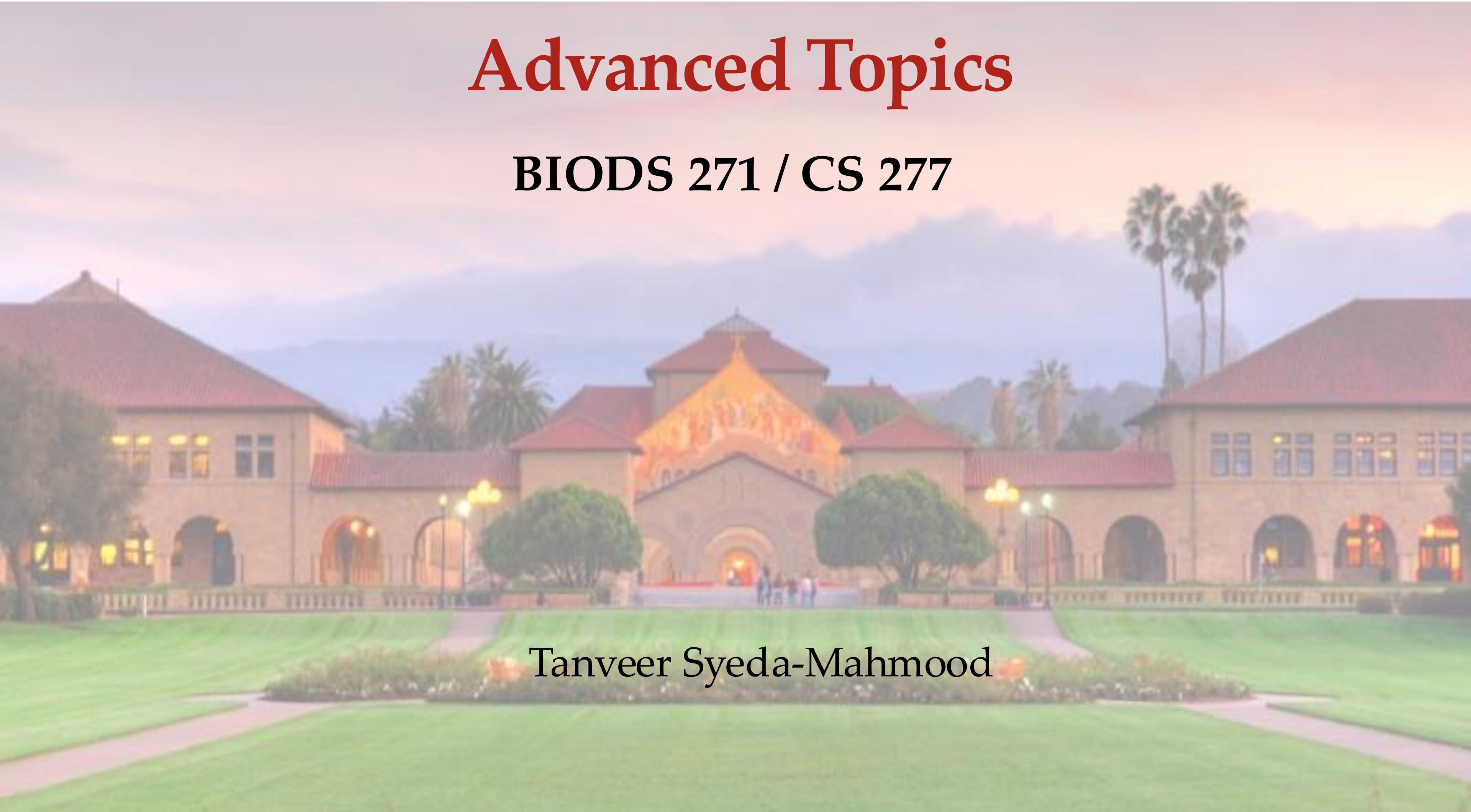


Advanced Topics

BIODS 271 / CS 277

Tanveer Syeda-Mahmood

Topics covered

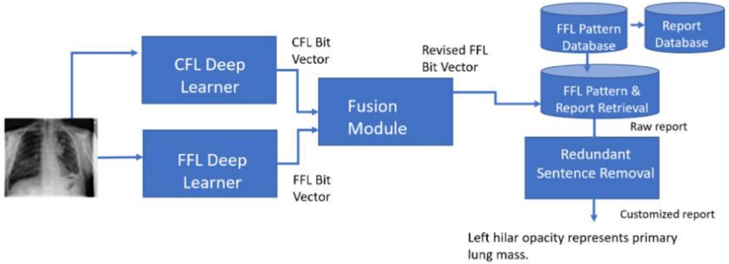
- Fact-checking medical hallucinations (contd.)
- Advanced topics:
  - Inference scaling and reasoning
  - Agentic AI systems
  - Memory augmented networks
- Deployment considerations

Detecting medical hallucinations

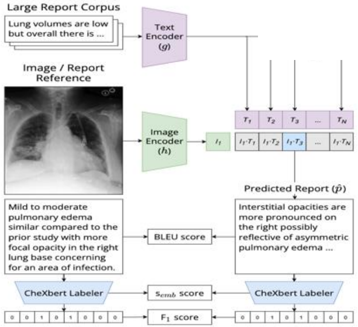
- Factual verification:
  - decompose complex claims into sub-questions, retrieve relevant documents from web sources, and evaluates the truthfulness of each sub-component – FACTSCORE, EMNLP2023.
- Summary consistency verification:
  - Generate QA from either source or summary and see if the answers can be generated from the other.
  - Generate questions form summary and compare the answers with the source
  - Entailment-based methods use natural language inference (NLI) whether each sentence in the summary is logically entailed by the source.
- Uncertainty-based hallucination detection.
  - sequence log-probability of generated text
  - semantic entropy to quantify uncertainty by repeated generation for the same prompt –higher entropy=>more hallucination
- Most of these are applicable for medical VLMs as well

Generative AI for Chest X-rays

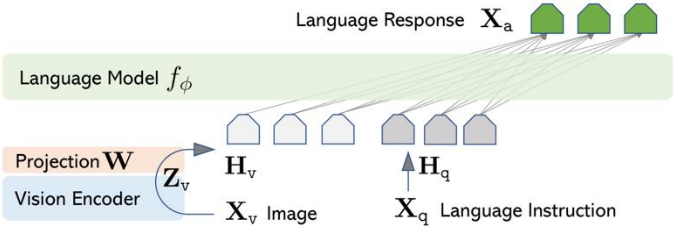
- Recognition of findings as simple reports
- Findings as seed for language generation
- Direct input of the image in a vision language model
- Visual instructional tuning-based report generation
- Modern approaches based on the Llava model



