Scaling ECG-Language Models: A Comparative Study of PEFT-Based Adaptation Across Backbone Architectures

Anonymous Authors[1]

Abstract

We present preliminary findings from an ongoing systematic study of scaling ECG language models via Parameter-Efficient Fine-Tuning (PEFT). We compare four language models spanning two families and two size tiers—Qwen3.5-9B and Gemma 4-E4B (4.5B effective) as small models, Qwen3.6-27B and Gemma 4-31B as large models—each connected to a pretrained wav2vec2 ECG encoder via a projection layer. We compare three PEFT strategies: *projection-only* (LLM frozen), *joint PEFT* (connector + LoRA on LLM), and *few-shot connector PEFT* (with in-context examples). Projection-only fine-tuning consistently outperforms joint PEFT for large quantized models, which destabilize under cosine LR restarts. Scaling is non-monotonic: the 9B model achieves **46.79%** exact-match accuracy on PTB-XL (+14.3% over prior ELMs), outperforming the 27B by 9.7 points, while Gemma 4-31B achieves **74.19** ROUGE-L (+32.3%). Crucially, 20K-scale results show in-context learning unlocks latent capabilities in quantized models: the 27B with few-shot outperforms its projection-only baseline across *all* metrics (e.g., +2.9% ACC, +4.2 F1 on PULSE), while the 9B with few-shot trades precision for fluency. The 31B with few-shot shows more modest gains (+0.6% ACC on PTB-XL), suggesting the effect is modulated by both quantization and architecture. Failure-mode analysis reveals a clinical dichotomy: Qwen excels at rule-out (93–95% specificity) while Gemma 4 excels at screening (66–75% sensitivity).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, contributing to over 17 million deaths annually (World Health Organization, 2024). The electrocardiogram (ECG) is the primary non-invasive diagnostic tool for assessing cardiac electrophysiology, providing critical temporal waveform information for diagnosing arrhythmias, myocardial ischemia, and conduction abnormalities.

Recent advances in deep learning have achieved expert-level performance on specific ECG classification tasks (Nolin-Lapalme et al., 2026; McKeen et al., 2025). However, most existing approaches are narrow classifiers trained for single-task outputs, limiting their clinical utility. The emergence of ECG-language models (ELMs) (Jin et al., 2025; Zhao et al., 2025; Lan et al., 2025) presents an opportunity to develop unified systems capable of generating interpretable clinical reports from raw signals.

In this work, we systematically investigate how the choice of language model backbone and fine-tuning strategy affect ECG interpretation quality across four models spanning two families and two size tiers. Our key findings are:

(1) Backbone matters, but not monotonically with size. The 9B Qwen3.5 model achieves higher exact-match accuracy than both the 27B Qwen3.6 and the 31B Gemma 4, while Gemma 4 produces richer, more comprehensive responses. The models exhibit opposite failure modes—terse precision vs. verbose recall—suggesting complementary clinical roles.

(2) Projection-only fine-tuning outperforms joint PEFT for large quantized models under current configurations. Updating only the projection layer (connector) between the ECG encoder and frozen LLM yields better and more stable results than co-training LoRA adapters on the LLM. The joint PEFT configuration peaks early and then degrades monotonically, suggesting that aggressive learning-rate restarts destabilize quantized LLM weights.

(3) Few-shot prompting improves 4-bit models across all metrics. Adding in-context examples (few-shot connector PEFT) substantially improves performance on generative tasks (BLEU-4 +5.7 on PULSE). Crucially, prelim-

inary 20K-scale results show this benefit is architecture-dependent: the 4-bit 27B model improves on *all* metrics including ACC (+2.9%), while the bf16 9B model trades precision for fluency—suggesting in-context examples help quantized models compensate for information loss.

(4) Clinical dichotomy: Screener vs. Rule-Out. Binary abnormality detection analysis reveals that Gemma 4 models are better *screeners* (66–75% sensitivity), while Qwen models are better *rule-out* tools (93–95% specificity). Neither is suitable for autonomous diagnosis, but together they suggest a dual-model clinical pipeline.

