SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Evaluating the Accuracy and Safety of Patient-friendly Complete Radiology Reports using a HIPAA-compliant Large Language Model

Juan Serna, University of California, San Francisco; Jae Ho Sohn, MD, MS; Yannan Yu, MD

Introduction/Background

The 21st Century Cures Act grants patients immediate electronic access to radiology reports. However, the reports are written with medical terminology, which can be a source of anxiety and confusion for patients. In previous studies, Large Language Models (LLM) have generated patient-friendly summaries; however, the accuracy and safety in translating the entire report has not been evaluated. We investigated whether LLMs could accurately simplify complex CT/MRI reports to a desired education level.

Methods/Intervention

We manually crafted twenty synthetic radiology reports with common outpatient findings using the UCSF mPower database. Three prompt strategies for generating patient-friendly reports at an eighth-grade education level were evaluated using OpenAI's o3-mini model via the HIPAA-compliant UCSF Versa API (Table 1). Each LLM output was assessed blindly by a radiologist using custom 3- and 5-point Likert scales for metrics including accuracy, understandability, and hallucinations (Tables 2&3). Readability was measured via Flesch-Kincaid Grade-Level. Statistical analysis was performed using R v4.5.1.

Results/Outcome

Twenty reports included 99 unique benign/malignant findings. The average accuracy for patient-friendly reports was 4.85±0.37 (prompt1), 4.9±0.31 (prompt2), and 4.95±0.22 (prompt3), with no significant difference between prompts. The understandability improved from 3.85±0.75 (prompt1), to 4.15±0.67 (prompt2) and 4.4±0.68 (prompt3), where the difference between prompt 1 and 3 was significant ($p=0.04$). Flesch-Kincaid grade-level dropped from 13.2±1.3 for original reports to 8.95±0.32 (prompt1), 7.60±0.32 (prompt2), and 7.64±0.24 (prompt3) ($p< 0.001$ for prompt2 and 3 vs 1). For each prompt, there was 1 finding that had clinically significant hallucinations. Across all findings, medically accurate descriptions were achieved for 89.2% (prompt1), 92.3% (prompt2), and 94.6% (prompt3), with no significant difference.

Conclusion

LLMs demonstrate high accuracy for converting full radiology reports into patient-friendly language, regardless of prompt techniques. More sophisticated prompts yield a more understandable report at the target education level. However, potentially clinically significant hallucinations warrant further evaluation before large-scale clinical application.

Statement of Impact

LLMs accurately simplify complex radiology findings into accessible language, thereby enhancing patient understanding and promoting health literacy.

Table 1. Candidate prompts for patient-friendly report translation

Prompt 1	*A simple prompt a patient might use on their personal reports.* "Explain this radiology report to me."
Prompt 2	*A more sophisticated prompt using few-shot, chain-of-thought, and structured output techniques.* "You are a medical professional and your task is to simplify radiology report to a patient who has education level of eighth-grade. These are the requirements of the task: 1. Think about the task step-by-step. 2. Extract all the medical terminologies in the report and explain them in plain language at patient's education level, but do not include this part in the final output. 3. Then using the info above, you should explain all the items mentioned in the findings and impression section, and should not miss any important information in the report. Be concise, but clearly explain medical terminologies. 4. Be aware of the level of confidence in the report language and deliver these nuances to the patient. For example, 'malignancy cannot be excluded' vs 'concern for malignancy' has a different level of confidence. 5. Make sure to deliver the next step recommendation only if mentioned in the report; otherwise, you can fill in with 'please discuss with your doctor regarding the next step plan'. 6. After simplifying the report, check for any factual errors and correct them. Check any hallucinations that did not existed in the original report and correct them. 7. Always advise the patient to contact the provider if they have any questions or concerns regarding the report. 8. Make sure that the simplified report also follows the format of the original report, which includes clinical history, technique, findings in each organ section and impressions. At the end of the simplified report, provide a concise summary. Example: Findings: 'There is a 2.5 cm LIRADS-5 lesion in the hepatic segment 8…' Impression: 'LIRADs-5 lesion, which is concerning for hepatocellular carcinoma.' Simplified report: Findings: 'There is a 2.5 cm lump in the right upper part of your liver. LIRADS is a grading system for liver ranging from 1 to 5, with 5 representing the most suspicious category for cancer.' Impression: 'The imaging features are suspicious for liver cancer.'"
Prompt 3	*A more sophisticated, highly-structured few-shot prompt that builds a "medical entity extractor" internally in the LLM API to define complex medical terminology prior to translating the report.* ***Medical Entity Extractor*** **"Part 1:** From the following radiology report, extract a comma-separated list of all potentially complex medical and anatomical terms that a layman with a eighth-grade reading level would not understand. **IMPORTANT**: You MUST add all terms with directional components (i.e. 'medial,' 'lateral,' 'inferior,' 'superior,' etc.) to the list. **IMPORTANT**: Ensure that multi-word anatomical or medical terms are kept TOGETHER. For example, for the term 'anterior thoracic osteophytes' you MUST list the three-word term rather than just 'osteophytes'. You MUST do this for all terms. **IMPORTANT**: Additionally, if a medical term is composed of words that are individually simple (i.e. 'connective tissue disease'), you MUST include that term in the list, as the complete term is likely to be a term that a layperson would NOT recognize. **Part 2:** entity_critique_prompt = You are a terminology auditor. From the following candidate list, create a new comma-separated list that includes ONLY the terms that are genuine medical jargon or specific anatomical locations a layperson with a eighth-grade reading level would not understand. **INCLUDE:** Specific findings, diseases, technical anatomical descriptions, and specific locators like vertebral levels (e.g., 'T4-T6'). For example, for the term 'anterior thoracic osteophytes' you MUST list the three-word term rather than just "osteophytes". You MUST do this for all terms. **CRITICAL RULE**: ALWAYS INCLUDE 'benign', 'cyst', and similar terms in the final list of entities if present in the original report. **CRITICAL RULE** **EXCLUDE:** Very common terms like 'lungs', 'heart', 'abdomen', 'kidney', etc. Your response must ONLY be the final, refined, comma-separated list. **Part 3:** In one simple sentence, what does '{entity}' mean in a medical context? **Part 4:** You are a senior medical editor reviewing definitions for a patient education guide. Below is a JSON object of medical terms and their AI-generated definitions. Your task is to review each definition for accuracy, simplicity, tone, and conciseness, and improve it if necessary.

