Towards expert-level AI-based classification of coronary angiography reports: a comparison of LLM approaches

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 invasive coronary angiography (ICA) reports are stored in electronic health records (EHRs) worldwide, but their free-text format limits 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 ICA reports, facilitating efficient structuring of unstructured clinical data.

Methods: ICA reports, written in Dutch, from patients with acute coronary syndrome were retrospectively collected from a local EHR system (2010-2022). A random subset of 1000 reports was manually annotated for occlusion, bypass graft presence, macrovascular coronary artery disease (CAD) (binary), intervention type (PCI, CABG or no intervention), and culprit vessel(s) (main and branch). Annotations were based on the cardiologists' final reports and were reviewed by two researchers, with discrepancies resolved through discussion. The data was randomly split in a training (n=700) and a testing (n=300) set. Two classification approaches were developed: (1) a few-shot prompt engineering (FS) method, using iteratively optimized prompts with a commercially available LLM and (2) a fine-tuning method (FT), trained on multiple state-of-the-art pretrained LLMs. Culprit vessel classification was correct only if predefined labels were exactly matched; partial matches were classified as incorrect. Model performance was assessed using accuracy, macro-averaged F1-scores, and recall. Statistical significance differences between models, were evaluated per label using bootstrap resampling and a paired t-test.

Results: All ICA reports belonged to unique patients. The Inter-observer agreement for manual annotation ranged from 81% (culprit branch) to 98% (graft label). The best fine-tuning results were obtained with RoBERTa, pretrained on multilingual datasets. (4) Average accuracy was 89% for the FT method and 88% for the FS method. Accuracy of culprit branch prediction and recall of the no CAD prediction were statistically significantly different between the models (FT outperformed FS in culprit branch, FS outperformed FT for no CAD). Most classification errors occurred in selecting the correct culprit branch(es).

Conclusion: Both methods approached expert-level accuracy, enabling rapid and standardized classification of ICA reports, significantly reducing the time required for dataset creation in large-scale clinical research. Interestingly, the FT method reached similar or better results compared to the much larger FS model. Meaning that with a relatively small annotated dataset, similar results can be achieved in a much less costly manner.

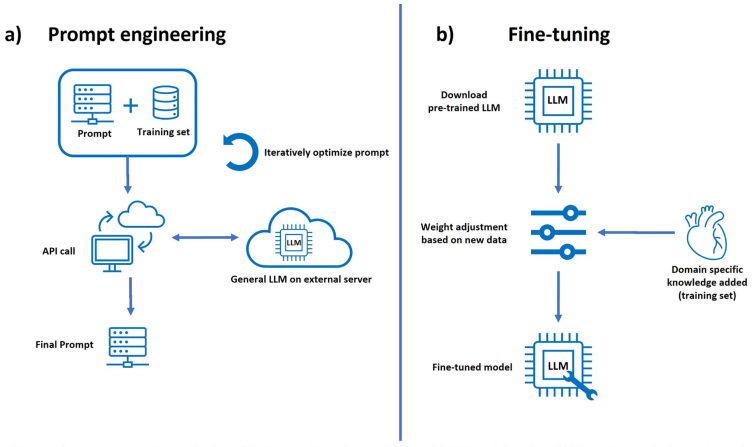

Figure 1. a) prompt engineering method, resulting in an optimized prompt, no model trainings takes place. **b)** Fine tuning method, a pre-trained LLM is downloaded and trained on a domain specific dataset, adjusting the weights of the network. Resulting in a fine-tuned model