2. Related Work

ECG Foundation Models. Recent work has shifted from narrow classifiers toward foundation models trained on millions of ECGs. DeepECG-SSL (Nolin-Lapalme et al., 2026) uses self-supervised contrastive learning on over 1 million ECGs. ECG-FM (McKeen et al., 2025) provides open weights with hybrid contrastive alignment. These models demonstrate strong generalization but require substantial computational resources for full fine-tuning.

ECG-Language Models. HeartLang (Jin et al., 2025) treats heartbeats as words and rhythms as sentences. ECG-Chat (Zhao et al., 2025) and GEM (Lan et al., 2025) explore instruction-tuning for ECG understanding. PULSE (Liu et al., 2026) provides large-scale instruction-following datasets. ELF (Han et al., 2026) demonstrates that encoder-free approaches with a single linear projection achieve competitive performance, raising questions about whether current benchmarks evaluate genuine signal understanding or primarily language priors.

PEFT for Medical AI. LoRA (Hu et al., 2021) enables efficient fine-tuning by learning low-rank updates. EnECG (Xu et al., 2026) freezes 99% of parameters while adapting to multiple ECG tasks. Our work extends PEFT to multimodal ECG-language modeling across multiple backbone families and size tiers.

3. Methods

3.1. Model Architecture

Our architecture follows the LLaVA (Liu et al., 2023) paradigm, connecting a pretrained ECG encoder to a language model through a projection layer.

ECG Encoder. We use wav2vec2 (Baevski et al., 2020), pretrained on ECG signals via self-supervised learning with Contrastive Multi-Segment Coding (CMSC). The encoder processes raw 12-lead ECG waveforms $X \in \mathbb{R}^{12\times2500}$ (10 seconds at 250 Hz) through convolutional downsampling and a Transformer, outputting contextualized embeddings $H \in \mathbb{R}^{N\times768}$, where N is the sequence length of the resulting temporal patches.

Language Models. We evaluate four backbones spanning two families and two size tiers:

- **Small tier:** Gemma 4-E4B-IT (Google, 2026) (4.5B effective, dense with Per-Layer Embeddings that offload parameters to flash storage, significantly reducing VRAM compared to a traditional 9B model, bf16; see Appendix A for PLE details) and Qwen3.5-9B (Qwen, 2026a) (9B, hybrid attention combining GatedDeltaNet with multi-head attention, bf16; see Appendix A for architecture details).
- **Large tier:** Qwen3.6-27B (Qwen, 2026b) (27B, hybrid attention, 4-bit quantized) and Gemma 4-31B-IT (Google, 2026) (31B, dense, 4-bit quantized).

Connector. A two-layer MLP with GELU activation pools and projects the ECG encoder outputs (768-dim) to the LLM embedding space. This process compresses the sequence into a single token (1 encoder token) to minimize LLM context window consumption.

3.2. PEFT Strategies

We compare three fine-tuning strategies of increasing complexity:

Projection-only. Only the connector MLP is updated; the LLM and ECG encoder remain frozen. Trainable parameters: $\sim$5–15M ($<$0.1% of total), varying with LLM embedding dimension. This is our best configuration for large quantized models.

Joint PEFT. The connector and LoRA (Hu et al., 2021) adapters on the LLM are co-trained. For Gemma 4 (4-bit), LoRA targets attention projections (r=16, α=32); for Qwen3.5-9B (bf16), we additionally target MLP layers. For Qwen3.6-27B (4-bit), we use r=8, α=16 to reduce gradient memory. Joint PEFT requires reducing `llm_input_len` from 2048 to 1536 for larger models to fit in 96GB VRAM.

Few-shot Connector PEFT. Same as projection-only training, but with one in-context QA demonstration per dataset type appended to the system prompt. This aligns the model's output format with the specific task. Because the LLM remains frozen, this strategy inherits projection-only stability while enabling in-context learning—particularly beneficial for 4-bit models where quantization information loss can be partially compensated by explicit formatting demonstrations.

