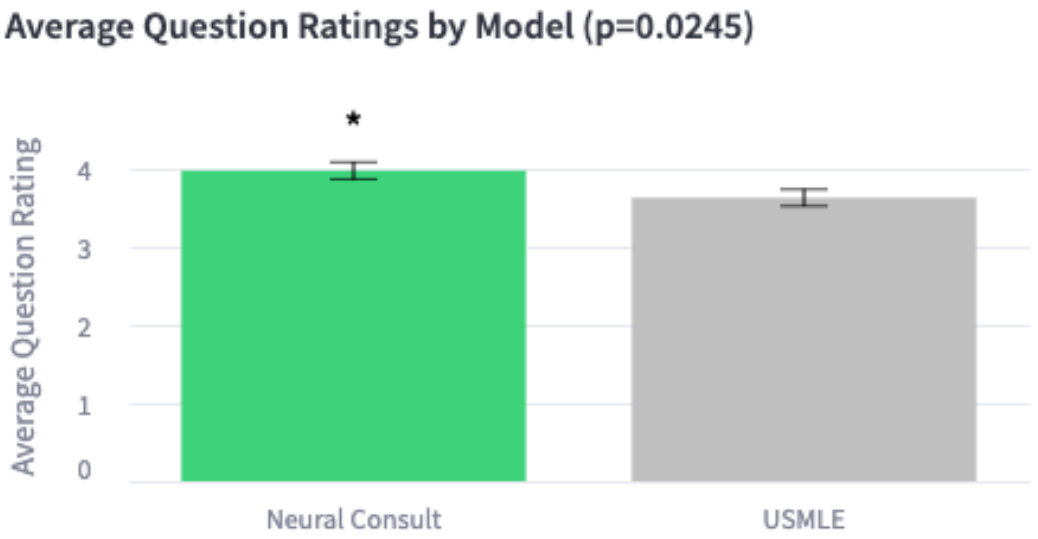
Reviewers could also flag questions for critical errors, defined as factual inaccuracy, two or more correct answers, or misleading construction. Flagged questions were independently reviewed by a group of two evaluators to confirm or reject final error status.

Results

A total of 163 questions were reviewed (77 AI-generated, 86 NBME). Initial screening flagged 3 AI questions (3.9%) and 2 NBME-written questions (2.3%) for possible critical errors. Upon flagged question re-review, 0 questions were confirmed to contain critical issues (0 AI, 0 NBME/USMLE), and thus, were excluded from analysis. Final analysis included 74 AI questions and 84 NBME questions.

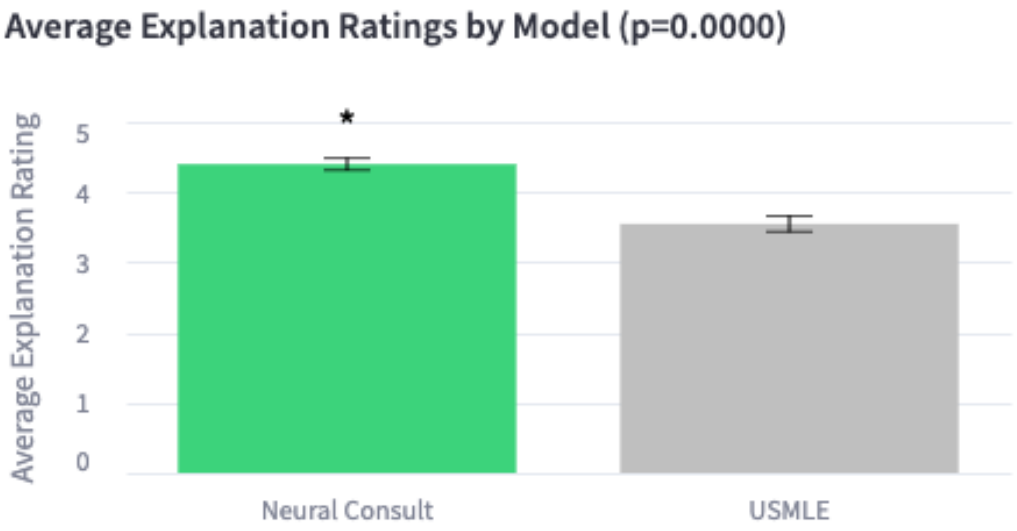
Mean question stem quality score:

- **Neural Consult AI**: 4.0 (SD = 0.99)
- **NBME/USMLE**: 3.64 (SD = 0.92)
- **p = 0.025**



Mean question explanation quality score:

- **Neural Consult AI**: 4.4 (SD = 0.78)
- **NBME/USMLE**: 3.5 (SD = 0.96)
- **p = 0.0000**



Conclusion

Neural Consult's AI-generated board questions demonstrated equal or superior performance in both question stem quality and explanation clarity when compared with publicly available NBME questions. The AI questions were non-inferior in perceived correctness and demonstrated a 0% error rate upon re-review in this sample. These findings support the validity and indicate educational utility of Neural Consult's AI in generating high-quality, clinically relevant board-prep content. Future directions include rigorous, independent, peer-reviewed analysis of more AI generated questions and other site features.