Critique Criteria:
1. **Accuracy:** Is the definition medically correct?
    - ***KEEP IN MIND anatomical context*** For instance 'thoracic' may refer simply to the chest, OR may refer to the thoracic SPINE if in the proper context.
2. **Simplicity:** Is it easy for a layperson to understand?
3. **Tone:** Is it reassuring and not overly alarming?
4. **Conciseness:** Is it a single, clear sentence?

Review the following definitions:
{draft_definitions}

Return your final, approved list as a single, valid JSON object where each key is the medical term and the value is the best possible simple definition. Your response MUST be ONLY the JSON object itself, starting with {{ and ending with }}.'

Translation Prompt
'You are an expert medical communicator. Your task is to translate a radiology report into a simple, clear, and reassuring summary for a patient who has a eighth-grade reading level.

Primary Goal: Create a patient-friendly translated report and a final summary section.

Step-by-Step Instructions:

1. **Analyze the Full Report:** First, read and understand the entire original report provided by the user.

2. **Use Verified Prior Knowledge:** You are given a pre-processed and verified list of complex medical terms and their simple definitions. You MUST use this information as the primary source of truth for your explanations.
- **Verified Medical Terms:** {list_of_entities}
- **Simple Definitions for these Terms:**
{definitions_for_prompt}

