ELF: A Family of Encoder-Free ECG-Language Models

William Jongwon Han* WJHAN@ANDREW.CMU.EDU
Carnegie Mellon University, USA

Tony Chen* TONYCHE2@ANDREW.CMU.EDU
Carnegie Mellon University, USA

Chaojing Duan CHAOJING.DUAN@AHN.ORG
Allegheny Health Network, USA

Xiaoyu Song XIAOYUSO@ANDREW.CMU.EDU
Carnegie Mellon University, USA

Yihang Yao YIHANGYA@ANDREW.CMU.EDU
Carnegie Mellon University, USA

Yuzhe Yang YUZHEY@UCLA.EDU
University of California, Los Angeles, USA

Michael A. Rosenberg MICHAEL.A.ROSENBERG@CUANSCHUTZ.EDU
University of Colorado, USA

Emerson Liu EMERSON.LIU@AHN.ORG
Allegheny Health Network, USA

Ding Zhao DINGZHAO@ANDREW.CMU.EDU
Carnegie Mellon University, USA

Abstract

ECG–Language Models (ELMs) extend recent advances in Multimodal Large Language Models (MLLMs) to automated ECG interpretation. However, most existing ELMs inherit Vision–Language Model (VLM) design choices and rely on pretrained ECG encoders, introducing substantial architectural and training complexity. Inspired by encoder-free VLMs, we introduce **ELF**, a family of three encoder-free ELMs that remain competitive with, and often outperform, prior state-of-the-art ELMs across two datasets despite substantially simpler architectures and training pipelines. All code and data are available at github.com/ELM-Research/ECG-Language-Models.

1. Introduction

The global volume of over 300 million ECGs recorded annually (Zhu et al., 2020), combined with the expertise required for accurate interpretation and the growing shortage of physicians (AAMC, 2024), has created demand for automated ECG analysis. Recently, the automation of ECG interpretation has advanced from classification-based models (Rajpurkar et al., 2017) to ECG–Language Models (ELMs) (Zhao et al., 2024).

1. * Denotes equal contribution.

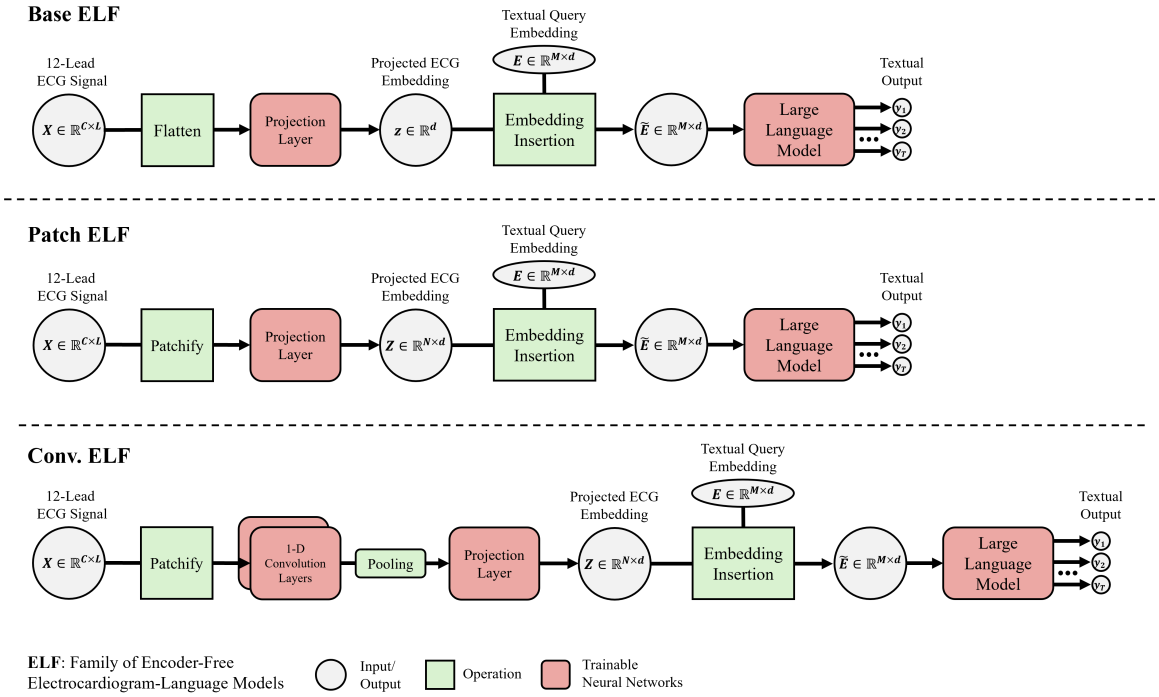

Figure 1: **The Family of ELF**. We present three encoder-free ELM architectures: Base ELF, Patch ELF, and Conv. ELF. Base ELF maps the full ECG into a single embedding token, Patch ELF represents the ECG as a sequence of non-overlapping patch tokens, and Conv. ELF augments patch-based tokenization with lightweight temporal convolutions before projection into the LLM hidden space.

In Vision-Language Models (VLMs), RGB images are typically processed by a large neural network (vision encoder), which is first pretrained on internet-scale image datasets with its own learning objective (Bai et al., 2025). The pretrained encoder is then frozen and paired with a projection layer that maps its output to a latent representation compatible with the Large Language Model (LLM) (Liu et al., 2023). Similarly, many ELM architectures include an ECG encoder that is separately pretrained on large collections of ECG data (Li et al., 2025). In practice, developing an effective encoder, whether for ECG or vision data, is a complex process that requires careful feature engineering and tuning of the architecture, learning objectives, and training procedures.

To address these complexities in the VLM domain, an encoder-free VLM called Fuyu-8B (Bavishi et al., 2023) was introduced. Unlike conventional VLMs, Fuyu-8B omits a dedicated vision encoder and instead adds a simple linear projection for patched images. The projection layer and decoder-only transformer are trained jointly in an end-to-end manner using an autoregressive objective. Fuyu-8B demonstrated that a vision encoder is not necessary to achieve competitive performance on VLM benchmarks compared to state-of-the-art (SOTA) VLMs at the time that relied on intricate vision encoders.

Taking inspiration from Fuyu-8B, we introduce ELF, a family of three encoder-free ELM architectures: (1) Base ELF, (2) Patch ELF, and (3) Conv. ELF. We compare ELF against four state-of-the-art ELMs that rely on more complex encoders and training pipelines: OpenTSLM (Langer et al., 2025), SLIP (Chen et al., 2026), ST-MEM (Na et al., 2024; Han et al., 2025), and MERL (Liu et al., 2024; Han et al., 2025). Across modern ECG-language benchmarks, including ECG-QA-CoT (Langer et al., 2025; Oh et al., 2023) and ECG-Instruct 45K (Zhao et al., 2024), ELF remains consistently competitive and often outperforms these baselines. To better understand the factors underlying ELF's performance, we conduct additional experiments varying the LLM backbone, training duration, number of tokens, and trainable components. We summarize our main contributions below:
1. We introduce **Base ELF**, the simplest member of the ELF family, which directly flattens a 12-lead, 10-second ECG and projects it as a single representative ECG token into the LLM input space alongside the textual query embedding.
2. We introduce **Patch ELF**, which extends Base ELF by partitioning the ECG into non-overlapping temporal patches and projecting each patch as a separate token, allowing us to study whether patch-based tokenization improves performance.
3. Lastly, we introduce **Conv. ELF**, which builds on Patch ELF by adding lightweight temporal convolutions before projection, enabling us to examine whether simple local inductive biases provide additional benefit.

Based on these contributions, we summarize our main findings below:

Summary of Main Findings

- All ELF architectures remain highly competitive with prior ELMs, often outperforming prior baselines despite using substantially simpler architectures and training pipelines. (§4.1)
- Pretrained ECG encoders yield only modest performance gains while requiring substantially more training compute and larger models than ELF. (§4.1)
- Increasing the number of encoder tokens from $N = 50$ to $N = 100$ for both Patch and Conv. ELF does not improve performance, indicating no measurable benefit from a larger token budget. (§4.2)
- Selective training results show that Patch ELF depends primarily on the pretrained LLM backbone rather than the projection layer. (§4.2)
- Patch ELF performance drops substantially under inference-time ECG perturbations, suggesting that coherent ECG inputs are important for strong performance. (§4.3)

Generalizable Insights about Machine Learning in the Context of Healthcare

This study suggests that, in ECG-Language modeling, added architectural and training complexity should not be assumed to improve performance. Across our experiments, ELF remain highly competitive with substantially more complex baselines (e.g., OpenTSLM

(Langer et al., 2025), SLIP (Chen et al., 2026)), while ELMs with pretrained ECG encoders (e.g., ST-MEM (Na et al., 2024), MERL (Liu et al., 2024)) yield only modest improvements despite requiring much greater compute.

2. