RSNA Press Release

Fine-tuned LLMs Boost Error Detection in Radiology Reports

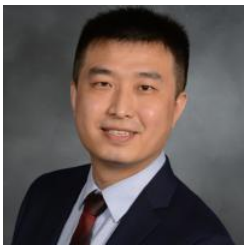
Released: May 20, 2025

OAK BROOK, Ill. — A type of artificial intelligence called fine-tuned large language models (LLMs) greatly enhances error detection in radiology reports, according to a new study published today in *Radiology*, a journal of the Radiological Society of North America (RSNA). Researchers said the findings point to an important role for this technology in medical proofreading.

Radiology reports are crucial for optimal patient care. Their accuracy can be compromised by factors like errors in speech recognition software, variability in perceptual and interpretive processes and cognitive biases. These errors can lead to incorrect diagnoses or delayed treatments, making the need for accurate reports urgent.

LLMs like ChatGPT are advanced generative AI models that are trained on vast amounts of text to generate human language. While they offer great potential in proofreading, their application in the medical field, particularly in detecting errors within radiology reports, remains underexplored.

download photo



Yifan Peng, Ph.D.

To bridge this gap in knowledge, researchers evaluated fine-tuned LLMs for detecting errors in radiology reports during medical proofreading. A fine-tuned LLM is a pre-trained language model that is further trained on domain-specific data.

"Initially, LLMs are trained on large-scale public data to learn general language patterns and knowledge," said study senior author Yifan Peng, Ph.D., from the Department of Population Health Sciences at Weill Cornell Medicine in New York City. "Fine-tuning occurs as the next step, where the model undergoes additional training using smaller, targeted datasets relevant to particular tasks."

To test the model, Dr. Peng and colleagues built a dataset with two parts. The first consisted of 1,656 synthetic reports, including 828 error-free reports and 828 reports with errors. The second part comprised 614 reports, including 307 error-free reports from MIMIC-CXR, a large, publicly available database of chest X-rays, and 307 synthetic reports with errors.

The researchers used the synthetic reports to boost the amount of training data and fulfill the data-hungry needs of LLM fine-tuning.

"Synthetic reports can also increase the coverage and diversity, balance out the cases and reduce the annotation costs," said the study's first author, Cong Sun, Ph.D., from Dr. Peng's lab. "In radiology, or more broadly, the clinical domain, synthetic reports allow safe data-sharing without compromising patient privacy."

The researchers found that the fine-tuned model outperformed both GPT-4 and BiomedBERT, a natural language processing tool for biomedical research.

"The LLM that was fine-tuned on both MIMIC-CXR and synthetic reports demonstrated strong performance in the error detection tasks," Dr. Sun said. "It meets our expectations and highlights the potential for developing lightweight, fine-tuned LLM specifically for medical proofreading applications."

The study provided evidence that LLMs can assist in detecting various types of errors, including transcription errors and left/right errors, which refer to misidentification or misinterpretation of directions or sides in text or images.

The use of synthetic data in AI model building has raised concerns of bias in the data. Dr. Peng and colleagues took steps to minimize this by using diverse and representative samples of real-world data to generate the synthetic data. However, they acknowledged that synthetic errors may not fully capture the complexity of real-world errors in radiology reports. Future work could include a systematic evaluation of how bias introduced by synthetic errors affects model performance.

The researchers hope to study fine-tuning's ability to reduce radiologists' cognitive load and enhance patient care and find out if fine-tuning would degrade the model's ability to generate reasoning explanations.

"We are excited to keep exploring innovative strategies to enhance the reasoning capabilities of fine-tuned LLMs in medical proofreading tasks," Dr. Peng said. "Our goal is to develop transparent and understandable models that radiologists can confidently trust and fully embrace."

"Generative Large Language Models Trained for Detecting Errors in Radiology Reports." Collaborating with Drs. Peng and Sun were Kurt Teichman, M.S., Yiliang Zhou, M.S., Brian Critelli, B.S., David Nauheim, M.D., Graham Keir, M.D., Xindi Wang, Ph.D., Judy Zhong, Ph.D., Adam E. Flanders, M.D., and George Shih, M.D.

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 how to read a radiology report, visit *RadiologyInfo.org*.