3. **Create the Translated Report:**
- Generate a new, patient-friendly version of the report. This new version MUST contain these four sections, in this order: 'Indication/Clinical History', 'Technique', 'Findings', and 'Impression'.
- Use the rules and examples below to guide your writing style for MAXIMUM clarity and conciseness.
- **Rule 3a (Conciseness Guideline):** Strive for conciseness. As a guideline, aim for each 'Findings' sub-section to be under 50 words (or ~200 words for the total 'Findings' section if there are no sub-sections), and each 'Impression' point to be no more than two sentences. **However, this is only a guideline. Your top priority is to follow all other rules, especially the requirement to explain every medical term. If explaining all terms in a section requires you to exceed the word count, that is acceptable.**
- **Rule 3b (Jargon Integration):** Your goal is to replace complex medical jargon with simpler language, using the verified 'Prior Knowledge' from Step 2 to integrate explanations naturally.
- **Combining Terms:** When a finding and a location are combined (like 'compression deformities in T4-T6 vertebrae'), use the simple explanation for both and combine them gracefully into one flowing description. For example, a good output would be: 'There is squishing (compression deformities) in several spinal bones in your middle back (T4-T6).'
- **Important rule for vertebrae**: All cervical (C1,C2,C3,C4,C5,C6,C7,C8), thoracic (T1,T2,T3,T4,T5,T6,T7,T8,T9,T10,T11,T12), sacral (S1,S2,S3,S4,S5) vertebrae must be explained in PLAIN language to the patient. For instance, you may NOT simply refer to C3-C4 vertebrae as 'C3-C4 vertebrae' in the translated report, but would instead write 'the spinal bones in your neck (C3-C4 vertebrae)'. This rule **SIMILARLY** would apply to any numbered anatomy elsewhere in the body.
- **Rule 3c (Handling Acronyms/Sub-terms):** If a term is part of a longer term (e.g., 'pulmonary fibrosis' is part of 'idiopathic pulmonary fibrosis'), **only define the longest, most specific term.** Do not provide separate definitions for the shorter, nested terms.
- **Rule 3d (No Repetition Within a Section):** Within any single section, once you have defined a term, do not define it again. You may refer to it by its simple name thereafter in that section.
- **Rule 3e (Preserve Nuance and Detail):** Be mindful of confidence levels (e.g., 'concern for' vs. 'cannot be excluded') and preserve the location of all findings, describing them in simple terms based on the provided definitions. For example, 'T4-T6 vertebrae' would be explained as 'the vertebrae in your middle back'
- **Rule 3f (No Editorializing):** DO NOT add qualitative descriptors (such as 'small,' 'slight,' 'a bit,' 'large,' 'mild,' etc.) UNLESS that EXACT descriptor is present in the original report. ONLY IF that EXACT descriptor is present in the original report, then you are allowed to use that EXACT SAME descriptor to describe the SAME finding it referred to in the original report.

- **Rule 3g (Clarity Check for Impression):** This rule overrides others if needed for clarity. The 'Impression' section is the most important. **Ensure every single medical term mentioned in the translated 'Impression' section is clearly and simply explained in that same sentence.** For example, ‘Findings are suggestive of cholecystitis’ should be written as either ‘The findings suggest cholecystitis, which is an inflammation of the gallbladder’ or ‘The findings suggest inflammation of the gallbladder.’

- **High-Quality Example to Follow:**

<example>
<original_report>
Findings: There is a new 2.5 cm lesion with arterial enhancement, capsule, and washout in hepatic segment 8. LIRADS-5.
Impression: New LIRADS-5 lesion in segment 8, which is concerning for hepatocellular carcinoma.
</original_report>
<patient_friendly_translation>
Findings: There is a 2.5 cm LIRADS-5 lump in the right upper part of your liver. LIRADS is a grading system for liver imaging findings that ranges from 1 to 5, with 5 representing the most suspicious category for cancer.
Impression: The imaging features are suspicious for liver cancer.
</patient_friendly_translation>
</example>

<example_2>
<original_report>
Impression: 1. Moderate hepatic steatosis, suggestive of nonalcoholic fatty liver disease (NAFLD).
</original_report>
<patient_friendly_translation>
Impression: 1. The scan shows a moderate amount of fat in your liver, which may be a sign of a condition called nonalcoholic fatty liver disease (NAFLD).
</patient_friendly_translation>
</example_2>
---

4. **Create the Final Summary Section:** After the translated report, add a final section with two sub-headings: 'Patient-Friendly Summary' and 'Next Steps'.

5. **'Patient-Friendly Summary' Rules:**
- The summary must be a high-level overview based **primarily** on the 'Impression' section, but may include one single sentence broadly summarizing other findings.
- It must be concise (ideally 3-4 sentences, max 5). The sentences should be short and simple.
- It must re-explain key diagnostic terms for clarity.
- It must conclude with: ‘Please contact your provider if you have any questions or worries regarding this report.’

6. **'Next Steps' Rules (VERY STRICT):**
- If the 'Impression' section in the original report has an explicit recommendation, rephrase it simply.
- If not, you MUST use ONLY the phrase: ‘Please discuss this report with your doctor to determine the next steps.’
- **DO NOT INFER or add any information not explicitly stated in the original report's recommendations.**

7. **Final Quality Audit (Self-Critique):** Before producing your final output, you MUST perform one last review of the text you have generated. Adopt the persona of a meticulous clinical editor whose only job is to ensure the final text is 100% factual based on the source report.
- **A) Factual Accuracy:** Check that all numbers, findings, and **anatomical locations** match the source report. For example, 'medial right lower lobe' must be translated as 'inner part of the lower right lung', not just 'right lung'.
- **B) REMOVE ALL EDITORIALIZING** **CRITICAL RULE**: Scrutinize your generated draft for any qualitative or size-describing adjectives that were not explicitly present in the original report. This is a STRICT check for words like ‘small,’ ‘slight,’ ‘a bit,’ ‘little,’ ‘minor,’ ‘harmless,’ etc.**
- **Important**: This rule applies to the explanations you provide for terms as well. Do not add qualitative terms like 'small' to your medical term explanations UNLESS that term was used in the original report.
- **Correction Example:**
- **If original report says:** ‘Calcified pulmonary granuloma.’
- **And your translated draft says:** ‘a small, harmless, calcified scar in your lung.’
- **You MUST correct it to:** ‘a harmless, calcified scar in your lung.’

