RSNA Press Release

Researchers Test Large Language Model that Preserves Patient Privacy

Released: October 10, 2023

At A Glance

- Researchers tested the feasibility of using a locally run large language model (LLM) to label key findings from chest X-ray reports, while preserving patient privacy.
- The LLM was asked to identify and label the presence or absence of 13 specific findings on the chest X-ray reports.
- The LLM's performance was comparable to the current reference standard.

OAK BROOK, Ill. — Locally run large language models (LLMs) may be a feasible option for extracting data from text-based radiology reports while preserving patient privacy, according to a new study from the National Institutes of Health Clinical Center (NIH CC) published in *Radiology*, a journal of the Radiological Society of North America (RSNA). LLMs are deep-learning models trained to understand and generate text in a human-like way.

download photo

Ronald M. Summers, M.D., Ph.D.

Recently released LLM models such as ChatGPT and GPT-4 have garnered attention. However, they are not compatible with healthcare data due to privacy constraints.

"ChatGPT and GPT-4 are proprietary models that require the user to send data to OpenAI sources for processing, which would require de-identifying patient data," said senior author Ronald M. Summers, M.D., Ph.D., senior investigator in the Radiology and Imaging Sciences Department at the NIH. "Removing all patient health information is labor-intensive and infeasible for large sets of reports."

In this study, led by Pritam Mukherjee, Ph.D., staff scientist at the NIH CC, researchers tested the feasibility of using a locally run LLM, Vicuna-13B, to label key findings from chest X-ray reports from the NIH and the Medical Information Mart for Intensive Care (MIMIC) Database, a publicly available dataset of de-identified electronic health records.

"Preliminary evaluation has shown that Vicuna, a free publicly available LLM, approaches the performance of ChatGPT in tasks such as multi-lingual question answering," Dr. Summers said.

The study dataset included 3,269 chest X-ray reports obtained from MIMIC and 25,596 reports from the NIH.

Using two prompts for two tasks, the researchers asked the LLM to identify and label the presence or absence of 13 specific findings on the chest X-ray reports. Researchers compared the LLM's performance with two widely used non-LLM labeling tools.

A statistical analysis of the LLM output showed moderate to substantial agreement with the non-LLM computer programs.

"Our study demonstrated that the LLM's performance was comparable to the current reference standard," Dr. Summers said. "With the right prompt and the right task, we were able to achieve agreement with currently used labeling tools."

Dr. Summers said LLMs that can be run locally will be useful in creating large data sets for AI research without compromising patient privacy.

"LLMs have turned the whole paradigm of natural language processing on its head," he said. "They have the potential to do things that we've had difficulty doing with traditional pre-large language models."

Dr. Summers said LLM tools could be used to extract important information from other text-based radiology reports and medical records, and as a tool for identifying disease biomarkers.

"My lab has been focusing on extracting features from diagnostic images," he said. "With tools like Vicuna, we can extract features from the text and combine them with features from images for input into sophisticated AI models that may be able to answer clinical questions.

"LLMs that are free, privacy-preserving, and available for local use are game changers," he said. "They're really allowing us to do things that we weren't able to do before."

"Feasibility of Using the Privacy-preserving Large Language Model Vicuna for Labeling Radiology Reports." Collaborating with Drs. Summers and Mukherjee were Benjamin Hou, Ph.D., and Ricardo B. Lanfredi, Ph.D.

In 2023, *Radiology* is celebrating its 100th anniversary with 12 centennial issues, highlighting *Radiology*'s legacy of publishing exceptional and practical science to improve patient care.

Radiology is edited by Linda Moy, M.D., New York University, New York, N.Y., and owned and published by the Radiological Society of North America, Inc. (https://pubs.rsna.org/journal/radiology)

RSNA is an association of radiologists, radiation oncologists, medical physicists and related scientists promoting excellence in patient care and health care delivery through education, research and technologic innovation. The Society is based in Oak Brook, Illinois. (RSNA.org)

For patient-friendly information on chest X-rays, visit *RadiologyInfo.org*.

