Evaluating Large Language Models for Automated TNM Classification of Lung Cancer Radiology Reports

GIG CYMRU NHS WALES | Bwrdd Iechyd Prifysgol Bae Abertawe | Swansea Bay University Health Board

Dr Dafydd Ap Emyr, Dr Khattab Alkhedairy, Dr Emma Kirby

Swansea Bay University Health Board

Purpose

AI integration in clinical radiology aims to improve diagnostic accuracy and workflow efficiency to address increasing workloads for radiologists. While large language models (LLMs) have shown promise in text analysis, their application to specialised tasks like lung cancer TNM classification remains understudied.

This study evaluates how three leading LLMs – Open AI's ChatGPT-4.0, Anthropic's Claude 3.5 Sonnet, and Google's Gemini Advanced 1.5 Pro - perform in determining TNM classification from radiology reports, using criteria from the International Association for the Study of Lung Cancer (IASLC) 3rd Edition Staging Manual.

The study also examines whether report complexity affects accuracy and analyses the reasons for incorrect classifications.

Method

Synthetic reports

40 synthetic reports were created using simple unambiguous language. All reports used consistent terminology. They had a range of TNM classifications from T1a N0 M0 to T4 N3 M1c2. The reports did not have conclusions or impressions e.g. 'Findings are in keeping with a lung malignancy with contralateral mediastinal lymphadenopathy'.

Real-world reports

20 real repots were chosen from recent lung cancer Multidisciplinary Team (MDT) meetings, with reports chosen due to perceived clear and simple language used with lack of ambiguous terms (e.g. "indeterminate" / "borderline"). We excluded the indication, technique, conclusion and impression sections of the report.

Ground Truth

A Consultant Thoracic Radiologist analysed both the synthetic and assigned the definitive TNM classification.

Model Evaluation

The reports were input into three LLMs: *OpenAI ChatGPT-4.0*, *Anthropic Claude Sonnet 3.5* and *Google Gemini Advanced 1.5 Pro.*

Prompt (Scan QR code to view the complete prompt)

The same prompt was used for each LLM. It included a brief task explanation, asked to provide reasoning for the assigned classification and comment on how confident it was with the answer. We uploaded relevant sections from the IASLC 3rd Edition Staging Manual in Thoracic Oncology[1]. We also provided 'tips' for some of the nuances of interpreting the document and general method of classifying.

Results

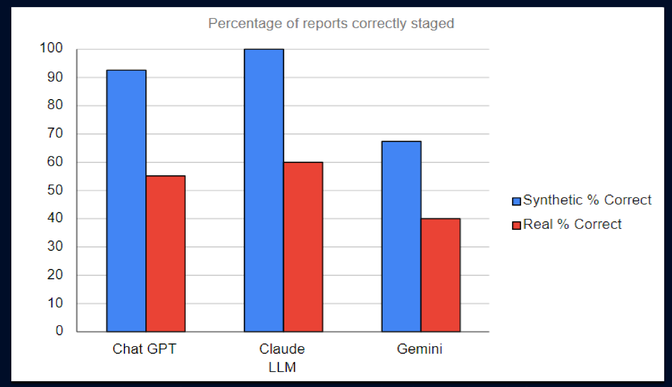


Fig 1. Percentage of reports correctly staged per LLM, comparing the synthetic and real report classification accuracy

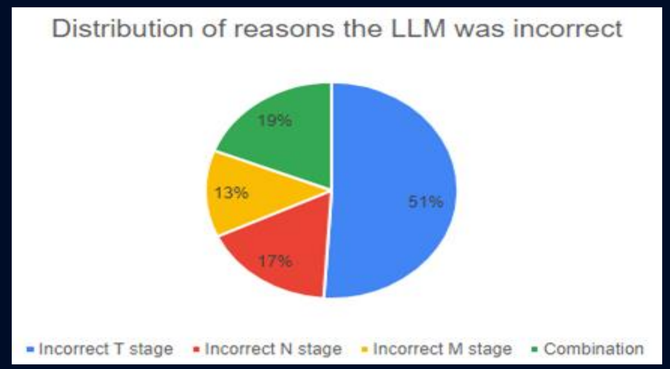


Fig 2. Distribution of the main reasons the classification provided by the LLM was incorrect

LLM	Complexity Level 1	Complexity Level 2	Complexity Level 3
ChatGPT	100% Correct	62.5% Correct	64% Correct
Claude	96.3% Correct	100% Correct	64% Correct
Gemini	81.5% Correct	37.5% Correct	32% Correct

Table 1. Comparison of the accuracy of each LLM depending on the complexity of the report

Discussion

Synthetic Vs Real Reports

All models show higher accuracy on synthetic reports than real reports (Fig. 1), likely reflecting challenges with increased variability in vocabulary, wording and structuring of real-world reports. *Claude* outperforms the others in both synthetic and real scenarios, while *Gemini* performed the worst. This emphasizes the importance of robust training on diverse real-world data to improve accuracy.

Incorrect staging for T, N or M

Figure 2 shows that the most errors (51%) were due to incorrect classification of the T stage indicating that tumour size and involvement are the most challenging for the models to classify, which is surprising as we predicted that the objective numerical nature of this would play to the strength of the models. We postulate that in some real reports, which measured a lesion in 3 planes separated with an 'x' (e.g. 3 x 2 x 1cm), an average value was taken, even though our prompt advised not to do this. However, *Gemini* demonstrated poor performance with this error even with synthetic reports where only one measurement was given. For N stage, the main inaccuracies were either classifying contralateral nodes as N2 (rather than N3) or incorrectly including hilar nodes as a N2 station, therefore upstaging from N2a to N2b. Inaccuracies in M staging were mainly related to either not including a pleural effusion as an M stage, incorrectly classifying contralateral pulmonary nodules or incorrectly classifying extra thoracic metastases.

Complexity

Both the synthetic and real reports were categorised into different levels of complexity. A report with 1 or 2 'positive' findings was assigned Level 1, 3 positive findings as Level 2 and ≥4 positive findings Level 3.

ChatGPT and *Claude* handle simpler tasks (Level 1) very well, with *Claude* performing the best for better for Level 2. This is unsurprising given that there is fewer positive findings to interpret to assign a final stage. *ChatGPT* and *Claude* performed equally for more complex reports (Level 3), whilst *Gemini* performed poorly with this class.

Conclusion

Whilst current LLMs demonstrate strong potential for TNM classification, their performance appears to be hindered by the complexity and variability of real-world radiology reports. Future work should focus on fine-tuning custom GPT models using diverse, real-world medical datasets against expert agreed final classifications to improving their ability to parse nuanced clinical language. We should also consider greater use of standardised reporting terminology to aid this.

These tailored solutions may bridge the gap between synthetic and real-world accuracy, paving the way for effective integration of LLMs into clinical radiology workflows to improve efficiency, augment accuracy and provide consistency. The hope is that these AI solutions will help with the ever-increasing demands placed on human radiologists.

Reference

1. International Association for the Study of Lung Cancer (IASLC). (2024). *IASLC Staging Manual in Thoracic Oncology* (3rd ed.). Springer.