Towards Clinically Faithful ECG Reports via Quantization-Based Tokenization

Rohan Banerjee[1,2], **Jacques Delfrate**[1,2] **& Robert Avram**[1,2,3]
[1]Heartwise (heartwise.ai), Montreal Heart Institute, Montreal, Quebec, Canada
[2]Montreal Heart Institute, Department of Medicine, Montreal, Quebec, Canada
[3]Department of Biochemistry and Molecular Medicine, Faculty of Medicine,
University of Montreal, Montreal, Quebec, Canada
banerjee.rohan98@gmail.com

Abstract

Effective tokenization is a critical barrier to bridging continuous electrocardiogram (ECG) signals and discrete language models for automated report generation. We introduce a novel ECG tokenizer based on an adaptive residual vector quantization framework, QINCo, that learns a high-fidelity, discrete representation of raw 12-lead signals. The clinical utility of these tokens is demonstrated through a downstream classification task, where their frozen embeddings achieve performance comparable to specialized supervised and self-supervised methods. Leveraging this tokenizer with an attention-based adapter, our approach to report generation outperforms byte-level tokenizer baselines and establishes a benchmark on a large-scale clinical dataset. Our work presents an effective and computationally efficient tokenization framework, enabling a more powerful integration of complex biosignals into generative models for clinical applications.

1 Introduction

Electrocardiography (ECG) is the most frequently performed non-invasive diagnostic tool in cardiology, renowned for its accuracy in detecting a broad range of cardiovascular diseases [1]. The increasing volume of ECG recordings, coupled with the limitations of manual analysis which demand considerable expertise of at least four years of medical training and six years of cardiology training and can be prone to variability and error rates reaching 25 percent [2], highlights the need for accurate, automated clinical report generation. Artificial intelligence (AI) has been shown to outperform cardiologists in ECG interpretation [3]; however, most existing ECG-AI algorithms are limited to classification tasks and fail to capture the full variability of clinical reports. To address this, Large Language Models (LLMs) could be used for report generation, however effective methods for representing raw ECG signals in a discrete, tokenized format are essential in order to do so.

Our work focuses on exploring tokenization methods to convert continuous ECG signals into discrete tokens. By learning meaningful discrete representations directly from raw ECG data, we aim to bridge the gap between complex waveform patterns and automated LLM applications such as clinical report generation and decision support in diagnostic assistance.

Tokenization converts complex data into discrete units, facilitating alignment of continuous signals with LLMs. In language, sentences are tokenized into discrete units enabling efficient processing by LLMs. In the audio domain, tokenization has revolutionized tasks such as speech recognition and synthesis, allowing for the representation of continuous audio waveforms as sequences of continuous or discrete acoustic tokens [4, 5] leading to the development of multi-modal LLMs [6]. Similarly, in the physiological signals space including ECG, EEG, EHR, there have been works aligning continuous

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

representations with LLMs [7, 8, 9, 10]. If we specifically focus on ECG, the literature exploring discrete tokenization remains limited. A recent work ECGByte [11] explored byte-pair encoding for tokenizing ECG signals by taking inspiration from how tokenization is done in LLMs [12]. We draw inspiration from audio codecs [13, 14] and use vector quantization (VQ) [15], specifically QINCo [16] to tokenize ECG signals. QINCo has an adaptive residual quantization nature which would dynamically tailor its codebooks at each training epoch in order to capture the subtle and time-varying patterns present in ECG data.

2 Methods

2.1 Data

The development and evaluation of our models were conducted using two large-scale clinical ECG datasets: the public MIMIC-IV-ECG [17] database with 551,306 ECGs, and the Montreal Heart Institute (MHI) dataset with 1,453,937 ECGs, previously used in the work of [18]. To augment the initial pretraining of our ECG tokenizer, we also incorporated the unlabeled CODE-15 dataset, which comprises 345,779 ECGs. The MIMIC-IV-ECG dataset consists of 10-second, 12-lead ECGs sampled at 500Hz. The MHI dataset contains recordings with the same duration and lead configuration but sampled at 250Hz; for this cohort, we exclusively selected ECGs with clinician-verified reports to ensure a high standard of ground truth. Both datasets were partitioned into training and testing sets using an 80:20 ratio. This partitioning was performed at a patient level to prevent data leakage and was stratified across 77 clinical condition labels to ensure a balanced diagnostic distribution.