	- **C) Terminology** **CRITICAL RULE:** ENSURE ALL MEDICAL TERMS ARE DEFINED. If there is a medical term that appears in the refined 'Prior Knowledge' medical term list, but has not been explained in the translated report, you MUST explain the term simply and naturally. - However, AVOID repetitiveness--do **NOT** repeat a definition of a term in the same sentence, even in parentheses. - **CRITICAL**: Additionally, if a medical term is composed of words that are individually simple (i.e. 'connective tissue disease'), you MUST still explain that medical term, as the complete term is likely to be a term that a layperson would NOT recognize. - Additionally, if the translated report contains the word 'benign', you MUST specify that this means non-cancerous. - **Your final output MUST be the polished version *AFTER* you have completed this strict audit.**"

Table 1. Candidate prompts for patient-friendly report translation

Table 2. Likert scales for evaluation of the accuracy, understandability, and safety of LLM generated patient-friendly radiology reports

Rating scale for overall quality of generated report		
Accuracy of Key findings	1	Fails to state the key findings in the "Impressions" section or grossly misrepresents finding (i.e. interval DECREASE in size of mass vs INCREASE).
	2	Misrepresents the most important/urgent/actionable findings in the "Impressions" section (i.e. gives a wrong location or wrong size for a major finding).
	3	Captures all urgent clinically significant findings in the "Impressions section," but may miss a non-urgent finding (i.e. captures "focal consolidation" but misses "unchanged chronic rib fracture").
	4	Captures all findings in the "Impression section," but is not fully specific in non-critical descriptors (i.e. for location, "right lower lobe" is simplified to "right lung") or the level of concern/confidence is inaccurate.
	5	Perfectly captures all findings in the "Impression Section", including size (if mentioned), location (if mentioned) and delivers the level of concern/confidence accurately (i.e. concern for malignancy vs malignancy cannot be excluded are different levels of confidence)
Hallucinations	1	Contains clinically significant hallucination (such as “no significant concern” when there is new severe disease)
	2	Contains clinically insignificant hallucination (such as adding “small” or “little” as descriptors for minor findings, but does not change overall conclusion)
	3	Contains no hallucination
Understandability	1	The report is essentially unsimplified, with the vast majority still being comprised of medical jargon. The reader would not be able to understand the report without the aid of a medical professional.
	2	The output contains significant amounts of medical jargon; it is only slightly simplified from the original report, and the reader is unable to comprehend a majority of the report.
	3	The output contains multiple instances of medical jargon (i.e. mediastinal) that is not defined, and that a patient would likely need to search themselves. The reader feels like they got some important information, but thinks they may not have understood all that was intended
	4	The output uses mostly plain, every-day language. Only a few words may need clarification/simplification (i.e. calcification vs calcium deposit). The reader feels that they understand the gist of the report
	5	The output uses plain, every-day language, with all medical terms explained. It is fully understandable and the reader feels confident they received the intended knowledge from the report.
Rating scale for findings mentioned within the generated report		

Accuracy	1	Medically inaccurate description/explanation (such as wrong explanation of medical concept)
	2	Medically partially accurate description/explanation (minor errors in explanation that do not change overall conclusion)
	3	Medically accurate description/explanation of the diagnosis

Table 2. Likert scales for evaluation of the accuracy, understandability, and safety of LLM generated patient-friendly radiology reports

Table 3. Example of large language model output using prompt (3)

