European Heart Journal (2025) 46 (Suppl 1)
e-Cardiology/Digital Health – Large Language Model

LLM powered echocardiography report structuring: a step towards precision medicine

W. Van Der Loo[1], V.O. Van Der Valk[2], T.J. Van Den Broek[3], D.E. Atsma[1], M. Staring[2], R.W.C. Scherptong[1]

[1]Leiden University Medical Center, Department of Cardiology, Leiden, Netherlands (The)
[2]Leiden University Medical Center, Department of Radiology, Leiden, Netherlands (The)
[3]TNO Research Institute, Leiden, Netherlands (The)

Funding Acknowledgements: Type of funding sources: Public grant(s) – EU funding. Main funding source(s): iCARE4CVD

Introduction: Large volumes of transthoracic echocardiography (TTE) reports are stored in electronic health records (EHRs) worldwide, but their free-text format limits utility for research and machine learning applications. (1) Manual annotation remains the primary structuring method, a time-consuming and subjective process. (2) Recent advancements in natural language processing, particularly large language models (LLMs), offer a promising solution. LLMs excel in processing human language, making them valuable for automated medical data annotation. (3)

Purpose: Two LLM-based methods were developed and evaluated for the automated classification of TTE reports, facilitating efficient structuring of unstructured clinical data.

Methods: TTE reports, written in Dutch, from post-myocardial infarction patients were retrospectively collected from a local EHR system (2010-2024). A random subset of 1000 reports was manually annotated for left ventricular (LV) function and valvular pathology. LV function was categorized as normal, mildly, moderately, severely impaired or unknown. Valve dysfunction was classified by type (none, regurgitation, stenosis, both, or unknown) and severity (none, mild, moderate, severe or unknown) based on cardiologists' final reports. Reports were reviewed by two researchers, with discrepancies resolved through discussion. Data was randomly split into training (n=700) and test (n=300) sets. Two classification approaches were developed: (1) a few-shot prompt engineering method (FS); Iteratively optimized, category-specific prompts with a commercially available LLM, (2) a fine-tuning method (FT) trained on multiple state-of-the-art pretrained LLMs. Model performance was assessed using accuracy, F1-scores and recall. Statistical significance per label was evaluated using bootstrap resampling and a paired t-test.

Results: TTE reports included 501 unique patients. The dataset exhibited class imbalance, reflecting real-world distributions, with most cases classified as mildly impaired LV function and no/mild valvular dysfunction. Inter-observer agreement was high (98%-100%). The best fine-tuning results were obtained with RoBERTa, pretrained on multilingual datasets. (4) FS outperformed FT in all categories, achieving 99% overall accuracy vs. 93%. All recall and F1-scores were statistically significantly better for the FS method. Most misclassifications involved distinguishing mildly impaired from normal valve function.

Conclusion: The FS method approached expert-level accuracy, enabling rapid and standardized classification of TTE reports, significantly reducing the time required for dataset creation in large-scale clinical research. The FS method achieved superior classification accuracy over fine-tuning, likely due to model scale and class imbalance, based on the differences in recall and F1-score. Most errors occurred in distinguishing mild impairment from normal valve function, a distinction with limited clinical impact.

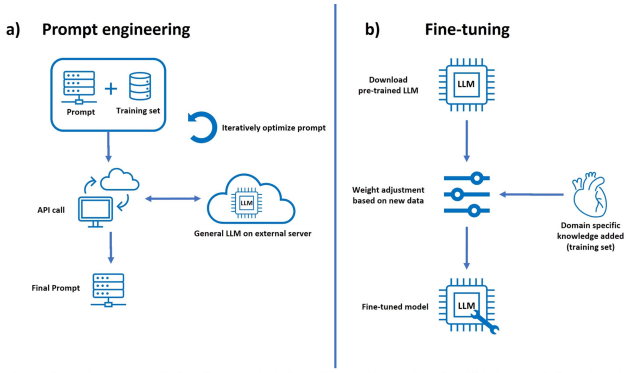

Figure 1. **a)** prompt engineering method, resulting in an optimized prompt, no model trainings takes place. **b)** Fine tuning method, a pre-trained LLM is downloaded and trained on a domain specific dataset, adjusting the weights of the network. Resulting in a fine-tuned model