Preprocessing To standardize the MIMIC-IV and CODE-15 datasets against the MHI reference, we applied a multi-stage preprocessing pipeline to each recording. First, all the ECGs were resampled to 250Hz. Then, to correct for baseline wander, a zero-phase 1 Hz high-pass filter was applied only if the spectral energy below 1 Hz exceeded that of the 1–30 Hz band by more than tenfold. Next, narrowband powerline noise at 50/60 Hz and their harmonics were suppressed by identifying spectral peaks that exceeded the local mean by over two standard deviations and flattening them with a LOESS fit. Finally, all signal amplitudes were rescaled to millivolts (mV) to match the dynamic range of the MHI dataset, ensuring harmonized units and spectral properties across all cohorts.

Clinical Labels Following the methodology of [18], we established a multi-label ground truth of 77 diagnostic labels. These labels span six clinical categories, namely, Rhythm Disorder, Conduction Disorder, Chamber Enlargement, Pericarditis, Infarction or Ischemia, and Other, in accordance with American Heart Association recommendations [19]. The label set was created by two expert cardiologists, each with over four years of experience, who independently annotated a curated subset of 10,075 reports from the MHI, MIMIC-IV, and UK BioBank datasets to ensure comprehensive coverage. Any of the 77 conditions were marked as present if observed within the 10-second ECG recording. The high reliability of this process was validated through an inter-rater variability analysis, which yielded Cohen's kappa coefficients exceeding 0.80 across all diagnostic categories.

2.2 ECG Tokenization

The foundation of this method is to convert continuous 12-Lead ECG signals into a sequence of discrete tokens. For this task, we employ QINCo, a state-of-the-art residual vector quantization (RVQ) method. This approach is distinguished by its ability to dynamically generate codebooks using neural networks, enabling an adaptive and efficient quantization process tailored to the intricate patterns of ECG data. Our tokenizer is an autoencoder-style network where the bottleneck contains the quantization module which we discuss below.

Encoder The encoder serves as a feature extractor, mapping a raw 12-lead ECG signal into a sequence of lower-dimensional continuous vectors. It consists of a deep convolutional neural network with subsequent sequential blocks employing residual connections. Each block systematically reduces the temporal resolution via strided convolutions and max-pooling while expanding the feature representation by increasing the channel depth.

Quantizer Our model's core discretization module transforms the continuous latent vectors $z_t \in R^D$ from the encoder into discrete tokens using an L-stage RVQ scheme. A fundamental limitation of standard RVQ is its reliance on static codebooks, which are suboptimal for the evolving residual error across stages. To overcome this, we employ the QINCo paradigm, where each codebook is generated dynamically. At each stage l, a dedicated neural network f_{θ_l} takes the partial reconstruction from the previous stage, $\hat{z}_t^{(l-1)}$, and a set of base centroids to produce a context-aware codebook. The index of the nearest codeword, $s_t^{(l)} = \text{argmin}_k ||r_t^{(l)} - c_k^{(l)}||^2$, is then selected to represent the current residual.

Decoder The reconstruction of the quantized latent vector, $\widetilde{z}_t$, is performed iteratively, a direct consequence of the QINCo framework's conditional nature. Since the generation of an adaptive codeword $c^l_{s_t^{(l)}}$ at the stage l depends on the previous partial reconstruction $\hat{z}_t^{(l-1)}$, the vector is progressively synthesized using the update rule: $\hat{z}_t^{(l)} = \hat{z}t^{(l-1)} + c^{(l)}_{s_t^{(l)}}$. The final synthesized vector, $\hat{z}_t := \hat{z}_t^{(L)}$, is then passed to a signal decoder composed of transposed convolutional layers, which upsamples the representation to reconstruct the final ECG signal $\hat{x}$.

