Do General Purpose Large Language Models Outperform Domain-Specific NLP Methods for Radiology Report Label Extraction?

Cody H. Savage[1], MD, Hyoungsun Park[1], MS, Kijung Kwak[1], Steven Rothenberg[2], MD, Florence X. Doo[1], MD, Vishwa S Parekh[1], MD, Paul Yi[1], MD

[1]*University of Maryland Medical Intelligent Imaging (UM2ii) Center, Department of Diagnostic Radiology & Nuclear Medicine, University of Maryland School of Medicine*
[2]*Department of Radiology, University of Alabama at Birmingham Heersink School of Medicine*

University of Maryland Medical Intelligent Imaging

Visit our website

UNIVERSITY of MARYLAND SCHOOL OF MEDICINE

Introduction

- Traditional natural language processing (NLP) methods for radiology report labeling require fine-tuning on reports that have been manually annotated – an extremely time-consuming task.
- Large language models (LLMs) such as Generative Pre-trained Transformer (GPT) have demonstrated domain-agnostic non-medical language task abilities [1].
- We evaluated GPT's ability to classify radiology reports across multiple domains with and compare its performance to state-of-the-art NLP methods (CheXpert).

Methods

- Radiology reports from Indiana chest Xray dataset (n = 3665) [2] were processed by CheXpert [3] to diagnose 14 disease labels (including 'no finding').
- The reports were also processed by OpenAI's ChatGPT-3.5 and GPT-4 models.
- GPT was integrated with ConversationalRetrievalQA chain, to allow for merging of chat histories with contemporary queries, termed context-GPT (C-GPT).
- A subset of reports (n = 200) were classified for any pathological imaging finding by a radiology resident, which was used to compare GPT performance in classifying all pathological imaging findings present in each radiology report

Findings	Pathological Finding	Normal vs Abnormal	Total
Cholecystectomy clips are present. Small T-spine osteophytes. There is biapical pleural thickening, unchanged from prior. Mildly hyperexpanded lungs.	Cholecystectomy clips, osteophyte, pleural thickening, hyperexpanded lungs	Abnormal	4
There are extremely low lung volumes. There is a right basilar opacity. There is no pneumothorax. There is no large pleural effusion. Cardiac silhouette and mediastinal contours are within normal limits.	Lung hypoinflation, lung opacity	Abnormal	2
Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Lungs are clear. No pneumothorax or pleural effusion. No acute osseous findings.	None	Normal	0

Table 1. Pathological imaging finding reference standard process. The findings section from three reports from the Indiana Chest X-ray dataset are shown with the pathological imaging findings highlighted in red. Reports with no imaging findings listed in the reference standard were classified as "normal". Pathological imaging findings were defined as any abnormal imaging finding, support device (e.g. endotracheal tube, pacemaker), or post-surgical changes/hardware. Terms that indicated uncertainty about a pathological imaging finding were not counted in the reference standard, as previously described by Demner-Fushman et al [2].