Hyperparameters. We train on mixed datasets (PTB-XL + MIMIC-IV-ECG + ECG-Instruct-PULSE, $\sim$1.9M samples) for 1 epoch. Learning rate 2×10^{-5}, cosine restarts with 25

cycles, 2000 warmup steps, effective batch size 8. Training uses a single GPU with 96GB VRAM. Checkpoints contain only trainable parameters (160–280MB).

3.3. Datasets

PTB-XL (Wagner et al., 2022): 21,837 diagnostic-quality 12-lead ECGs with 71-class labels, converted to QA pairs via ECG-QA (Oh et al., 2023). **MIMIC-IV-ECG** (Gow et al., 2023): ∼800K diagnostic ECGs with clinical reports, converted to QA pairs via ECG-QA (Oh et al., 2023). **ECG-Instruct-PULSE** (Liu et al., 2026): 803K instruction-following samples from the PULSE training set. The combined training mixture contains ∼1.9M QA pairs. All ECGs are preprocessed to 12-lead, 2500 samples at 250 Hz.

4. Experiments

4.1. Experimental Setup

Evaluation Metrics. We report BLEU-4, ROUGE-L, METEOR, BERTScore-F1, token-level F1, and exact-match accuracy (ACC).

Evaluation Protocol. Full-scale evaluation uses 20,000 test examples (PTB-XL, MIMIC-IV) or 5,000 (ECG-Instruct-PULSE), mean ± std over 3 seeds, matching the ELF protocol (Han et al., 2026).

Baselines. We compare against ELF (Han et al., 2026) (Llama-3.2-1B-Instruct with multiple encoder configurations).

4.2. Main Results

Table 1 presents full-scale results for projection-only models. Both large models substantially outperform published baselines, but on different metrics.

Qwen3.5-9B wins on exact-match accuracy–+14.3% on PTB-XL, +7.5% on MIMIC-IV over ELF. **Gemma 4-31B wins on response richness**–ROUGE-L exceeds ELF by +32.3% on PTB-XL. Few-shot connector PEFT (†) yields a striking architecture-dependent effect: the 4-bit Qwen3.6-27B† improves over its projection-only baseline on *all* metrics (ACC +2.9, F1 +3.7, BLEU +4.2 on PTB-XL), while the bf16 Qwen3.5-9B† trades ACC for generative quality (ACC −3.3, but PULSE BLEU-4 +5.7). The 4-bit Gemma 4-31B† shows more modest gains (ACC +0.6 on PTB-XL, +1.8 on MIMIC-IV), suggesting the few-shot effect is modulated by both quantization and architecture.

4.3. Scaling Across Model Sizes

Table 2 compares projection-only and few-shot connector PEFT results, revealing that few-shot effectiveness is modulated by quantization.

Table 1. Full-scale evaluation (20K/5K, 3 seeds). †Few-shot connector PEFT. Best in **bold**.

