MSR113

From Unstructured Text to Enriched Data: Evaluating Large Language Models for Clinical Information Extraction from Echocardiogram Reports

Yu-Tung Huang, BPharm[1], Sandy Hsu, BA[1], Chia-Chang Wen, PhD[2], Chih-Fan Yeh, MD, PhD[3], Wen-Li Kuan, RPh, MS[1], Fang-Ju (Irene) Lin, RPh, PhD[1,4]

1. Graduate Institute of Clinical Pharmacy, College of Medicine, National Taiwan University, Taipei, Taiwan
2. Information Technology Office, National Taiwan University Hospital, Taipei, Taiwan
3. Division of Cardiology, Department of Internal Medicine and Cardiovascular Center, National Taiwan University Hospital, Taipei, Taiwan
4. Department of Pharmacy, National Taiwan University Hospital, Taipei, Taiwan

BACKGROUND

- Unstructured echocardiogram reports provide valuable diagnostic information for research; however, their lack of standardized terminology, inconsistent format, and poor integration with structured data limit their scalability for real-world studies.[1,2]
- While large language models (LLMs) show promise in clinical text processing, the influence of model type and prompt complexity on their ability to extract cardiac features from echocardiogram reports remains poorly understood.[3]

OBJECTIVES

- To compare LLM performance in extracting key clinical information from echocardiogram reports across different (1) **model sizes**, (2) **prompt complexities**, and (3) **cardiac feature types**.

METHODS

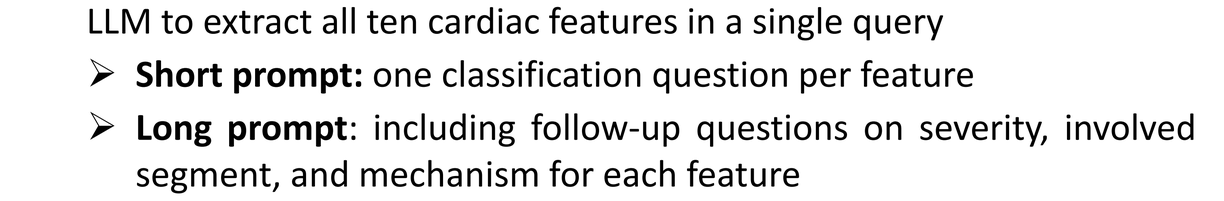

Figure 1. Study workflow overview

1. Task definition:

- To identify 10 predefined cardiac features from unstructured echocardiogram reports (systolic dysfunction, diastolic dysfunction, hypertrophy, chamber dilation, valvular stenosis, regurgitation, leaflet and cusp abnormality, thrombosis, pericardial effusion, and mass or tumor)

2. Data curation:

- **Data source**: unstructured echocardiogram reports from patients with lung cancer who visited National Taiwan University Hospital (NTUH)
- **Gold standard dataset**: 50 annotated cases per feature (30 positives + 20 negatives); 435 unique reports were selected in total

3. Model selection:

- 3 open-source LLMs (from small to large: **LLaMA3.1:8B**, **Mistral-Small:24B**, and **LLaMA3.3:70B**) tested in a zero-shot setting

4. Prompt design:

2 prompt formats were compared, each asking the LLM to extract all ten cardiac features in a single query

- **Short prompt:** one classification question per feature
- **Long prompt**: including follow-up questions on severity, involved segment, and mechanism for each feature

5. Performance evaluation:

- **Metrics**: accuracy, precision, recall, and F1-score
- **Analysis level**: feature-wise and model-wise
- **Evaluation approach**: compare the LLM outputs with the gold standard answers