Learning Objective The tokenizer is trained end-to-end by minimizing a composite loss function: $\mathcal{L} = \mathcal{L}_{rec} + \alpha\mathcal{L}_{commit}$. The primary component, $\mathcal{L}_{rec}$, is the Mean Absolute Error (L1 loss) between the original signal x and its reconstruction $\hat{x}$, ensuring high-fidelity tokenization. The auxiliary term, $\mathcal{L}_{commit}$, is a commitment loss that regularizes the encoder by minimizing the squared Euclidean distance between its output $E(x_t)$ and the chosen codeword $c^{(l)}_{s_t^{(l)}}$. A stop-gradient operator (sg) is applied to the codeword, as shown in $||E(x_t) - \text{sg}(c^{(l)}_{s_t^{(l)}})||_2^2$, to ensure that gradients from this loss only update the encoder. This joint optimization learns discrete tokens that are both informative for reconstruction and arise from a stable latent space.

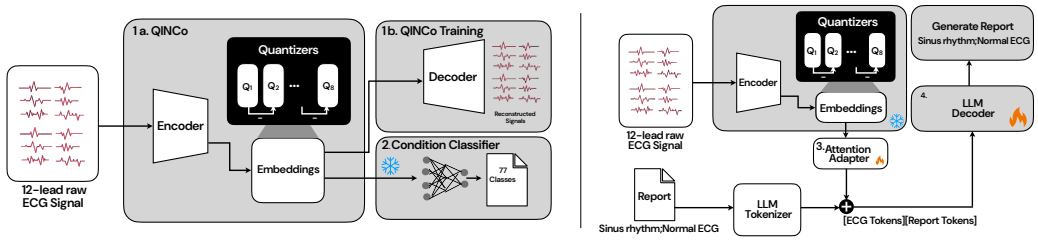

Figure 1: **Overview of the Proposed Methodology.** The **left panel** illustrates the tokenizer pre-training and validation: (1a) a QINCo-based model converts a raw ECG into embeddings, which are (1b) trained via a signal reconstruction objective. (2) The frozen embeddings are then used to train a classifier for intrinsic validation. The **right panel** details the LLM fine-tuning for report generation: (3) the frozen ECG embeddings are aligned with text tokens via a trainable attention adapter, and (4) the combined representation is used to fine-tune the LLM.

2.3 Report Generation

The primary application of our ECG tokenization method is its integration into a system for automated clinical report generation. The objective of this task is to translate the information encoded in our discrete ECG representations into fluent and clinically accurate diagnostic text. To this end, we fine-tune pre-trained LLMs using a multimodal architecture. The subsequent sections describe this architecture and the corresponding fine-tuning methodology.

Attention-Based ECG Adapter To align our ECG representations with the LLM's embedding space, we introduce a trainable adapter module. This adapter's function is to map the sequence of quantized ECG vectors, $Z_{ecg} = (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_T)$, to a fixed-size context vector, $c_{ecg} \in \mathbb{R}^{D_{llm}}$, where D_{llm} is the decoder-only LLM's embedding dimension. The transformation begins by projecting the input sequence to an intermediate dimension and adding learnable positional embeddings. This sequence is then contextualized using a multi-head self-attention layer with a residual connection.

Finally, a global average pooling operation aggregates the sequence into a vector, which is passed through a final projection layer to produce the context vector c_{ecg}.

LLM Finetuning The context vector c_{ecg} is prepended to the sequence of the report's word embeddings, $(w_1, w_2, ..., w_N)$, to form the complete input to the LLM. The system is then fine-tuned by updating the parameters of the ECG adapter and the pre-trained LLM, while the ECG tokenizer remains frozen. This training follows a standard auto-regressive, next-token-prediction objective. The model is optimized to minimize the cross-entropy loss between its predicted token distribution and the ground-truth report, with the learning objective formally defined as: $\mathcal{L}_{LM} = -\sum_{i=1}^{N} \log P(w_i|\mathbf{c}_{ecg}, w< i)$ where w_i is the target token at position i, conditioned on the ECG context vector and all preceding tokens.