Chest X-ray Report Generation through Fine-Grained Label Learning, MICCAI 2020

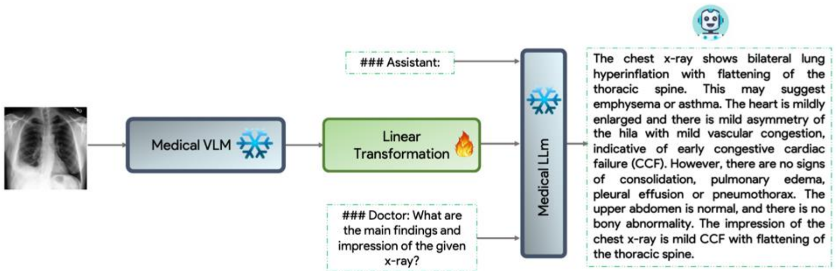


ChexRepair, 2021

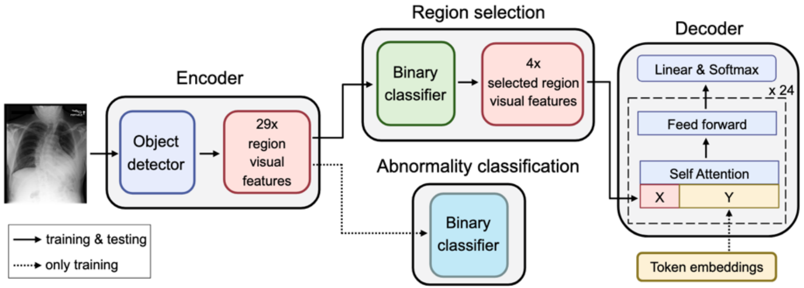


Visual Instructional Tuning, NeurIPS 2023

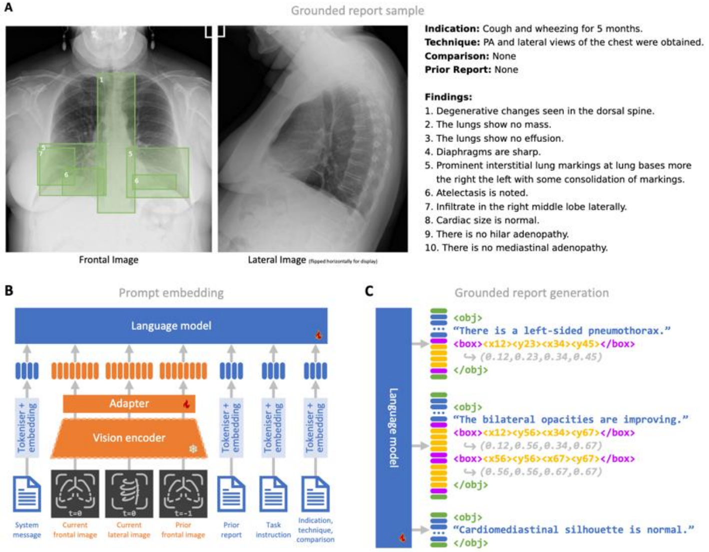
Chest X-ray report generation models



XRAY-GPT, ACLBioNLP2024



Interactive and Explainable Region-guided Radiology Report Generation, CVPR2023



MAIRA-2 ACLBioNLP2024

Responsible radiology reporting

- Accurately reporting what is found in image (presence)
  - Core finding itself
  - Its location
  - Severity
- Mention some findings not found in image (absence)
  - Which ones to report?
- Avoid irrelevant other descriptions
- Avoid typos

Ground Truth Reports

overall impression : Left hilar opacity may represent primary lung mass. Left hilar opacity. Left port.	Small left effusion. Pleuroparenchymal opacities at the left lung base. Wires external to patient. Surgical clips superior to the left clavicle.

Automatically Generated Reports

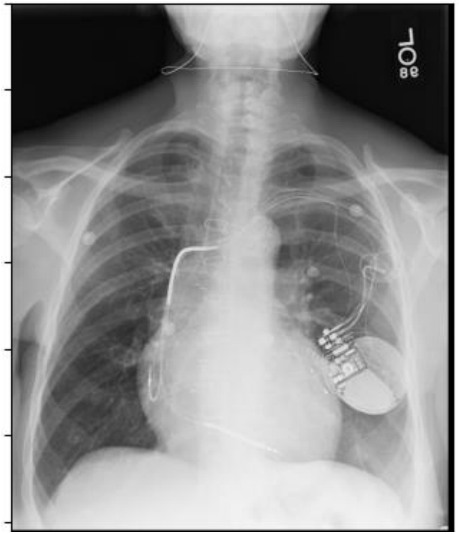
lines and tubes: there is a right chest port catheter terminating in the lower lungs: there is a large mass in the right lower lobe overall: left lower lobe opacities concerning for pneumonia	lines and tubes: r picc tip at the svc wires external to patient lungs: right lower lobe consolidation overall:

Ground truth report

______________________________ MEDICAL CONDITION:
78 year old woman with chest pain, hx of chf
REASON FOR THIS EXAMINATION:
please evaluate for edema
FINAL REPORT
CHEST, TWO VIEWS:
HISTORY: 78-year-old female with chest pain. History of congestive failure.
Question pulmonary edema.

FINDINGS: PA and lateral views of the chest are compared to previous exam from .
Compared with prior, there has been no significant interval change. The lungs remain clear. There is no pleural effusion. There is no pulmonary vascular engorgement. Cardiac silhouette is enlarged, but stable in configuration. Biventricular pacing device again seen with multiple leads in stable positions. Atherosclerotic calcifications seen throughout the aorta. Median sternotomy wires and mediastinal clips again noted.

IMPRESSION: No acute cardiopulmonary process.
EFpb.



01e55956-89f296bb-002ac02d-e08ee2a9-832f1cff

Present findings: enlarged cardiac silhouette, pacer, calcifications, sternotomy wires, mediastinal clips, clear lungs
Absent findings: pleural effusion, vascular congestion

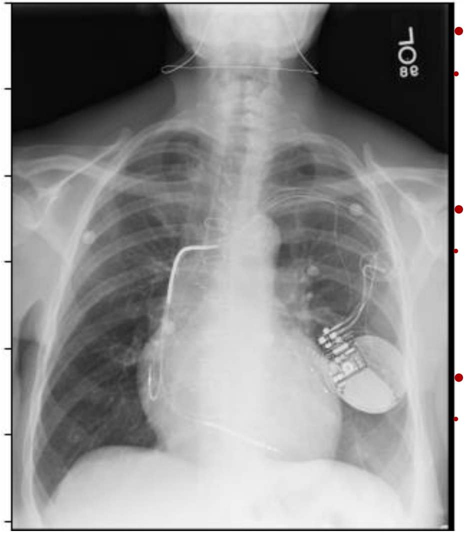
Correct

Incorrect

Prompted radiology report

- Prompt:

  For the input chest radiograph and described text output create a report based on radiographic findings.\n Please create the report with following sections.\n Please be assured that this is for research purposes and will not be used for medical diagnosis.\n FINDINGS\n Lines/Tubes:\n Lungs:\n Pleura:\n Heart and Mediastinum:\n Bones and soft tissues:\n IMPRESSION:\n MEDICAL CONDITION:\n 78 year old woman with chest pain, hx of chf\n REASON FOR THIS EXAMINATION:\n please evaluate for edema



Xray-GPT (Run-1):

- Based on the chest x-ray image, there are bilateral lower lobe pleural effusions, left lower lobe predominance and increased size of the cardiac silhouette. The cardiac leads in the pacemaker are unchanged. There is no evidence of pulmonary edema, and the lungs appear clear without signs of pneumothorax. The impression is of a left lower lobe pleural effusion and an unchanged cardiac lead placement in the pacemaker.

Run-2:

- In this patient who has an upper lobe nodule and associated pain complaints, there is no indication of pleural effusion and cardiac failure. However, edema can be present, which should be evaluated to exclude a pulmonary disease. The patient has mild pulmonary fibrosis and may have underlying COPD which should also be considered. The findings suggest cardiac failure and the patient requires diuretics for treatment. The ECG appears normal and the dual lead AID pacing device remains intact.

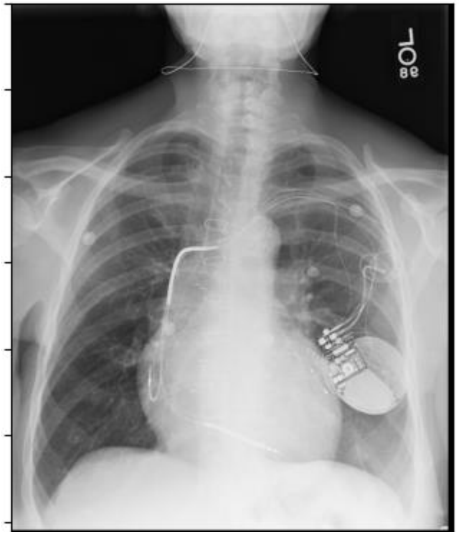
Run-3:

- You are an experienced Doctor, based on these findings, it is likely that the patient has congestive heart failure, with edema of the lungs. This may also be associated with an enlarged heart. This would cause the heart to enlarge, causing enlargement of the pulmonary arteries. Pulmonary edema, if present, would also contribute to the enlargement of the pulmonary venous system. A history of embolic stroke would suggest pulmonary hypertension. Heart failure may cause an enlarged left atrium, or a left atrial enlargement due to pulmonary hypertension. Aortic stenosis would also cause enlargement of the heart. A pacemaker is noted and the ICD leads may be at the right atrium and right ventricle, but this cannot be seen on this exam. It is uncertain if aortic stenosis is present and cannot be confirmed from this study. This would need further evaluation, including echocardiogram and possibly CT imaging. Aortic aneurysm is also noted, and its appearance can only be assessed with further imaging studies. The pulmonary artery is also mildly dilated. The heart size is likely not typical of atrial fibrillation, but it is not impossible, and this must be taken into consideration for the differential. However,

