SIIM CMIMI24
Oct 21-22 | Boston, MA
Conference On Machine Intelligence In Medical Imaging

Enhancing Radiology Report Comprehension: A Study on GPT-4's Identification of Key Radiological Terms

Jad Alsheikh, Creighton University School of Medicine; Ali Memon; Daniel Spalinski; Kimberly Mendez, PhD; Sherif Zineldine; Dorina Pinkhasova; Michael Fei

Introduction/Background

This study evaluates GPT-4's accuracy in identifying common radiological terms from chest X-ray (CXR) reports, comparing it to terms derived from actual radiology reports. The objective is to assess GPT-4's potential in creating a database for highlighting and defining medical terms to aid patient comprehension.

Methods/Intervention

This was a retrospective analysis of CXR reports. Two lists of the top 40 most common radiological CXR findings and phrases were generated. The first list was derived from 3,999 reports from the Open-i service of the NLM, covering a wide array of pathologies. The second list was generated by GPT-4, identifying what it believed to be the 40 most common findings and phrases. We compared GPT-4’s performance against terms derived from the sample reports by analyzing the overlap between the two lists and assessing the frequency of GPT-4 terms in actual reports, considering exact and similar matches. Additionally, we evaluated how well GPT-4 can account for variations in radiologist terminology by examining the coefficient of variation (CV) for the term frequencies.

Results/Outcome

GPT-4 demonstrated the ability to identify frequently used terms such as "effusion" and "pneumothorax" with high accuracy. Terms with high exact match proportions included "pleural effusion" (100%), "pulmonary edema" (100%), and "pneumothorax" (73.7%). The precision, recall, and F1 score were all 0.30, indicating moderate overlap between the terms identified by GPT-4 and those derived from the CXR reports. The Spearman's rank correlation was 0.32, suggesting a weak correlation between the ranks of term frequencies in GPT-4’s list and the actual reports. The Chi-Square Test ($chi2=840.00$, $p=7.53e-04$, $dof=714$) indicated that the differences between the observed frequencies of terms in actual reports and those identified by GPT-4 were statistically significant.

Conclusion

GPT-4 demonstrated reasonable accuracy in identifying common radiological terms in CXR reports and can effectively account for variations in terminology. While it successfully identified frequently used terms, its performance varied for less common terms.

Statement of Impact

This study underscores the potential of GPT-4 in enhancing patient understanding of radiological reports by providing a reliable database of terms. Incorporating AI tools like GPT-4 could improve patient communication and engagement in radiology, ultimately contributing to better healthcare outcomes.

Term	Frequency	TF-IDF Score
pneumothorax	2603	0.291836
effusion	2528	0.277761
consolidation	1200	0.144364
pleural effusion	2202	0.132574
opacity	961	0.113767
heart size	1804	0.108534
acute cardiopulmonary	1493	0.089786
mediastinum	625	0.069575
atelectasis	503	0.060493
chest pain	884	0.053194
edema	405	0.048723
nodule	368	0.044131
cardiomediastinal silhouette	712	0.042828
cardiopulmonary abnormality	647	0.038918
infiltrate	318	0.038256
airspace disease	585	0.035209
focal consolidation	536	0.032261
pleural effusions	387	0.024076
mediastinal contours	385	0.023239
thoracic spine	359	0.021615
lung volumes	339	0.020452
thorax	160	0.019249
osseous structures	316	0.019008
fracture	152	0.018266
pulmonary vasculature	301	0.018106
cardiopulmonary disease	298	0.018005
mediastinal contour	276	0.017123
pulmonary findings	246	0.015384
lower lobe	235	0.014801
pulmonary edema	218	0.013654
focal airspace disease	353	0.013446
structures intact	199	0.012445
upper lobe	185	0.01157
bony structures	184	0.011528
pleural disease	174	0.01134
cardiac silhouette	165	0.010423
degenerative thoracic spine	160	0.010277
calcified granuloma	145	0.009131

This table presents the most frequently occurring radiological terms identified from a sample of 3,999 chest X-ray reports obtained from the Open-i service of the National Library of Medicine (NLM). The terms are ranked by their TF-IDF scores to highlight their relevance and frequency within the reports.

Term	Exact Matches	Similar Matches	Total Matches	CV
effusion	2983	1505	4488	0.2328662706
pneumothorax	2669	950	3619	0.3358708364
pleural effusion	2585	0	2585	0.7071067812
opacity	1487	979	2466	0.1456651439
atelectasis	1287	661	1948	0.2272324666
cardiomegaly	1063	624	1687	0.1840070403
granuloma	939	600	1539	0.1557564645
consolidation	1234	137	1371	0.5657885769
pneumonia	555	247	802	0.2715572177
interstitial	493	297	790	0.1754340875
emphysema	410	378	788	0.02871499619
airspace disease	732	0	732	0.7071067812
nodule	349	367	716	0.01777642746
infiltrate	394	264	658	0.1397019477
calcification	123	235	358	0.2212177639
pulmonary edema	350	0	350	0.7071067812
hemidiaphragm	137	69	206	0.2334138889
hyperinflation	73	110	183	0.1429669448
mass	101	78	179	0.09085729591
cardiac silhouette	165	0	165	0.7071067812
fibrosis	98	51	149	0.2230471055
perihilar	62	59	121	0.01753157309
costophrenic angle	106	0	106	0.7071067812
pleural thickening	82	0	82	0.7071067812
interstitial markings	68	0	68	0.7071067812
vascular congestion	58	0	58	0.7071067812
lesion	23	30	53	0.09339146167
bronchiectasis	22	12	34	0.2079725827
interstitial lung disease	27	0	27	0.7071067812
cavity	8	2	10	0.4242640687
mediastinal shift	7	0	7	0.7071067812
peribronchial cuffing	4	0	4	0.7071067812
hilar enlargement	3	0	3	0.7071067812
bronchovascular markin	1	0	1	0.7071067812
tracheal deviation	0	0	0	
lung fields	0	0	0	
silhouette sign	0	0	0	
ground-glass opacity	0	0	0	
reticular pattern	0	0	0	
honeycombing	0	0	0	

This table presents the comparison of GPT-4 generated top 40 radiological terms with their exact and similar match frequencies, total matches, and coefficient of variation (CV) based on 3,999 chest X-ray reports from the Open-i service of the National Library of Medicine (NLM).

Keywords

GPT-4; Chest X-ray Reports; Natural Language Processing; Term Identification