2.4 Evaluation of ECG Representations

To evaluate the clinical utility of our learned ECG representations, we performed an intrinsic validation on a multi-label cardiac condition classification task. For this, the pre-trained ECG tokenizer was kept frozen, and its output embedding served as the direct input to a downstream classifier. To ensure a fair comparison with competitive baselines, we employed an EfficientNetV2 [20] as the classifier.

3 Experiments

Tokenizer training We evaluated three VQ methods for ECG tokenization: a vanilla VQ, RVQ, and QINCo. To ensure a fair comparison, all models were trained on the MIMIC train set, MHI train set and Code-15 for 10 epochs with a consistent codebook configuration totaling 512 codewords. A key quantitative difference emerged in codebook utilization: the percentage of active codes used after training was 38.9% for VQ, 88.41% for RVQ, and 99.99% for QINCo. This suggests that the vanilla VQ and, to a lesser extent, RVQ suffered from codebook collapse. This quantitative finding was corroborated by a qualitative assessment of reconstructed MIMIC and MHI test-set signals. We observed that both VQ and RVQ frequently failed to accurately reconstruct critical features such as the QRS complex and introduced artifacts in the T-waves. In contrast, QINCo consistently produced the highest fidelity reconstructions, a result we attribute to its efficient and near-total utilization of its available codebook capacity.

Condition Classification To assess the clinical utility of our ECG representations, we benchmarked them on a multi-label condition classification task. Following the protocol of DeepSL and DeepSSL [18], we attached an EfficientNetV2 classifier head to the frozen embeddings from our QINCo tokenizer, as well as those from vanilla VQ and RVQ baselines, using identical hyperparameters. Reporting the micro-averaged Area Under the Curve, we observe that our QINCo-based method performs on par with the state-of-the-art DeepSL and DeepSSL. Notably, our training set is smaller due to a stricter criterion of using only clinician-verified reports, which we attribute to a slight performance decrease in specific classes like Infarction. The significantly lower performance of the vanilla VQ representations further validates that a multi-stage, residual quantization approach is critical for capturing clinically relevant ECG features.

Report Generation For the report generation task, we first established a direct comparison against our primary baseline, ECG-Byte. To ensure a fair evaluation, both our method and the baseline were trained on an identical subset of the MIMIC-IV dataset (n_train=350,997, n_test=100,031), which excludes signals that were filtered out during our preprocessing due to quality control issues. We conducted experiments in two settings: a standard format where the report is generated directly from ECG tokens, and a question-answering format where question tokens are also provided as input. For the standard and instruction-based settings, we LoRA [21] fine-tuned the Llama3.2-1B and Llama3.2-1B-Instruct [22] models, respectively. Critical hyperparameters, including batch size of 2, 1 training epoch, and the learning rate scheduler, were kept consistent with the ECG-Byte methodology. We observed that our method using the base Llama3.2-1B model, without any instruction-tuning, achieved the best overall performance in this comparison. This suggests that our discrete tokens carry more structured, clinically salient information than the byte-level embeddings used by the baseline.

Building on the finding that direct generation was most effective, we conducted a final experiment to assess performance on a more diverse, real-world clinical dataset. For this, we fine-tuned the

Table 1: **Report Generation Performance on MIMIC-IV.** Comparison of our proposed tokenizer (ECG-Tok) against the ECG-Byte baseline. We report standard NLP metrics and clinical finding F1-scores for models trained with and without instruction-tuning ('inst.').

Method	ROUGE-1	ROUGE-L	METEOR	BLEU-4	Precision	Recall	F1
ECG-Byte (inst.)	0.171	0.160	0.204	0.021	0.807	0.877	0.840
ECG-Byte	0.216	0.207	0.212	0.038	0.847	0.883	0.864
ECG-Tok (inst.)	0.429	0.427	0.409	0.234	0.771	0.864	0.812
ECG-Tok	0.548	0.542	0.518	0.293	0.928	0.913	0.920

Llama3.2-1B model without LoRA on a combined dataset of both MIMIC and MHI train set and inference on the MIMIC and MHI test set, using the standard ECG-to-report generation format. This model achieved good performance, yielding ROUGE-1, ROUGE-L, BLEU-4, and METEOR scores of 0.671, 0.667, 0.415, and 0.640, respectively. These results quantitatively surpass the baseline and, upon qualitative review, produce the most clinically coherent reports.