Model	ACC	F1	BL	RG	MT	BS
PTB-XL						
ELF Best	32.5	–	16.7	41.9	27.3	92.5
Gemma 4-E4B	31.4	39.8	11.8	72.0	25.6	95.2
Gemma 4-E4B†	32.7	42.1	12.5	72.4	27.3	95.2
Gemma 4-31B	35.4	44.1	14.8	**74.2**	28.1	**95.6**
Gemma 4-31B†	36.0	43.9	13.3	74.7	27.6	95.7
Qwen3.6-27B	37.1	45.1	12.4	45.0	27.9	91.6
Qwen3.6-27B†	**40.0**	**48.8**	**16.6**	48.6	**31.3**	92.2
Qwen3.5-9B	46.8	53.4	10.6	53.2	31.6	92.4
Qwen3.5-9B†	43.5	48.0	3.5	48.2	25.9	91.7
MIMIC-IV						
ELF Best	26.6	–	21.9	41.7	30.7	91.5
Gemma 4-E4B	24.2	35.4	11.4	65.7	24.1	93.9
Gemma 4-E4B†	24.6	36.9	13.7	66.2	25.3	94.0
Gemma 4-31B	26.5	37.6	12.3	**67.3**	24.8	**94.2**
Gemma 4-31B†	28.4	40.5	16.3	69.2	27.0	94.5
Qwen3.6-27B	27.3	39.3	13.1	39.6	26.1	90.4
Qwen3.6-27B†	31.4	43.8	14.4	44.3	29.2	91.0
Qwen3.5-9B	**34.1**	**47.9**	18.8	48.1	**32.1**	91.3
Qwen3.5-9B†	31.7	41.8	6.7	42.4	24.5	90.5
PULSE (5K)						
ELF Best	14.6	–	25.7	66.8	63.6	95.6
Gemma 4-E4B	9.9	38.6	7.1	44.6	26.2	90.9
Gemma 4-E4B†	10.4	40.1	8.2	45.7	27.8	91.2
Gemma 4-31B	14.4	47.0	11.7	**50.9**	33.5	92.2
Gemma 4-31B†	16.3	47.8	12.3	52.1	34.3	92.4
Qwen3.6-27B	13.6	47.1	14.7	39.2	33.7	90.6
Qwen3.6-27B†	15.8	51.3	**16.3**	43.1	37.0	91.3
Qwen3.5-9B	**17.0**	**53.4**	16.5	45.1	**37.7**	91.5
Qwen3.5-9B†	16.3	52.0	16.3	43.3	36.4	91.2

Table 2. Projection-only vs. few-shot connector PEFT, 20K/5K, 3 seeds. Δ=Few-shot−Proj-only. Few-shot improves all metrics for 4-bit models but trades ACC on bf16.

Model	PTB ACC	Δ	MIM ACC	Δ	PUL F1	Δ
bf16 models						
Qwen3.5-9B	**46.8**		**34.1**		**53.4**	
Qwen3.5-9B†	43.5	−3.3	31.7	−2.4	52.0	−1.4
4-bit models						
Qwen3.6-27B	37.1		27.3		47.1	
Qwen3.6-27B†	40.0	+2.9	31.4	+4.1	51.3	+4.2
Gemma 4-31B	35.4		26.5		47.0	
Gemma 4-31B†	36.0	+0.6	28.4	+1.8	47.8	+0.8

For bf16 models, few-shot examples shift the model toward more verbose, format-aligned outputs at the cost of exact-match precision (9B†: ACC −3.3). For 4-bit models, the same examples compensate for quantization-induced information loss, improving both precision and fluency (27B†: ACC +2.9, PULSE F1 +4.2). This suggests that in-context demonstrations help quantized models recover representational capacity that is lost during 4-bit compression.

4.4. Projection-Only vs. Joint PEFT

We compare projection-only and joint PEFT for Gemma 4-31B. Joint PEFT peaks at a specific checkpoint then collapses dramatically (Table 3).

Joint PEFT peaks at step 155,960 (near a cosine-restart

Table 3. Gemma 4-31B: projection-only vs. joint PEFT across training steps (200-sample eval). Joint PEFT collapses after peak.

Strategy	Step	PTB ACC	PTB ROUGE	MIMIC ACC
Proj-only	97,475	36.5	74.5	30.3
Proj-only	136,465	36.0	73.8	27.5
Joint PEFT	38,990	17.5	47.6	10.2
Joint PEFT	155,960	**34.0**	**72.9**	**25.3**
Joint PEFT	194,950	16.5	44.3	11.8

trough, where LR is low) then collapses at step 194,950 (near an LR restart peak). This instability pattern suggests that co-training quantized LLM weights with aggressive LR restarts is fundamentally unstable. **For 4-bit quantized models, projection-only training is safer and performs better.**

4.5. Failure Mode Analysis

The metric divergence between Qwen (high ACC, low ROUGE-L) and Gemma 4 (low ACC, high ROUGE-L) reflects different failure modes: **Qwen when wrong** defaults to "none" (60% of errors), preserving exact-match on normals but missing abnormalities. **Gemma 4 when wrong** hallucinates related conditions (79% contain semantically related tokens), creating partial ROUGE-L overlap but never matching exactly.