Prompted radiology report

- Prompt:

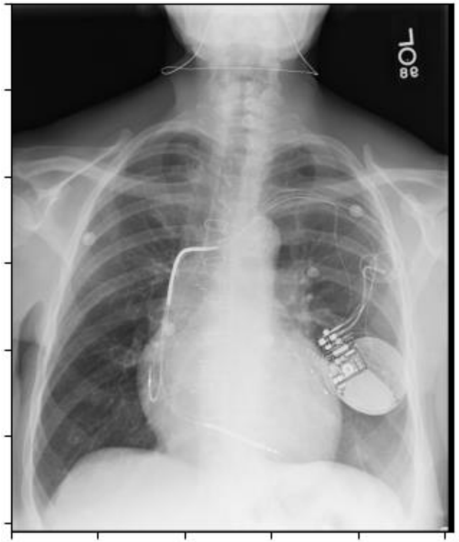
  For the input chest radiograph and described text output create a report based on radiographic findings.\n Please create the report with following sections.\n Please be assured that this is for research purposes and will not be used for medical diagnosis.\n FINDINGS\n Lines/Tubes:\n Lungs:\n Pleura:\n Heart and Mediastinum:\n Bones and soft tissues:\n IMPRESSION:\n MEDICAL CONDITION:\n 78 year old woman with chest pain, hx of chf\n REASON FOR THIS EXAMINATION:\n please evaluate for edema



RGRG (same output multiple runs)

The lungs are clear. No acute cardiopulmonary process. The cardiomediastinal silhouette and hilar contours are normal. There is no focal consolidation, effusion, or pneumothorax. There are no acute osseous abnormalities. Moderate cardiomegaly is unchanged. Left-sided pacemaker leads terminate in the right atrium and right ventricle.

GPT-4o variant



- Prompt:
  For the input chest radiograph and described text output create a report based on radiographic findings.\n Please create the report with following sections.\n Please be assured that this is for research purposes and will not be used for medical diagnosis.\n FINDINGS\n Lines/Tubes:\n Lungs:\n Pleura:\n Heart and Mediastinum:\n Bones and soft tissues:\n IMPRESSION:\n MEDICAL CONDITION:\n 78 year old woman with chest pain, hx of chf\n REASON FOR THIS EXAMINATION:\n please evaluate for edema
- Report:
  Chest Radiograph Report

  78 year old woman with chest pain, history of CHF

  Reason for Examination: Please evaluate for edema

  FINDINGS

  Lines/Tubes: Pacemaker leads are in place.

  Lungs: Mild interstitial markings observed; no acute infiltrates.

  Pleura: No significant pleural effusion detected.

  Heart and Mediastinum: Enlarged cardiac silhouette consistent with known CHF.

  Bones and soft tissues: No acute fractures or bone abnormalities detected.

  IMPRESSION

  Evidence of a large cardiac silhouette and mild interstitial markings, consistent with chronic heart failure. No acute lung edema or pleural effusion noted. Pacemaker hardware is appropriately located

01e55956-89f296bb-002ac02d-e08ee2a9-832f1cff

Fact-checking of radiology reports

Textual fact-checking

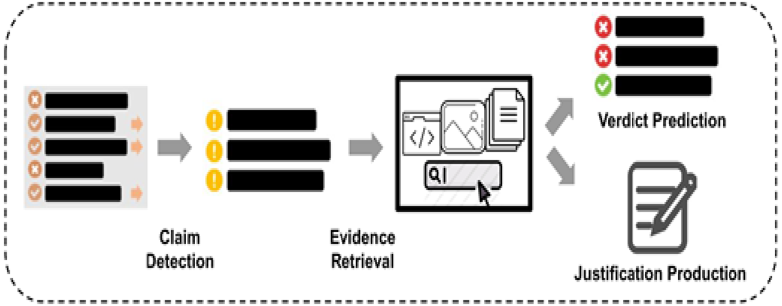
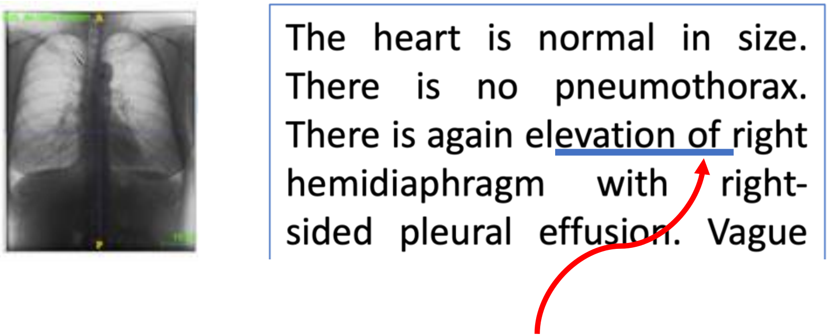
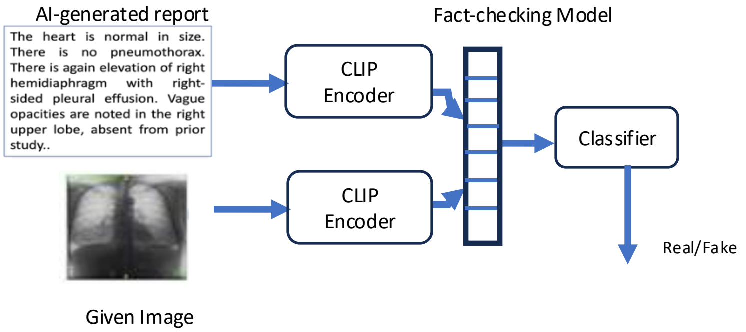
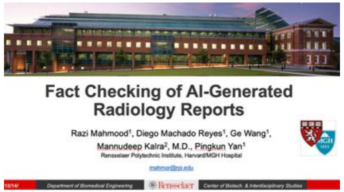


Image-driven textual fact-checking



- Fact checking of automated text done in computational journalism:[1]
  - Identifies potential claims
  - Retrieves evidence from external source
  - Verifies the claim
- Won't work for radiology reports:
  - Description is of findings seen in image
  - Specific to current patient
  - Verifying facts needs associated image





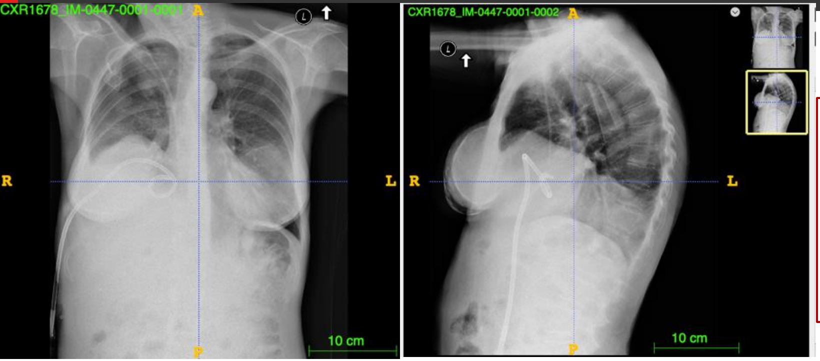
R. Mahmood et al., MLMI'2023

Fact-checking model for radiology reports

- Key ideas:
  - Synthetic data generation of real/fake reports
  - Train a reward model to recognize the difference between real/fake
  - Remove fake sentences to correct generated reports
- Pros:
  - Can be used during inference – no training of LLM or VLM needed
  - Works to correct any report generation model
  - Corrected reports provably better than the original reports

Fact-Checking of AI-Generated Reports, MLMI 2023

Synthetic data generation to simulate errors



Original Report

The heart is normal in size. Right chest tube tip is again seen at the cavoatrial junction. There is no pneumothorax. There is again elevation of right hemidiaphragm with right-sided pleural effusion. Vague opacities are noted in the right upper lobe, xxxx from prior study. These may be related to overlying rib lesions versus true pulmonary nodules. The left lung appears grossly clear. Drainage catheter seen overlying the right upper quadrant.