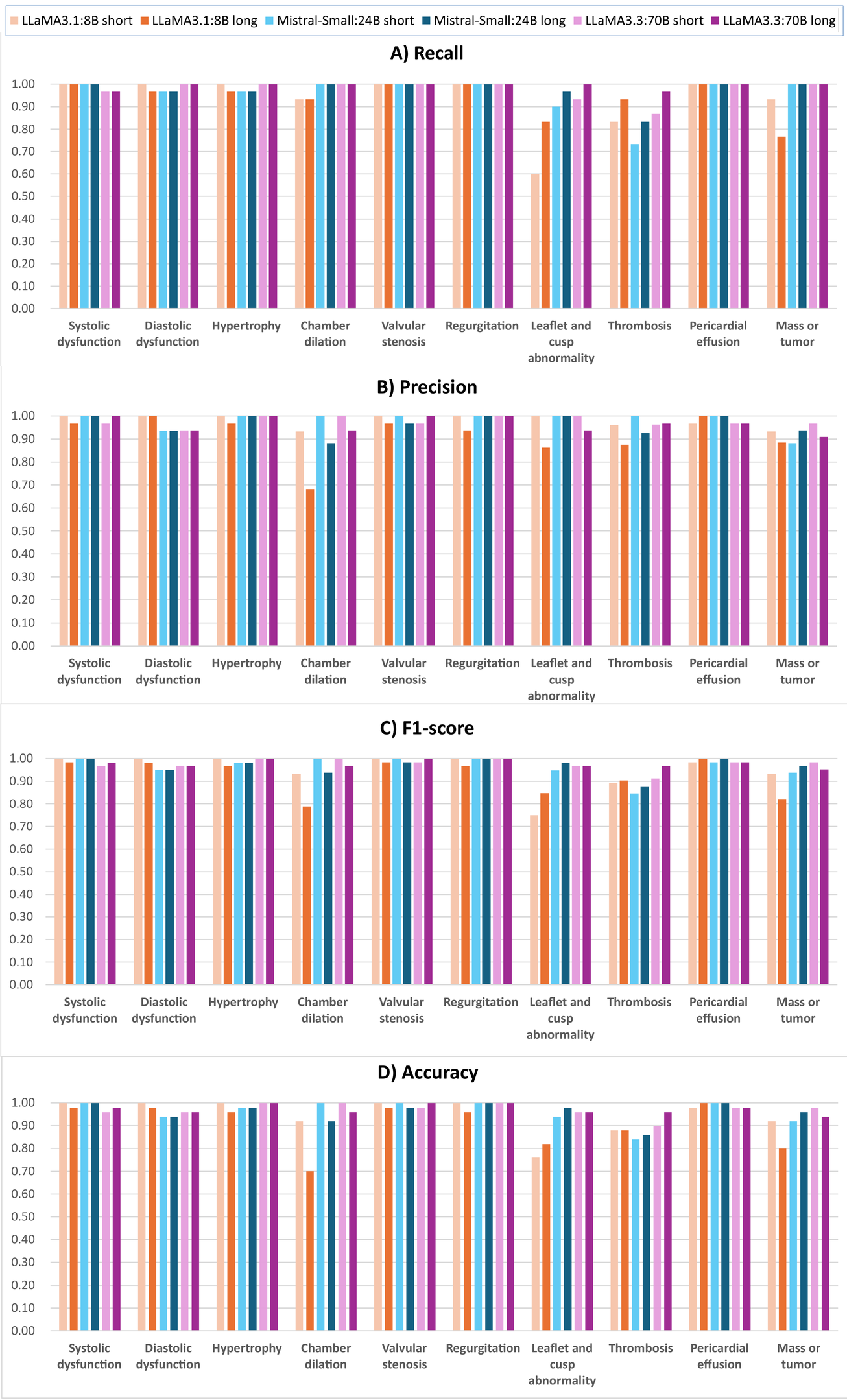

Figure 2. Performance comparison of large language models across prompt formats and cardiac features: A) Recall, B) Precision, C) F1-score, and D) Accuracy

RESULTS

- **Model comparison:** The small model (LLaMA3.1:8B) exhibited greater variability in accuracy (range: 0.70–1.00), whereas the large model (LLaMA3.3:70B) consistently scored above 0.90 for all cardiac features. (**Figure 2**)
- **Feature comparison**: Chamber dilatation and leaflet or cusp abnormalities showed the lowest accuracy and F1-score.
- **Effect of using the longer prompt**:
  - **Small model (LLaMA3.1:8B):** More features with accuracy and F1-score <90%. The most significant decline was observed in chamber dilatation, with accuracy dropping from 0.92 to 0.70 and F1-score from 0.93 to 0.79.
  - **Mid-scale (Mistral-Small:24B) and large-scale (LLaMA3.3:70B) models:** Performance remained stable across prompt formats.

CONCLUSION

- Open-source LLMs demonstrated strong utility in extracting information from unstructured echocardiogram reports.
- However, the performance of smaller models was more affected by the complexity of prompts and variability in clinical text descriptions.
- These results underscore the importance of tailoring LLMs and prompt strategies to facilitate data enrichment for real-world evidence.

References:

1. Writing Committee Members; ACC/AHA Joint Committee Members. 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure. *J Card Fail*. 2022;28(5):e1-e167. doi:10.1016/j.cardfail.2022.02.010
2. Nath C, Albaghdadi MS, Jonnalagadda SR. A Natural Language Processing Tool for Large-Scale Data Extraction from Echocardiography Reports. *PLoS One*. 2016;11(4):e0153749. doi:10.1371/journal.pone.0153749
3. Jia YY, Pang LY, Bi MM, Yang XL, Song JP. Dependability of Large Language Models in Cardiovascular Medicine: A Scoping Review. *J Cardiothorac Vasc Anesth*. Published online July 22, 2025. doi:10.1053/j.jvca.2025.07.026

For further information, please contact, please contact yth.donna@gmail.com