4.6. Binary Abnormality Detection

Exact-match ACC obscures the clinically critical question: *can the model detect abnormalities?* A binary re-analysis (Table 4 in Appendix) reveals a clinically meaningful dichotomy: Gemma 4 is the better *screener* (66–75% sensitivity), while Qwen3.5 is the better *rule-out* tool (93–95% specificity). Together they suggest a dual-model clinical pipeline, though neither approaches the sensitivity required for autonomous diagnosis.

5. Discussion

Non-monotonic scaling and the few-shot correction. The 9B Qwen3.5 outperforms the 27B and 31B models on exact-match accuracy, confirming that hybrid attention architectures are more parameter-efficient for ECG QA than dense models. However, few-shot connector PEFT partially corrects this non-monotonicity: the 27B† narrows the ACC gap from 9.7 to 6.8 on PTB-XL while improving ROUGE-L by +3.6, and the 31B† improves MIMIC-IV ACC by +1.8 and PULSE ROUGE-L by +1.2. The effect is architecture-dependent: the 27B† (hybrid attention) shows strong across-the-board gains, while the 31B† (dense) shows more modest improvements—suggesting the scaling plateau of large quantized models is partly an artifact of suboptimal prompting, not an inherent capacity limit. This establishes few-shot connector PEFT as the baseline for the next phase: targeted LoRA on MLP layers, layer-wise rank allocation, and DoRA, which may further close the ACC gap with the 9B model.

Signal Understanding vs. Language Priors. ELF (Han et al., 2026) raised whether ELM benchmarks evaluate genuine signal understanding or language priors. Qwen3.5 achieves high ACC by defaulting to "none," exploiting the prevalence of normal ECGs. Gemma 4's higher sensitivity suggests it attends to signal features, but its false-alarm rate confirms that current benchmarks alone are insufficient for clinical validation.

Efficiency and Deployability. All models train on a single 96GB GPU with $\sim$5–15M trainable parameters ($<$0.1% of total). Checkpoints range from 160–280MB (connector-only), enabling rapid sharing across institutions without transferring base models—aligning with the workshop's focus on deployable, privacy-preserving health AI.

Limitations. (1) Binary analysis is preliminary (200 samples). (2) Neither model approaches clinical-grade sensitivity. (3) Current evaluation relies on text-generation metrics; expert clinician review is essential.

6. Conclusion

We present preliminary findings from an ongoing systematic comparison of multiple LLM backbones for ECG-language modeling across four models spanning two families and two size tiers. Three key findings emerge. First, projection-only fine-tuning consistently outperforms joint PEFT for large quantized models under current configurations, which destabilize under cosine LR restarts. Second, model choice creates a clinically meaningful precision-recall tradeoff: Qwen models excel at high-precision rule-out (93–95% specificity), while Gemma 4 models excel at abnormality screening (66–75% sensitivity). Third, few-shot connector PEFT improves 4-bit quantized models across *all* metrics (27B†: +2.9% ACC, +4.2 F1 on PULSE), while trading precision for fluency on bf16 models—suggesting in-context examples help quantized models compensate for information loss. The success of our few-shot formulation serves as the primary building block for the next phase of this ongoing study: targeted LoRA updates on MLP layers (currently skipped for large models), layer-wise rank allocation, DoRA, and exploration of smaller architectures (Qwen3.5-4B) to test whether hybrid attention preserves ECG QA capability at consumer-hardware scale.

References

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech

representations. *arXiv preprint arXiv:2006.11477*, 2020.

Google. Gemma 4. https://huggingface.co/google/gemma-4-e4b-it, 2026.