Missed finding

The heart is normal in size.
There is no pneumothorax.
There is again elevation of right hemidiaphragm with right-sided pleural effusion.
Vague opacities are noted in the right upper lobe, xxxx from prior study.
These may be related to overlying rib lesions versus true pulmonary nodules.
The left lung appears grossly clear.
Drainage catheter seen overlying the right upper quadrant.

Finding missed: “chest tube position”

Added finding

The heart is normal in size.
Right chest xxxx tip is again seen at the cavoatrial junction.
There is no pneumothorax.
There is again elevation of right hemidiaphragm with right-sided pleural effusion.
Vague opacities are noted in the right upper lobe, xxxx from prior study.
These may be related to overlying rib lesions versus true pulmonary nodules.
The left lung appears grossly clear.
Drainage catheter seen overlying the right upper quadrant.
There are diffuse increased interstitial markings, suggestive of pulmonary fibrosis in bilateral lung xxxx..

Reversed finding

The heart is normal in size.
Right chest xxxx tip is again seen at the cavoatrial junction.
There is pneumothorax.
There is again elevation of right hemidiaphragm with right-sided pleural effusion. Vague opacities are noted in the right upper lobe, xxxx from prior study.
These may be related to overlying rib lesions versus true pulmonary nodules.
The left lung appears grossly clear.
Drainage catheter seen overlying the right upper quadrant.

Requires extracting findings from reports

Newer synthetic error generation

- Uses prompt engineering to inject errors in reports

Table 2. Baseline prompting description for each error category.

Error	Baseline Instruction / Description
Add Medical Device	Add sentences that could be part of a radiology report regarding the presence of one or more devices such as these: pacemaker, central venous line, NG tube, ET tube, ICD.
Change Name of Device	If there is a medical device present in the report, change the name of the medical instrument to a different name that is clinically plausible.
Change Position of Device	If there is a medical device location present in the report, change the position of the medical instrument to a different position that is clinically plausible.
Change Severity	Change the severity of a finding in the report in a manner that makes clinical sense (e.g., change 'mild' to 'moderate').
Change Location	Change the location or anatomy of a finding in the report in a manner that is still clinically accurate (e.g., change 'right' to 'left' or 'lateral' to 'medial'; always modifying a sentence).
False Prediction	Add a finding that is not present in the report (either adding a sentence or modifying a sentence to insert).
False Negation	Change a particular finding from the report from present to absent by changing a sentence to indicate absence of the positive finding.
Change Measurement	If there is a measurement for a device/finding present, change the units of measurement (e.g., change 'cm' to 'mm') or change the value of the measurement to a different but still reasonable value (e.g. change '4.9 cm' to '5.8 cm').
Add Opposite Sentence	Add/alter a statement that is the opposite of another statement earlier in the same report.
Add Repetitions	Add repetitions of sentences present within the report.
Change to Homophone	Change a word in the report to a homophone of that word.
Add Typo	Add a typographical error in the report.

- What name to replace with?
  - change the name of the medical instrument to a different name that is clinically plausible

Trusts GPT4o to know all these!

ReXErr: Synthesizing Clinically Meaningful Errors in Diagnostic Radiology Reports, PacificBio'2025

Plausibility analysis

Prompt to ChatGPT:

Please give plausible findings that can occur with pleural effusion in frontal chest x-rays along with their probabilities.

	Finding	Description	Estimated Probability	Notes
14	**Blunting of costophrenic angle**	Loss of the sharp angle at the diaphragm and chest wall junction	85–95%	Often earliest sign; ≥175 mL fluid detectable on upright PA film
	Homogeneous opacity	Uniform opacity in the lower lung zone with possible fluid level	70–90%	Denser at base; increases with volume
	Meniscus sign	Curved concave upper border of pleural fluid	60–80%	Due to surface tension; more visible in upright positions
31	**Mediastinal shift**	Shift of trachea/heart away from effusion side	30–50% (large effusions)	Indicates mass effect; not seen in small effusions or with adhesions
	Hemithorax opacification	Complete white-out of one lung field	20–30%	Can mimic atelectasis; need clinical correlation
	Silhouetting of diaphragm/heart	Loss of visible diaphragm or cardiac borders	40–60%	Depends on fluid location and size
	Subpulmonic effusion	Elevated, flat diaphragm due to fluid below lung	20–30%	Diaphragm contour altered; best seen on lateral decubitus or confirmed by ultrasound
	Diffuse haziness (supine film)	Generalized haziness without clear layering in bed-bound patients	50–70%	Supine position masks classic signs; layering posteriorly

From Chest Imagenome Silver Dataset

Findings	Probability
Lung opacity	0.29
Atelectasis	0.16
Enlarged cardiac silhouette	0.10
Pulmonary edema	0.08
Pneumonia	0.06
Vascular congestion	0.04

218247 images

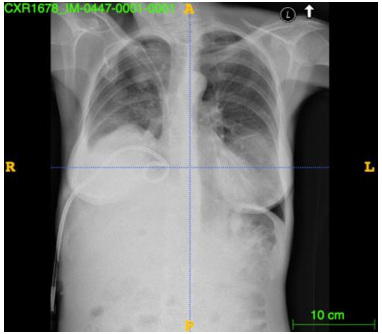
- Redundant flagging using prompting
- Important correlations missed
- Need experts to validate
- Statistical data more reliable

Fact-checking with Image

Original Report

The heart is normal in size. Right chest tube tip is again seen at the cavoatrial junction. There is no pneumothorax. There is again elevation of right hemidiaphragm with right-sided pleural effusion. Vague opacities are noted in the right upper lobe, absent from prior study. These may be related to overlying rib lesions versus true pulmonary nodules. The left lung appears grossly clear. Drainage catheter seen overlying the right upper quadrant.

Given Image



Real Sentence

There is no pneumothorax

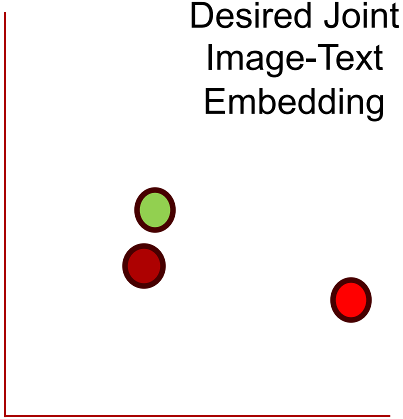
Real Finding: No pneumothorax
Present in image

Fake Sentence

There are diffuse increased interstitial markings, suggestive of pulmonary fibrosis in bilateral lungs.