Video (MP4):

Video 1. Ronald M. Summers, M.D., Ph.D., discusses his research on testing large language models that preserve patient privacy.
Download

Images (JPG, TIF):

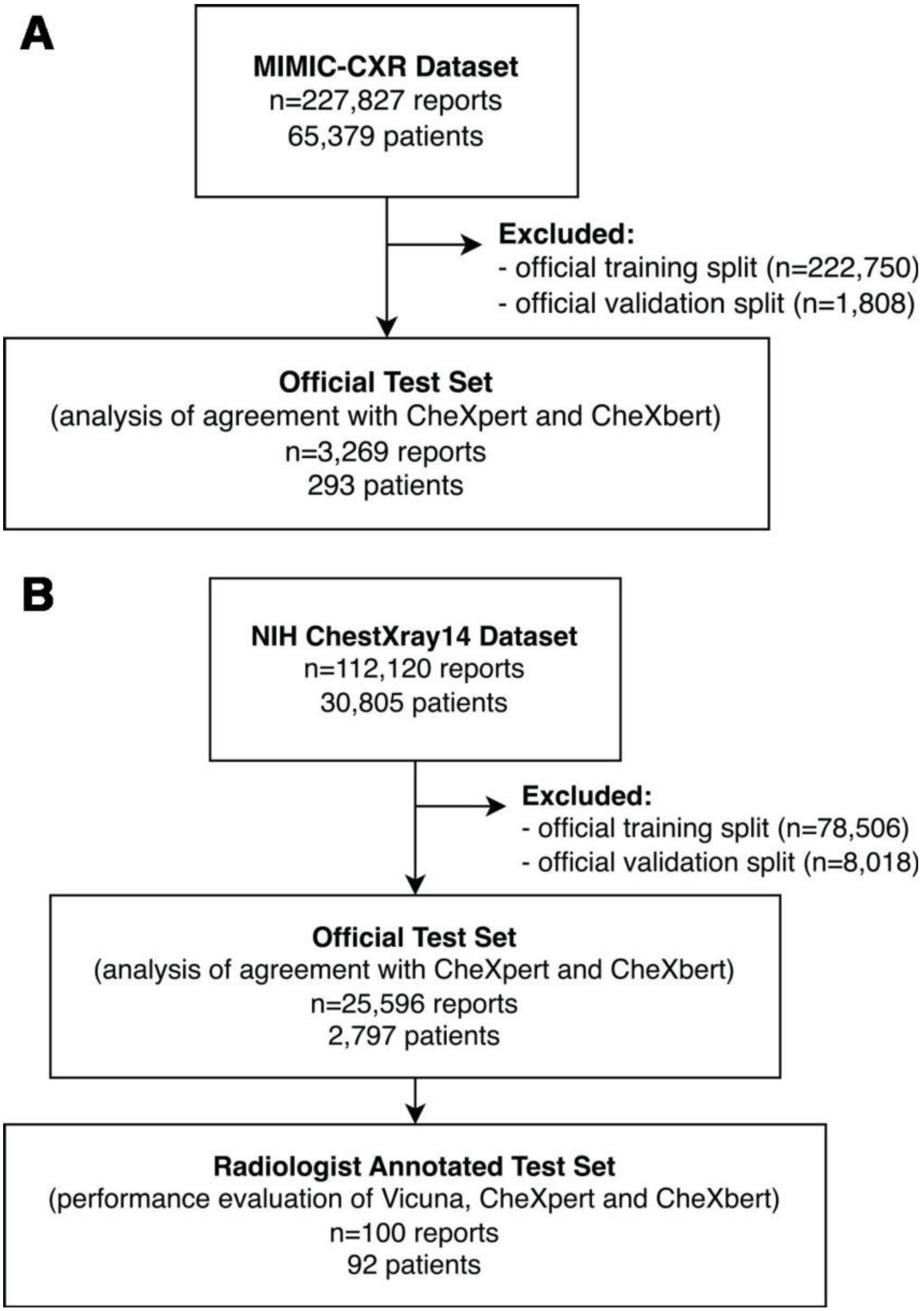
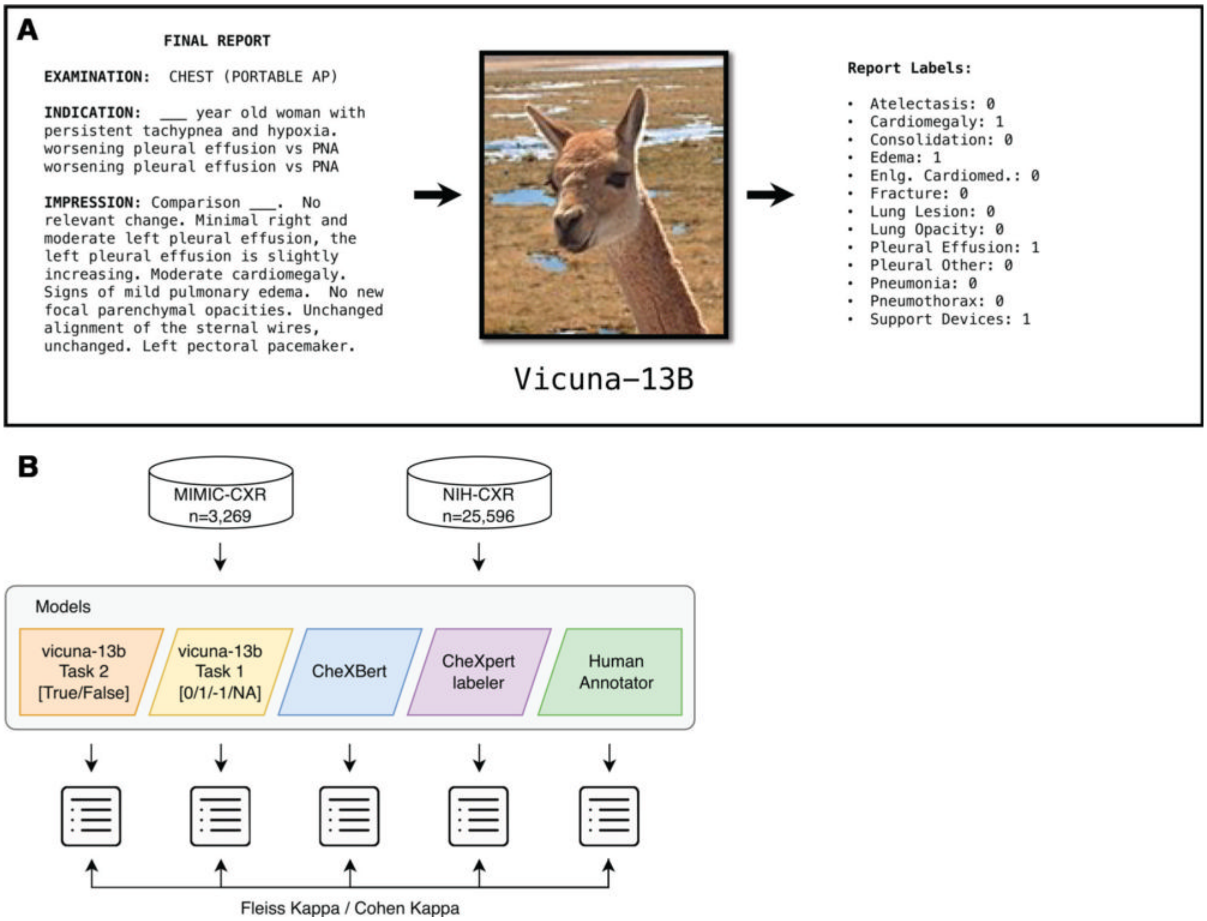