Images (JPG, TIF):

<Report Template>
CLINICAL HISTORY: []
TECHNIQUE: AP chest and abdomen radiograph. []
COMPARISON: [Radiograph from one day prior].
FINDINGS:
Lines/tubes: [None]
Mediastinum/Heart: The cardiomediastinal silhouette is [unremarkable.]
Lungs/Airways/Pleura: [No focal opacities, pneumothorax or pleural effusion.]
Bones/Soft Tissue: [No acute osseous abnormalities.]
Upper abdomen: [Unremarkable]
IMPRESSION:
[]
</Report Template>

You are a radiologist assistant helping to create synthetic reports using the <Report Template> above.

Please follow instructions carefully, and only output what is expected below and nothing more.

Radiology reports can have a variety of errors and issues. 'Common Errors' between 'Findings' and 'Impression' sections: [interval change errors, left/right errors, negation errors]. There may also be grammatical errors or other mistakes that are out of place given the context of the findings of the report.

Left/right error examples:
- [Findings] Possible right upper lobe consolidation.
[Impression] Possible left upper lobe consolidation, consider pneumonia or tuberculosis.
- [Findings] Possible rib fracture on the right.
[Impression] Possible left-sided rib fracture.

Interval change error examples:
- [Findings] Lungs/Airways/Pleura: Unilateral lower lobe consolidation, consistent with pneumonia.
[Impression] Bilateral pneumonia.
- [Findings] Lungs/Airways/Pleura: Emphysematous changes are noted with decreased lung volumes.
[Impression] Lung volumes increased in size.

Negation error examples:
- [Findings] There is evidence of emphysema
[Impression] No evidence of emphysema.
- [Findings] Appendicitis is not seen
[Impression] Appendicitis is seen.

Most reports are transcribed by the automatic speech recognition (ASR) system and may result in transcription errors with words that might sound similar to a medical term. Some of these are single words and some may be multi-word terms. Examples:
- 'chloral effusion' should be 'pleural effusion'
- 'costal phrenic' should be 'costophrenic'

{Clinical Conditions}
- Aortic enlargement
- Atelectasis
- Cardiomegaly
- Atherosclerotic Calcification
- Clavicle fracture
- Consolidation
- Edema
- Emphysema
- Enlarged pulmonary artery
- Interstitial lung disease (ILD)
- Infiltration
- Lung cavity
- Lung cyst
- Lung opacity
- Nodule/Mass
- Pulmonary fibrosis
- Pneumothorax
- Pleural thickening
- Pleural effusion
- Rib fracture
- Other lesion
- Lung tumor
- Pneumonia
- Tuberculosis
- COPD

Synthetic Report Formatting Rules:
- Words after a colon':' in the template and at the beginning of a new line should be capitalized, but not all capital letters.
- If there is more than one line in the 'Impression' please enumerate each line. A single line in the 'Impression' should not be enumerated
- Do not reorder the report – keep the template fields in the same order
- Do not show the output wrapped in the triple quotes

