SCALING ECG-LANGUAGE MODELS TO 12 MILLION TRAINING EXAMPLES

Anonymous authors
Paper under double-blind review

ABSTRACT

ECG-language models (ELMs) show promise for automated ECG interpretation, but the effects of data scale and training strategies remain underexplored. We introduce ORAH, an ELM combining a SigLIP2-adapted ECG encoder (SigLEP) with a 4 billion parameter large language model (LLM). ORAH is trained on approximately 12 million examples across 12 datasets, the largest ELM training corpus to our knowledge. We develop a multi-stage training recipe spanning SigLEP pretraining, ORAH pretraining, and supervised fine-tuning (SFT). We investigate the contributions of data scale and ORAH pretraining across three benchmarks. ORAH exceeds the strongest reported baselines in all metrics on ECG-QA-CoT and the ECG-R1 Benchmark. On the ECG-Reasoning Benchmark, adding approximately 3,000 task-specific examples from the opposing ECG source improves Initial Diagnosis Accuracy (IDA) by 27.26 points on PTB-XL and 30.45 points on MIMIC-IV-ECG, highlighting the value of small amounts of targeted supervision. Furthermore, our ablations show that increasing training data and computation under our recipe consistently improves performance across all three benchmarks. These results establish ORAH as a strong ELM and provide empirical guidance for scaling data and training ELMs. Upon acceptance, we will release all publicly available datasets, model weights, and code for preprocessing, training, and evaluation used throughout this study.

1 INTRODUCTION

More than 300 million electrocardiograms (ECGs) are recorded worldwide each year (Zhu et al., 2020). This demand, together with the expertise required for reliable interpretation and a shortage of physicians (AAMC, 2024), motivates the development of automated ECG interpretation systems. Recent works have progressed beyond classification-based models (Rajpurkar et al., 2017) toward ECG-language models (ELMs) (Pham et al., 2025).

Like vision-language models (VLMs) (Liu et al., 2023), ELMs generate textual responses conditioned on ECG signals and natural-language queries (Wan et al., 2024b). However, existing ECG datasets typically only provide diagnostic statements and metadata rather than detailed signal descriptions. Prior work therefore uses large language models (LLMs) to expand these annotations into conversational and reasoning data (Liu et al., 2024), supporting open-ended ECG interpretation (Jin et al., 2026).

Recent work investigates dataset construction, ECG representations, model architectures, and learning objectives (Jin et al., 2025), but the contributions of training data scale and training strategies remain underexplored. We advance ELMs in this study by investigating the following three questions: (1) how performance changes when

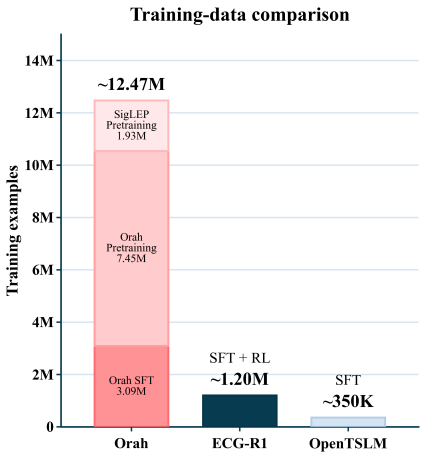


Figure 1: **Training Data Comparison.** We compare the total number of training examples used by ORAH with those used by ECG-R1 (Jin et al., 2026) and OpenTSLM (Langer et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

training data and computation increase, (2) which ELM capabilities benefit from pretraining, and (3) whether a small amount of benchmark-specific supervision transfers across ECG sources. To investigate these questions, we introduce ORAH, an ELM that connects a SigLIP2-adapted (Tschannen et al., 2025) ECG encoder (SigLEP) to a pretrained Qwen3.5-4B LLM through an MLP projector. We collect and preprocess approximately 12 million ECG-language examples from 12 datasets, forming, to our knowledge, the largest training corpus used by any ELM to date (Figure 1). Using this corpus, we develop a multi-stage training recipe comprising SigLEP pretraining, ORAH pretraining, and supervised fine-tuning (SFT). We summarize our main contributions below:

Summary of Contributions

- We collect and preprocess $\sim$12 million ECG-language examples across 12 datasets, the largest ELM training corpus to date.
- We introduce SigLEP, an ECG encoder, and ORAH, an ELM, together with a multi-stage training recipe comprising SigLEP pretraining, two ORAH pretraining stages, and two SFT stages. Using this recipe, ORAH outperforms all reported baselines on ECG-QA-CoT (§4.1) and the ECG-R1 Benchmark (§4.2).
- On the ECG-Reasoning Benchmark, adding augmented task-specific examples from the opposing ECG source, comprising only $\sim$0.15% of the second SFT stage's data, improves initial diagnosis accuracy (IDA) by 27.26 points on PTB-XL and 30.45 points on MIMIC-IV-ECG (§4.3).
- Comparing full-data training against 25% and 50% dataset subsets per stage shows that scaling data and compute under our recipe consistently improves performances across all three benchmarks (§4.4).
- ORAH pretraining substantially improves ECG Feature Grounding, Evidence-Based Reasoning, and Clinical Diagnostic Fidelity, while its effects on Diagnosis Accuracy and Depth are smaller and mixed (§4.5).

2 Related Work

ECG-Language Models Many ELMs follow the encoder-projector-LLM architecture widely used in VLMs. ECG-Chat (Zhao et al., 2024) adopts a LLaVA-style architecture (Liu et al., 2023), connecting a pretrained ECG encoder to an LLM for instruction tuning, while ELF (Han et al., 2026) introduces an encoder-free approach that maps ECG signals directly into the LLM embedding space through a single projection layer. Other work explores joint processing of ECG signals and rendered RGB images through ECG and vision encoders (Lan et al., 2025; Jin et al., 2026), as well as training on ECG alongside other time-series modalities (e.g., EEG and accelerometry) using multi-stage curriculum learning (Langer et al., 2026). More recently, ECG-R1 incorporated reinforcement learning (RL) for clinical reasoning (Jin et al., 2026). Building on these advances, we prioritize architectural simplicity by minimally adapting a SigLIP2 vision encoder (Tschannen et al., 2025) and connecting it to a strong pretrained LLM, Qwen3.5 (Team, 2026), through a standard MLP projector. This straightforward design allows us to focus on the effects of data scale and training strategies.

ECG-Language Datasets The public ECG-language datasets considered in this study are from five ECG databases: PTB-XL (Wagner et al., 2020), MIMIC-IV-ECG (Johnson et al., 2023), CODE-15% (Ribeiro et al., 2021), HEEDB (Koscova et al., 2026), and EchoNext (Elias & Finer, 2026). Most ECG-language datasets derive from MIMIC-IV-ECG and PTB-XL. ECG-QA (Oh et al., 2023) provides question-answer pairs derived from PTB-XL and MIMIC-IV-ECG, which Langer et al. (2026) augment with chain-of-thought (CoT) rationales (ECG-QA-CoT). Other studies use LLMs to construct a more verbose interpretation than diagnostic reports (Liu et al., 2024). However, ELMs can exploit benchmark-specific shortcuts rather than use the ECG input meaningfully (Han et al., 2026). The ECG-Reasoning Benchmark (Oh et al., 2026) examines this issue through chains of clinically grounded questions requiring information specific to the given ECG (e.g., QRS interval). In this study, we collect and preprocess the largest ELM training corpus to our knowledge. We use this large-scale data to investigate how scaling training data and computation affects performance and whether small amounts of benchmark-specific supervision transfer across ECG sources.

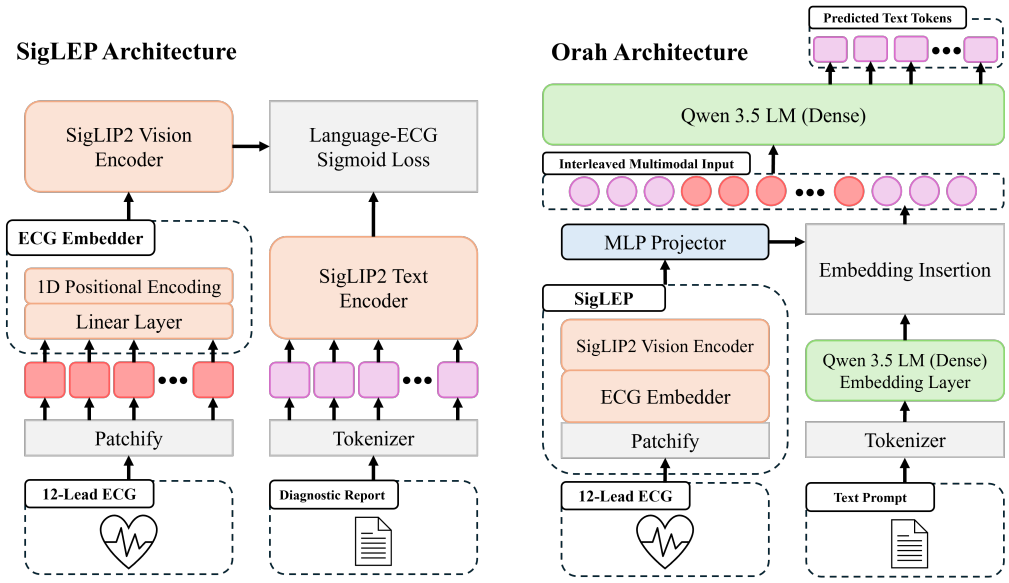


Figure 2: **Architecture of SigLEP and ORAH. Left:** SigLEP adapts a pretrained SigLIP2 (Tschannen et al., 2025) to 12-lead ECGs using a lightweight embedder that linearly projects ECG patches and adds 1D positional encodings. During pretraining, ECG and diagnostic-report representations produced by the vision and text encoders are aligned using a language-ECG sigmoid loss. **Right:** ORAH encodes the ECG with SigLEP, projects the resulting ECG features into the LLM embedding space using an MLP, and interleaves them with the embedded text prompt for autoregressive generation.

3 METHODS

3.1 DATASETS

We collect 12 ECG-language datasets (9 publicly available and 3 private), comprising approximately 12.47 million examples: 1.93 million for SigLEP pretraining, 7.45 million for ORAH pretraining, and 3.09 million for SFT. These counts refer to training examples rather than unique ECG recordings. We exclude every ECG recording in each benchmark's test split from every training source to prevent record-level overlap. Consequently, no benchmark test ECG appears in any SigLEP pretraining, Orah pretraining, or SFT training split. Further dataset details are in Appendix A.

Given an ECG signal $X \in \mathbb{R}^{C \times L}$, where C is the number of leads and L is the signal length, we set $C = 12$ and $L = 2500$, corresponding to a 10-second, 12-lead ECG resampled to 250 Hz. We reorder the leads of all ECGs to match the PTB-XL ordering (Wagner et al., 2020). Beyond resampling and normalization, we apply no additional preprocessing. We independently normalize each ECG to $[0, 1]$ using min-max (Uwaechia & Ramli, 2021):

$$\widetilde{X} = \text{clip}\left(\frac{X - \min(X)}{\max(X) - \min(X) + \epsilon_{\text{norm}}}, 0, 1\right),$$

where the min-max is computed across all leads and time points and $\epsilon_{\text{norm}} = 10^{-6}$ ensures numerical stability. Global scaling preserves relative peak-to-peak amplitudes across leads, which independent per-lead normalization would remove, but discards absolute voltage calibration. We leave a systematic comparison of normalization strategies, including per-lead min-max scaling and z-score normalization, to future work.

3.2 ORAH COMPONENTS

SigLEP ECG Encoder We adapt SigLIP2 (Tschannen et al., 2025) to process ECG signals, which we call Sigmoid Loss for Language-ECG Pretraining (SigLEP). We select SigLIP2 for its strong

visual representations and demonstrated utility as a vision encoder in VLMs such as Qwen3-VL (Bai et al., 2025). Our adaptation replaces its image patch embedding and positional embeddings with a lightweight 1D ECG embedder while retaining the pretrained transformer backbone. This design processes ECG signals as continuous patch representations without requiring image rendering or a discrete ECG vocabulary, providing a straightforward, simple way to transfer pretrained vision parameters to ECG-language modeling.

Given a normalized ECG signal $\widetilde{X} \in \mathbb{R}^{C\times L}$, we divide it into $N = L/P$ non-overlapping temporal patches of length $P = 25$. Let $\widetilde{X}^{(i)} \in \mathbb{R}^{C\times P}$ denote the i-th patch. We flatten each patch as $\widetilde{x}_i = \mathrm{vec}(\widetilde{X}^{(i)}) \in \mathbb{R}^{CP}, \quad i = 1, \ldots, N$. Each patch is linearly projected to the SigLIP2 hidden dimension and combined with a learned positional embedding: $h_i = W_{\text{patch}}\widetilde{x}_i + b_{\text{patch}} + p_i$. We replace the original image patch embedding layer of SigLIP2 with this ECG patch embedding layer denoted as ECG Embedder (Figure 2). The patch projection and positional embeddings are randomly initialized, while the remaining vision tower and text tower are initialized from the SigLIP2 checkpoint[1]. During SigLEP pretraining, we retain the pretrained SigLIP2 pooling heads. ECG patch representations are aggregated into a single vector using multi-head attention pooling, with a learned probe attending to the N patch tokens. The text representation is obtained from the text tower's final-layer hidden state at the last sequence position, followed by layer normalization and linear projection. Both vectors are ℓ_2-normalized, and their dot product is used in the original SigLIP loss (Zhai et al., 2023). Using this same loss, we update all SigLEP parameters during pretraining. We *do not* use the additional LocCa (Wan et al., 2024a), SILC (Naeem et al., 2023), or TIPS (Maninis et al., 2025) objectives from the SigLIP2 training recipe.

When training ORAH, we discard the text tower and ECG attention pooling head. We only keep the ECG Embedder and vision encoder to obtain patch-level representations. Let $H_{\text{enc}} = [h_1^{\text{enc}}, \ldots, h_N^{\text{enc}}] \in \mathbb{R}^{N\times d_{\text{enc}}}$ denote the vision encoder's outputted last hidden state patch representation, where d_{enc} is the encoder hidden dimension. We adaptively pool the sequence along the token dimension to obtain $\bar{H}_{\text{enc}} = [\bar{h}_1^{\text{enc}}, \ldots, \bar{h}_{N^*}^{\text{enc}}] \in \mathbb{R}^{N^*\times d_{\text{enc}}}$. These representations are then passed through the MLP projector. We use adaptive average pooling with $N^* = N = 100$, which leaves the encoder representations unchanged ($\bar{H}_{\text{enc}} = H_{\text{enc}}$). We retain this pooling operation to allow future work to vary the number of ECG tokens passed to the LLM without changing the encoder's patch size.