A
MIMIC-CXR Dataset
n=227,827 reports
65,379 patients
Excluded:
- official training split (n=222,750)
- official validation split (n=1,808)
Official Test Set
(analysis of agreement with CheXpert and CheXbert)
n=3,269 reports
293 patients
B
NIH ChestXray14 Dataset
n=112,120 reports
30,805 patients
Excluded:
- official training split (n=78,506)
- official validation split (n=8,018)
Official Test Set
(analysis of agreement with CheXpert and CheXbert)
n=25,596 reports
2,797 patients
Radiologist Annotated Test Set
(performance evaluation of Vicuna, CheXpert and CheXbert)
n=100 reports
92 patients

Figure 1. Flowchart shows the data selection process from the **(A)** MIMIC-CXR and **(B)** National Institutes of Health (NIH) ChestX-ray14 data sets.
High-res (TIF) version



FINAL REPORT

EXAMINATION: CHEST (PORTABLE AP)

INDICATION: ___ year old woman with persistent tachypnea and hypoxia. worsening pleural effusion vs PNA worsening pleural effusion vs PNA

IMPRESSION: Comparison ___. No relevant change. Minimal right and moderate left pleural effusion, the left pleural effusion is slightly increasing. Moderate cardiomegaly. Signs of mild pulmonary edema. No new focal parenchymal opacities. Unchanged alignment of the sternal wires, unchanged. Left pectoral pacemaker.

Report Labels:

- Atelectasis: 0
- Cardiomegaly: 1
- Consolidation: 0
- Edema: 1
- Enlg. Cardiomed.: 0
- Fracture: 0
- Lung Lesion: 0
- Lung Opacity: 0
- Pleural Effusion: 1
- Pleural Other: 0
- Pneumonia: 0
- Pneumothorax: 0
- Support Devices: 1

Figure 2. Overview of the study. **(A)** The open-access large language model Vicuna-13B, which can be run on a local computer without the need for de-identification of patient data, was prompted to examine unstructured, free-text chest radiography (CXR) reports and generate an output file reporting the results of 13 specific findings. AP = anteroposterior, Enlg. Cardiomed. = enlarged cardiomediastinum, PNA = pneumonia. **(B)** Reports from the MIMIC-CXR data set (*n* = 3269) and the National Institutes of Health (NIH) data set (*n* = 25 596) were used in this study. Vicuna was given two independent tasks that generated two different output files, one in which the 13 possible findings were labeled as positive or negative (task 2, orange model) and the other in which the 13 possible findings were labeled as positive, negative, unsure, or not mentioned (task 1, yellow model). The agreement between Vicuna model outputs and the CheXbert labeler, CheXpert labeler, and human annotations were compared using Fleiss or Cohen κ as appropriate.