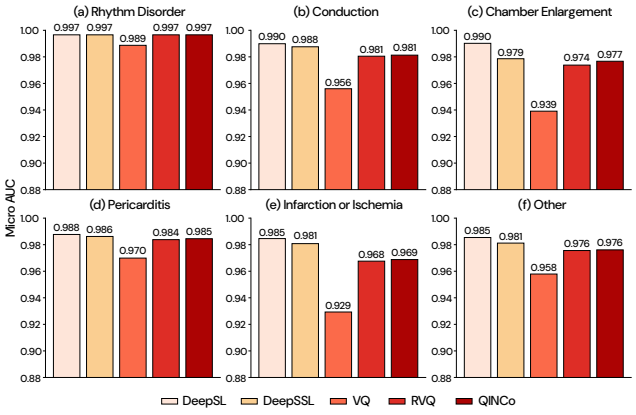

Figure 2: **Performance on Cardiac Condition Classification.** Micro-averaged AUC comparing our QINCo-based representations against vanilla VQ, RVQ, and the DeepSL (supervised) and DeepSSL (self-supervised) baselines across six clinical diagnostic categories.

4 Discussion

In this work, we introduced a novel method for the discrete tokenization of 12-lead ECG signals using the QINCo framework. We demonstrated that this approach learns clinically relevant representations without direct supervision. This was evidenced by its performance on a downstream cardiac condition classification task, which was comparable to that of specialized supervised and self-supervised methods. We then showed that these high-fidelity representations significantly boost the performance of our primary application: automated clinical report generation.

While our system can generate coherent reports from real-world clinical data, we acknowledge that crucial steps must be taken before it can be considered for clinical deployment. We observed that classical NLP metrics like ROUGE, BLEU, and METEOR are often insensitive to critical clinical errors. For example, one generated report received high quantitative scores (ROUGE-L of 0.652, F1 of 0.897) despite introducing a highly significant and misleading finding "*** CONSIDER ACUTE ST ELEVATION MI ***" that was absent from the ground truth. Such a clinical "hallucination" should be penalized, but current metrics fail to do so adequately. This highlights a need for clinically-aware metrics. To address this, we plan to supplement our quantitative analysis with a qualitative validation of generated reports by expert cardiologists, which is essential for any rigorous clinical evaluation.

Another key observation relates to the trade-off between token efficiency and instruction-tuning. Our use of a small, fixed number of tokens per ECG significantly reduces training and inference time. However, this compactness appeared to hinder the effectiveness of instruction-tuning, consistently worsening its performance. We hypothesize that when the number of ECG tokens is small, the generic text of the instruction prompts consumes a disproportionate amount of the model's capacity and training gradients, thereby diluting the supervision signal for the actual report generation.

Looking ahead, we contend that simply scaling the decoder size of the LLM is not the primary path to improving performance in this domain. Instead, we believe future gains will be driven by several key factors: increasing the number of ECG tokens to provide a richer and more detailed conditional signal, designing more specific and directed questions for instruction-tuning, developing novel alignment losses that better correlate with clinical accuracy, and further scaling the diversity and volume of pre-training data.

References

[1] Nikita Rafie, Anthony H. Kashou, and Peter A. Noseworthy. Ecg interpretation: Clinical relevance, challenges, and advances. *Hearts*, 2021.

[2] David A. Cook, So Young Oh, and Martin V. Pusic. Accuracy of physicians' electrocardiogram interpretations: A systematic review and meta-analysis. *JAMA internal medicine*, 2020.

[3] J. Weston Hughes, Jeffrey E. Olgin, Robert Avram, Sean Abreau, Taylor Sittler, Kaahan Radia, Henry H. Hsia, Tomos E. Walters, Byron K. Lee, Joseph E. Gonzalez, and Geoffrey H. Tison. Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation. *JAMA cardiology*, 2021.