Results

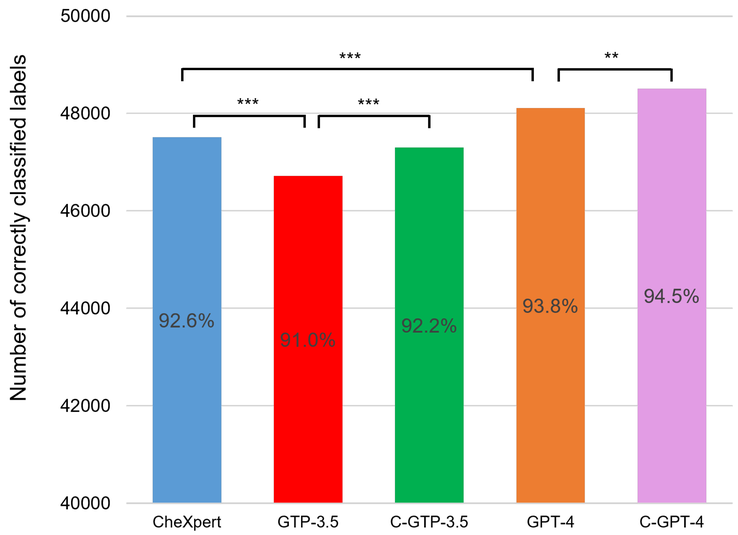

Figure 1. Total disease label classification performance. CheXpert correctly classified 92.6% (47514/51310) of the disease labels, outperforming GPT-3.5 at 91.0% (46717/51310)[P < .001]. However, GPT-4 had an accuracy of 93.8% (48114/51310), greater than CheXpert (P < .001). Allowing GPT to have context prior to its response improved performance for GPT-3.5 (91% to 92.2% for C-GPT-3.5, *P* < .0001) and GPT-4 (93.8% to 94.5% for C-GPT-4, *P* = .006).
**P < .01
***P < .0001

Classification Task	CheXpert	GTP-3.5	GPT-4
Normal vs abnormal reports	183 (91.5%)[a]	198 (99.0%)[a,b]	198 (99.0%)[a,b]
Pathological imaging findings	-	346 (94.0%)	354 (96.2%)

Table 2. Abnormal report and pathological findings classification task performance. [a]GPT-3.5 and GPT-4 achieved 99% (198/200) accuracy in classifying reports as normal versus abnormal, outperforming CheXpert at 91.5% (183/200)[*P* = .003]. [b]GPT-4 correctly identified 96.2% of the 368 pathological imaging findings in the 142 abnormal reports, similar to GPT-3.5 at 94.0% (*P* = .08).

Results

Disease Label	CheXpert	GTP-3.5	C-GTP-3.5	GPT-4	C-GPT-4
No finding (n=1406)	84.8%	62.1%	64.7%	79.4%	80.5%
Enlarged Cardiom. (n=380)	86.8%	90.4%	93.5%	92.5%	91.0%
Cardiomegaly (n=326)	91.5%	97.9%	95.2%	98.2%	97.7%
Lung Lesion (n=1174)	71.1%	73.1%	76.7%	76.0%	81.1%
Lung Opacity (n=1151)	83.3%	75.5%	77.0%	80.4%	84.2%
Edema (n=44)	98.4%	97.9%	98.9%	99.3%	99.0%
Consolidation (n=26)	99.0%	95.8%	99.3%	99.6%	99.7%
Pneumonia (n=36)	97.0%	98.7%	98.9%	99.7%	98.7%
Atelectasis (n=296)	97.1%	98.1%	96.8%	98.6%	98.3%
Pneumothorax (n=25)	99.2%	99.6%	99.8%	99.9%	99.9%
Pleural Effusion (n=145)	98.7%	99.1%	99.2%	99.4%	99.2%
Pleural Other (n=72)	98.3%	94.5%	98.2%	97.5%	98.7%
Fracture (n=84)	98.5%	98.0%	98.9%	96.6%	98.5%
Support Devices (n=286)	92.9%	94.1%	93.5%	95.6%	97.1%

Table 3. Individual disease label classification performance. Data shown is percent of reports that the disease label was correctly classified. GPT-3.5 notably showed poor performance on identifying normal reports (i.e. "no finding") in comparison to CheXpert despite outperforming CheXpert on most disease label classifications. This gap in performance decreased for GPT-4 and even further for C-GPT-4. GPT outperforms CheXpert for labels that exhibit greater diversity in the language used to describe them by radiologists (e.g. enlarged cardiomediastinum, cardiomegaly, lung lesions).

Conclusions

- GPT-4 surpasses CheXpert in classifying pre-defined disease labels from radiology reports, despite CheXpert being trained specifically for this task.
- GPT's performance can be further enhanced when the context of the dataset is provided.
- GPT can classify pathological imaging findings present in reports without pre-defined labels with high accuracy, suggesting that the time-consuming task of manual annotation of reports can potentially be entirely automated.

References

1. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. Radiology. 2023;307(5):e230922.
2. Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, et al. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc. 2016;23(2):304-10.
3. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al., editors. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI conference on artificial intelligence; 2019.
4. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176-82.