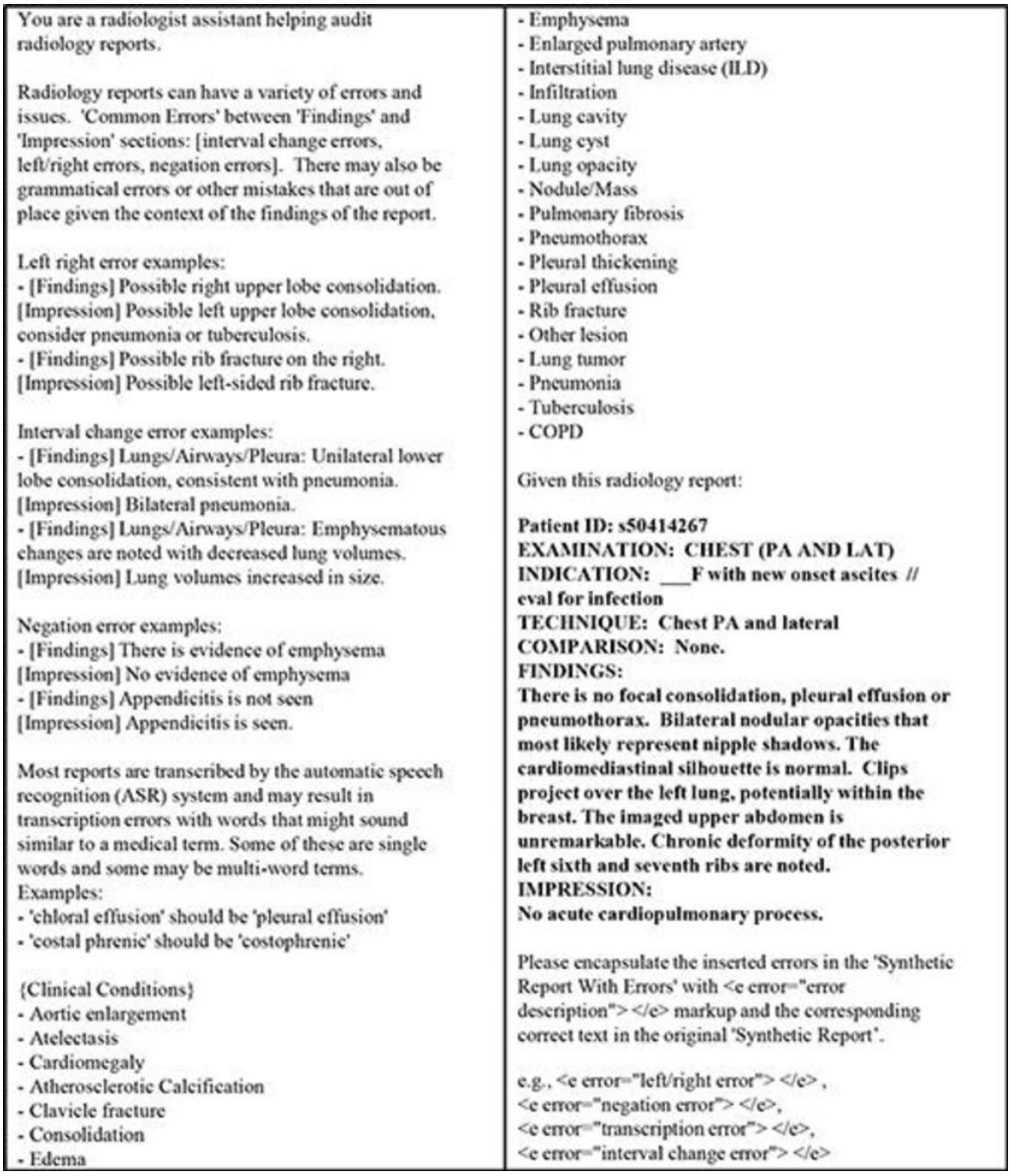
{Task 1 Instructions} Synthetic Report Generation
- Enumerate the 'Patient ID: []' using 5-digits with zero padding: Patient ID: [00001]
- Format the 'History []' with patient age, gender, and one or more clinical conditions or symptoms like this. Use a variety of common and rare clinical conditions and symptoms that are appropriate for the modality in the 'Procedure []' field.
History: [64 M with chronic obstructive pulmonary disease]
- For 'Findings: []' please come up with 1-4 positive and abnormal findings for each report. Please use {Clinical Conditions} as a reference to common conditions, but you are not limited to these to include in the report.
- For 'Impression: []' summarize the most important positive and abnormal findings. Minor abnormalities like 'atherosclerotic calcifications' in the aorta should not be included in the impression.

- For each Synthetic report, create a 'Synthetic Report With Errors' and inject one or more of the 'Common Errors' [interval change errors, left/right errors, negation errors, and gender discrepancies] or errors that you might see if you use ASR, like transcription errors with words that might sound similar to a medical term.

- Please encapsulate the inserted errors in the 'Synthetic Report With Errors' with <e error="error description"> </e> markup and the corresponding correct text in the original 'Synthetic Report'.
e.g., <e error="left/right error"> </e> ,
<e error="negation error"> </e>,
<e error="transcription error"> </e>,
<e error="interval change error"> </e>

Please generate 828 reports using {Task 1 Instructions}
- Use a markdown table format: [Report Number, Synthetic Report, Synthetic Report With Errors, Error Type(s)]

Figure 1. Prompts for GPT-4–0125-Preview used to generate 828 error-free synthetic reports and 828 synthetic reports with errors. AP = anteroposterior, COPD = chronic obstructive pulmonary disease.
High-res (TIF) version



You are a radiologist assistant helping audit radiology reports.

Radiology reports can have a variety of errors and issues. 'Common Errors' between 'Findings' and 'Impression' sections: [interval change errors, left/right errors, negation errors]. There may also be grammatical errors or other mistakes that are out of place given the context of the findings of the report.

Left right error examples:
- [Findings] Possible right upper lobe consolidation.
[Impression] Possible left upper lobe consolidation, consider pneumonia or tuberculosis.
- [Findings] Possible rib fracture on the right.
[Impression] Possible left-sided rib fracture.