High-res (TIF) version

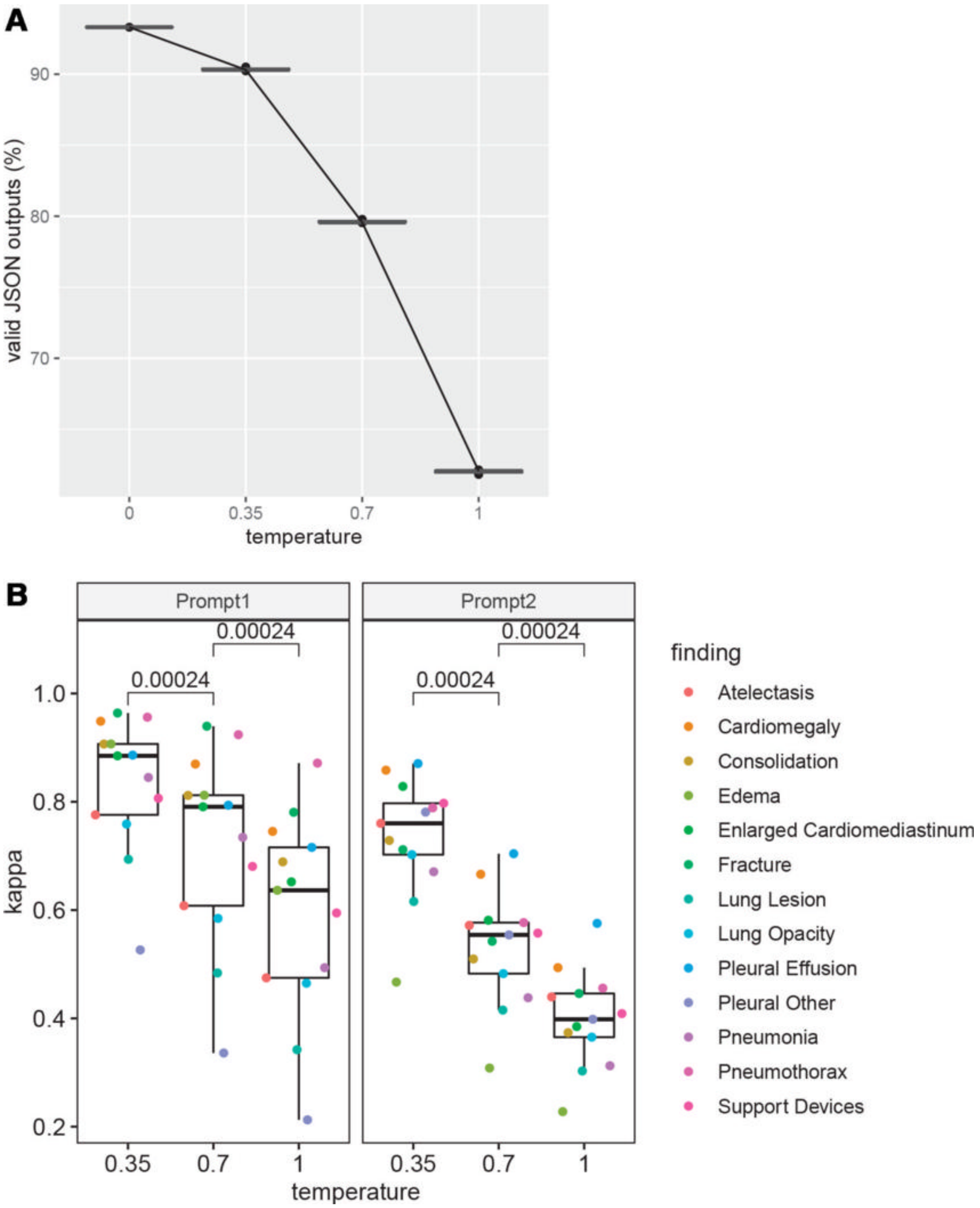