MLP Projector The MLP projector independently maps each ECG token $\bar{h}_i^{\text{enc}}$ into the LLM hidden space:

$$z_i = W_2\ \mathrm{GELU}\left(W_1\bar{h}_i^{\text{enc}} + b_1\right) + b_2, \qquad i = 1, \ldots, N^*,$$

where $W_1 \in \mathbb{R}^{d_{\text{mid}}\times d_{\text{enc}}}, W_2 \in \mathbb{R}^{d\times d_{\text{mid}}}$, and d is the LLM hidden dimension. Unless otherwise specified, we set $d_{\text{mid}} = d$. The projected sequence of ECG tokens is given as $Z = [z_1, \ldots, z_{N^*}] \in \mathbb{R}^{N^*\times d}$.

Large Language Model Backbone We use the base version of the pretrained, dense, 4B-parameter Qwen3.5[2] (Team, 2026) as the LLM backbone. Let $E = [e_1, \ldots, e_M] \in \mathbb{R}^{M\times d}$ denote the embeddings of the textual input sequence, which contains N^* consecutive signal placeholder positions within M total tokens. During training, we limit sequences to 1,024 tokens, truncating longer sequences and dynamically right-padding shorter sequences to the longest sequence in each batch. This limit exceeds the highest mean untruncated sequence length across pretraining and SFT training stages (359.62 tokens; Table 3). During evaluation, we do not truncate input sequences and allow up to 1,024 newly generated tokens. Further details are provided in Appendix B.4.

Let s_i denote the position of the i-th placeholder token, denoted as `<signal>`. We replace each placeholder embedding with the corresponding ECG token z_i:

$$\tilde{e}_j = \begin{cases} z_i, & j = s_i, \\ e_j, & \text{otherwise.} \end{cases}$$

The resulting sequence $\tilde{E} = [\tilde{e}_1, \ldots, \tilde{e}_M] \in \mathbb{R}^{M\times d}$ contains the ECG tokens at the designated signal positions and the original text-token embeddings at all remaining positions. Given $\tilde{E}$, the LLM is

[1] `google/siglip2-base-patch16-naflex`

[2] `Qwen/Qwen3.5-4B-Base`

trained autoregressively to generate the response tokens $y_{1:T}$:

$$\mathcal{L}_{\mathrm{AR}} = -\sum_{t=1}^{T} \log P_\theta \left(y_t \mid y_{<t}, \tilde{E} \right),$$

where θ denotes the trainable model parameters. During autoregressive training, positions corresponding to the context, including ECG placeholders, system and user prompts, are assigned a label of -100 and excluded from direct loss computation. Padding positions are masked in the same manner. The target tokens included in $\mathcal{L}_{\mathrm{AR}}$ are specified for each training stage below.

3.3 Training Stages

Each training stage runs as a separate job with stage-specific hyperparameters detailed in Table 4 of the appendix. Figure 8 and Table 1 in the appendix report the datasets and number of examples used at each training stage. We use the final-epoch checkpoint for SigLEP. We also use the final-epoch checkpoint from each stage to initialize the subsequent stage of ORAH training. All ORAH stages use bfloat16 mixed precision with float32 master parameters and optimizer states on six NVIDIA H100 NVL GPUs using FSDP2 (Zhao et al., 2023). SigLEP is pretrained with float32 parameters and optimizer states using distributed data parallelism (DDP). SigLEP pretraining uses AdamW, while all ORAH pretraining and SFT stages combine Muon (Jordan et al., 2024; Liu et al., 2025) and AdamW using disjoint parameter groups. Additional details about the optimizer are in Appendix B.2. All groups share the stage-specific learning rate and schedule in Table 4.

SigLEP Pretraining We train SigLEP for 30 epochs on a held-out subset of HEEDB comprising 1,927,353 paired 12-lead, 10-second ECGs and diagnostic reports (Koscova et al., 2026). This stage's primary goal is to adapt the pretrained SigLIP2 to ECG signals and diagnostic statements.

ORAH Pretraining 1 After pretraining SigLEP, we construct ORAH by connecting SigLEP to a pretrained Qwen3.5 LLM through an MLP projector. To align the SigLEP and Qwen3.5 representation spaces, we follow Qwen-VL's training strategy (Bai et al., 2025) and update only the MLP projector during this first stage of ORAH pretraining to align the modality gap between the LLM and ECG encoder. This stage uses 882,588 ECG–diagnostic report pairs sampled from four sources: (1) EchoNext (Elias & Finer, 2026), (2) HEEDB (Koscova et al., 2026), and (3–4) two privately curated datasets from Allegheny General Hospital, denoted Internal Datasets 1–2. We do not apply a system prompt, chat template, or introduce special reasoning tokens at this stage. Instead, ORAH is trained using standard autoregressive loss ($\mathcal{L}_{\mathrm{AR}}$) to generate the diagnostic report conditioned on the projected ECG tokens. More precisely, we compute $\mathcal{L}_{\mathrm{AR}}$ over the diagnostic report tokens and a trailing document-separator token (<|endoftext|>). ECG-placeholder (<signal>) and padding positions are excluded. This stage does not include any beginning- or end-of-sequence tokens in the input.

ORAH Pretraining 2 In the second ORAH pretraining stage, all aspects are the same as the first ORAH pretraining stage, except that we jointly update the MLP projector and Qwen3.5 LLM on 6,567,989 ECG-diagnostic report pairs drawn from five sources: (1) EchoNext (Elias & Finer, 2026), (2) HEEDB (Koscova et al., 2026), and (3-5) three privately curated datasets from Allegheny General Hospital, denoted Internal Datasets 1-3. All five datasets in this stage use separate samples from those used in all preceding training stages (including SigLEP pretraining). As in Pretraining 1, we compute $\mathcal{L}_{\mathrm{AR}}$ over the diagnostic report tokens and a trailing document-separator token. The motivation for this second pretraining stage is to adapt the LLM to ECG-conditioned report generation beyond the initial alignment achieved through projector-only training. Jointly updating the projector and LLM on a larger corpus allows both components to learn associations between ECG representations and diagnostic language, providing a foundation for the next SFT stage.

ORAH SFT 1 In the first SFT stage, we train ORAH on 914,847 question-answer (QA) and multi-turn conversation examples from four sources: (1) ECG-QA MIMIC-IV (Oh et al., 2023), (2) Pretrain MIMIC (Zhao et al., 2024), (3) ECG-Instruct 45K (Zhao et al., 2024), and (4) ECG-Grounding (Lan et al., 2025). We also introduce a custom system prompt and Qwen3.5's chat template (Appendix D). During this stage, we jointly update the MLP projector and LLM. We compute $\mathcal{L}_{\mathrm{AR}}$

over the assistant response, including the <|im_end|> token that terminates the response. The motivation for this first SFT stage is to extend ORAH beyond ECG-conditioned diagnostic statement generation to instruction following.

ORAH SFT 2 In the second SFT stage, we continue updating the MLP projector and LLM. We train on unseen examples from ECG-QA MIMIC-IV and ECG-Grounding while introducing three additional datasets for a total of 2,173,907 examples: (1) ECG-Instruct (Liu et al., 2024), (2) ECG Protocol-Guided Grounding CoT (Jin et al., 2026), and (3) ECG-QA-CoT (Langer et al., 2026). ECG Protocol-Guided Grounding CoT and ECG-QA-CoT contain explicit reasoning and answer tags (e.g., <think>, </think>, <answer>, </answer>). For the ECG-Reasoning Benchmark, we train ORAH under three separate settings: (1) without including any augmented ECG-Reasoning data, (2) with an augmented PTB-XL split, and (3) with an augmented MIMIC-IV-ECG split (Table 7). This experiment tests whether targeted, benchmark-specific supervision from one ECG source (e.g., PTB-XL) can improve performance on another (e.g., MIMIC-IV-ECG). In both augmented settings, we train on one source split and evaluate on the opposing source split, so no ECG from any test set is included during training, preventing data contamination. Further ECG-Reasoning augmented data preparation details are provided in Appendix B.1. We use the same loss masking as in SFT 1. Reasoning and answer tags are part of the assistant response and included in $\mathcal{L}_{\text{AR}}$. The second SFT stage's purpose is to broaden ORAH's instruction-following capabilities and introduce explicit reasoning.

3.4 EVALUATION

We evaluate ORAH on three benchmarks: (1) ECG-QA-CoT (Langer et al., 2026), (2) ECG-R1 Benchmark (Jin et al., 2026), and (3) ECG-Reasoning Benchmark (Oh et al., 2026). We follow the same evaluation splits and protocols of prior work and compare against baseline results reported in the corresponding studies. For ECG-QA-CoT and ECG-R1 Benchmark, we use the checkpoint obtained after SFT 2 with the augmented ECG-Reasoning PTB-XL split included during training. For ECG-Reasoning Benchmark, checkpoints trained with benchmark-specific examples from one source are evaluated on the other source. Benchmark and dataset details, including evaluation metrics and ECG sources, are provided in Appendix A.

4 RESULTS

4.1 ECG-QA-COT BASELINES

We evaluate ORAH on ECG-QA-CoT and compare its performance with ELF (Han et al., 2026), OpenTSLM (Langer et al., 2026), and SLIP (Chen et al., 2026) in Figure 3. ELF does not use clinical context or CoT rationales during training and generates only the answer, whereas all other baselines and ORAH generate both the CoT and answer. Despite these differences, ORAH outperforms all baselines, exceeding the strongest baseline, ELF with Gemma-2-2B-it, by 1.56 and 1.95 points in accuracy and F1, respectively. Among models that generate CoT rationales, ORAH exceeds the strongest baseline, OpenTSLM Flamingo with Llama3.2-3B, by 13.86 and 23.53 points in accuracy and F1, respectively (Table 5). These results demonstrate strong ECG question-answering performance while retaining the ability to generate CoT. We include a qualitative example in Figure 13.

4.2 ECG-R1 BENCHMARK BASELINES

In Figure 4, we compare ORAH with the strongest baselines on the ECG-R1 Benchmark: PULSE (Liu et al., 2024), GEM (Lan et al., 2025), and ECG-R1 (Jin et al., 2026). Results for additional models are provided in Table 6 in the appendix. As observed, relative to ECG-R1 (RL), ORAH achieves deltas of $+0.02$, $+0.21$, $+0.99$, $+1.53$, $+4.50$, $+0.90$, and $+1.77$ on Diagnosis Accuracy, Analysis Completeness, Analysis Relevance, Lead Evidence Validity, ECG Feature Grounding, Evidence-Based Reasoning, and Clinical Diagnostic Fidelity, respectively. Although ORAH has half as many parameters as ECG-R1 (4B vs. 8B), it outperforms the strongest baseline on all metrics. We include qualitative examples in Figures 17 through 19.

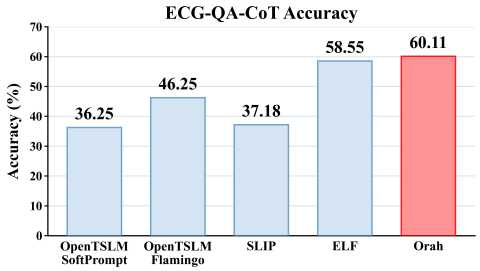


Figure 3: **ECG-QA-CoT Baselines.** We report the accuracy after evaluating ORAH on the ECG-QA-CoT dataset. Table 5 in the appendix provides a tabular version of the results, including additional OpenTSLM variants and F1 scores where available.

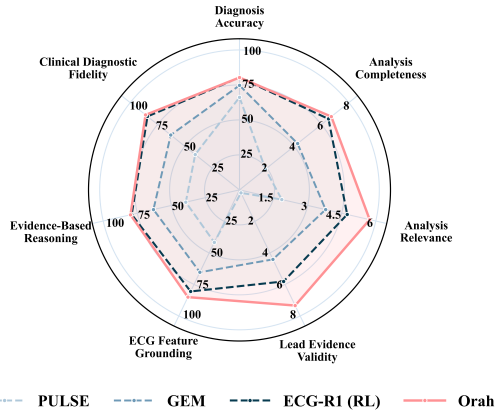


Figure 4: **ECG-R1 Benchmark Baselines.** We report ORAH's performance on the 7 metrics used by the ECG-R1 Benchmark (Jin et al., 2026) as a radar plot. Each spoke uses an independent scale. Table 6 in the appendix reports additional baselines, including Gemini-3-Pro and GPT-5.1-Instant.

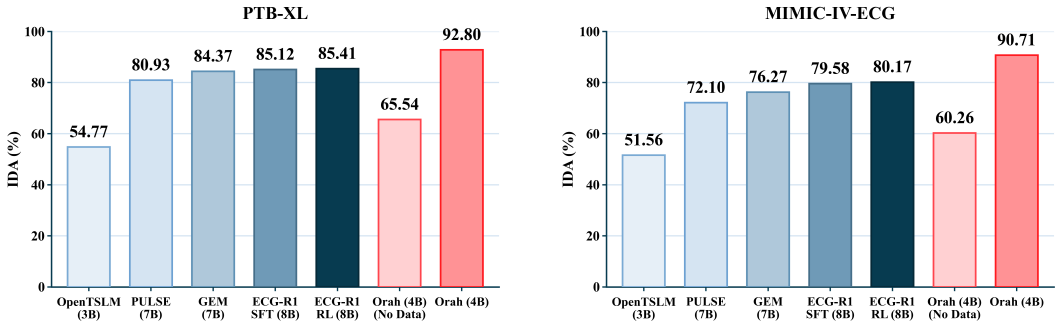


Figure 5: **Cross-Dataset Transfer.** We compare ORAH with and without benchmark-specific supervision on the PTB-XL (**Left**) and MIMIC-IV-ECG (**Right**) splits of the ECG-Reasoning Benchmark (Oh et al., 2026). Table 7 in the appendix reports all four metrics (IDA, Completion, Depth, and GT-RDA) for a larger set of models.

4.3 CROSS-DATASET TRANSFER ON THE ECG-REASONING BENCHMARK

Despite achieving strong results on ECG-QA-CoT and the ECG-R1 Benchmark, ORAH shows limited performance on the ECG-Reasoning Benchmark without benchmark-specific supervision. We therefore investigate whether a small amount of targeted supervision transfers across ECG sources. We separately augment SFT 2 with 3,359 MIMIC-IV-ECG examples for PTB-XL evaluation or 3,084 PTB-XL examples for MIMIC-IV-ECG evaluation, comprising approximately 0.14-0.15% of the resulting SFT 2 training examples (Table 2). Augmentation increases PTB-XL IDA from 65.54% to 92.80% and MIMIC-IV-ECG IDA from 60.26% to 90.71%, gains of 27.26 and 30.45 points, respectively. Completion also increases by 8.50 and 7.56 points, respectively (Figure 5 and Table 7). ORAH with the augmented data during SFT 2 surpasses the strongest reported baselines in IDA. These results are achieved with a 4B LLM backbone, smaller than every baseline with a reported parameter count except OpenTSLM (3B). All other baselines range from 7B to 90B.

However, even the gains from augmentation do not extend uniformly across reasoning metrics. ORAH's Depth of 1.55 and 1.49 and GT-RDAs of 40.43% and 34.62% on PTB-XL and MIMIC-IV-ECG, respectively, remain below those of Hulu-Med (32B) and Gemini-3-Flash. Its GT-RDA is also substantially lower than its IDA, whereas evaluated models without ECG specialization show

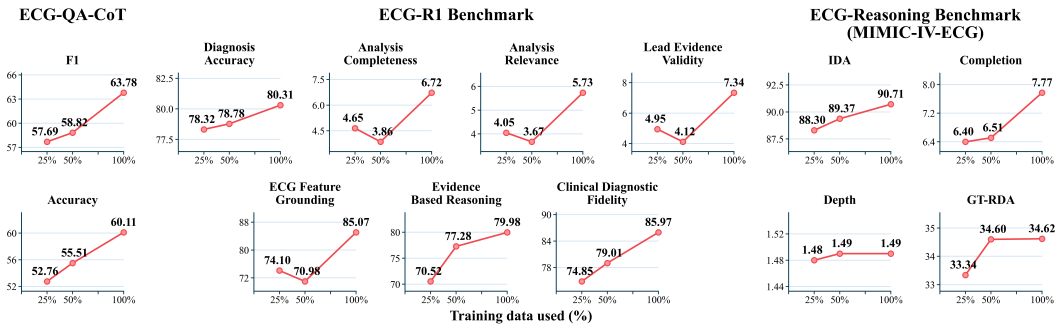


Figure 6: **Effects of Data Scaling.** We evaluate ORAH across three benchmarks after training on randomly sampled 25% and 50% subsets of each dataset at every training stage. Tabular results are provided in Tables 5, 6, and 7.

the opposite pattern. This is consistent with Oh et al. (2026)'s observation that ECG-specific models struggle to use ground-truth reasoning trajectories despite strong IDA. Although ECGs from both PTB-XL and MIMIC-IV-ECG are present in other SFT datasets, adding only ∼3,000 benchmark-specific examples from the opposing source substantially improves performance. These findings suggest that training on a large-scale dataset alone does not adequately prepare ORAH for this benchmark, motivating future work on curating small amounts of task-specific supervision for specialized ECG interpretation tasks. We include qualitative examples in Figures 14 through 16.

4.4 Effects of Scaling Data

To examine data scaling under our training recipe, we train two ORAH variants using randomly sampled 25% and 50% subsets of each dataset at every training stage, preserving the relative proportions of the source datasets. Figure 6 reports performance across these data scales. Increasing training data from 25% to 50% improves all metrics on ECG-QA-CoT and the MIMIC-IV-ECG split of the ECG-Reasoning Benchmark, but yields mixed results on the ECG-R1 Benchmark, where four metrics decline. Further increasing training data to 100% improves every evaluated metric except Depth, which remains unchanged. Relative to the 50% variant, ECG-QA-CoT F1 and accuracy increase by 4.96 and 4.60 points, respectively. On the ECG-R1 Benchmark, ECG Feature Grounding, Evidence-Based Reasoning, and Clinical Diagnostic Fidelity increase by 14.09, 2.70, and 6.96 points, respectively. On the ECG-Reasoning Benchmark, IDA, Completion, and GT-RDA increase by 1.34, 1.26, and 0.02 points, respectively. Overall, full-data training improves nearly every evaluated metric over the 25% and 50% variants, although gains are not consistently monotonic, motivating further investigation into how data scaling affects different ELM capabilities.

4.5 Which Capabilities Benefit from ORAH Pretraining?

The two ORAH pretraining stages use approximately 413 million and 1.07 billion tokens across all epochs, respectively, together accounting for 31.8% of the total training token budget (Figure 8 and Table 3 in the appendix). This large investment motivates examining which downstream capabilities benefit from pretraining. We compare the full training recipe with a variant that omits both ORAH pretraining stages and proceeds directly to SFT 1 and 2. Figure 7 reports the performance differences. On the ECG-QA-CoT benchmark, pretraining improves F1 and accuracy by 3.14 and 3.72 points, respectively. On the ECG-R1 Benchmark, pretraining improves ECG Feature Grounding, Evidence-Based Reasoning, and Clinical Diagnostic Fidelity by 17.98, 15.15, and 20.35 points, respectively, while Diagnosis Accuracy decreases by 1.03 points. On the MIMIC-IV-ECG split of the ECG-Reasoning Benchmark, pretraining improves GT-RDA by 9.33 points. IDA and Completion also improve, while Depth decreases by 0.12. These results highlight substantial benefits from pretraining for QA, ECG Feature Grounding, Evidence-Based Reasoning, and Clinical Diagnostic Fidelity, with mixed effects on Diagnosis Accuracy and Depth.

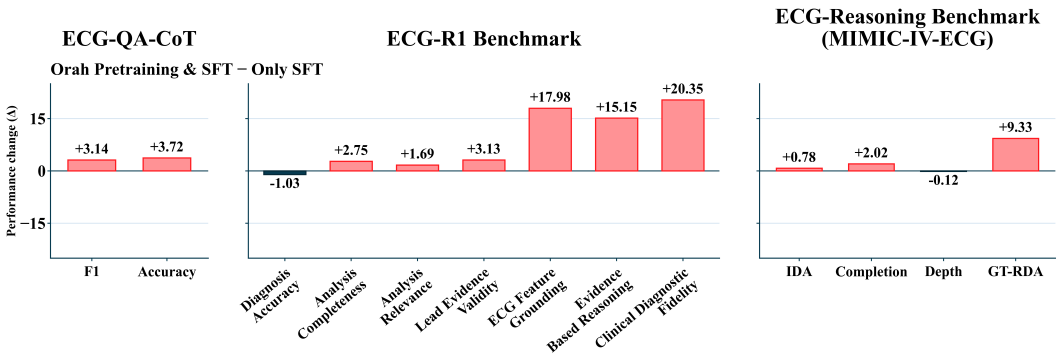


Figure 7: **Effects of ORAH Pretraining.** Performance differences across three benchmarks when training with both pretraining and SFT stages minus training with only the two SFT stages. Tabular results are provided in Tables 5, 6, and 7.

5 DISCUSSION AND CONCLUSION

In this study, we introduce ORAH, an ELM combining SigLEP with a pretrained 4B LLM, trained on approximately 12 million ECG-language examples across 12 datasets. After this large-scale training, we evaluate ORAH on three benchmarks and study the effects of data scale, ORAH pretraining, and targeted supervision. Our experimental results provide four main takeaways:

Summary of Main Takeaways

- Under our training recipe, ORAH outperforms all reported baselines on ECG-QA-CoT and the ECG-R1 Benchmark, including models with larger LLM backbones.
- Scaling data and compute from the 25% and 50% subsets to the full dataset improves nearly every metric across the three benchmarks.
- ORAH pretraining, which accounts for 31.8% of the training token budget, especially improves performances in QA, ECG Feature Grounding, Evidence-Based Reasoning, and Clinical Diagnostic Fidelity, with mixed effects on Diagnosis Accuracy and Depth.
- Large-scale ECG-language training does not eliminate the need for targeted supervision on the ECG-Reasoning Benchmark.

To this end, we express several limitations of this study and corresponding directions for future work.

5.1 LIMITATIONS

First, our training data is skewed towards healthy ECGs (Figure 10), and multiple SFT datasets derive from the same underlying ECG sources. Collecting data from a broader distribution of diseases and ECG dataset sources is an important direction for future work.

Second, the ECG-R1 and ECG-Reasoning Benchmarks rely on LLM-based evaluation, which may introduce evaluator bias and scoring variability. Clinician-reviewed, ECG-grounded evaluation is needed to distinguish clinically meaningful errors from differences in terminology and reporting preferences.

Lastly, our scaling comparison increases both data and computation, while the pretraining ablation removes both ORAH pretraining stages and their associated training exposure. These experiments therefore do not isolate data quantity, individual stages, or stage ordering. We also use a single 4B LLM backbone, leaving the effects of LLM parameter scaling unexplored. Future work should examine these factors under matched compute budgets and investigate how larger LLMs and training budgets affect ECG-language capabilities.

AI USE STATEMENT

In this work, we use generative AI tools for code generation, paper-writing, and proofreading. Specifically, generative AI was used to produce initial versions of the code for all visualizations and the first implementation of SigLEP. We also used generative AI to correct grammatical errors and improve the clarity of the manuscript. We did not use generative AI to develop the research ideas, conduct the literature search, or formulate the hypotheses. We also did not use generative AI to create new data for this study. Any AI-generated data included in our experiments were produced previously as part of their respective studies. Theoretical modeling, mathematical claims, and proofs are not part of this work. Every line of code and all manuscript text were reviewed by at least one author. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

ECGs are derived from human participants and therefore warrant consideration of privacy, data governance, and potential clinical misuse. Most ECGs used in this study are obtained from publicly available datasets and are used in accordance with their respective access requirements and data-use agreements. Our private internal datasets consist exclusively of de-identified ECGs that were collected retrospectively. Finally, ORAH is developed solely for research purposes and is not intended for clinical use, diagnosis, treatment, or deployment in real-world healthcare settings.

REPRODUCIBILITY STATEMENT

We believe reproducibility is important and therefore plan to release our codebase, preprocessing pipelines for all public datasets, and model weights upon acceptance. We also provide extensive implementation and experimental details throughout the paper. We describe datasets and preprocessing pipelines in Tables 1, 2, Section 3.1, and Appendix A. We report training details and hyperparameters in Tables 4, Section 3.3, and Appendix B.

ACKNOWLEDGMENTS

We will add acknowledgments upon acceptance.

REFERENCES

AAMC. The complexities of physician supply and demand: Projections from 2021 to 2036, 03 2024. URL https://www.aamc.org/media/75236/download.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Yuliang Chen, Arvind Pillai, Yu Yvonne Wu, Tess Z. Griffin, Lisa Marsch, Michael V. Heinz, Nicholas C. Jacobson, and Andrew Campbell. Learning transferable sensor models via language-informed pretraining, 2026. URL https://arxiv.org/abs/2603.11950.

Pierre Elias and Joshua Finer. EchoNext: A Dataset for Detecting Echocardiogram-Confirmed Structural Heart Disease from ECGs. *PhysioNet*, April 2026. doi: 10.13026/7cfw-d091. URL https://doi.org/10.13026/7cfw-d091. Version 1.1.1.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization, 2025. URL https://arxiv.org/abs/2511.20347.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj,

Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

William Han, Tony Chen, Chaojing Duan, Xiaoyu Song, Yihang Yao, Yuzhe Yang, Michael A. Rosenberg, Emerson Liu, and Ding Zhao. Elf: A family of encoder-free ecg-language models, 2026. URL https://arxiv.org/abs/2601.18798.

Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. Reading your heart: Learning ecg words and sentences via pre-training ecg language model, 2025. URL https://arxiv.org/abs/2502.10707.

Jiarui Jin, Haoyu Wang, Xingliang Wu, Xiaocheng Fang, Xiang Lan, Zihan Wang, Deyun Zhang, Bo Liu, Yingying Zhang, Xian Wu, Hongyan Li, and Shenda Hong. Ecg-r1: Protocol-guided and modality-agnostic mllm for reliable ecg interpretation, 2026. URL https://arxiv.org/abs/2602.04279.

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 A. Celi, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10, 01 2023. doi: 10.1038/s41597-022-01899-x.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Zuzana Koscova, Qiao Li, Chad Robichaux, Valdery Moura Junior, Manohar Ghanta, Aditya Gupta, Jonathan Rosand, Aaron D. Aguirre, Erik Reinertsen, Steven Song, Shenda Hong,

David E. Albert, Joel Xue, Aarya Parekh, Reza Sameni, Matthew A. Reyna, M. Brandon Westover, and Gari D. Clifford. The harvard-emory ECG database. *Scientific Data*, 13(1):516, February 2026. doi: 10.1038/s41597-026-06861-9. URL https://doi.org/10.1038/s41597-026-06861-9.

Xiang Lan, Feng Wu, Kai He, Qinghao Zhao, Shenda Hong, and Mengling Feng. Gem: Empowering mllm for grounded ecg understanding with time series and images, 2025. URL https://arxiv.org/abs/2503.06073.

Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A. Xu, Winnie Chow, Martin Maritsch, Robert Jakob, Ning Wang, Juncheng Liu, Aradhana Verma, Brian Han, Daniel Seung Kim, Henry Chubb, Scott Ceresnak, Aydin Zahedivash, Alexander Tarlochan Singh Sandhu, Fatima Rodriguez, Daniel McDuff, Elgar Fleisch, Oliver Aalami, Filipe Barata, and Paul Schmiedmayer. Opentslm: Time-series language models for reasoning over multivariate medical text- and time-series data, 2026. URL https://arxiv.org/abs/2510.02410.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for llm training, 2025. URL https://arxiv.org/abs/2502.16982.

Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. Teach multimodal llms to comprehend electrocardiographic images, 2024. URL https://arxiv.org/abs/2410.19008.

Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, Dan Gnanapragasam, Mojtaba Seyedhosseini, Howard Zhou, and Andre Araujo. Tips: Text-image pretraining with spatial awareness, 2025. URL https://arxiv.org/abs/2410.16512.

Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation, 2023. URL https://arxiv.org/abs/2310.13355.

Jungwoo Oh, Gyubok Lee, Seongsu Bae, Joon myoung Kwon, and Edward Choi. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram, 2023. URL https://arxiv.org/abs/2306.15681.

Jungwoo Oh, Hyunseung Chung, Junhee Lee, Min-Gyu Kim, Hangyul Yoon, Ki Seong Lee, Youngchae Lee, Muhan Yeo, and Edward Choi. Ecg-reasoning-benchmark: A benchmark for evaluating clinical reasoning capabilities in ecg interpretation, 2026. URL https://arxiv.org/abs/2603.14326.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan

Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Hung Manh Pham, Jialu Tang, Aaqib Saeed, and Dong Ma. Q-heart: Ecg question answering via knowledge-informed multimodal llms, 2025. URL https://arxiv.org/abs/2505.06296.

Pranav Rajpurkar, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y. Ng. Cardiologist-level arrhythmia detection with convolutional neural networks, 2017. URL https://arxiv.org/abs/1707.01836.

Antônio H. Ribeiro, Gabriela M.M. Paixao, Emilly M. Lima, Manoel Horta Ribeiro, Marcelo M. Pinto Filho, Paulo R. Gomes, Derick M. Oliveira, Wagner Meira Jr, Thömas B Schon, and Antonio Luiz P. Ribeiro. Code-15%: a large scale annotated dataset of 12-lead ecgs, 06 2021. URL https://zenodo.org/records/4916206.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language

encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Anthony Ngozichukwuka Uwaechia and Dzati Athiar Ramli. A comprehensive survey on ecg signals as new biometric modality for human authentication: Recent advances and future challenges. *IEEE Access*, 9:97760–97802, 2021. doi: 10.1109/ACCESS.2021.3095248.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, May 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0495-6. URL https://www.nature.com/articles/s41597-020-0495-6. Number: 1 Publisher: Nature Publishing Group.

Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim Alabdulmohsin, Xiao Wang, André Susano Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. Locca: Visual pretraining with location-aware captioners, 2024a. URL https://arxiv.org/abs/2403.19596.

Zhongwei Wan, Che Liu, Xin Wang, Chaofan Tao, Hui Shen, Zhenwu Peng, Jie Fu, Rossella Arcucci, Huaxiu Yao, and Mi Zhang. Meit: Multi-modal electrocardiogram instruction tuning on large language models for report generation, 2024b. URL https://arxiv.org/abs/2403.04945.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. Pytorch fsdp: Experiences on scaling fully sharded data parallel, 2023. URL https://arxiv.org/abs/2304.11277.

Yubao Zhao, Tian Zhang, Xu Wang, Puyu Han, Tong Chen, Linlin Huang, Youzhu Jin, and Jiaju Kang. Ecg-chat: A large ecg-language model for cardiac disease diagnosis, 2024. URL https://arxiv.org/abs/2408.08849.

Hongling Zhu, Cheng Cheng, Hang Yin, Xingyi Li, Ping Zuo, Jia Ding, Fan Lin, Jingyi Wang, Beitong Zhou, Yonge Li, Shouxing Hu, Yulong Xiong, Binran Wang, Guohua Wan, Xiaoyun Yang, and Ye Yuan. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. *The Lancet Digital Health*, 2:e348–e357, 07 2020. doi: 10.1016/s2589-7500(20)30107-2.

Table 1: Tabular form of training stages, trainable modules, datasets, and sample counts in Figure 8. When a dataset appears in multiple stages, each occurrence represents a distinct, non-overlapping subset.

Stage	Trainable Modules	Epochs	Dataset	Samples
SigLEP Pretraining	Encoder	30	HEEDB (Koscova et al., 2026)	1,927,353
			Total	**1,927,353**
ORAH Pretraining 1	Connector	3	EchoNext (Elias & Finer, 2026)	24,763
			HEEDB (Koscova et al., 2026)	642,451
			Internal Dataset 1	149,929
			Internal Dataset 2	65,445
			Total	**882,588**
ORAH Pretraining 2	Connector, LLM	1	EchoNext (Elias & Finer, 2026)	57,780
			HEEDB (Koscova et al., 2026)	5,996,208
			Internal Dataset 1	349,835
			Internal Dataset 2	152,706
			Internal Dataset 3	11,460
			Total	**6,567,989**
ORAH SFT 1	Connector, LLM	3	ECG-QA MIMIC-IV (Oh et al., 2023)	352,382
			Pretrain MIMIC (Zhao et al., 2024)	502,687
			ECG-Instruct 45K (Zhao et al., 2024)	44,778
			ECG-Grounding (Lan et al., 2025)	15,000
			Total	**914,847**
ORAH SFT 2	Connector, LLM	3	ECG-QA MIMIC-IV (Oh et al., 2023)	822,226
			ECG-Grounding (Lan et al., 2025)	15,000
			ECG-Instruct (Liu et al., 2024)	1,147,368
			ECG Protocol-Guided Grounding CoT (Jin et al., 2026)	30,000
			ECG-QA-CoT (Langer et al., 2026)	159,313
			Total	**2,173,907**
ORAH RL	Connector, LLM	2	RL ECG-R1 (Jin et al., 2026)	3,948
			Total	**3,948**
			Grand Total	**12,470,632**

A Dataset Details

In this section, we describe each dataset used in this study.

Table 2: ECG-Reasoning Benchmark training splits used during ORAH SFT 2.

Split	Dataset	Samples
Both	ECG-QA MIMIC-IV (Oh et al., 2023)	822,226
	ECG-Grounding (Lan et al., 2025)	15,000
	ECG-Instruct (Liu et al., 2024)	1,147,368
	ECG Protocol-Guided Grounding CoT (Jin et al., 2026)	30,000
	ECG-QA-CoT (Langer et al., 2026)	159,313
PTB-XL	ECG-Reasoning (Oh et al., 2026)	3,084
MIMIC-IV-ECG	ECG-Reasoning (Oh et al., 2026)	3,359
PTB-XL	**Total**	**2,176,991**
MIMIC-IV-ECG	**Total**	**2,177,266**

A.1 Publicly Available

HEEDB (Koscova et al., 2026) The Harvard-Emory ECG Database (HEEDB) (Koscova et al., 2026) is a large-scale collection of standard 12-lead, 10-second ECG recordings acquired at Massachusetts General Brigham and Emory University Hospital. The recordings span several decades and were sampled at either 250 or 500 Hz. We resample all ECGs to 250 Hz. Version 3 contains 10,608,417 ECGs from 1,818,247 patients at Massachusetts General Hospital and 998,844 ECGs from 349,548 patients at Emory University Hospital. After filtering for valid data entries, we arrived at 8,566,012 ECGs. The database includes diagnostic reports, which we use for pretraining SigLEP and the 2 ORAH pretraining stages.

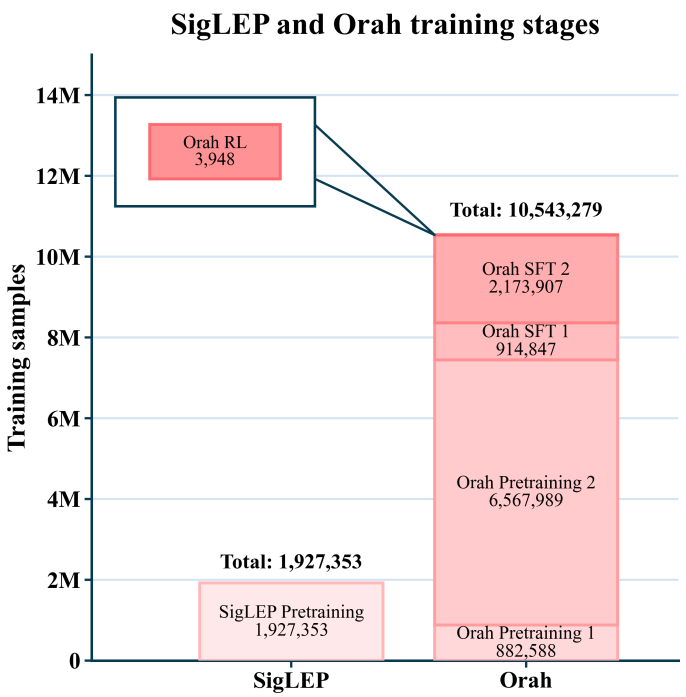


Figure 8: **Training Stage Breakdown.** We report the number of ECG-language examples used during each stage of training for SigLEP and ORAH. The ORAH RL stage is magnified.

EchoNext (Elias & Finer, 2026) EchoNext (Elias & Finer, 2026) is a de-identified collection of 100,000 standard 12-lead, 10-second ECG recordings acquired from adult patients at Columbia University Irving Medical Center. Each ECG was sampled at 250 Hz and paired with demographic and ECG-specific metadata. The metadata contain binary labels for left ventricular systolic dysfunction, left ventricular hypertrophy, moderate or greater valvular disease, right ventricular systolic dysfunction, moderate or large pericardial effusion, pulmonary hypertension, and elevated tricuspid regurgitation velocity. We convert each positive binary label into its corresponding diagnostic statement. ECGs without any positive labels are assigned the report "No significant structural heart disease." We use the ECGs and constructed reports during the 2 ORAH pretraining stages.

ECG-QA MIMIC-IV (Oh et al., 2023) ECG-QA MIMIC-IV comprises question-answer pairs associated with ECGs from the MIMIC-IV-ECG dataset (Johnson et al., 2023). Domain experts created the question-answer templates, and GPT-4o (OpenAI, 2024) generated paraphrases of the questions and answers using a clinically validated corpus of question templates.

Pretrain MIMIC (Zhao et al., 2024) Pretrain MIMIC (Zhao et al., 2024) is derived from the MIMIC-IV-ECG dataset (Johnson et al., 2023). For each ECG, its diagnostic report serves as the answer to a user instruction generated by GPT-4 (OpenAI et al., 2024), forming an additional question-answer dataset.

ECG-Instruct 45K ECG-Instruct 45K (Zhao et al., 2024) is an instruction-tuning dataset of physician-patient conversations generated using GPT-4o (OpenAI, 2024). The dialogues address a variety of topics, including ECG interpretation, heart rate, waveform morphology, cardiac rhythm, and electrical axis. The dataset is constructed from ECG recordings in MIMIC-IV-ECG (Johnson et al., 2023).

ECG-Grounding Lan et al. (2025) ECG-Grounding Lan et al. (2025) consists of 30,000 instruction-response pairs curated from MIMIC-IV-ECG Johnson et al. (2023). They also extract ECG features and use these features to build cardiology-specific prompts that guide GPT-4o to generate fine-grained interpretations. The resulting dataset is intended to produce more feature-

Table 3: Total token budget across all epochs at each stage of ORAH training, with average token lengths.

Training Stage	Epochs	Total Token Budget	Average Token Length
Pretraining 1	3	412,845,867	155.92
Pretraining 2	1	1,073,096,339	163.38
SFT 1	3	827,039,985	301.34
SFT 2	3	2,345,341,305	359.62
RL 1	2	8,580,674	1,086.71
Total	–	**4,666,904,170**	–

grounded and clinically plausible responses. Following ECG-R1 Jin et al. (2026), we exclusively use the MIMIC-IV-ECG test split from ECG-Grounding for evaluation, comprising 2,381 samples.

ECG-QA-CoT ECG-QA-CoT (Langer et al., 2026) augments ECG-QA (Oh et al., 2023) with Chain-of-Thought (CoT) rationales (Wei et al., 2023) generated by GPT-4o (OpenAI, 2024). Each rationale is generated from a plotted ECG image together with its associated clinical context and question-answer pair. The clinical context comprises non-diagnostic metadata from PTB-XL (Wagner et al., 2020), including patient demographics, basic recording details, and signal-quality annotations. This context is also provided during the training and evaluation of OpenTSLM (Langer et al., 2026). When training ORAH, we include both the clinical context and CoT rationales. We use the same ECG-QA-CoT training and test splits adopted by OpenTSLM (Langer et al., 2026) and SLIP (Chen et al., 2026).

ECG-Instruct (Liu et al., 2024) ECG-Instruct consists of 1.15 million instances of ECG-Language pairs derived from four sources: 1) PTB-XL Wagner et al. (2020), 2) MIMIC-IV-ECG Johnson et al. (2023), 3) CODE-15% Ribeiro et al. (2021), and 4) ECG-QA Oh et al. (2023). They first used PTB-XL and MIMIC-IV-ECG as seed data to synthesize visually grounded graduate-level ECG tasks and corresponding responses with Llama-3-70B-Instruct Liu et al. (2024); Grattafiori et al. (2024). They then used an independent Llama-based model, although its architecture was not specified, to evaluate the synthesized text for quality control. This dataset is used for models like GEM (Lan et al., 2025), PULSE (Liu et al., 2024), and ECG-R1 (Jin et al., 2026).

ECG Protocol-Guided Grounding CoT (Jin et al., 2026) ECG Protocol-Guided Grounding CoT Jin et al. (2026) consists of 30,000 instruction-response pairs curated from MIMIC-IV-ECG Johnson et al. (2023) using a protocol-guided generation pipeline. They first extract ECG features with a deterministic, non-trainable extractor, and use them to construct protocol-aware prompts based on a standardized ECG interpretation workflow from an ECG textbook. Given this prompt, DeepSeek-V3.1-Terminus generates responses containing a `<think>` block, a brief summary, and an `<answer>` block with the final diagnosis.

RL ECG-R1 (Jin et al., 2026) The RL ECG-R1 dataset is derived from a curated subset of the ECG Protocol-Guided Grounding CoT dataset (Jin et al., 2026). Specifically, the 500 most frequent unique diagnostic reports are identified, and up to 10 samples are uniformly selected without replacement for each report. ECG-R1 exclusively uses this dataset for RL training.

ECG-R1 Benchmark (Jin et al., 2026) The ECG-R1 Benchmark is derived from the ECG-Grounding test set (Jin et al., 2026; Lan et al., 2025). Note that this is separate from the 30,000 instruction-response pairs from the ECG-Grounding training set. It evaluates model outputs across seven metrics: 1) Diagnosis Accuracy, 2) Analysis Completeness, 3) Analysis Relevance, 4) Lead Evidence Validity, 5) ECG Feature Grounding, 6) Evidence-Based Reasoning, and 7) Clinical Diagnostic Fidelity. Each metric is scored by DeepSeek-V3.1-Terminus using a predefined instruction template. When evaluating ORAH, we follow the same evaluation protocol and instruction template used in ECG-R1 (Jin et al., 2026). We provide the full instruction template below. `GroundedECGUnderstanding` and `RealisticDiagnosticProcess` are mapped to ECG Feature Grounding and Clinical Diagnostic Fidelity, respectively.

DeepSeek-V3.1-Terminus Instruction Template for ECG-R1 Benchmark
Task: Evaluate the alignment and quality of a generated ECG interpretation by comparing it to a ground truth clinician's interpretation.
Evaluation Criteria:
1. DiagnosisAccuracy: Evaluates whether the generated diagnosis is correct, specific, and supported by ECG findings.
Scoring: +2 per diagnosis: correctly identified key diagnosis with supporting ECG features; +1: mostly correct but lacking key supporting details; +0: incorrect or vague diagnosis unsupported by ECG features.
2. AnalysisCompleteness: Checks if all key ECG components (rhythm, intervals, waveforms, and lead-specific findings) are discussed.
Scoring: +1 per correctly addressed key ECG feature (e.g., rhythm, PR interval, QRS duration, ST segment, T-wave morphology); +0 per missing or inaccurately described feature.
3. AnalysisRelevance: Assesses whether each provided explanation directly supports the diagnosis.
Scoring: +2 per feature/lead strongly supporting the diagnosis with clear ECG evidence; +1 if relevant but not fully justified; +0 if unrelated or misleading.
4. LeadEvidenceValidity: Evaluates whether lead-related statements are diagnostically necessary, correctly grounded, and free of unsupported lead-wise claims.
Scoring: +2 per key lead/region correctly referenced with explicit matching ECG evidence; +1 for partially correct evidence or minor inconsistencies; +0 if omitted, incorrect, or non-grounded.
5. GroundedECGUnderstanding: Determines if the interpretation references actual ECG features rather than generic terms.
Scoring (0–100): 100: comprehensive findings linked to diagnoses; 80: findings explicitly cited and linked to diagnoses; 50: incomplete ECG references; 0: no specific waveform references.
6. EvidenceBasedReasoning: Evaluates whether the diagnosis follows logical, evidence-supported steps.
Scoring (0–100): 100: thorough and clear evidence-supported reasoning; 80: clear reasoning; 50: incomplete reasoning; 0: unclear or unsupported reasoning.
7. RealisticDiagnosticProcess: Assesses whether the model follows a realistic clinician-style ECG interpretation process.
Scoring (0–100): 100: structured approach considering all relevant factors; 80: structured clinical approach; 50: incomplete clinical reasoning; 0: no structured clinical reasoning.
Note: Each score must be calculated using the criteria above.
Generated ECG Interpretation:
`{{generated}}`
Ground Truth Clinician's Interpretation:
`{{groundtruth}}`
Response: Organize the output in JSON format for each criterion, including a brief explanation for each score, strictly following the specified scoring structure.

ECG-Reasoning Benchmark (Oh et al., 2026) The ECG Reasoning Benchmark is derived from the PTB-XL (Wagner et al., 2020) and MIMIC-IV-ECG datasets (Johnson et al., 2023). It is designed to more rigorously evaluate ELMs on grounded clinical reasoning and granular ECG features. The benchmark consists of chains of multiple-choice questions that progressively build on one another to reach a final diagnostic decision for a given ECG. Performance is measured using four metrics: Initial Diagnosis Accuracy (IDA), which evaluates diagnostic accuracy before reasoning; Completion, which measures the proportion of samples for which the entire reasoning chain is completed correctly; Depth, which measures how far the model progresses through the reasoning sequence; and Ground-Truth Reasoning-Based Diagnosis Accuracy (GT-RDA), which evaluates the final diagnosis when the model is provided with the correct reasoning trajectory. Using this benchmark, Oh et al. (2026) found that ELMs and general VLMs possess the medical knowledge required to identify important ECG findings for diagnosis, but often lack the ability to ground these findings in the underlying ECG signal.

A.2 Private

Internal Dataset 1-3 Internal Dataset 1-3 refers to our private ECG data collected at Allegheny General Hospital (AGH) in Pittsburgh, Pennsylvania. 729,375 12-lead, 10 second ECG recordings sampled at 250 Hz were retrospectively collected and de-identified prior to this study. Each ECG recording is also paired with a diagnostic report that has been verified by a physician at AGH. We will not be open-sourcing Internal Dataset 1-3.

B Additional ORAH Training Details

In this section, we describe additional training details for ORAH.

B.1 Preparing Data for the ECG-Reasoning Benchmark

The ECG-Reasoning Benchmark (Oh et al., 2026) consists of separate PTB-XL and MIMIC-IV-ECG test sets. We augment both sets by adding static, naturalistic instructions and formulating each step as a turn in a multi-turn conversation. We provide an example of a conversation created from the PTB-XL split of the ECG-Reasoning Benchmark in Figure 9 of the appendix. As shown in Figure 11, we train ORAH under three settings: without including any ECG-Reasoning data, with the PTB-XL split, and with the MIMIC-IV-ECG split. We evaluate the first setting on both

Table 4: Training hyperparameters for all training stages. SigLEP pretraining used 8× NVIDIA H100 NVL GPUs, whereas the ORAH stages used 6× NVIDIA H100 NVL GPUs. Batch size denotes the per-GPU micro-batch size.

Pretraining SigLEP

Hyperparameter	Value
Optimizer	AdamW
Learning rate	3×10^{-4}
LR schedule	Cosine
Weight decay	0.01
Batch size	512
Epochs	30

Pretraining ORAH Stage 1

Hyperparameter	Value
Optimizer	Muon + AdamW
Learning rate	5×10^{-4}
LR schedule	Cosine
Warmup ratio	0.03
Weight decay	0.01
Batch size	12
Gradient accumulation	1
Epochs	3

Pretraining ORAH Stage 2

Hyperparameter	Value
Optimizer	Muon + AdamW
Learning rate	2×10^{-4}
LR schedule	Cosine
Warmup ratio	0.03
Weight decay	0.01
Batch size	12
Gradient accumulation	1
Epochs	1

SFT ORAH Stage 1

Hyperparameter	Value
Optimizer	Muon + AdamW
Learning rate	1×10^{-4}
LR schedule	Cosine
Warmup ratio	0.03
Weight decay	0.01
Batch size	12
Gradient accumulation	1
Epochs	3

SFT ORAH Stage 2

Hyperparameter	Value
Optimizer	Muon + AdamW
Learning rate	5×10^{-5}
LR schedule	Cosine
Warmup ratio	0.03
Weight decay	0.01
Batch size	12
Gradient accumulation	1
Epochs	3

RL ORAH Stage 1

Hyperparameter	Value
Optimizer	Muon + AdamW
Learning rate	5×10^{-6}
LR schedule	Constant with warmup
Warmup ratio	0.03
Weight decay	0.01
Batch size	1
Gradient accumulation	3
Epochs	2
Group size	12
Updates per rollout	1
Maximum new tokens	1024
Temperature	1.0
τ_{pos}	1.0
τ_{neg}	1.05

splits of the ECG-Reasoning Benchmark and each of the latter two settings on the opposing split. Because training and evaluation use disjoint datasets, no ECG from the test set is included during training, preventing data contamination. If ORAH improves under this setting, it would suggest that performance on this challenging benchmark can be enhanced through the collection of targeted training data.

B.2 Optimizer Details

For ORAH pretraining and SFT, we partition trainable parameters into disjoint Muon and AdamW groups within a single optimizer. Muon updates two-dimensional parameters, excluding parameters belonging to embedding layers or the LLM output head. AdamW updates all remaining trainable parameters. Depending on the training stage, frozen parameters are excluded from optimization.

Muon uses Nesterov momentum with coefficient 0.95 and five Newton–Schulz iterations per update. For an $m \times n$ matrix, the resulting update is scaled by $0.2\sqrt{\max(m, n)}$. The paired AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$ and $\epsilon_{\text{opt}} = 10^{-10}$. All parameter groups share the stage-specific learning rate and schedule in Table 4. For ORAH pretraining and SFT, the reported learning rates and schedules apply to both Muon and its companion AdamW. The weight decay applies only to tensors with two or more dimensions. Scalars and vectors receive no weight decay.

User:
Question: Does this ECG suggest the presence of first degree AV block?
This question has one of the following options as the correct answer:
- Yes
- No
Your response must be **ONLY** the full text of the selected option. Do not include any uncertainty, explanation, reasoning, or extra words.

Assistant: Yes

User:
Question: To accurately diagnose first degree AV block, which of the following diagnostic criteria should be evaluated?
This question has one of the following options as the correct answer:
- Presence of grouped QRS complexes separated by pauses
- Presence of non-conducted P waves
- Evidence of consistent 1:1 atrioventricular conduction
- Presence of pathological Q waves in inferior leads
- Presence of premature atrial complexes with compensatory pauses
Your response must be **ONLY** the full text of the selected option. Do not include any uncertainty, explanation, reasoning, or extra words.

Assistant: Evidence of consistent 1:1 atrioventricular conduction

⋮

User:
Question: For the selected segment, which range does the measured PR interval fall into?
This question has one of the following options as the correct answer:
- [200ms - 210ms]
- [210ms - 220ms]
- [220ms - 230ms]
- [230ms - 240ms]
- Above 240ms
Your response must be **ONLY** the full text of the selected option. Do not include any uncertainty, explanation, reasoning, or extra words.

Assistant: [230ms - 240ms]

User:
Question: Based on all the findings identified so far, does this ECG suggest the presence of first degree AV block?
This question has one of the following options as the correct answer:
- Yes
- No
- Further findings are required to confirm the diagnosis
Your response must be **ONLY** the full text of the selected option. Do not include any uncertainty, explanation, reasoning, or extra words.

Assistant: Yes

Figure 9: An example of a conversation created out of the PTB-XL split of the ECG-Reasoning Benchmark (Oh et al., 2026) for training. All added text is shown in red.

B.3 RL Stage Details

We further optimize ORAH on the same RL dataset used in ECG-R1 (Jin et al., 2026) using Soft Adaptive Policy Optimization (SAPO) (Gao et al., 2025) to investigate the effects of RL on ORAH.

For each prompt q, comprising an ECG and textual instruction, we sample $G = 12$ candidate responses $\{y_i\}_{i=1}^{G}$ from the behavior policy $\pi_{\theta_{\text{old}}}$. We use stochastic decoding with a sampling temperature of 1.0 and top-p sampling with $p = 0.95$.

Each response receives a scalar reward:

$$R_i = R_i^{\text{format}} + R_i^{\text{tag}} + R_i^{\text{answer}}.$$

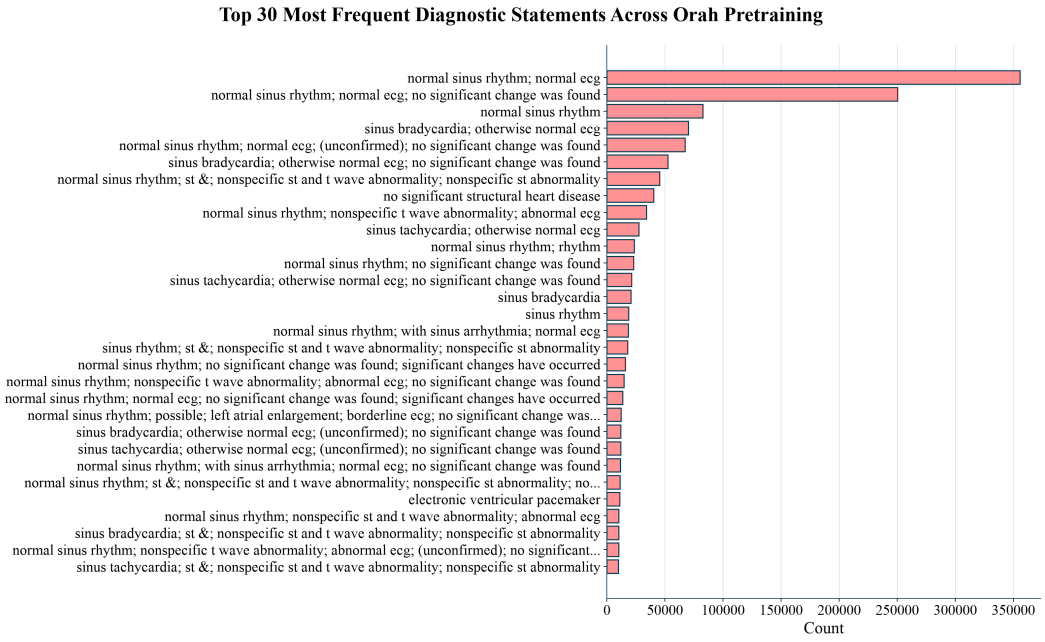


Figure 10: We plot the frequencies of unique diagnostic statements across ORAH pretraining stages 1 and 2. The distribution is long-tailed, with statements describing normal, healthy ECGs occurring most frequently. Note how cardiac electrophysiologists describe normal ECGs in several ways, such as "normal sinus rhythm; normal ecg" and "normal sinus rhythm."

Here, R_i^{format} evaluates whether the response follows the expected reasoning and answer structure, and R_i^{tag} measures the fraction of required thinking and answer tags that appear exactly once. The answer reward R_i^{answer} equally combines exact matching and set-level F1 between the predicted and ground-truth diagnostic labels, represented as semicolon-separated label sets. All three reward components are equally weighted.

We normalize rewards within each sampled group to obtain the advantage:

$$A_i = \frac{R_i - \mu_R}{\sigma_R + \epsilon_{\text{adv}}},$$

where μ_R and σ_R are the mean and standard deviation of the group rewards, respectively, and $\epsilon_{\text{adv}} = 10^{-6}$. The same sequence-level advantage A_i is assigned to every generated token in response y_i. Groups in which all responses receive the same reward are excluded from optimization because they provide no relative preference signal.

For each response token $y_{i,t}$, we compute the importance ratio between the current policy and the behavior policy:

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid q, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid q, y_{i,<t})}.$$

SAPO (Gao et al., 2025) replaces the hard clipping used in GRPO (Shao et al., 2024) with a temperature-controlled soft gate. We minimize

$$\mathcal{L}_{\text{SAPO}} = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\frac{4}{\tau_{i,t}}\sigma(\tau_{i,t}\,(r_{i,t}(\theta)-1))\,A_i,$$

where $\sigma(\cdot)$ is the sigmoid function and

$$\tau_{i,t} = \begin{cases} \tau_{\text{pos}}, & A_i > 0, \\ \tau_{\text{neg}}, & A_i \le 0. \end{cases}$$

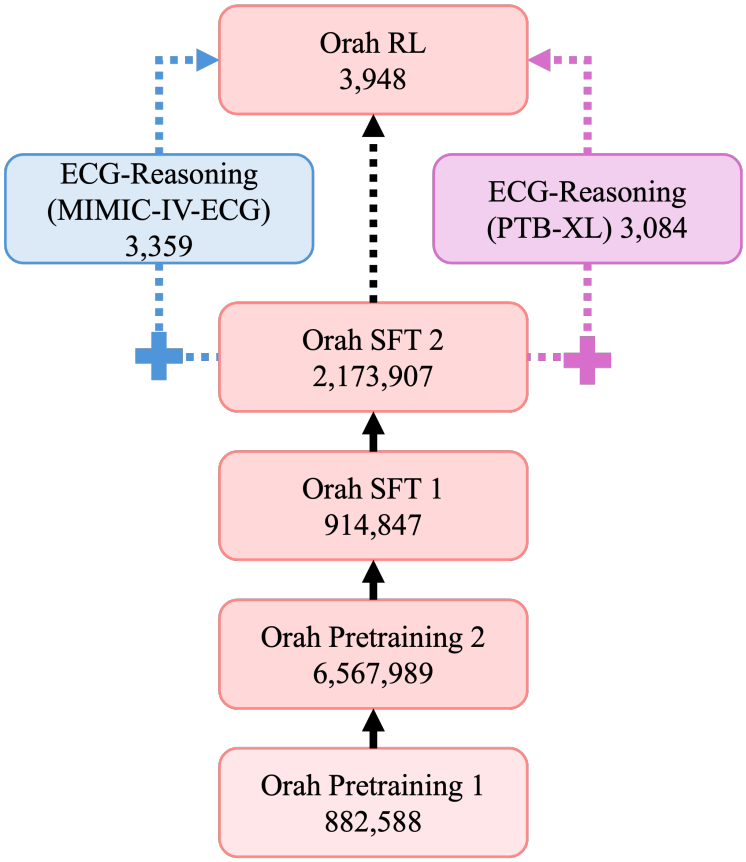


Figure 11: **Data Preparation for the ECG-Reasoning Benchmark.** We illustrate the ECG-Reasoning training splits added to ORAH SFT 2. ⋯▶ indicates the path in which no ECG-Reasoning data are included. ⋯▶ and ⋯▶ indicate the paths in which the PTB-XL and MIMIC-IV-ECG splits are added, respectively. ✚ and ✚ indicate the addition of their corresponding data splits.

We set $\tau_{\text{pos}} = 1.0$ and $\tau_{\text{neg}} = 1.05$ (Gao et al., 2025). The token sums and response lengths include generated reasoning and answer tags and, when present, the first generated termination token. Prompt and padding positions are excluded.

B.4 DYNAMIC PADDING/TRUNCATION

To reduce unnecessary computation, we dynamically pad each token sequence to the longest sequence in its batch rather than to the maximum context length. Let $\mathcal{B}$ denote a batch of B examples, and let

$$\mathbf{u}^{(b)} = \left[u_1^{(b)}, \ldots, u_{M_b}^{(b)}\right], \qquad b \in \mathcal{B},$$

denote the token sequence for example b after any required truncation, where M_b is its length. The sequence length used for the batch is

$$M_{\mathcal{B}} = \max_{b \in \mathcal{B}} M_b.$$

Each sequence is right-padded by

$$\Delta_b = M_{\mathcal{B}} - M_b$$

positions. Specifically, we construct

$$\widetilde{\mathbf{u}}^{(b)} = \left[u_1^{(b)}, \ldots, u_{M_b}^{(b)}, \underbrace{u_{\text{pad}}, \ldots, u_{\text{pad}}}_{\Delta_b \text{ positions}}\right],$$

Table 5: Comparing performances on ECG-QA-CoT. Reported baselines are taken from Langer et al. (2026) and Han et al. (2026).

Method	LLM Backbone	F1	Accuracy
OpenTSLM SoftPrompt	Llama3.2-1B	32.84	35.49
	Llama3.2-3B	33.67	36.25
	Gemma3-270M	1.29	1.11
	Gemma3-1B-pt	27.86	34.76
OpenTSLM Flamingo	Llama3.2-1B	34.62	38.14
	Llama3.2-3B	40.25	46.25
	Gemma3-270M	32.71	35.50
	Gemma3-1B-pt	35.31	37.79
SLIP	Gemma3-270M	–	37.18
ELF	Gemma2-2B-it	61.83	58.55
ORAH (SFT)	Qwen3.5-4B-Base	**63.78**	**60.11**
ORAH (SFT, 25%)	Qwen3.5-4B-Base	57.69	52.76
ORAH (SFT, 50%)	Qwen3.5-4B-Base	58.82	55.51
ORAH (SFT, No Pretraining)	Qwen3.5-4B-Base	60.64	56.39
ORAH (RL)	Qwen3.5-4B-Base	63.41	59.75

Table 6: Comparing performances on the ECG-R1 benchmark. Reported baselines are taken from Jin et al. (2026).

Method	Diagnosis Accuracy	Analysis Completeness	Analysis Relevance	Lead Evidence Validity	ECG Feature Grounding	Evidence-Based Reasoning	Clinical Diagnostic Fidelity
Gemini-3-Pro	13.40	2.41	0.97	0.74	30.13	21.47	27.48
GPT-5.1-Instant	31.48	3.03	1.48	1.92	47.29	40.33	43.46
PULSE	66.13	1.90	1.86	0.19	41.60	39.42	40.53
GEM	74.70	4.25	3.79	4.41	65.34	63.15	62.90
ECG-R1 (SFT)	79.33	6.36	4.58	5.53	79.92	78.08	83.51
ECG-R1 (RL)	80.29	6.51	4.74	5.81	80.57	79.08	84.20
ORAH (SFT)	80.31	**6.72**	**5.73**	**7.34**	**85.07**	**79.98**	85.97
ORAH (SFT, 25%)	78.32	4.65	4.05	4.95	74.10	70.52	74.85
ORAH (SFT, 50%)	78.78	3.86	3.67	4.12	70.98	77.28	79.01
ORAH (SFT, No Pretraining)	81.34	3.97	4.04	4.21	67.09	64.83	65.62
ORAH (RL)	**84.63**	5.83	5.38	5.73	78.49	77.52	**87.22**

such that every sequence in the batch has length $M_{\mathcal{B}}$. Any truncation to the maximum context length of 1024 tokens is performed before batching; the collation function only pads sequences to $M_{\mathcal{B}}$.

C Additional Experiments and Results

We provide additional experiments and results in tabular form. SFT indicates ORAH was trained up to the second SFT stage. RL indicates ORAH was trained with the RL stage. 25% and 50% means ORAH was trained on the randomly sampled 25% and 50% subset of each dataset at every training stage respectively. No Pretraining denotes all pretraining ORAH stages were skipped. No Augmented Data means we did not include the augmented ECG-Reasoning data during SFT 2.

C.1 Tabular Results on ECG-QA-CoT

We provide the tabular results of Figures 3, 6, and 7 with additional baselines and the F1 metric when available in Table 5.

C.2 Tabular Results on ECG-R1 Benchmark

We provide the tabular results of Figures 4, 6, and 7 with additional baselines in Table 6.

C.3 Tabular Results on ECG-Reasoning Benchmark

We provide the tabular results of Figures 5, 6, and 7 with additional metrics and baselines in Table 7.

Table 7: Performance comparison across PTB-XL and MIMIC-IV-ECG splits of the ECG-Reasoning benchmark. Reported baselines are taken from Oh et al. (2026).

Source Dataset	Model	IDA (%)	Completion (%)	Depth (0–4)	GT-RDA (%)
PTB-XL	PULSE (7B)	80.93	5.45	1.24	35.18
	GEM (7B)	84.37	6.13	1.34	65.66
	ECG-R1-SFT (8B)	85.12	5.45	1.80	21.76
	ECG-R1-RL (8B)	85.41	5.90	1.79	22.70
	OpenTSLM (3B)	54.77	0.65	0.17	37.58
	ORAH (SFT, 4B, No Augmented Data)	65.54	0.06	0.14	35.75
	ORAH (SFT, 4B)	**92.80**	**8.56**	1.55	40.43
	Hulu-Med (32B)	57.49	3.21	1.62	**99.42**
	MedGemma (27B)	55.19	1.20	1.51	76.46
	Qwen3-VL (32B)	59.73	5.71	1.63	89.46
	Llama-3.2-Vision (90B)	56.84	4.47	1.34	91.21
	Gemini-3-Flash	65.86	6.26	**2.09**	91.76
	GPT-5.2	67.80	5.84	1.97	85.70
MIMIC-IV-ECG	PULSE (7B)	72.10	4.76	1.22	30.01
	GEM (7B)	76.27	5.30	1.30	57.01
	ECG-R1-SFT (8B)	79.58	5.45	1.71	22.12
	ECG-R1-RL (8B)	80.17	5.81	1.71	22.95
	OpenTSLM (3B)	51.56	0.36	0.13	35.43
	ORAH (SFT, 4B, No Augmented Data)	60.26	0.21	0.14	31.17
	ORAH (SFT, 4B)	**90.71**	7.77	1.49	34.62
	ORAH (SFT, 4B, 25%)	88.30	6.40	1.48	33.34
	ORAH (SFT, 4B, 50%)	89.37	6.51	1.49	34.60
	ORAH (SFT, 4B, No Pretraining)	89.93	5.75	1.61	25.29
	ORAH (RL, 4B)	90.04	**8.07**	1.50	35.31
	Hulu-Med (32B)	51.62	3.36	1.53	**97.20**
	MedGemma (27B)	50.88	1.22	1.45	78.24
	Qwen3-VL (32B)	54.30	5.03	1.59	88.00
	Llama-3.2-Vision (90B)	52.52	4.38	1.27	88.39
	Gemini-3-Flash	60.64	5.72	**2.07**	91.25
	GPT-5.2	63.41	5.27	1.93	83.83

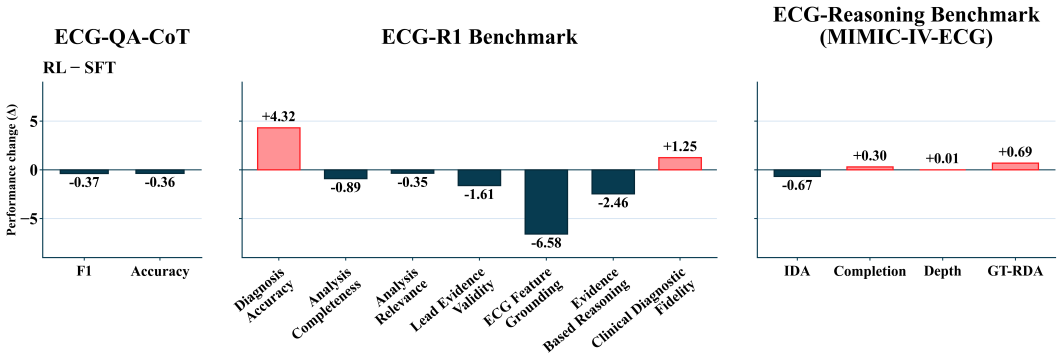


Figure 12: **Effects of RL.** Performance differences between RL and SFT across three benchmarks.

With and without Augmented ECG-Reasoning We evaluate the contribution of augmented ECG-Reasoning data by training ORAH with and without these data during SFT 2. As shown in Table 7, including augmented data improves all four metrics, increasing IDA by 30.45 percentage points, Completion by 7.56 percentage points, Depth by 1.35, and GT-RDA by 3.45 percentage points. These results highlight the importance of augmented reasoning data in our training recipe.

C.4 DOES RL IMPROVE PERFORMANCE?

We evaluate the effect of RL across three benchmarks in Figure 12. On ECG-QA-CoT, RL decreases F1 and accuracy by 0.37 and 0.36 points, respectively. On the MIMIC-IV-ECG split of the ECG-Reasoning Benchmark, Completion, Depth, and GT-RDA increase slightly by 0.30, 0.01, and 0.69, respectively, while IDA decreases by 0.67 points. The ECG-R1 Benchmark shows larger

trade-offs: Diagnosis Accuracy and Clinical Diagnostic Fidelity improve by 4.32 and 1.25 points, respectively, but the other five metrics decline, most notably ECG Feature Grounding (6.58 points) and Evidence-Based Reasoning (2.46 points). Although ECG-R1 reports modest gains from RL on its own benchmark (Table 6), these benefits do not consistently extend to ORAH under our training setup. These results suggest that diagnostic gains from RL do not necessarily translate into improved grounding or reasoning under our training setup, motivating further investigation into RL data curation and reward design.

D PROMPT TEMPLATES

We provide all the prompt templates used throughout the study.

D.1 SYSTEM PROMPT

ELM System Prompt
You are an expert cardiac electrophysiologist. You interpret 12-lead ECG signals and answer any questions about them.
Deliver responses that are precise and grounded in electrophysiology and clinical reasoning. For general, non-ECG questions, respond appropriately.
If the response opens with `<think>`, close it with `</think>`, then give a brief interpretation if necessary and end with `<answer>...</answer>`. Otherwise answer the question directly, with no tags. Never repeat a tag that has already been written.

D.2 QWEN3.5 TEMPLATE

We provide the conversation templates for Qwen-3.5-4B-Base with and without thinking (Team, 2026). q_{sys}, q_1, y_1 refers to the system prompt, textual query and response respectively.

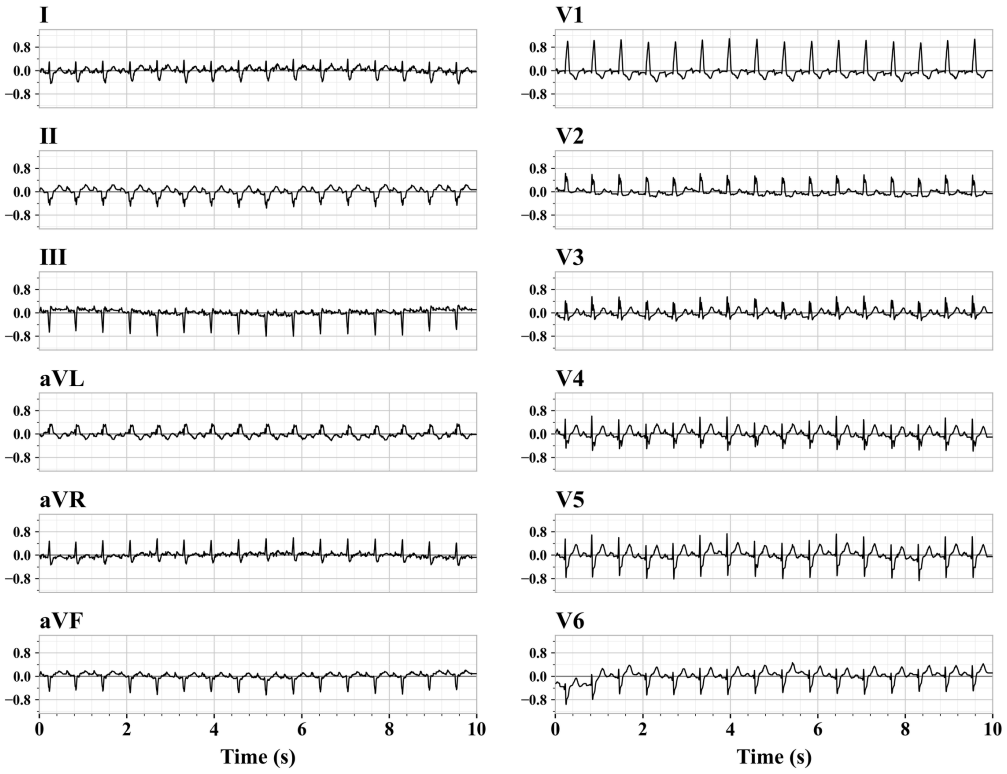
Qwen-3.5-4B-Base Conversation Template without Thinking

$<|\text{im_start}|>$ system
q_{sys} $<|\text{im_end}|>$
$<|\text{im_start}|>$ user
`<signal>`
q_1 $<|\text{im_end}|>$
$<|\text{im_start}|>$ assistant
`<think>`

`</think>`
y_1 $<|\text{im_end}|>$

Qwen-3.5-4B-Base Conversation Template with Thinking
$<|\text{im_start}|>$ system
q_{sys} $<|\text{im_end}|>$
$<|\text{im_start}|>$ user
`<signal>`
q_1 $<|\text{im_end}|>$
$<|\text{im_start}|>$ assistant
`<think>`

{Thinking Content} `</think>`
`<answer>`y_1`</answer>` $<|\text{im_end}|>$



1. Clinical Context: 85-year-old female patient. 12-lead ECG. clinical recording. recorded with AT-6 C 5.5. Signal quality: baseline drift noted, static noise present, burst noise present, electrode artifacts present. extra beats detected during recording. pacemaker present.
Question: Which range does the p duration of this ECG fall in, above the normal range or within the normal range?

Incorrect

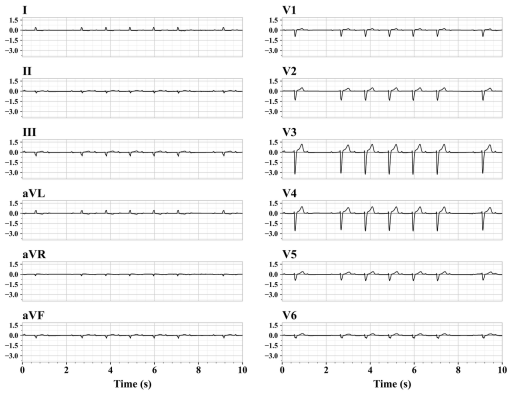
Orah:
<think>
To determine the P wave duration, we must first consider the standard normal range, which is typically between 0.08 to 0.11 seconds. In this ECG, despite the presence of baseline drift, static noise, burst noise, and electrode artifacts, the P waves can be evaluated for duration. The presence of a pacemaker may influence the rhythm, but it does not directly affect the intrinsic P waveduration. The P waves appear consistent across multiple leads, suggesting that the atrial depolarization is uniform. Given the clinical context of an 85-year-old female with a pacemaker, it is essential to consider any atrial enlargement or conduction abnormalities that could prolong the Pwave duration. However, the ECG does not show evidence of prolonged P wave duration, which wouldindicate a different set of atrial depolarization. Therefore, the P wave duration in this ECG iswithin the normal range
</think>

<answer>none</answer>

Ground Truth:
<think>
To determine the P wave duration on this ECG, we must first consider the clinical context and the technical aspects of the recording. The presence of baseline drift, static noise, burst noise, and electrode artifacts can complicate the interpretation, but careful analysis of the P wave morphology and duration is still possible. In a typical ECG, the normal P wave duration is up to 120milliseconds. Given that this patient has a pacemaker, it is crucial to differentiate between native atrial activity and any pacemaker-induced activity. The P waves should be examined across multiple leads, particularly leads II and V1, where they are often most prominent. Despite the noise and artifacts, the P wave duration appears prolonged when measured against the standard ECG grid, indicating atrial enlargement or conduction delay. This is consistent with the clinical presentation of an elderly patient, where atrial enlargement or conduction system disease is more prevalent. Thus, the P wave duration in this ECG falls above the normal range
</think>

<answer>above the normal range</answer>

Figure 13: **ECG-QA-CoT Example 1.** A qualitative comparison of ORAH's responses to questions with ground-truth answers from ECG-QA-CoT. We also include the generated and ground truth thinking traces when available.



1. Does this ECG suggest the presence of second degree AV block?
A. Yes
B. No

Correct
Orah: Yes
Ground truth: Yes

2. To accurately diagnose second degree AV block, which of the following diagnostic criteria should be evaluated?
A. Presence of compensatory pauses following premature beats
B. Presence of left axis deviation
C. Occurrence of intermittent dropped beats
D. Rapid, regular ventricular response
E. Pattern of alternating long and short RR intervals

Correct
Orah: Occurrence of intermittent dropped beats
Ground truth: Occurrence of intermittent dropped beats

3. Regarding the criterion you selected, does this ECG show any possibility of dropped beats, where an RR interval is significantly longer than the preceding RR intervals?
A. Yes
B. No

Correct
Orah: Yes
Ground truth: Yes

4. Which of the following intervals between two consecutive QRS waves represents this significantly longer interval? To ensure clear indexing, the '1st QRS wave' refers to the QRS wave located at [0.53s - 0.69s] in the ECG signal.
A. [2nd QRS wave - 3rd QRS wave]
B. [3rd QRS wave - 4th QRS wave]
C. [5th QRS wave - 6th QRS wave]
D. [6th QRS wave - 7th QRS wave]

Correct
Orah: [6th QRS wave - 7th QRS wave]
Ground truth: [6th QRS wave - 7th QRS wave]

5. Based on the finding identified above, does this ECG suggest the presence of second degree AV block?
A. Yes
B. No
C. Further findings are required to confirm the diagnosis

Correct
Orah: Further findings are required to confirm the diagnosis
Ground truth: Further findings are required to confirm the diagnosis

6. In addition to the finding you just identified, which other diagnostic criterion should be evaluated to diagnose second degree AV block?
A. Presence of prominent U waves mimicking P waves
B. Presence of ST-segment elevation in anterior leads
C. Presence of an early, ectopic P wave with different morphology
D. Presence of a non-conducted P wave within the dropped interval
E. Complete absence of atrial activity during the pause

Correct
Orah: Presence of a non-conducted P wave within the dropped interval
Ground truth: Presence of a non-conducted P wave within the dropped interval

7. Regarding the criterion you selected, is a non-conducted P wave, which is not caused by a premature beat, visible within the prolonged RR interval that you identified?
A. Yes
B. No

Incorrect
Orah: No
Ground truth: Yes

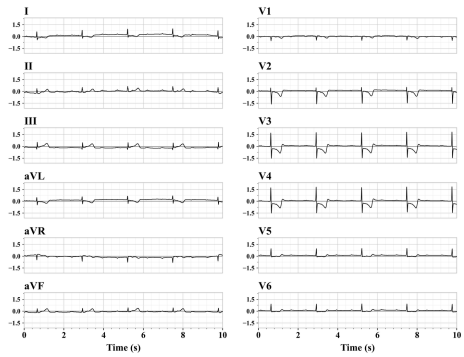
8. In which of the following segments can you observe the non-conducted P wave, specifically within the prolonged RR interval? The options below refer to time ranges on the ECG signal, provided in seconds.
A. [4.81s - 4.97s]
B. [5.02s - 5.32s]
C. [7.53s - 7.66s]
D. [8.67s - 8.75s]

Incorrect
Orah: [5.02s - 5.32s]
Ground truth: [7.53s - 7.66s]

9. Based on all the findings identified so far, does this ECG suggest the presence of second degree AV block?
A. Yes
B. No
C. Further findings are required to confirm the diagnosis

Incorrect
Orah: Further findings are required to confirm the diagnosis
Ground truth: Yes

Figure 14: **ECG-Reasoning Benchmark Example 1.** A qualitative comparison of ORAH's responses to step-by-step questions with ground-truth answers from the MIMIC-IV-ECG split of the ECG-Reasoning Benchmark.



1. Does this ECG suggest the presence of anterior ischemia?
A. Yes
B. No

Correct
Orah: Yes
Ground truth: Yes

2. To accurately diagnose anterior ischemia, which of the following diagnostic criteria should be evaluated?
A. Presence of an RSR' pattern in leads V1 and V2
B. Presence of J-point elevation in anterior leads
C. Presence of ST-segment depression in leads II, III, and aVF
D. Presence of ST-segment depression in anterior leads
E. Presence of ST-segment elevation in anterior leads

Correct
Orah: Presence of ST-segment depression in anterior leads
Ground truth: Presence of ST-segment depression in anterior leads

3. Regarding the criterion you selected, does this ECG show ST-segment depression in at least two of the anterior leads, including leads V2, V3, and V4? Note that ST depression is defined as a depression of the J-point greater than 0.1mV (1mm) in lead V2, and greater than 0.07mV (0.7mm) in leads V3 and V4.
A. Yes
B. No

Correct
Orah: Yes
Ground truth: Yes

4. Which of the following leads show ST-segment depression? Select all possible leads from the options below.
A. Lead II
B. Lead aVF
C. Lead V2
D. Lead V3
E. Lead V4

Incorrect
Orah: Lead V3, Lead V4
Ground truth: Lead V2, Lead V3, Lead V4

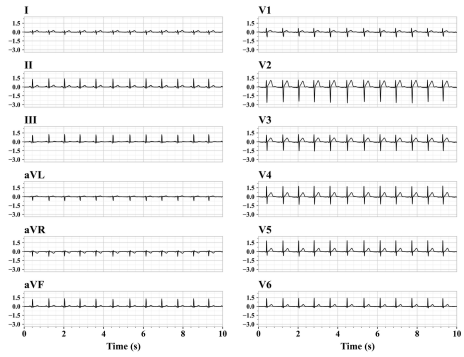
5. Within the selected leads, in which of the following waves can you observe the QRS complex with ST-segment depression? The options below refer to time ranges on the ECG signal, provided in seconds.
A. [1.02s - 1.29s]
B. [2.86s - 2.98s]
C. [3.30s - 3.52s]
D. [4.62s - 4.76s]

Incorrect
Orah: [3.30s - 3.52s]
Ground truth: [2.86s - 2.98s]

6. Based on the finding identified above, does this ECG suggest the presence of anterior ischemia?
A. Yes
B. No
C. Further findings are required to confirm the diagnosis

Correct
Orah: Yes
Ground truth: Yes

Figure 15: **ECG-Reasoning Benchmark Example 2.** A qualitative comparison of ORAH's responses to step-by-step questions with ground-truth answers from the MIMIC-IV-ECG split of the ECG-Reasoning Benchmark.



1. Does this ECG suggest the presence of anterior myocardial infarction?
A. Yes
B. No

Incorrect
Orah: No
Ground truth: Yes

2. To accurately diagnose anterior myocardial infarction, which of the following diagnostic criteria should be evaluated?
A. Presence of ST-segment elevation in anterior leads
B. Presence of diffuse concave ST-segment elevation across precordial leads
C. Presence of spiked pacemaker artifacts preceding the QRS complex
D. Evidence of consistent 1:1 atrioventricular conduction
E. Presence of PR-segment depression in anterior leads

Correct
Orah: Presence of ST-segment elevation in anterior leads
Ground truth: Presence of ST-segment elevation in anterior leads

3. Regarding the criterion you selected, does this ECG show ST elevation in at least two of the anterior leads, including leads V1, V2, V3, and V4? Note that ST elevation is defined as an elevation of the J-point greater than 0.1mV (1mm) in these leads.
A. Yes
B. No

Incorrect
Orah: No
Ground truth: Yes

4. Which of the following leads show ST elevation? Select all possible leads from the options below.
A. Lead aVF
B. Lead V1
C. Lead V2
D. Lead V3
E. Lead V4

Correct
Orah: Lead V1, Lead V2, Lead V3, Lead V4
Ground truth: Lead V1, Lead V2, Lead V3, Lead V4

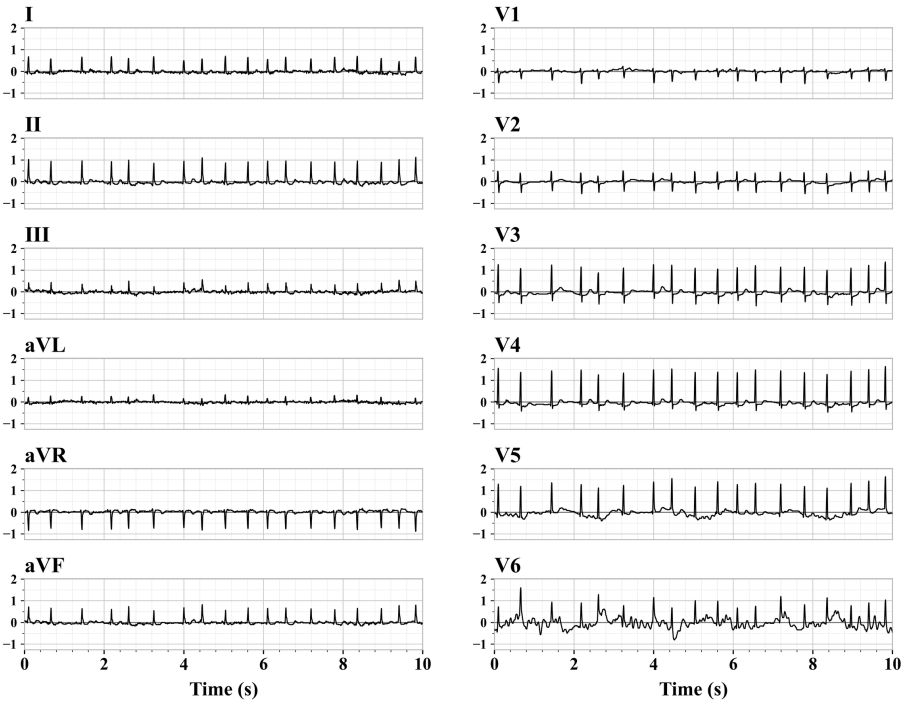
5. Within the selected leads, in which of the following segments can you observe the QRS complex with ST elevation? The options below refer to time ranges on the ECG signal, provided in seconds.
A. [1.84s - 1.95s]
B. [3.57s - 3.66s]
C. [7.84s - 8.07s]
D. [8.13s - 8.42s]

Correct
Orah: [3.57s - 3.66s]
Ground truth: [3.57s - 3.66s]

6. Based on the finding identified above, does this ECG suggest the presence of anterior myocardial infarction?
A. Yes
B. No
C. Further findings are required to confirm the diagnosis

Correct
Orah: Yes
Ground truth: Yes

Figure 16: **ECG-Reasoning Benchmark Example 3.** A qualitative comparison of ORAH's responses to step-by-step questions with ground-truth answers from the MIMIC-IV-ECG split of the ECG-Reasoning Benchmark.



1. Interpret the provided ECG signal, identify key features and abnormalities in each lead, and generate a clinical diagnosis that is supported by the observed evidence.

Not scored

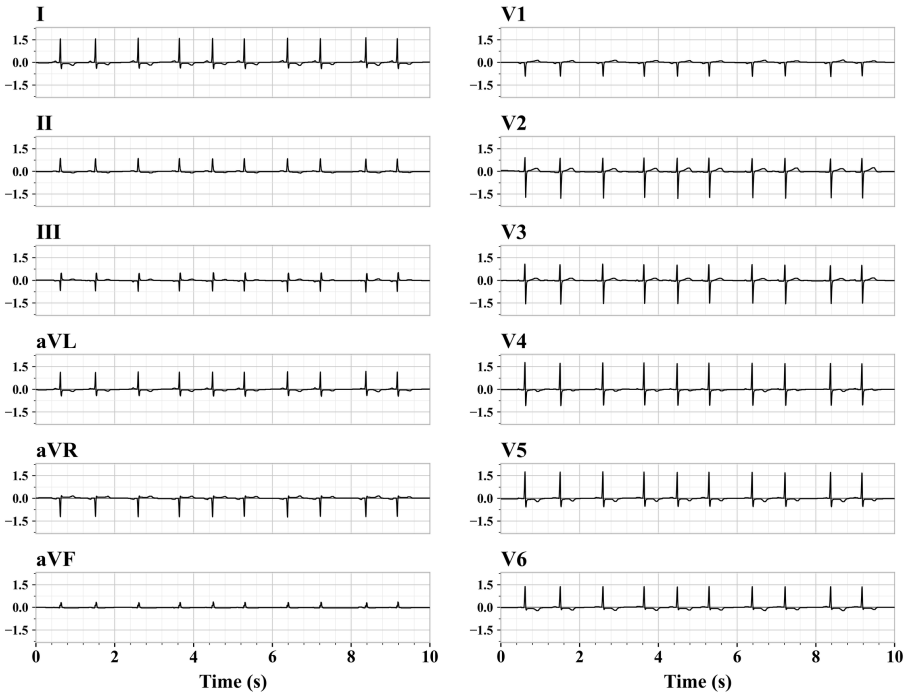
Orah:

<think>
Step 1: Technical, Rate & Rhythm
The ECG demonstrates an irregularly irregular rhythm with a mean ventricular rate of 103 bpm, consistent with tachycardia. There is a complete absence of organized P waves across all leads, including Lead II and V1, where no consistent atrial activity is visible. The baseline shows some undulation but no significant artifact that would preclude interpretation. The irregularity and lack of P waves confirm the underlying rhythm as atrial fibrillation, ruling out sinus rhythm, atrial flutter, and other tachycardias dependent on atrial activity.
Step 2: Conduction, Axis & Intervals
The QRS complexes are narrow with a duration of less than 0.10 seconds in most leads, ruling outbundle branch blocks and ventricular rhythms. The frontal plane axis is normal, as evidenced by positive QRS complexes in both Lead I and Lead II. The PR interval is not measurable due to the absence of P waves, which is consistent with the diagnosis of atrial fibrillation. There is no evidence of pre-excitation or AV block.
Step 3: Chamber Hypertrophy & Voltage
There is no evidence of left ventricular hypertrophy, as the S wave in V1 and the R wave in V6 do not meet the Sokolow-Lyon voltage criteria. The R wave progression from V1 to V4 is normal, with a transition zone occurring appropriately by lead V3, which rules out conditions like poor R wave progression or reversed progression suggestive of right ventricular hypertrophy or posterior infarction. The QRS voltage in the limb and precordial leads is within normal limits.
Step 4: Ischemia, Infarction & Mimics
A systematic analysis of all lead groups reveals no evidence of acute ischemia or infarction. There is no significant ST-segment elevation or depression meeting diagnostic thresholds in the inferior (II, III, aVF), lateral (I, aVL, V5, V6), or anterior (V1-V4) leads. Lead aVR shows no ST elevation to suggest left main disease. There are no pathologic Q waves in any territory. The ST segments and T waves are nonspecifically abnormal but do not show the convex ST elevation or reciprocal changes required to diagnose an acute STEMI.
Step 5: Electrolytes & QT
The QT interval is within normal limits when corrected for rate, with a QTc value below 440 ms. The T wave morphology in the precordial leads is not characteristically tented to suggest hyperkalemia, nor are there prominent U waves to indicate hypokalemia. The T waves are of normal configuration for the given rhythm.
Step 6: Final Medical Reasoning
The primary and definitive abnormality is atrial fibrillation with a controlled ventricular response. The absence of P waves and the irregularly irregular rhythm are the hallmark findings. There is no concomitant evidence of acute coronary syndrome, chamber hypertrophy, conduction abnormalities, or electrolyte disturbances. The nonspecific ST-T changes observed are not diagnostic of ischemia in this context.
</think>
This ECG demonstrates atrial fibrillation with a ventricular rate of 103 bpm, characterized by an irregularly irregular rhythm and the absence of organized P waves. The QRS axis is normal, and the QRS complex duration is narrow, ruling out bundle branch blocks or ventricular tachycardia. There is no voltage criteria for left ventricular hypertrophy, and R wave progression across the precordium is normal.

Ground Truth:

Upon examining the provided ECG signal, several key features indicate the presence of atrial fibrillation (AF). The irregularly irregular rhythm, as evidenced by varying RR intervals, supports this observation. Notably, the RR intervals range widely, which is a hallmark of AF due to inconsistent atrial impulses reaching the ventricles. Additionally, the absence of distinct P waves on leads with traditionally observable P waves is another pivotal indicator; this suggests a loss oforganized atrial electrical activity, consistent with AF.
The heart rate approximates 107 bpm, indicative of a moderate ventricular rate, aligning with a presentation of AF with a controlled rate. No consistent pattern in the PR interval is observed, further corroborating the diagnosis of AF, as atrial activity is chaotic and not leading to a predictable conduction pattern through the AV node.
Analyzing leads such as lead II, where P wave characteristics are more observable under normal conditions, we see minimal or absent P wave formations. The variability and inconsistency of PR intervals signify typical conduction during AF episodes, contributing to its erratic rhythm.
In terms of additional abnormalities, minor QRS and T wave deviations across various leads were noted, but they did not form a pathological pattern significant enough to suggest additional diagnoses such as ischemia or infarction. The ST segment and QTc intervals are within normal limits given the context of AF. This ECG analysis supports a diagnosis of atrial fibrillation, underpinning the clinical signal with evidenced irregular RR intervals and absent P waves. My interpretation remains focused on the features aligning with AF without assuming further conditions not supported by the ECG findings.

Figure 17: **ECG-R1 Benchmark Example 1.** A qualitative comparison of ORAH's responses to questions with ground-truth answers from the ECG-R1 Benchmark. We also include the generated and ground truth thinking traces when available.



1. Interpret the provided ECG signal, identify key features and abnormalities in each lead, and generate a clinical diagnosis that is supported by the observed evidence.

Not scored

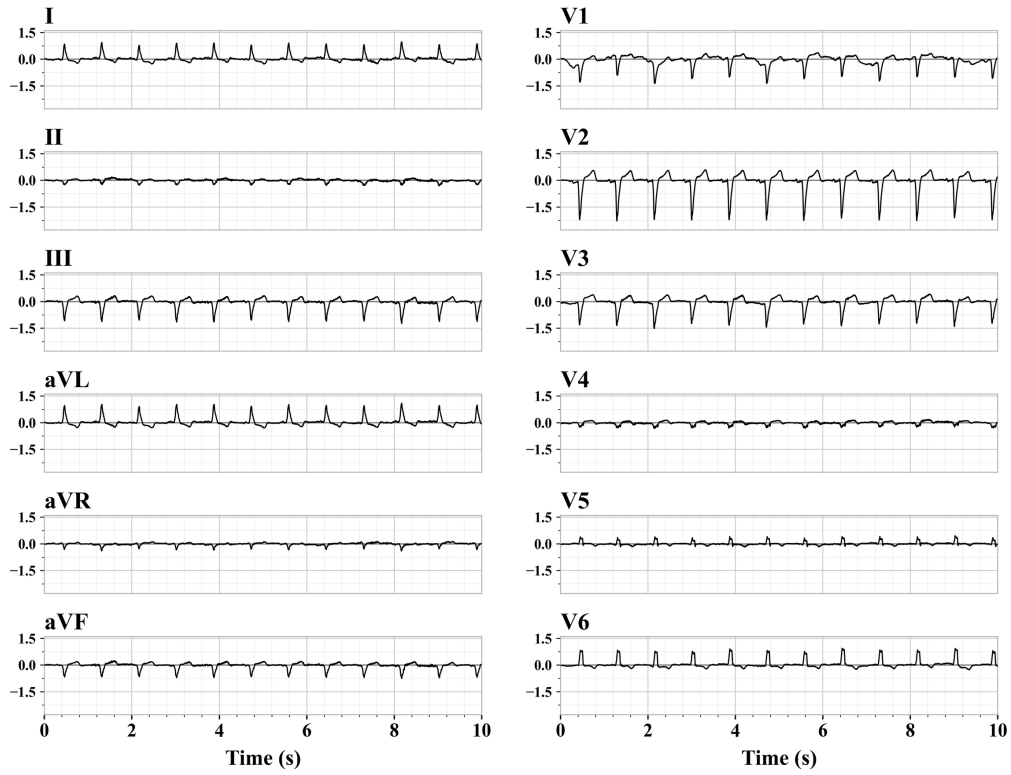
Orah:

<think>
Step 1: Technical, Rate & Rhythm
The ECG demonstrates good technical quality with a stable baseline, standard 25 mm/s speed, and 10 mm/mV calibration; there is no evidence of lead reversal as P waves are upright in lead I. Theheart rate is 63 bpm, indicating a normal sinus rate. The rhythm is sinus, confirmed by the presence of upright P waves in leads II, III, and aVF; however, there is sinus arrhythmia evidenced by the variable RR intervals. The P wave morphology in lead II is normal in duration and amplitude, ruling out significant atrial enlargement.
Step 2: Conduction, Axis & Intervals
The frontal plane axis is normal, as leads I and II are both positive. The PR interval is withinnormal limits at approximately 160 ms, ruling out AV block or pre-excitation syndromes. The QRS duration is normal at less than 100 ms, which excludes bundle branch blocks; specifically, there is no rSR' pattern in V1 to suggest RBBB and no broad monophasic R wave in V6 to suggest LBBB.
Step 3: Chamber Hypertrophy & Voltage
There is evidence of left ventricular hypertrophy based on voltage criteria, with an R wave amplitude in lead aVL measuring greater than 11 mm and the S wave in V1 plus the R wave in V6 exceeding 35 mm. R wave progression across the precordial leads V1 to V4 is normal, with a smooth transition, which rules out poor R wave progression or reversed progression suggestive of right ventricular hypertrophy or posterior infarction.
Step 4: Ischemia, Infarction & Mimics
Analysis of the inferior leads (II, III, aVF) shows no significant ST segment elevation or depression, and there are no pathologic Q waves, ruling out acute inferior infarction or ischemia. Thelateral leads (I, aVL, V5, V6) demonstrate secondary repolarization abnormalities, including ST-T wave changes concordant with the QRS complex in the setting of LVH, but no primary ischemic patterns. The anterior leads (V1-V4) show no ST elevation meeting STEMI criteria; however, there aresubtle ST-T wave changes in V3 and V4. Lead aVR shows no significant ST elevation to suggest left main disease.
Step 5: Electrolytes & QT
The T waves in the precordial leads are not characteristically tented or peaked, which rules outhyperkalemia, and there are no prominent U waves to suggest hypokalemia. The QTc interval is within normal limits, calculated at approximately 450 ms, indicating no significant QT prolongation.
Step 6: Final Medical Reasoning
The ECG findings are consistent with sinus arrhythmia, left ventricular hypertrophy with secondary repolarization abnormalities, and no evidence of acute ischemia or infarction.
</think>
This ECG demonstrates sinus rhythm at 63 bpm with sinus arrhythmia. There is evidence of left ventricular hypertrophy based on voltage criteria in lead aVL and the Sokolow-Lyon index. Secondaryrepolarization abnormalities are present in the lateral leads, consistent with the strain pattern of LVH. No findings suggestive of acute ischemia, infarction, or conduction disease are identified.
<answer>Sinus arrhythmia; LVH with secondary repolarization abnormality; Abnormal R-wave progression, early transition</answer>

Ground Truth:

Upon analysis of the provided ECG signal, several key features are noticeable that align with specific clinical implications. The rhythm is regular with a heart rate of 62 bpm, indicating a sinus rhythm, which is a normal finding. In evaluating lead I, there is a moderate increase in QRS amplitude and a longer QRS duration, with some ST segment horizontal declination. This suggests the possibility of left ventricular hypertrophy (LVH), and further supports the suspicion of LVH seen in the prolonged QTc interval across the leads. Lead II displays minor alterations with slightly prolonged PR intervals, but no significant ST or T wave changes are observed, suggesting no acute inferior wall ischemia. Similar stability is noted in leads III and aVF, with consistent QRS amplitude and duration, indicating a stable inferior wall status.
A pronounced finding is seen in leads V2 to V4, where a high QRS amplitude is consistent with LVH. Additionally, the T wave durations in these leads show slight elongation, hinting at secondaryrepolarization abnormalities, which can accompany hypertrophy. Lead aVR and the anterior leads V1 to V3 display some upslope ST forms, typically not indicating significant pathology in isolation. Leads V5 and V6 show high QRS amplitude with no acute ST alterations, aligning with the assessment of hypertrophy rather than ischemia.
In conclusion, the ECG findings, particularly the augmented QRS voltage and extended durations in various leads, alongside unspecific ST-T changes, support a likely diagnosis of left ventricular hypertrophy with secondary repolarization abnormalities. The absence of prominent ST elevation or depression dismisses acute ischemic events, reinforcing that the identified hypertrophy is theprimary pathological observation.

Figure 18: **ECG-R1 Benchmark Example 2.** A qualitative comparison of ORAH's responses to questions with ground-truth answers from the ECG-R1 Benchmark. We also include the generated and ground truth thinking traces when available.



1. Interpret the provided ECG signal, identify key features and abnormalities in each lead, and generate a clinical diagnosis that is supported by the observed evidence.

Not scored

Orah:

<think>

</think>

The provided ECG signal and computed measurements reveal several key features that guide the clinical diagnosis. The sinus rhythm is evident from the regular P waves preceding each QRS complex,and the heart rate is approximately 69 bpm, which is within normal limits. Notably, there is a left axis deviation, which can be inferred from the QRS complexes in leads I and aVL. The presenceof a left bundle branch block is suggested by the prolonged QRS duration, especially in lead I where durations range from 170 to 180 ms. This finding is consistent across multiple leads, including leads V1 to V6, where a broadened QRS complex is observed.

Additionally, the possibility of an inferior infarct is suggested by the abnormal Q waves and STsegment changes in leads II, III, and aVF. These leads show a lack of expected R-wave progression, indicating possible necrosis or ischemia in the inferior wall of the heart. The ST segment andT wave changes in these leads further support this hypothesis, as they can reflect underlying ischemic processes.

The ECG also presents abnormalities in the anterior leads V2 to V4, where the presence of abnormal QRS complexes and ST segment deviations could suggest an anteroseptal infarction. The lateral leads, such as I and aVL, also show changes that may be indicative of lateral wall ischemia. These findings collectively suggest an abnormal ECG with evidence of a left bundle branch block, leftaxis deviation, and possible inferior and anteroseptal infarcts. These conclusions are drawn directly from the observed ECG features and the computed measurements from the leads.

Ground Truth:

The ECG analysis reveals several important findings. Firstly, there is a sinus rhythm present, indicated by the regular P waves and consistent rhythm across the leads. Notably, there is evidence of left axis deviation, which is supported by the findings in lead I, where the QRS complex is positive, and in lead aVF, where the QRS complex is negative. This pattern is characteristic of left axis deviation. Moreover, the ECG displays a left bundle branch block (LBBB) as evidenced by the prolonged QRS duration, particularly noticeable in leads I, aVL, V5, and V6. The QRS complexes here are broad and notched, which are classic for LBBB.

In terms of ischemic changes, there are indications that suggest a possible inferior infarct. This is inferred from the findings in leads III and aVF, where there are abnormal Q waves. AbnormalT wave inversions in these leads could also point towards an old inferior infarction. Furthermore, the ECG is classified as abnormal due to these observed irregularities. It's crucial to correlate these findings with the clinical context of the patient to consider the likelihood of additional conditions that may not be covered by this ECG alone. Comprehensive assessment through patient history, physical examination, and potentially further cardiac investigations would be prudent to validate the interpretation and diagnosis.

Figure 19: **ECG-R1 Benchmark Example 3.** A qualitative comparison of ORAH's responses to questions with ground-truth answers from the ECG-R1 Benchmark. We also include the generated and ground truth thinking traces when available.