Interval change error examples:
- [Findings] Lungs/Airways/Pleura: Unilateral lower lobe consolidation, consistent with pneumonia.
[Impression] Bilateral pneumonia.
- [Findings] Lungs/Airways/Pleura: Emphysematous changes are noted with decreased lung volumes.
[Impression] Lung volumes increased in size.

Negation error examples:
- [Findings] There is evidence of emphysema
[Impression] No evidence of emphysema
- [Findings] Appendicitis is not seen
[Impression] Appendicitis is seen.

Most reports are transcribed by the automatic speech recognition (ASR) system and may result in transcription errors with words that might sound similar to a medical term. Some of these are single words and some may be multi-word terms.
Examples:
- 'chloral effusion' should be 'pleural effusion'
- 'costal phrenic' should be 'costophrenic'

{Clinical Conditions}
- Aortic enlargement
- Atelectasis
- Cardiomegaly
- Atherosclerotic Calcification
- Clavicle fracture
- Consolidation
- Edema
- Emphysema
- Enlarged pulmonary artery
- Interstitial lung disease (ILD)
- Infiltration
- Lung cavity
- Lung cyst
- Lung opacity
- Nodule/Mass
- Pulmonary fibrosis
- Pneumothorax
- Pleural thickening
- Pleural effusion
- Rib fracture
- Other lesion
- Lung tumor
- Pneumonia
- Tuberculosis
- COPD

Given this radiology report:

Patient ID: s50414267
EXAMINATION: CHEST (PA AND LAT)
INDICATION: ___F with new onset ascites // eval for infection
TECHNIQUE: Chest PA and lateral
COMPARISON: None.
FINDINGS:
There is no focal consolidation, pleural effusion or pneumothorax. Bilateral nodular opacities that most likely represent nipple shadows. The cardiomediastinal silhouette is normal. Clips project over the left lung, potentially within the breast. The imaged upper abdomen is unremarkable. Chronic deformity of the posterior left sixth and seventh ribs are noted.
IMPRESSION:
No acute cardiopulmonary process.

Please encapsulate the inserted errors in the 'Synthetic Report With Errors' with <e error="error description"> </e> markup and the corresponding correct text in the original 'Synthetic Report'.

e.g., <e error="left/right error"> </e> ,
<e error="negation error"> </e>,
<e error="transcription error"> </e>,
<e error="interval change error"> </e>

Figure 2. Prompts for GPT-4–0125-Preview used to generate synthetic radiology reports using reports from the MIMIC chest radiograph (MIMIC-CXR) database. The bold text indicates the original reports from the MIMIC-CXR database that were used to create these synthetic reports with errors. COPD = chronic obstructive pulmonary disease, LAT = lateral, PA = posteroanterior.
High-res (TIF) version

A Prompts for zero-shot prompting:

USER: You are a medical assistant. Please classify the input report and respond in JSON format {'Classification': 'Negation error, Left/Right error, Interval Change error, Transcription error, or No error'}. Note that the input report may belong to multiple errors.
ASSISTANT: Sure! Please give the input report.
USER: INPUT_REPORT

B Prompts for one-shot prompting:

USER: Here is an example of Example_Label.
ASSISTANT: Example_Text
USER: Please classify the input report and respond in JSON format {'Classification': 'Negation error, Left/Right error, Interval Change error, Transcription error, or No error'}. Note that the input report may belong to multiple errors.
ASSISTANT: Sure! Please give the input report.
USER: INPUT_REPORT

C Prompts for four-shot prompting:

USER: Here is an example of Transcription errors.
ASSISTANT: Example_Tran
USER: Here is an example of Interval Change errors.
ASSISTANT: Example_Int
USER: Here is an example of Left/Right errors.
ASSISTANT: Example_LR
USER: Here is an example of Negation errors.
ASSISTANT: Example_Neg
USER: Please classify the input report and respond in JSON format {'Classification': 'Negation error, Left/Right error, Interval Change error, Transcription error, or No error'}. Note that the input report may belong to multiple errors.
ASSISTANT: Sure! Please give the input report.
USER: INPUT_REPORT

Figure 3. Example prompt designs. **(A)** A zero-shot prompt. **(B)** A one-shot prompt. **(C)** A four-shot prompt. INPUT_ REPORT refers to an input report from the test set. Example_Text is an example report from the training set, Example_Label is the label of the example report, Example_Tran is an example report containing a transcription error from the training set, Example_Int is an example report containing an interval change error, Example_LR is an example report containing a left/right error, and Example_Neg is an example report containing a negation error.
High-res (TIF) version