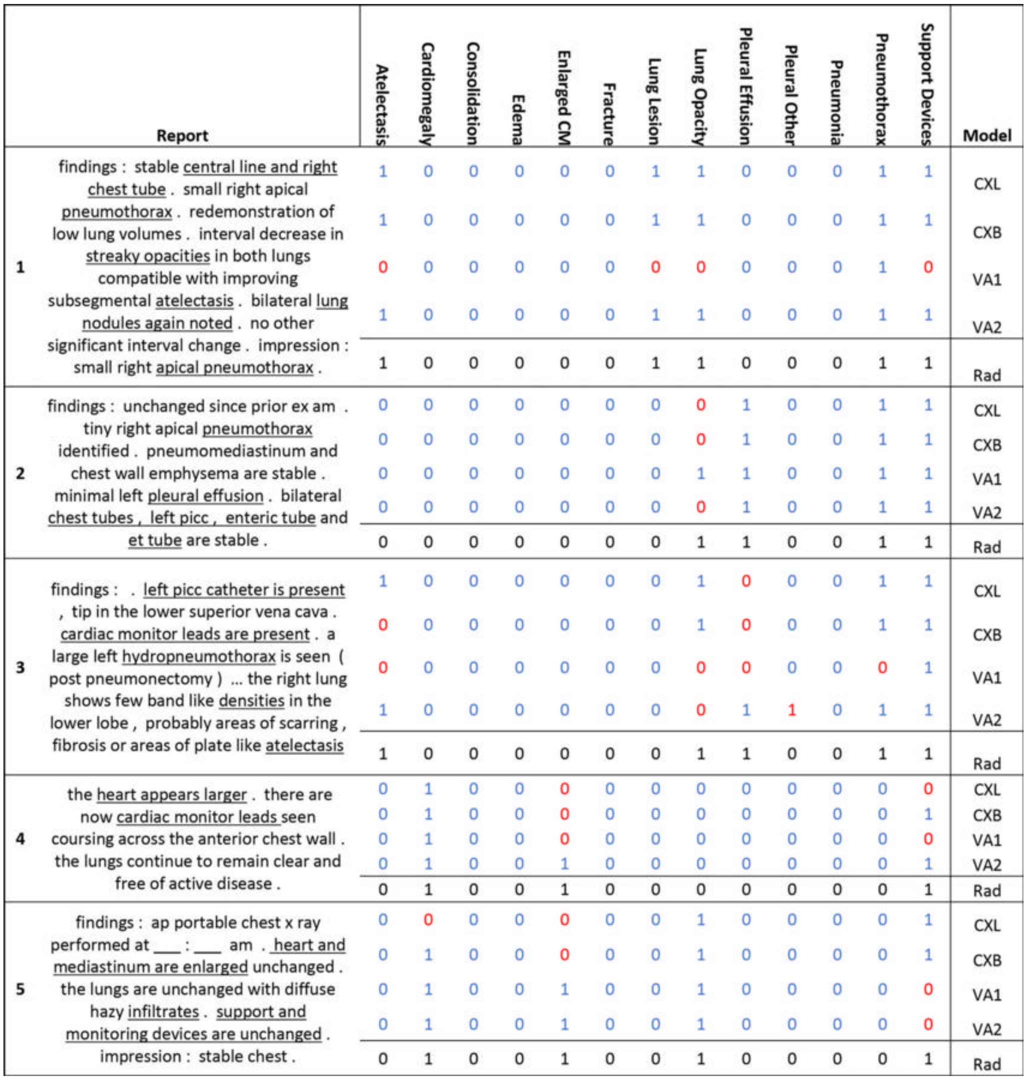


