AI vs Humans: predicting PE based on CTPA requests

Dr Sheikh Ruhul Amin, Dr John Ferrier
York and Scarborough Teaching Hospitals NHS Foundation Trust

NHS York and Scarborough Teaching Hospitals NHS Foundation Trust

Purpose

To compare two large language models (LLMs) against a group of human experts on their ability to predict the likelihood of a computed tomography pulmonary angiography (CTPA) scan being positive for a pulmonary embolism (PE), based on the clinical details of the scan request.

Methods and materials

A sample of 25 CTPA requests were selected of which 5 were positive for a PE on subsequent imaging. Two LLMs (Gemini Advanced and GPT-4) were compared against 31 human experts consisting of consultants, radiology trainees and radiographers. They were asked to select the 5 requests most likely to have had a PE identified on the subsequent CTPA scan, as well as the 5 least likely requests. The LLM's were allowed 31 attempts. The performance of each group was determined based on the number of correctly identified positive scans.

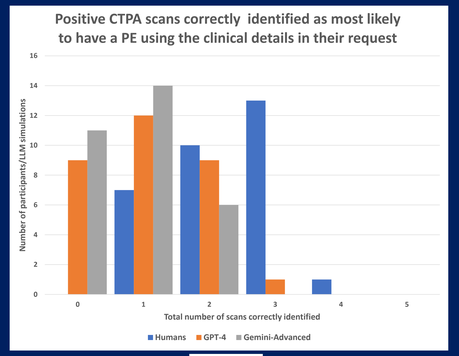


Figure 1

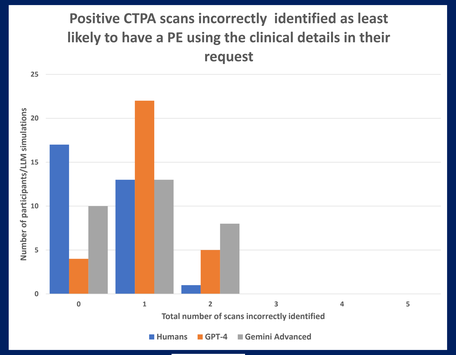


Figure 2

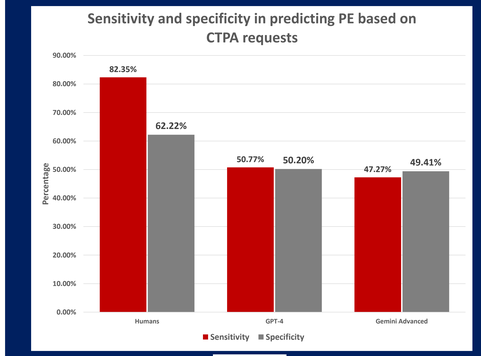


Figure 3

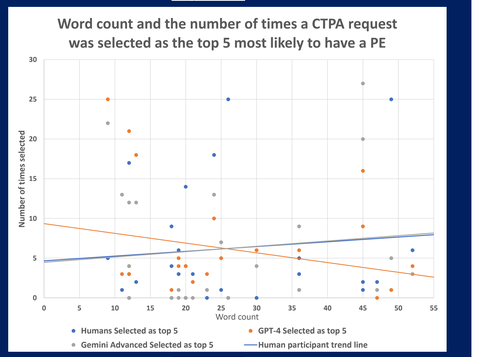


Figure 4

Results

The human experts correctly identified a mean number of 2.3 requests as compared to 1.1 for GPT-4 and 0.84 for Gemini Advanced (Fig.1). The humans rarely misidentified a positive scan request as least likely to have a PE with the mean number of 0.50 incorrectly identified positive scan requests. For GPT-4 and Gemini-Advanced, it was 1.0 and 0.94 respectively (Fig. 2). The positive likelihood ratio for humans was 2.2 whilst for GPT-4 it was 1.0 and 0.93 for Gemini Advanced. The negative likelihood ratios were 0.28, 0.98 and 1.1 for humans, GPT-4 and Gemini Advanced, respectively. Overall, the performance of both LLMs was equivalent to random chance as they had an approximate sensitivity and specificity of 50% in predicting whether a CTPA request would be positive for a PE on the subsequent scan. For humans, their overall sensitivity and specificity was 82% and 62% respectively (Fig. 3).