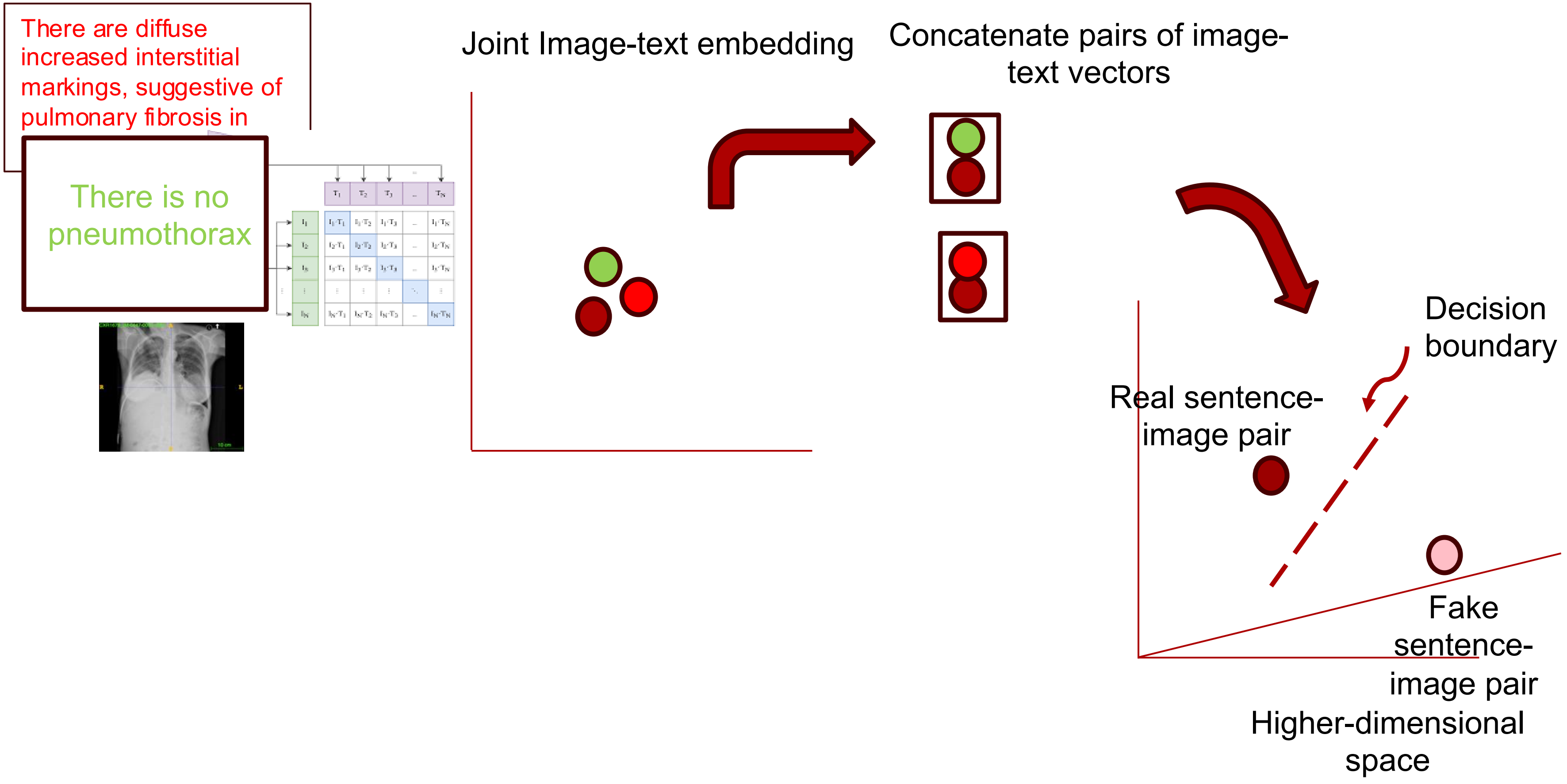
Fake Finding: interstitial markings
Not present in image