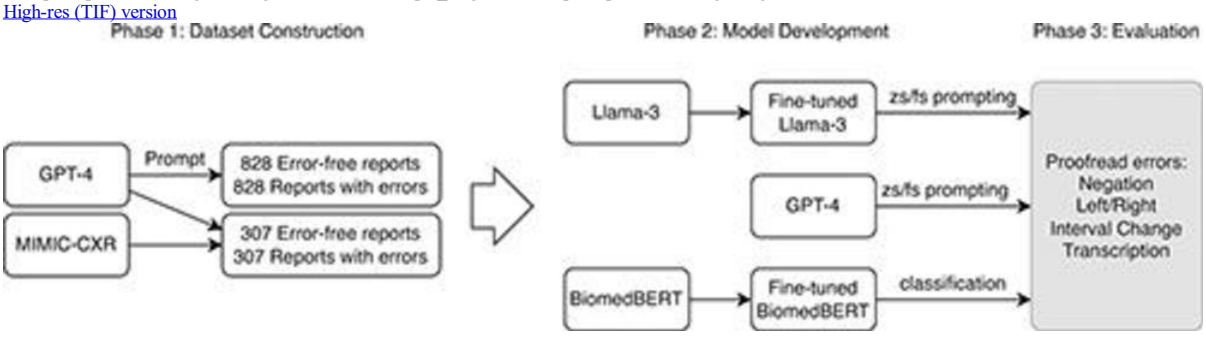


Figure 4. The overall workflow of large language models (LLMs). A dataset was constructed by combining synthetic radiology reports with a small subset of reports from the MIMIC chest radiograph (MIMIC-CXR) database, LLMs such as Llama-3 (Meta AI [31]) and GPT-4 (OpenAI [29]) were refined using zero-shot (zs) or few-shot (fs) prompting strategies, and the models' performance on the constructed dataset was evaluated.
High-res (TIF) version