Figure 3. Effect of the temperature hyperparameter. **(A)** Line graph shows that the proportion of valid JavaScript object notation (JSON) outputs obtained from Vicuna using prompt 1 on the National Institutes of Health data set decreased as the temperature hyperparameter increased. At each of the three temperatures of 0.35, 0.7, and 1, three points are plotted showing the actual proportion of the valid JSON outputs for the three runs. The horizontal lines mark the medians of the three points for each temperature. **(B)** Box and whisker plots of the degree of agreement measured with use of Fleiss κ among three runs of Vicuna show significant decreases as the temperature increases for both prompt 1 and prompt 2. The numbers near the top of the plot represent P values computed using the Wilcoxon signed-rank test. In this plot, the midline represents the median, and box edges represent the first and third quartiles. The whiskers represent the range of the data, excluding outliers. The outliers are points that are beyond.

High-res (TIF) version



	Report	Atelectasis	Cardiomegaly	Consolidation	Edema	Enlarged CM	Fracture	Lung Lesion	Lung Opacity	Pleural Effusion	Pleural Other	Pneumonia	Pneumothorax	Support Devices	Model
1	findings : stable central line and right chest tube . small right apical pneumothorax . redemonstration of low lung volumes . interval decrease in streaky opacities in both lungs compatible with improving subsegmental atelectasis . bilateral lung nodules again noted . no other significant interval change . impression : small right apical pneumothorax .	1	0	0	0	0	0	1	1	0	0	0	1	1	CXL
		1	0	0	0	0	0	1	1	0	0	0	1	1	CXB
		0	0	0	0	0	0	0	0	0	0	0	1	0	VA1
		1	0	0	0	0	0	1	1	0	0	0	1	1	VA2
		1	0	0	0	0	0	1	1	0	0	0	1	1	Rad
2	findings : unchanged since prior ex am . tiny right apical pneumothorax identified . pneumomediastinum and chest wall emphysema are stable . minimal left pleural effusion . bilateral chest tubes , left picc , enteric tube and et tube are stable .	0	0	0	0	0	0	0	0	1	0	0	1	1	CXL
		0	0	0	0	0	0	0	0	1	0	0	1	1	CXB
		0	0	0	0	0	0	0	1	1	0	0	1	1	VA1
		0	0	0	0	0	0	0	0	1	0	0	1	1	VA2
		0	0	0	0	0	0	0	1	1	0	0	1	1	Rad
3	findings : . left picc catheter is present , tip in the lower superior vena cava . cardiac monitor leads are present . a large left hydropneumothorax is seen (post pneumonectomy) ... the right lung shows few band like densities in the lower lobe , probably areas of scarring , fibrosis or areas of plate like atelectasis	1	0	0	0	0	0	0	1	0	0	0	1	1	CXL
		0	0	0	0	0	0	0	1	0	0	0	1	1	CXB
		0	0	0	0	0	0	0	0	0	0	0	0	1	VA1
		1	0	0	0	0	0	0	0	1	1	0	1	1	VA2
		1	0	0	0	0	0	0	1	1	0	0	1	1	Rad
4	the heart appears larger . there are now cardiac monitor leads seen coursing across the anterior chest wall . the lungs continue to remain clear and free of active disease .	0	1	0	0	0	0	0	0	0	0	0	0	0	CXL
		0	1	0	0	0	0	0	0	0	0	0	0	1	CXB
		0	1	0	0	0	0	0	0	0	0	0	0	0	VA1
		0	1	0	0	1	0	0	0	0	0	0	0	1	VA2
		0	1	0	0	1	0	0	0	0	0	0	0	1	Rad
5	findings : ap portable chest x ray performed at ___ : ___ am . heart and mediastinum are enlarged unchanged . the lungs are unchanged with diffuse hazy infiltrates . support and monitoring devices are unchanged . impression : stable chest .	0	0	0	0	0	0	0	1	0	0	0	0	1	CXL
		0	1	0	0	0	0	0	1	0	0	0	0	1	CXB
		0	1	0	0	1	0	0	1	0	0	0	0	0	VA1
		0	1	0	0	1	0	0	1	0	0	0	0	0	VA2
		0	1	0	0	1	0	0	1	0	0	0	0	1	Rad

Figure 4. Example chest radiography reports and findings identified by CheXpert, CheXbert, Vicuna, and a radiologist. Both prompt 1 (a single-step prompt) and prompt 2 (a multistep interactive rule-based prompt) were used to run task 2 (positive or negative finding) on the large language model Vicuna. A value of 1 indicates a positive finding, while 0 indicates a negative finding. Values in blue represent outputs that agreed with the findings of the radiologist (Rad) (numbers in black), and red values represent outputs that disagreed with the findings of the radiologist. Key phrases in the reports are underlined. ap = anteroposterior, CM = cardiomediastinum, CXB = CheXbert labeler, CXL = CheXpert labeler, et = endotracheal, exam = examination, picc = peripherally inserted central catheter, VA1 = Vicuna run with prompt 1, VA2 = Vicuna run with prompt 2.
High-res (TIF) version

Resources:

Editorial
Study abstract