

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.

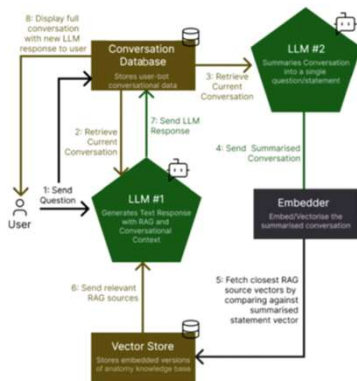


Figure 1. Architecture of Anatbuddy

RESULTS

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

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

Factual accuracy

- Overall, Anatbuddy outperformed in providing factually accurate information compared to ChatGPT3.5
- 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

RESULTS

Please evaluate the differences between the 1st rib and 12th rib

Anatbuddy

The first rib is considered atypical because it is wide and short compared to the other ribs. It has two costal grooves and one articular facet. The two costal grooves provide space for the subclavian artery and vein, while the single articular facet allows for articulation with the corresponding vertebra.

On the other hand, the twelfth rib is also atypical because it lacks a neck but has only one articular facet. This means that it attaches directly to the corresponding vertebra without a distinct neck region.

ChatGPT 3.5

The first rib attaches to the sternum (breastbone) at the front of the chest through the costal cartilage. It also articulates with the **vertebrae of the thoracic spine** at the back. The twelfth rib, on the other hand, does not attach to the sternum but rather has a 'floating' or 'free' end. It **attaches only to the vertebrae of the thoracic and lumbar spine**.

The first rib is relatively flat and **curves gently**. The twelfth rib is longer and **more curved than the first rib**.

Responses by Anatbuddy were much more accurate, whereas ChatGPT 3.5's output had more **factual inaccuracies**. Anatbuddy was statistically significantly better than ChatGPT 3.5 in factual accuracy.

Name the broncho pulmonary segments present in the lingula of the left lobe

Anatbuddy

The lingula of the left lobe has two bronchopulmonary segments; they are the superior lingula and the inferior lingula.

ChatGPT 3.5

These two segments, the **superior and inferior lingular segments**, make up the lingula of the left lung. They are named based on their anatomical position and relation to the lingula.

It's important to note that anatomical variations may exist, and there can be differences in the classification or subdivision of the lingular segments among different sources. The segmental anatomy of the lungs is of clinical importance for surgical procedures, diagnostic imaging, and the management of lung diseases.

Responses by Anatbuddy were much more accurate, whereas ChatGPT 3.5's output had areas which were **irrelevant and lacked coherence**. It is important to note, however, that the difference in relevance and coherence between the two chatbots did not reach statistical significance.

Please identify the auscultatory areas of the heart valves

Anatbuddy

The auscultatory areas of the heart valves are specific locations on the chest where the sounds produced by each heart valve can be heard using a stethoscope. There are four main auscultatory areas:

- Aortic area: located in the second intercostal space, to the right of the sternum
- Pulmonary area: located in the second intercostal space, to the left of the sternum
- Tricuspid area: located in the fourth intercostal space, to the left of the sternum
- Mitral area: located in the fifth intercostal space, in the midclavicular line

ChatGPT 3.5

The auscultatory areas of the heart valves are specific locations on the chest where the sounds produced by the closure of the heart valves can be best heard. **These areas correspond to the regions where the valves are closest to the surface of the chest wall.**

Aortic valve: best heard in the second intercostal space, to the right of the sternum

Pulmonary valve: best heard in the second intercostal space, to the left of the sternum

Tricuspid valve: best heard at the lower left sternal border, in the fourth and fifth intercostal spaces

Mitral valve: best heard at the apex of the heart, which is the fifth intercostal space at the midclavicular line

As seen above, responses by ChatGPT 3.5 had **factual inaccuracies**. However, the response by Anatbuddy did not mention which side of the chest for mitral valve auscultation, and is lacking in completeness compared to ChatGPT 3.5. There were no statistical difference in completeness between the two chatbots.

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.