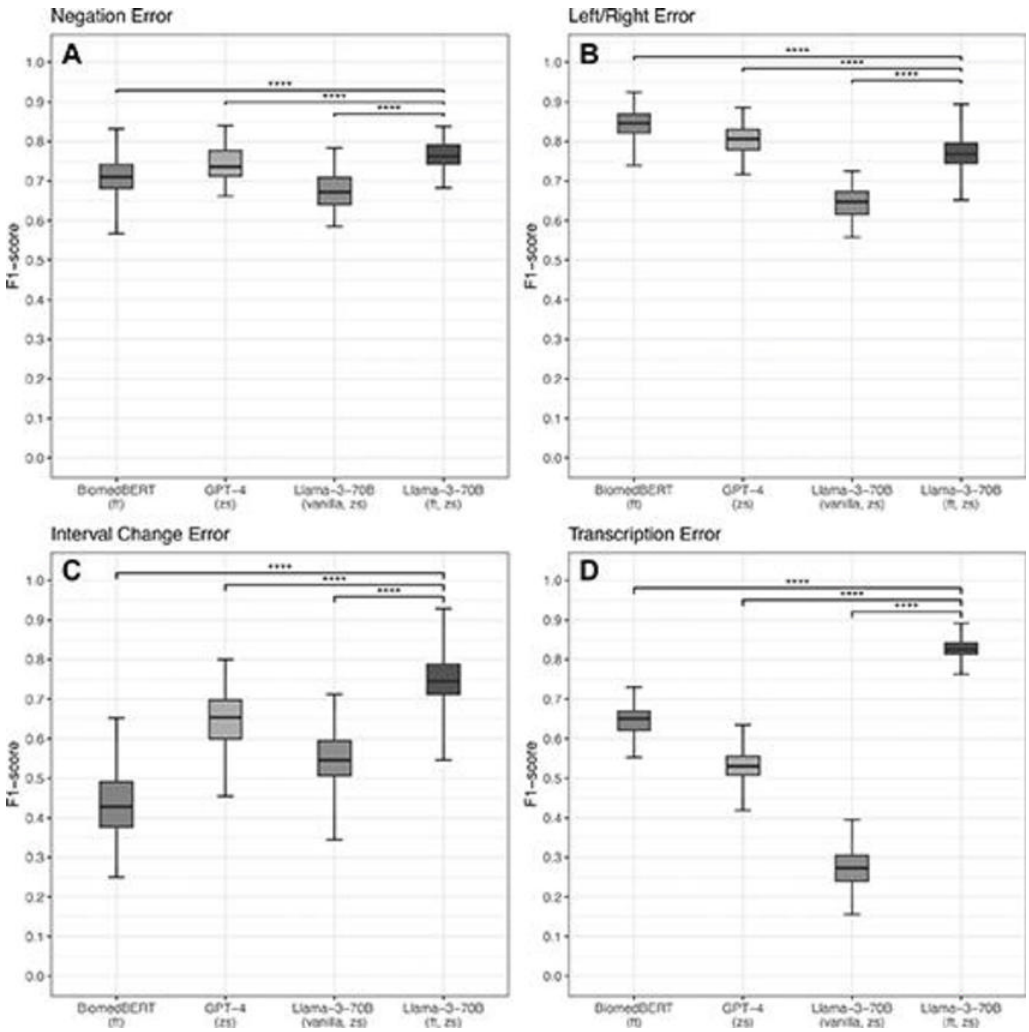


Figure 5. Box and whisker plots show F1 scores of different models, with boxplots showing the range (whiskers), median (box midline), and distribution (box edges). **(A)** Negation error, **(B)** left/right error, **(C)** interval change error, and **(D)** transcription error. The error bars are 95% CIs.
High-res (TIF) version