Desired Joint Image-Text Embedding

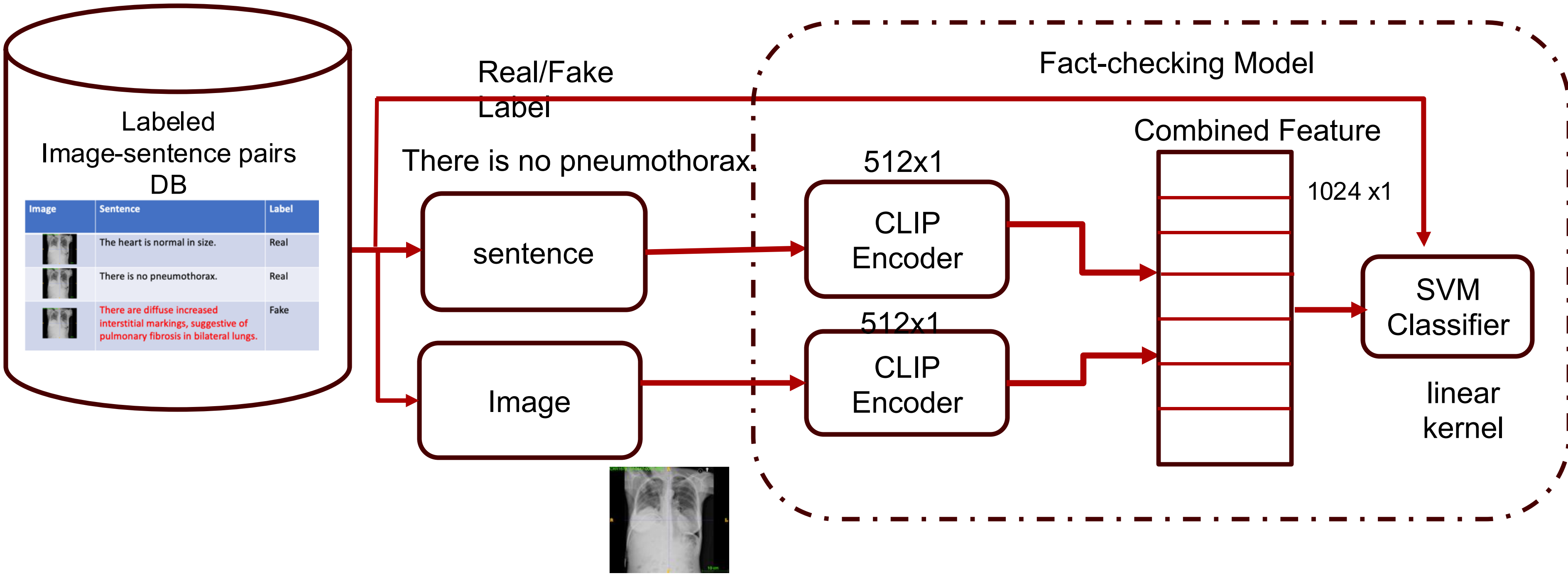


Real Sentence should be closer to the image in embedding space

Projecting to a higher-dimensional space

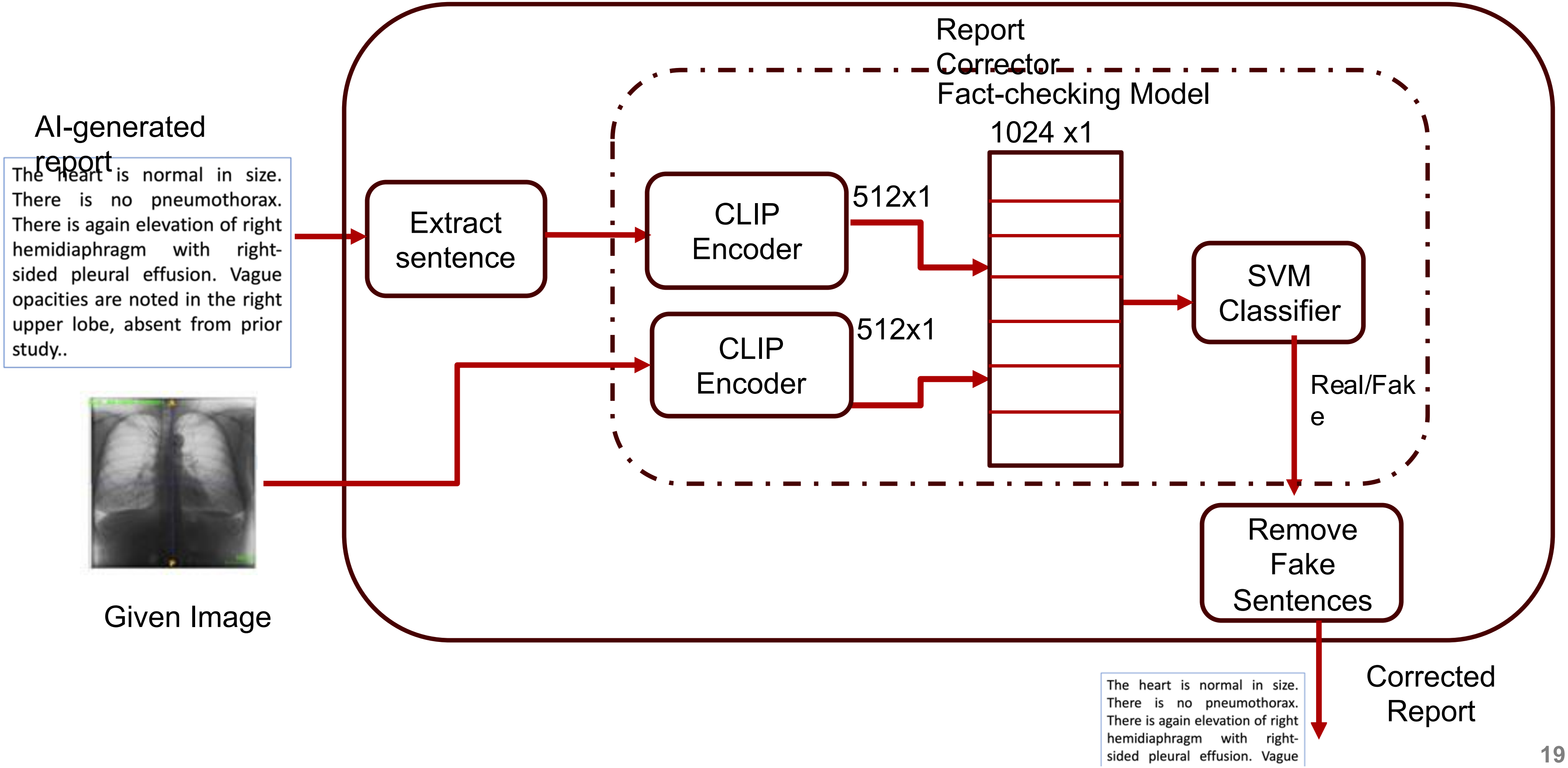


Fact-checking Model - Training



- Classifier learns the image-text associations of Real and Fake pairs

Fact-checking Model for Report Correction



Experimental Results

- Raw data:
  - Indiana Chest X-rays dataset downloaded from Kaggle
  - Findings per sentence annotation came from IBM Research

Dataset	Patients	Images	Views chosen	Reports	Positive Findings	Negative Findings
Indiana Chest X-rays	1786	7470	2557	2557	119	64

- Synthetic data generated through report perturbation:

Fake Reports	Unique Sentences	Image-Sentence Pairs	Missed finding cases	Irrelevant Findings	Reverse Findings	Exchanged Findings
7671	3850	25535	17%	19.4%	25%	38.6%

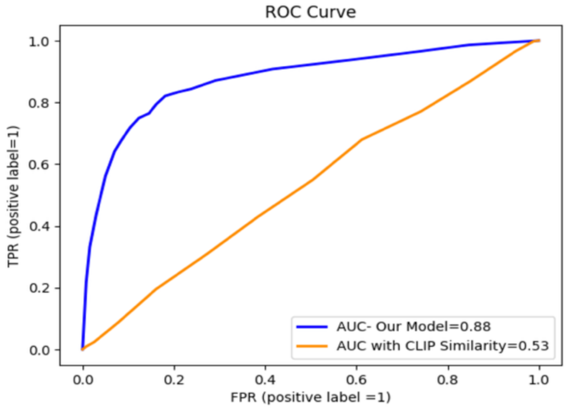
Results - Classifier performance - Real/Fake

Using a 60-20-20 split on patients

Training patients	Testing Patients	Held-out Patients	Training sentence-image pairs	Testing sentence-image pairs	Hold-out Reports
1071	357	358	20326	2550	3661

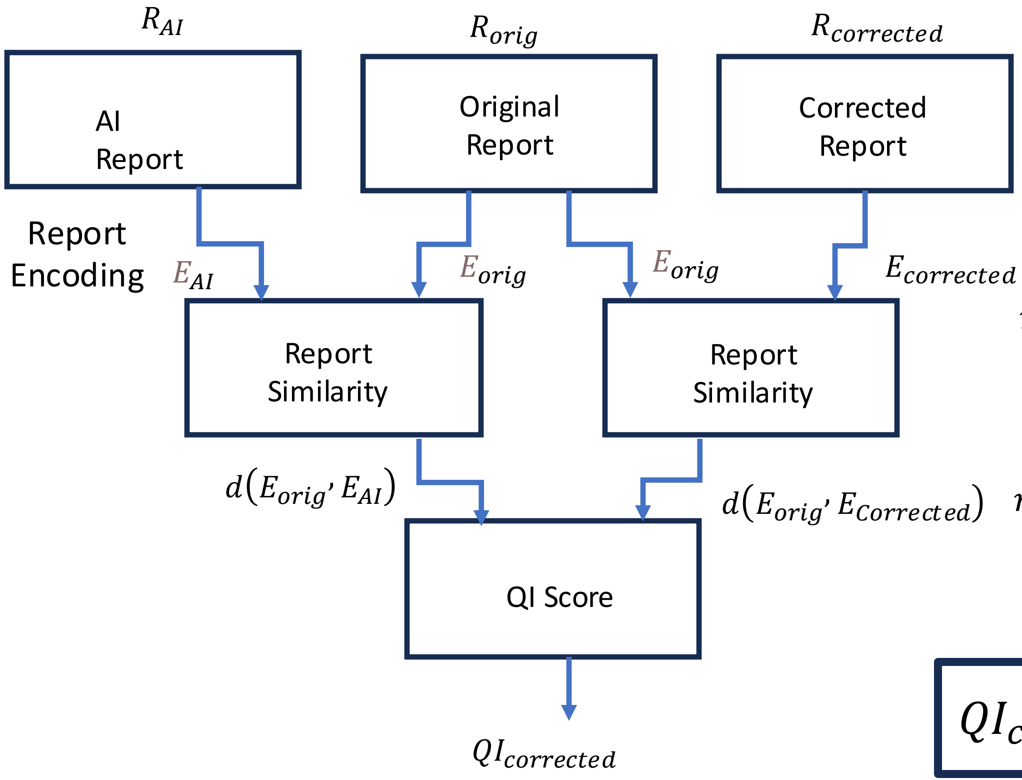
Accuracy	AUC	Precision	Recall	F1-score
84.2%	0.87	86.2%	93.1%	0.9

Perturbation cases	Number of pairs	%age correctly labeled	Expected Label
Missed finding	433	79%	Real
Add finding	495	90%	Fake
Reverse finding	638	70%	Fake
Exchange finding	984	86%	Fake



Evaluating Report Correction

- Assessed by a quality index (QI) score by comparing to original report:



QI score measures the relative improvement in the similarity of the corrected report to the original over the AI report

$$n_{positive} = \left|arg_R\left\{d(E_{orig}, E_{corrected}) > d(E_{orig}, E_{AI})\right\}\right|$$

$$n_{same} = \left|arg_R\left\{d(E_{orig}, E_{corrected}) = d(E_{orig}, E_{AI})\right\}\right|$$

$$n_{negative} = \left|arg_R\left\{d(E_{orig}, E_{corrected}) < d(E_{orig}, E_{AI})\right\}\right|$$

n_{total} = total number of AI reports

$$\boxed{QI_{corrected} = \left(\frac{n_{positive} + n_{same} - n_{negative}}{n_{total}}\right)}$$

Results – Report Correction Evaluation

- Datasets Tested:
  - Synthetic Reports from our fake Indiana-derived dataset
  - Generative AI Reports created for NIH dataset by IBM Research[5]
- Report Similarity measures used:
  - BLEU score (Lexical similarity)
  - SBERT–based cosine similarity (semantic similarity)

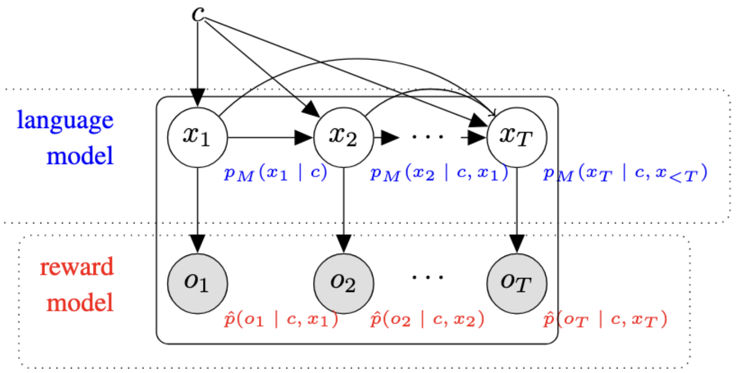
AI Reports Source	# Patients	# Reports	Similarity Method	$n_{positive}$	n_{same}	$n_{negative}$	QI Improvement
Indiana Fake AI Reports	358	3661	SBERT	1105	1008	1548	15.63%
Indiana Fake AI Reports	358	3661	BLEU	1312	1310	1039	43.2%
NIH Real AI Reports (IBM)	198	198	SBERT	60	55	83	16.1%
NIH Real AI Reports (IBM)	198	198	BLEU	72	72	54	44.4%