[4] Pooneh Mousavi, Jarod Duret, Salah Zaiem, Luca Della Libera, Artem Ploujnikov, Cem Subakan, and Mirco Ravanelli. How should we extract discrete audio tokens from self-supervised models? *ArXiv*, abs/2406.10735, 2024.

[5] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models. *ArXiv*, abs/2308.16692, 2023.

[6] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2022.

[7] Yifu Cai, Arvind Srinivasan, Mononito Goswami, Arjun Choudhry, and Artur Dubrawski. Jolt: Jointly learned representations of language and time-series for clinical time-series interpretation (student abstract). In *AAAI Conference on Artificial Intelligence*, 2024.

[8] Jungwoo Oh, Hyunseung Chung, Joon myoung Kwon, Dongwoo Hong, and E. Choi. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. *ArXiv*, abs/2203.06889, 2022.

[9] Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu kai Wang, and Ching-Teng Lin. Dewave: Discrete eeg waves encoding for brain dynamics to text translation. *ArXiv*, abs/2309.14030, 2023.

[10] Yoonhyung Lee, Younhyung Chae, and Kyomin Jung. Leveraging vq-vae tokenization for autoregressive modeling of medical time series. *Artificial intelligence in medicine*, 154:102925, 2024.

[11] William Jongwon Han, Chaojing Duan, Michael A. Rosenberg, Emerson Liu, and Ding Zhao. Ecg-byte: A tokenizer for end-to-end generative electrocardiogram language modeling. *ArXiv*, abs/2412.14373, 2024.

[12] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. OpenAI Technical Report.

[13] Alexandre D'efossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *ArXiv*, abs/2210.13438, 2022.

[14] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

[15] C. P. Mammen and Bhaskar Ramamurthi. Vector quantization for compression of multichannel ecg. *IEEE Transactions on Biomedical Engineering*, 37:821–825, 1990.

[16] Iris Huijben, Matthijs Douze, Matthew Muckley, Ruud van Sloun, and Jakob Verbeek. Residual quantization with implicit neural codebooks. *ArXiv*, abs/2401.14732, 2024.

[17] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10, 2023.

[18] Alexis Nolin-Lapalme, Achille Sowa, Jacques Delfrate, Olivier Tastet, Denis Corbin, Merve Kulbay, Derman Ozdemir, Marie-Jeanne Noël, François-Christophe Marois-Blanchet, François Harvey, Surbhi Sharma, Minhaj Ansari, I-Min Chiu, Valentina Dsouza, Sam F. Friedman, Michaël Chassé, Brian J. Potter, Jonathan Afilalo, Pierre Elias, Gilbert Jabbour, Mourad Bahani, Marie-Pierre Dubé, Patrick M. Boyle, Neal A. Chatterjee, Joshua P Barrios, Geoffrey H. Tison, David Ouyang, Mahnaz Maddah, Shaan Khurshid, Julia Cadrin-Tourigny, Rafik Tadros, Julie G. Hussin, and Robert Avram. Foundation models for generalizable electrocardiogram interpretation: comparison of supervised and self-supervised electrocardiogram foundation models. *medRxiv*, 2025.

[19] Paul D. Kligfield, Leonard S. Gettes, James J. Bailey, Rory W. Childers, Barbara Deal, E. William Hancock, Gerard van Herpen, Jan A. Kors, Peter W. Macfarlane, David M. Mirvis, Olle Pahlm, Pentti M. Rautaharju, Galen S. Wagner, Mark E. Josephson, Jay W. Mason, Peter M. Okin, Borys Surawicz, and Hein Maarten Wellens. Recommendations for the standardization and interpretation of the electrocardiogram: part i: the electrocardiogram and its technology a scientific statement from the american heart association electrocardiography and arrhythmias committee, council on clinical cardiology; the american college of card. *Journal of the American College of Cardiology*, 49 10:1109–27, 2007.

[20] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, 2021.

[21] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

[22] Abhimanyu Dubey et al. The llama 3 herd of models. *ArXiv*, abs/2407.21783, 2024.