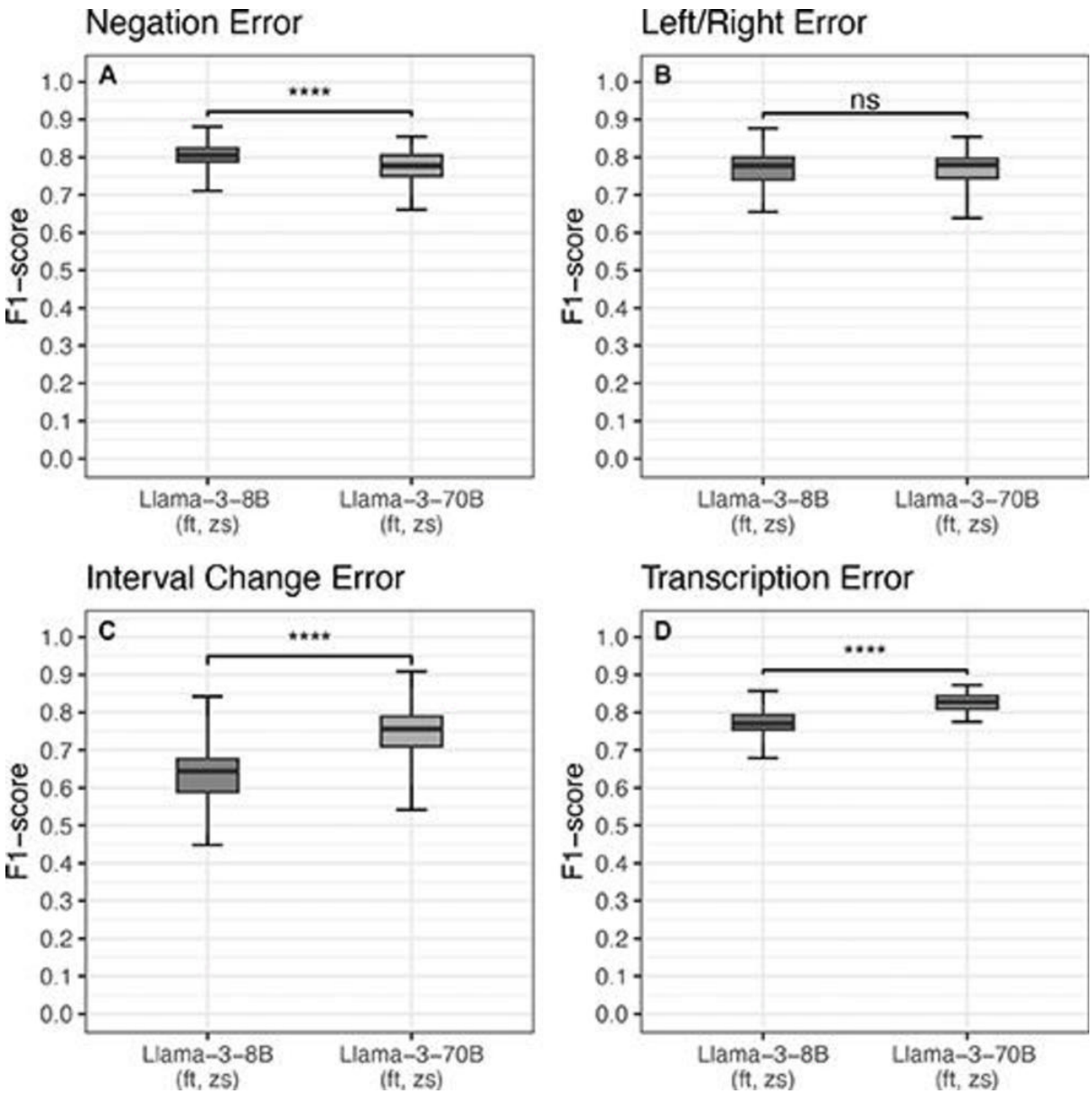


Figure 6. Box and whisker plots show F1 scores for different model parameter scales, with boxplots showing the range (whiskers), median (box midline), and distribution (box edges). The error bars are 95% CIs. **(A)** Negation error, **(B)** left/right error, **(C)** interval change error, and **(D)** transcription error. ft = fine-tuned, ns = not significant, zs = zero-shot prompting.
High-res (TIF) version

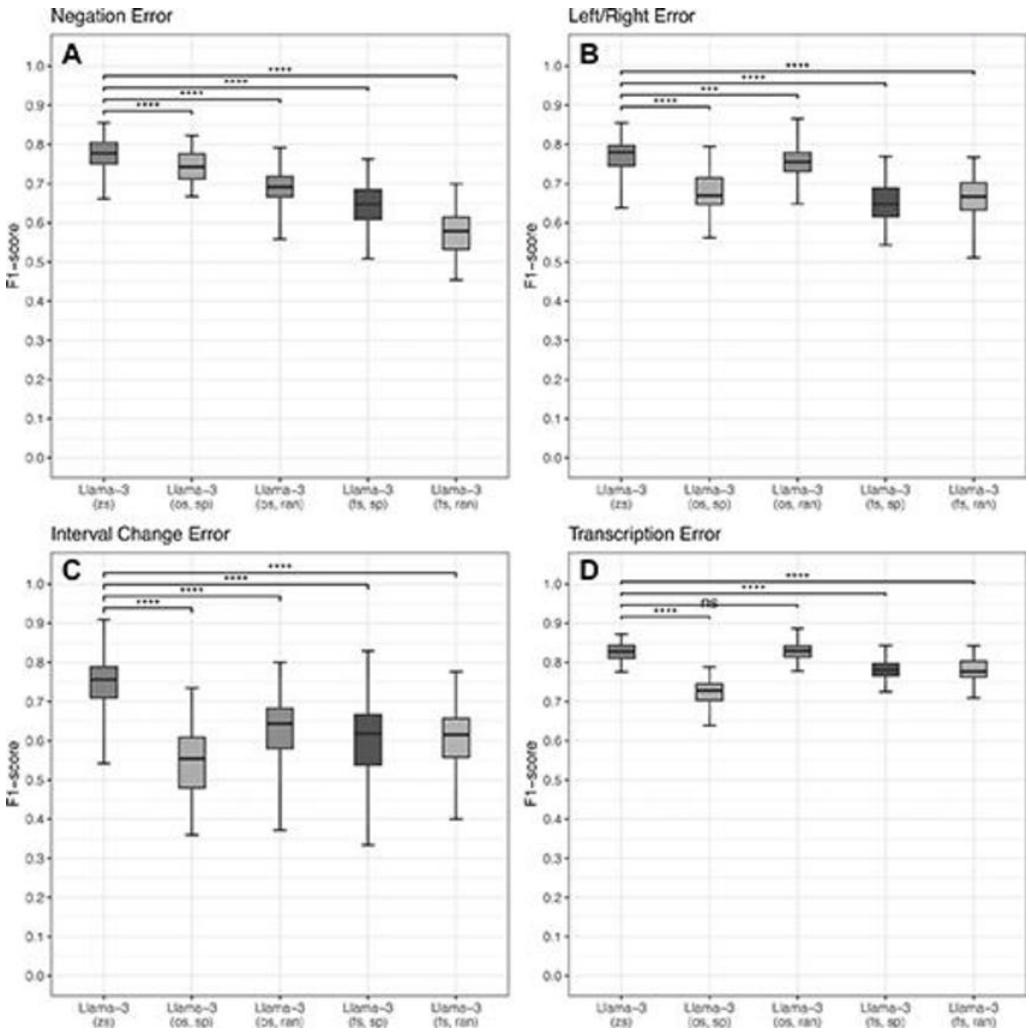


Figure 7. Box and whisker plots show F1 scores of the fine-tuned Llama-3-70B-Instruct model under different prompting strategies, with boxplots showing the range (whiskers), median (box midline), and distribution (box edges). The error bars are 95% CIs. **(A)** Negation error, **(B)** left/right error, **(C)** interval change error, and **(D)** transcription error. fs = four-shot prompting, ns = not significant, os = one-shot prompting, ran = random, sp = specified, zs = zero-shot prompting.
High-res (TIF) version