Original radiology report	Large language model generated report
PULMONARY ARTERIES: Findings concerning for chronic pulmonary embolism including eccentric thrombus within the right main pulmonary artery (5/94), pulmonary artery webs (5/100 and 5/75), and abrupt occlusion of segmental branches of the right lower lobe pulmonary artery (5/111). The main pulmonary artery is dilated, especially relative to the aorta, measuring up to 35 mm. Findings are slightly improved compared to prior examination.	**Pulmonary Arteries:** There are signs of long-term lung clots. A clot that attaches to one side of the large blood vessel carrying blood from your heart to your right lung (right main pulmonary artery) was seen, along with thin, web-like structures (pulmonary artery webs) in parts of these vessels. Some of the smaller blood vessels that supply your lower right lung (segmental branches) are suddenly blocked. Also, the main lung blood vessel is larger than usual (dilated), measuring up to 35 mm compared to your main artery (aorta). These findings are slightly improved from your last scan.
LUNGS: Persistent Mosaic perfusion of the lungs. Resolving right lower lobe nodular consolidation, now with areas of central clearing. No suspicious pulmonary nodules.	**Lungs:** Your lung images show an uneven pattern of blood flow (mosaic perfusion). An area in the lower right lung that was dense (nodular consolidation) is now showing central clearing, which suggests improvement. There are no other unusual lumps (nodules) in your lungs.
PLEURA: No pleural effusion or pneumothorax	**Pleura:** There is no extra fluid between the layers covering your lungs (pleural effusion) and no air is seen outside the lung that would cause it to collapse (pneumothorax).
MEDIASTINUM: No suspicious lymphadenopathy	**Mediastinum:** The lymph nodes in the center of your chest appear normal (no lymphadenopathy).
HEART/GREAT VESSELS: Mild cardiomegaly with right ventricular dilation and slight bowing of the intraventricular septum to the left. No pericardial effusion. No significant coronary artery atherosclerotic calcifications. No significant change in size of a thoracoabdominal aortic aneurysm originating at the level of the proximal descending thoracic aorta and extending throughout the entire abdominal aorta, partially visualized. No evidence of aneurysmal instability. Right PICC with tip terminating at the superior cavoatrial junction	**Heart/Great Vessels:** Your heart is mildly larger than normal (mild cardiomegaly) and the right chamber of your heart is enlarged (right ventricular dilation). The wall between the left and right heart chambers (intraventricular septum) is slightly bowed to the left. There is no extra fluid around the heart (pericardial effusion) and no concerning calcium buildup in the heart's arteries. A bulge (aneurysm) that starts in the upper part of your aorta in your chest (proximal descending thoracic aorta) and extends into the abdomen (thoracoabdominal aortic aneurysm) measures about 50 mm and does not show signs of instability. A tube (PICC) is seen on the right with its tip at the junction where a large upper vein (superior vena cava) meets the heart.
BONES/SOFT TISSUES:	**Bones/Soft Tissues:**

No suspicious osseous lesions. Postsurgical changes from prior anterior instrumented fusion of the cervical spine.	There are no abnormal areas in the bones (osseous lesions). Changes from previous surgery to fuse bones in your neck (anterior instrumented fusion of the cervical spine) are visible.
VISIBLE ABDOMEN: Limited evaluation of the upper abdomen is normal.	**Visible Abdomen:** The upper part of your abdomen that was seen appears normal.
IMPRESSION:	**Impression:**
1. Findings compatible with chronic thromboembolic disease with pulmonary hypertension including mosaic perfusion of the lungs, enlarged main pulmonary artery, eccentric thrombus/pulmonary artery webs, and abrupt occlusion of several right lower lobe pulmonary artery segmental branches. 2. Improving right lower lobe consolidation with central clearing, compatible with improving pulmonary infarct. 3. Unchanged thoracoabdominal aortic aneurysm measuring up to approximately 50 mm in diameter. No evidence of aneurysmal instability.	1. The scan shows signs of long-term blood clots in your lung arteries (chronic thromboembolic disease) that may be causing high pressure in those vessels (pulmonary hypertension). This is indicated by an uneven blood flow pattern (mosaic perfusion), an enlarged main lung artery, a clot attached to one side (eccentric thrombus) with web-like strands (pulmonary artery webs), and sudden blockages in several smaller vessels in your lower right lung. 2. The dense patch in the lower right lung is improving, with a central clearing pattern that is compatible with recovery from lung tissue damage due to a blocked blood vessel (pulmonary infarct). 3. The bulge in your main artery running from your chest into your abdomen (thoracoabdominal aortic aneurysm) remains unchanged at about 50 mm and does not show any signs of becoming unstable.
	Final Summary Patient-Friendly Summary: This scan shows evidence of long-term blood clots in your lung arteries, which may be causing higher blood pressure in these vessels. The uneven blood flow and an enlarged main lung vessel support this finding. The dense area in your lower right lung is getting better, and the bulge in your main artery remains stable. Please contact your provider if you have any questions or worries regarding this report. Next Steps: Please discuss this report with your doctor to determine the next steps.

Table 3. Example of large language model output using prompt (3)

Keywords

Large Language Model; Patient-Friendly Report; Report translation; Prompt engineering