Standardizing 12-Lead ECG Free-Text Reports to SNOMED-CT Using Mid-Sized Language Models

HyeonWoo Lee[1], Dillon J Dzikowicz[1], Junmo An[1]

[1] Philips, Cambridge, MA, USA

Large-scale electrocardiogram (ECG) datasets are increasingly available; however, their associated clinical interpretations remain largely unstructured free-text, limiting their utility for computational analysis. This study aims to develop and evaluate a scalable, privacy-preserving framework to automatically standardize 12-lead ECG free-text reports into Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) using open-weight, mid-sized language models.

We utilized the PTB-XL dataset and focused on 31 clinically relevant ECG interpretations aligned with expert-validated SNOMED-CT mappings. The task was formulated as a multi-label classification problem. Three mid-sized language models (GPT-OSS-20B, MedGemma-27B-text-it, DeepSeek-R1-Distill-Qwen-32B) were evaluated under three configurations: (1) zero-shot instruction prompting, (2) retrieval-augmented generation (RAG) with few-shot context retrieved via vector similarity search, and (3) parameter-efficient fine-tuning (PEFT) using Low-Rank Adaptation (LoRA). Performance was assessed on a held-out test set using precision, recall, accuracy, and F1-score.

In the zero-shot setting, DeepSeek-R1-Distill-Qwen-32B achieved the best performance (F1 = 0.704), outperforming MedGemma-27B (0.650) and GPT-OSS-20B (0.426). Incorporating RAG improved performance across models, with MedGemma-27B achieving the highest F1-score (0.813). PEFT via LoRA yielded the strongest overall results, with DeepSeek-R1-Distill-Qwen-32B achieving the highest F1-score (0.881), followed by MedGemma-27B (0.866) and GPT-OSS-20B (0.836). These findings demonstrate consistent gains from both retrieval-based augmentation and task-specific fine-tuning.

Mid-sized open-weight language models can effectively standardize free-text ECG reports into structured SNOMED-CT labels, enabling scalable and privacy-preserving clinical data annotation. While RAG provides strong performance without retraining, PEFT delivers the highest accuracy, highlighting the importance of task-specific adaptation. This framework facilitates the transformation of legacy ECG datasets into machine-readable formats and supports the scalable construction of high-quality labeled training databases for downstream AI model development, ultimately enabling more robust and generalizable cardiovascular AI systems.