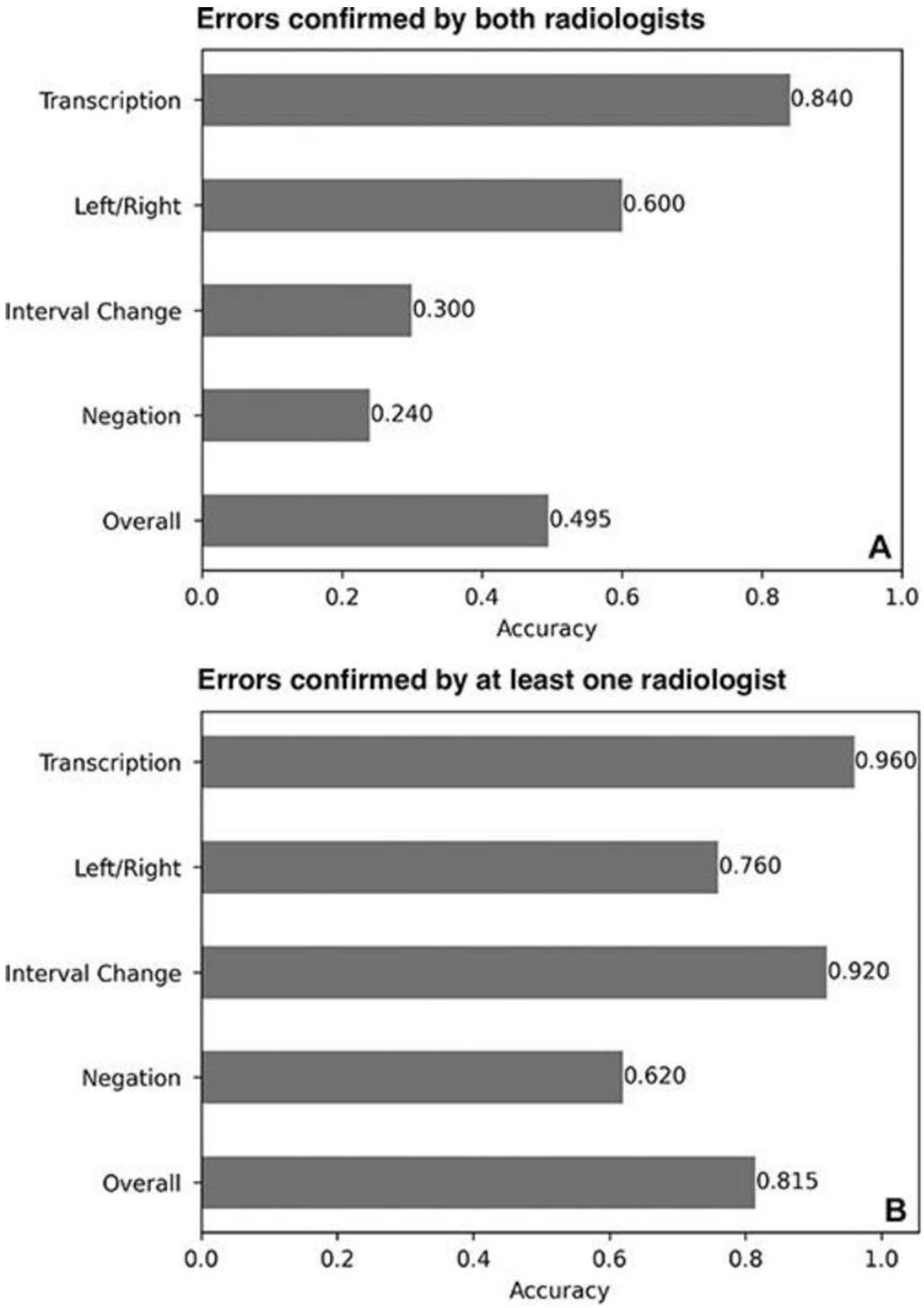


Figure 8. Bar graphs show radiologist-confirmed accuracy of error detection across different categories in radiology reports. **(A)** Errors confirmed by both radiologists. **(B)** Errors confirmed by at least one radiologist.
High-res (TIF) version

Resources:

Editorial
Study abstract