Further analysis showed performance between the consultants, radiology trainees and radiographers to be similar with the mean number of correctly identified requests ranging from 2.1 to 2.4 between the groups. There was also no correlation between the word count of a request and number of times it was selected as most likely to have had a PE confirmed on the subsequent CTPA for both the humans and LLMs (Fig.4). The inverse was also true with no statistically significant correlation between the word count and the number of times a CTPA request was selected as the least likely to have a PE.

The logic of AI

Gemini advanced

Keyword	Frequency	Weight
DVT	0.04	4
tachycardia	0.16	3
shortness of breath	0.12	3
chest pain	0.44	3
SOB	0.64	3
reduced mobility	0.08	2
pleurisy	0.04	2
infection	0.12	2
high BMI	0.04	1
smoking	0.08	1

1. **Absence of Specific PE Symptoms:** Patients X4LP, V2LP, E5LP, T0LP, and R8LP lack explicit mention of classic PE symptoms like pleuritic chest pain, hemoptysis, or a significantly elevated D-dimer. While some of them have shortness of breath or other non-specific symptoms, these could be attributed to their underlying conditions (e.g., COPD, cancer) or other causes.
2. **Alternative Explanations:** Patients X4LP and V2LP have clear alternative explanations for their symptoms (COVID-19 and metastatic cancer, respectively), making PE a less likely primary diagnosis.
3. **Focus on Ruling Out PE:** The clinical details for patients E5LP and V2LP specifically mention the need to "exclude" or rule out PE, suggesting that it is not the most likely diagnosis but rather a differential to be considered.

The following synonyms were accounted for:

- 'shortness of breath' and 'SOB'

The following misspellings were also accounted for:

- 'tachycardic' for 'tachycardia'
- 'tachy' for 'tachycardia'
- 'pleuritic' for 'pleuritic' (Note: this appears to be a duplicate in the original misspelling dictionary)

- **Synonyms:**
  - 'dd' (likely for 'd-dimer')
  - 'dvt' (deep vein thrombosis, a risk factor for PE)
  - 'dyspnoea' (shortness of breath)
- **Misspellings:**
  - 'tachy' (for 'tachycardia')

Figure 5

GPT-4



Figure 6A

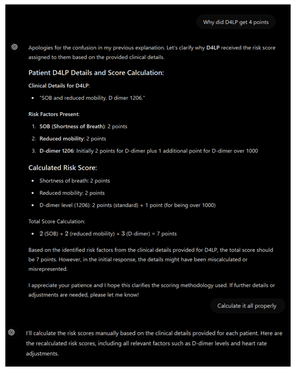


Figure 6B

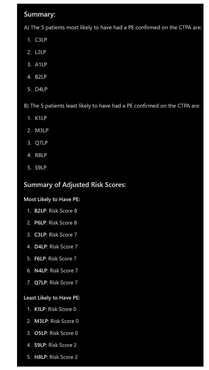


Figure 6C

Both LLMs consistently used a keyword model to derive a scoring system to stratify the requests in terms of likelihood of being positive. The models varied between each test and incorporated factors such as the presence of "DVT", "tachycardia", "SOB" and "reduced mobility" as well as their synonyms. The models were often complex and would account for alternative explanations if present, such as "COVID-19" (Fig. 5) or attempt to further stratify risk factors such as by awarding different weights to D-dimer values (Fig. 6A). However, despite the complexity of these models, they were shown not to out-perform human logic and intuition (see results section).

The LLMs were also proven to 'hallucinate' (the term for the observed phenomenon of AI producing false or inaccurate answers whilst stating them as a fact). The LLMs even misapplied their own model. This was apparent after interrogation of the dataset. In one instance, 3 out of 5 answers suggested by the LLM, changed after its error was noticed (by the researcher) and it was asked to "calculate it properly" (Fig. 6B,C). GPT-4 also invented new unique identifiers (which were allocate by the researchers to differentiate the CTPA requests) and then used these within its answers. This had the potential to confuse the analysis. Of note, there were no 'human errors' in the in the radiologist/ radiographer responses.

Conclusion

The findings of this study demonstrate the limitations of utilising current LLMs in determining the pretest probability of a CTPA being positive for a PE based on the clinical request details. Whilst human experts demonstrated moderate accuracy, the performance of GPT-4 and Gemini Advanced was comparable to random chance. This is despite both models creating and applying complex scoring criteria to derive their predictions, suggesting that human logic and intuition is superior in this task. The propensity for LLMs to hallucinate also limited their reliability and accuracy. These findings highlight the need for further research in improving the capabilities of artificial intelligence in this domain and to identify specific factors contributing to their poor performance.