Gow, B. et al. Mimic-iv-ecg: Diagnostic electrocardiogram matched subset (version 1.0). *PhysioNet*, 2023. doi: 10.13026/4nqg-sb35.

Han, W. et al. Encoder-free ecg-language models. *arXiv preprint arXiv:2601.18798*, 2026.

Hu, E. J. et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jin, J. et al. Reading your heart: Learning ecg words and sentences. In *ICLR*, 2025.

Lan, X., Wu, F., He, K., Zhao, Q., Hong, S., and Feng, M. Gem: Empowering mllm for grounded ecg understanding with time series and images. *arXiv preprint arXiv:2503.06073*, 2025.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Liu, R., Bai, Y., Yue, X., et al. Teaching multimodal llms to comprehend 12-lead electrocardiographic images. *npj Digital Medicine*, 2026. doi: 10.1038/s41746-026-02551-3.

McKeen, K., Masood, S., Toma, A., Rubin, B., and Wang, B. Ecg-fm: An open electrocardiogram foundation model. *arXiv preprint arXiv:2408.05178*, 2025.

Nolin-Lapalme, A. et al. Foundation models for electrocardiogram interpretation. *European Heart Journal*, 2026.

Oh, J., Lee, G., Bae, S., Kwon, J.-m., and Choi, E. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. *arXiv preprint arXiv:2306.15681*, 2023.

Qwen. Qwen3.5. https://huggingface.co/Qwen/Qwen3.5-9B, 2026a.

Qwen. Qwen3.6. https://huggingface.co/Qwen/Qwen3.6-27B, 2026b.

Wagner, P., Strodthoff, N., Bousseljot, R., Samek, W., and Schaeffter, T. Ptb-xl, a large publicly available electrocardiography dataset (version 1.0.3). *PhysioNet*, 2022. doi: 10.13026/kfzx-aw45.

World Health Organization. Cardiovascular diseases (cvds) key facts. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), 2024.

Xu, Y. et al. Enecg: Efficient ensemble learning for electrocardiogram multi-task foundation model. *arXiv preprint arXiv:2511.22935*, 2026.

Zhao, Y., Kang, J., Zhang, T., Han, P., and Chen, T. Ecg-chat: A large ecg-language model for cardiac disease diagnosis. *arXiv preprint arXiv:2408.08849*, 2025.

A. Small Model Architecture Details

Qwen3.5-9B (Hybrid Attention). Interleaves Gated Delta Networks (linear attention) with standard multi-head attention in a 3:1 pattern across 32 layers. DeltaNet layers provide $O(1)$ recurrent updates for local processing; full-attention layers enable global context. Hidden dimension 4096, 16 heads, 4 KV heads (GQA). This hybrid design likely explains its strong exact-match performance despite fewer parameters than the 27B and 31B models.

Gemma 4-E4B (Per-Layer Embeddings). Dense 30-layer architecture (hidden 2304, 8 heads, 4 KV heads) with sliding-window/global attention in 5:1 ratio. Per-Layer Embeddings (PLE): a $[262{,}144 \times 30 \times 256]$ lookup table ($\sim$2B params) storing per-layer residual embeddings per token, loaded from flash at inference—only 4.5B compute params in VRAM. Each layer retrieves its PLE slice, gates, projects 256$\rightarrow$2304, and adds as residual, "reminding" layers of token identity. PLE introduces 44 additional trainable projection weights during fine-tuning.

B. Binary Abnormality Detection

We re-analyzed predictions with a binary lens (abnormal vs. normal, where normal = "none/no/normal/n/a").

Table 4. Binary abnormality detection (200-sample).

Metric	PTB-XL		MIMIC-IV	
	Qwen3.5	Gemma 4	Qwen3.5	Gemma 4
Sensitivity	25.9%	**66.1%**	56.4%	**75.0%**
Specificity	**93.2%**	63.6%	**95.0%**	65.0%
Precision	**82.9%**	69.8%	**96.3%**	83.3%
F1 (abnormal)	39.5%	**67.9%**	71.2%	**78.9%**