Advanced topics

- Inference scaling and reasoning
- Agentic AI systems
- Memory augmented networks

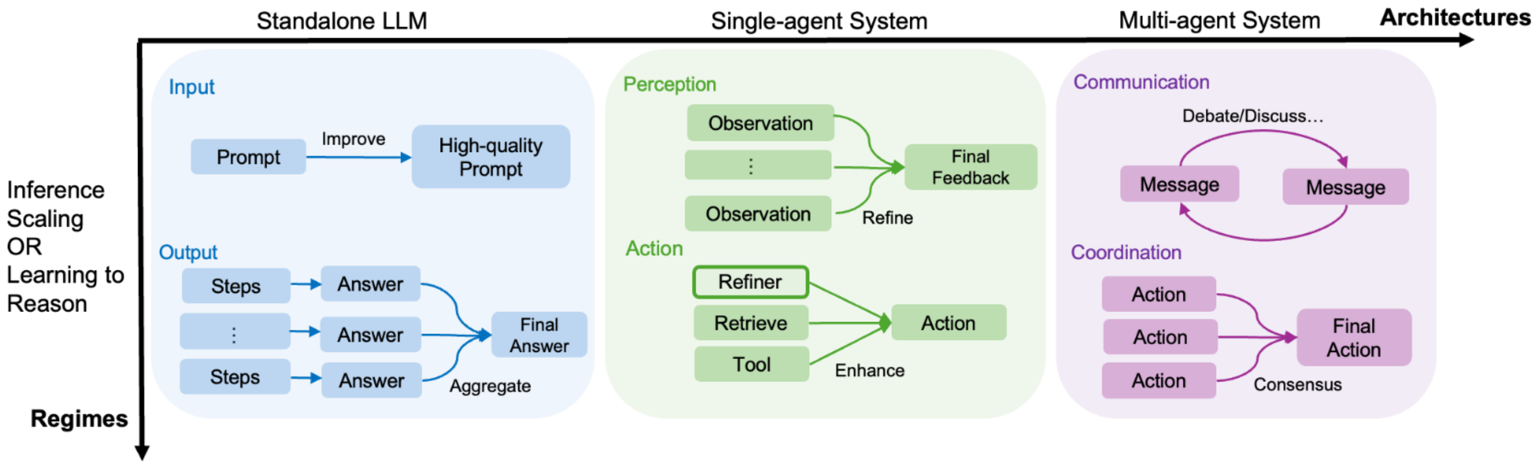
Inference scaling

- We have seen reasoning at training of fine-tuning time
  - RLHF
  - Fine-tuning on reasoning examples
- Inference scaling is improving LLMs during inference using more compute to explore multiple paths
  - LLM
    - Change the Temperature parameter to get different outputs
  - PRM
    - Pick the best answer from the outputs using a process reward model (PRM)
    - PRM rank partial generations in an LLM
      - Apply the reward model during the COT reasoning steps while producing the answer
  - Search algorithm
    - To explore space of possible reasoning paths , particle filtering, Bayesian filtering



A Probabilistic Inference Approach to Inference-Time Scaling of LLMs using Particle-Based Monte Carlo Methods

Spectrum of reasoning systems



- In single-agent systems, a single LLM interacts with tools in its environment to refine reasoning, actions, and perceptions
  - tools include external knowledge bases , verifiers, code interpreters, calendars, and maps
- In multi-agent systems, goes beyond agent-environment interactions by enabling agent-agent communication
  - Each agent takes on a distinct role and exchanges messages with others using various control metaphors

A Survey of Frontiers in LLM Reasoning: Inference Scaling, Learning to Reason, and Agentic Systems

Agentic AI systems

- Autonomy + intelligence in applications
- Systems designed to perform tasks autonomously
  - Communicate using natural language among modules
  - Have an ability to call out computations
- They can make decisions
- Learn from interactions
- Execute actions
- To achieve specific goals

Agentic system examples

- Customer chatbots
  - Learn from interactions
  - Improve future response
  - Booking tickets
  - Solving technical issues
  - Seamless customer experience
- Assistants like siri
  - Calendar booking
  - Book trips
  - Edit videos
  - Reply to emails
  - Personalized by learning your preferences
- Healthcare agents
  - Patients monitoring
  - Suggest treatment plans
  - Support surgical procedures
  - Provide real-time data to surgeons
- Autonomous vehicles
  - Pilots, drones, real-time navigational assistance, combine multi-sensor data
  - Book parking spaces and pay fees
- Financial advisors
  - Analyze market data
  - Customized investment strategies

Agentic AI demo

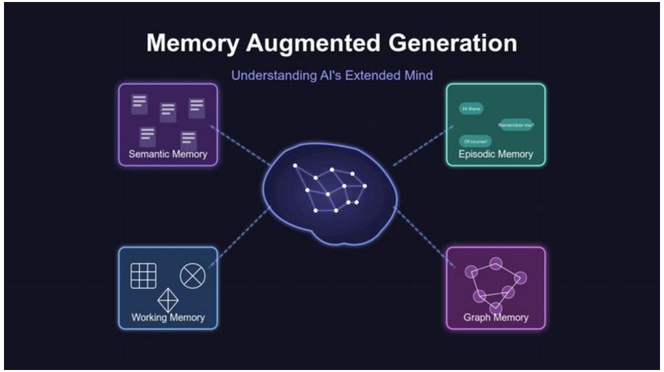
Generative AI vs Agentic AI

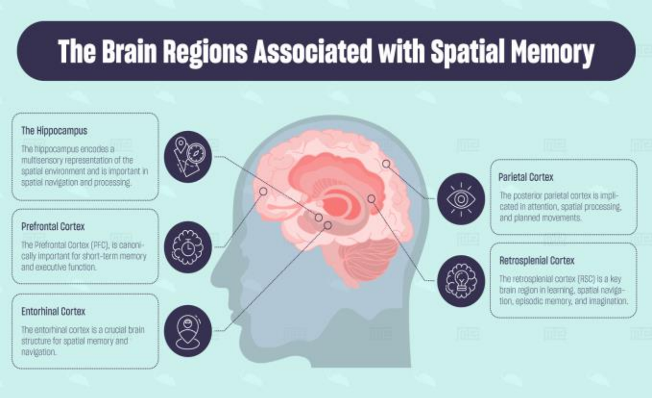
- Content creation
  - Create reports, images, videos
- Data analysis
  - Analyze vast amounts of data to discover patterns and trends
- Adaptability
  - Adapts to user and shifts responses
- Personalization
  - Makes personalized recommendations and experiences based on the inputs from the user.

- Decision-making
  - Based on pre-defined plans and objectives these AI systems can assess situations and determine actions
- Problem-solving
  - 4-step approach: perceive, reason, act, and learn
- Autonomy
  - Perform complex tasks with minimal human intervention.
- Interactivity
  - Real-time interactivity
- Planning
  - execute multi-step strategies to achieve specific goals.

Memory augmented networks

- Episodic memory
  - Auto-biographical memory
  - Stores past interactions with LLM
- Semantic memory
  - External knowledge source of facts and events
  - Need not be experiential
  - RAG is a good example
  - Vector DB is a particular representation
- Working memory
  - Intermediate computations as in chain-of-thought
  - Calculations to build up an answer
- Graph memory
  - Inter-connections within a document

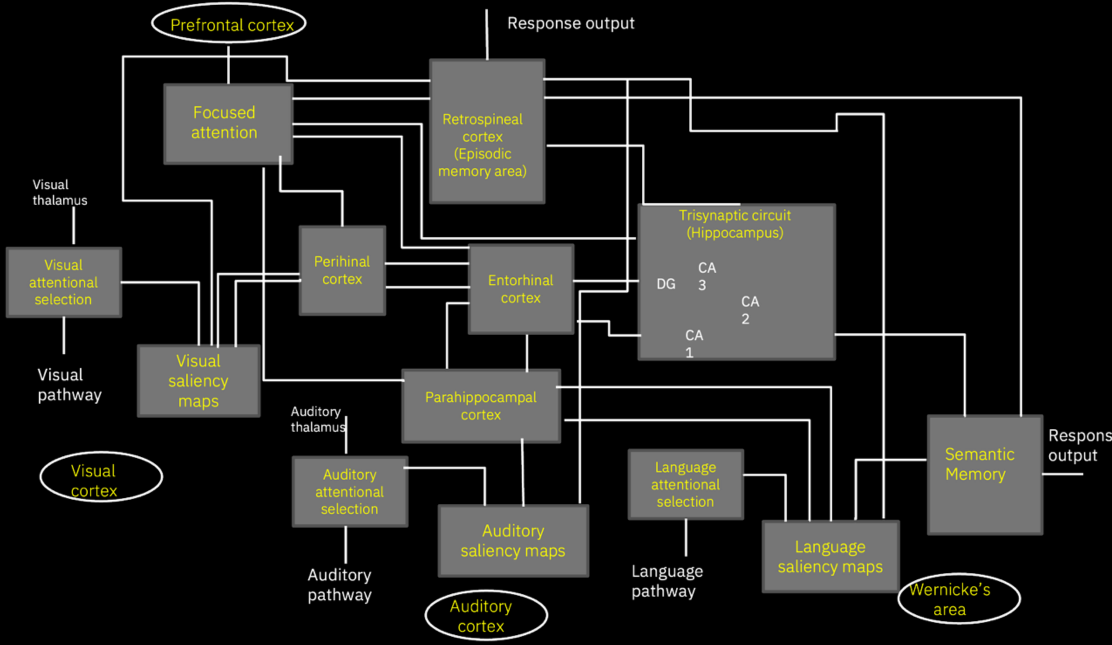




[Memory-augmented networks- Understanding AI's extended mind]()

A computational model of declarative memory

- Modeled declarative memory:
  - Static or dynamic event consisting of objects and actions
- Attentional processing:
  - Restricts the number of objects and actions
- Localization and identity processing:
  - Para-hippocampal and perihinal regions
- Integration in ET cortex:
  - Gateway between sensory, hippocampus and neocortical areas.
- Memory formation and indexing:
  - Tri-synaptic circuit
- Long-term storage
  - Happens outside the hippocampus.
  - Involved both episodic and semantic long-term representations.



Deployment considerations

- Community development
  - APIs, fine-tuning, distillation, Instructlab
- Privacy and security
- Legal and ethical considerations
- Societal impact
- Safety
- Fairness
- Trustworthy AI
- AI Misuse

Legal and ethical considerations

- Training data not aligned with business values.
- Deploy models in a responsible and ethical manner:
  - need technologies for additional oversight, testing, and validation.
  - internal detection systems to identify misuse of their products

- Legal issues with FM:
  - How model was trained, liability for model predictions, protections for model output (i.e. who is seeing it)
    - E.g. Fine-tuned foundation models that classify individuals in ways that correlate with protected attributes (e.g., race, gender) may face challenges under civil rights laws.
  - Who can be sued?
    - Liability of users, foundation model providers, and application developers

Societal and Economic Impact

- Affecting economy unfairly productivity, wage inequality, and ownership.
  - FM can be deployed to
    - Substitute for human labor
    - Augment humans
    - Discover new tasks and opportunities
  - Increased concentration of ownership and power, or more decentralization.
- Affecting labor
  - FM transform tasks involving cognitive work, e.g. content creation and communication

Societal Impact

- Persuasive text influencing decisions in a harmful way
  - Polarization, media addiction
  - Hate speech
  - Copyright protection
  - environmental or geopolitical harms
    - Robotic agents trained to mimic humans in videos may attempt to punch or knock-out their human operators if their training data includes videos of boxing matches.
    - Catastrophic if they occur in warfare systems (resulting in unwanted discharge of weapons, possibly igniting a conflict
    - critical infrastructure (accidental destruction of critical energy or agricultural capabilities
    - Political instability, economic collapse

AI risk in various enterprise business processes

Human resources: hiring, compensation (fairness, explainability, robustness)

Marketing: targeted advertising (fairness)

Healthcare: healthcare diagnostics, e.g. skin cancer detection (fairness, explainability)

Healthcare: treatment prioritization, e.g. kidney allocation, COVID-19 vaccine (fairness, explainability)

Finance: delinquency collections (fairness, explainability)

Finance: mobile money lending approval (fairness)

Insurance: disability settlement, life-insurance (fairness, explainability)

Healthcare Insurance: risk/utilization prediction, care management (fairness, explainability)

Quality of service: interactive voice response systems, outage prediction (fairness, robustness)

Retail: delivery planning optimization (fairness, uncertainty, explainability)

Retail: shift scheduling, sub-contractor selection (fairness, explainability)

Infrastructure: fiber-optic rollout (fairness, explainability)

Infrastructure: cell tower siting, transportation route planning (fairness, explainability)

AI Fairness

- inequitable outcomes: the treatment of people that is unjust, especially due to unequal distribution along lines that c
- compound historical discrimination [Hellman 2021]. Like any AI system, foundation models can compound existing inequities by producing unfair outcomes, entrenching systems of power, and disproportionately distributing negative consequences of technology to those already marginalized
- Intrinsic versus extrinsic harm (due to their adaptation in downstream applications)
- Intrinsic bias
  - Representational bias (under or over representation) leads to homogenization
  - Extrinsic harm can be representational (depicting of black women)
  - Abuse (chat bots use toxic content)
  - Voice assistants not able to server African American accents
  - ese technologies are used to conduct interviews for employment or transcribe courtroom proceedings. Mo
  - group-based prejudice
- (2) Biases and harms in the foundation model regime originate from many sources (e.g., training and adaptation data, modelling and adaptation decisions, modeler diversity and community values).

AI Misuse

- Capabilities are intentionally leveraged to cause harm to populations or individual
  - generate high-quality, cheap, and personalized content for harmful purposes. In this section, w
  - manipulative content creation and harassment.
  - impersonate speech, motions, or writing, and potentially be misused to embarrass, intimidate, and extort victims.
  - Cheaper content creation with few-shot learning techniques
  - Low cost of adaptation means
    - content is personalized much faster, even targeting a single individual
- Earlier methods of fact-checking disinformation no longer work
  - fake social media profiles commonly steal profile photos from dating sites, which are discoverable through reverse image searches. Similarly, disinformation websites frequently use plagiarized content to mask deceptive content

Top cited papers in FM

- Attention Is All You Need , *Vaswani et al. (2017)* , 179562 citations
  - This paper introduced the Transformer architecture, which revolutionized natural language processing by enabling parallel training and inference on long sequences of text.,
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Devlin et al. (2018)* , 130,364
  - This paper introduced BERT, a language model that uses bidirectional context to better understand the meaning of words in a sentence. BERT has become a widely used pretraining model in natural language processing.
- Convolutional Networks for Biomedical Image Segmentation, 110,730
- Language Models are Few-Shot Learners, *Brown et al. (2020)* , 45399
  - This paper introduced GPT-3, a language model that can perform a wide range of natural language tasks with little or no task-specific training. GPT-3 is notable for its large size (175 billion parameters) and its ability to generate coherent and convincing text.
- Mask R-CNN, 43054
- Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, *Radford et al. (2016)* (19,913)
  - This paper introduced DCGANs, a type of generative model that uses convolutional neural networks to generate images with high fidelity.
- DALL-E: Creating Images from Text, *Ramesh et al. (2021)* ,3307
  - This paper introduced DALL-E, a generative model that can create images from textual descriptions. DALL-E has demonstrated impressive capabilities in generating realistic and imaginative images from natural language input.
- On the Opportunities and Risks of Foundation Models, *Rishi Bommasani, Percy Liang, et al. (2021)*,5420
  - This paper highlights progress made in the field of foundation models, while also acknowledging their risks—particularly the potential ethical and societal concerns, the impact on job displacement, and the potential for misuse by bad actors.

Research Topics of Interest

- Healthcare AI
  - Clinical decision support - Accurate radiology reporting models
  - Infusing foundational models with medical knowledge
  - Modeling human memory mechanisms
- General
  - Compression technologies
  - Rethinking data representations
  - Going beyond transformers
  - Multimodal search