

ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: A comparative pilot study

Authors: Gautham Arun, Vivek Perumal, Francis Paul John Bato Urias, Yan En Ler, Bryan Wen Tao Tan, Ranganath Vallabhajosyula, Emmanuel Tan, Olivia Ng, Kian Bee Ng, Sreenivasulu Reddy Mogali

Presenters: Justin Chia, Lo Pei Qi, Dhruv Praveen

INTRODUCTION

Problem: Recent advancements in Artificial Intelligence (AI), especially conversational models like ChatGPT, are reshaping teaching and learning in health professions education. However, there are reliability, validity and ethical concerns with ChatGPT-based outputs, which will significantly affect students' anatomical knowledge.

Literature Gap: Currently, there is no application AI chatbot with accurate anatomy output for medical students, highlighting the necessity for tailored educational tools in anatomy

Purpose: To investigate whether Anatbuddy, a chatbot programmed using a Retrieval Augmented Generation (RAG) method on peer-reviewed learning resources, provides more accurate and high-quality information than ChatGPT 3.5.

METHODS

In this comparative pilot study, Anatbuddy, was developed using OpenAI's API in a secure cloud environment, leveraging the ChatGPT 3.5 model within the LangChain framework. Anatbuddy employs RAG, drawing information from curated knowledge base and feeding them into 2 language models to provide accurate responses. (Figure 1)

In order to evaluate Anatbuddy, 18 prompts covering thoracic anatomy were constructed according to Bloom's taxonomy, ranging from simple recall to complex application questions.

All the questions were fed into both ChatGPT 3.5 and Anatbuddy, and their responses were analyzed by three expert anatomists using a 5-point Likert scale to evaluate output accuracy and quality based on predefined criteria: factual accuracy, relevance, completeness, coherence, and fluency.

A panel discussion was conducted to score the outputs, with color coding highlighting errors. An independent expert later assessed the scoring for objectivity. Statistical analyses utilized SPSS version 29, using descriptive statistics and the Mann-Whitney U test for significance at alpha = 0.05, with Cohen's d calculated for effect size evaluation.

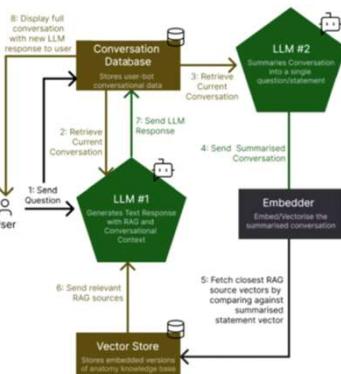
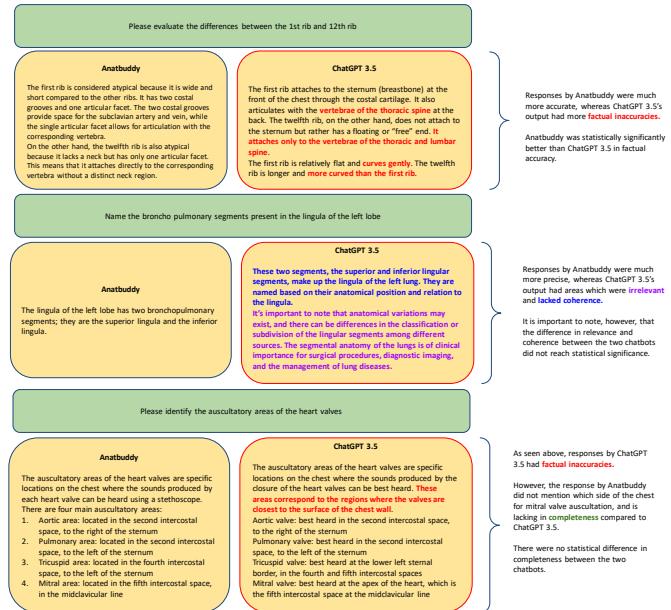


Figure 1. Architecture of Anatbuddy

RESULTS



DISCUSSION

Quality of training material

- Anatbuddy's knowledge base has a deficiency with regards to the topic of cardiac innervation; linked to quality of training data.
- Outputs are only as good as the quality and volume of their training material.

Novice learners

- The quality of AI outputs is limited by the user's ability to create relevant prompts; operator dependency limits output standardization.
- AI generative outputs are coherent and fluent, and users may be swayed to believe in its integrity even when outputs may be factually incorrect. Users need to maintain a certain degree of knowledge competency to discern fact from fiction.

Discipline Specific & Bespoke Chatbots

- Whilst outputs can be accurate and relatively well-organized, perceived relevance and acceptable detail levels vary depending on user expertise.
- Future advancements may enable discipline-specific customization and training-level adjustments.
- Need to mitigate model temperature to account for response contextuality.

FUTURE WORK

Visual Inputs

- Current chatbots cannot interpret or generate high-quality anatomical images, thus limiting their teaching potential. Implementation of the same (such as DALL-E) would improve application of anatomy to real-world scenarios.

Integration of Improved Learning Datasets

- Broader scope of learning resources are required to construct a more comprehensive chatbot that can cover a broader range of topics, with more body systems as well as histology and embryology.
- Most authoritative materials tend to be well-regarded textbooks that are copyright-protected.
- Need to explore use of institutional licenses for such applications or even creation of improved open-source resources to improve accessibility.

Table 1. Comparative Likert Scale evaluation of Anatbuddy vs ChatGPT 3.5

Criteria	ChatGPT 3.5 mean (\pm SD); median	Anatbuddy mean (\pm SD); median	Significance	Effect size (Cohen's d)
Factual accuracy	4.11 (\pm 0.83); 4.0	4.78 (\pm 0.43); 5.0	0.01	1.01
Relevance	4.50 (\pm 0.62); 5.0	4.39 (\pm 0.70); 4.5	0.70	0.17
Completeness	4.72 (\pm 0.46); 5.0	4.33 (\pm 0.69); 4.0	0.12	0.67
Coherence	4.56 (\pm 0.51); 5.0	4.78 (\pm 0.43); 5.0	0.27	0.47
Fluency	5.00 (\pm 0.00); 5.0	5.00 (\pm 0.00); 5.0	NA	NA

Factual accuracy

- Overall, Anatbuddy outperformed in providing factually accurate information compared to ChatGPT 3.5
- Comparable performance on application questions
- However, Anatbuddy was better than ChatGPT 3.5 at
 - Analytical** questions requiring higher-order thinking prompts
 - Recall-level** questions

No statistically significant differences detected for relevance, completeness, coherence, and fluency

However, panel discussion on coherence showed that:

- Anatbuddy's outputs were more **precise** and **specific** to users' queries, with minimal content repetition
- ChatGPT 3.5's output had extra details that were often deemed redundant.



Figure 2. Exemplar response of Anatbuddy interface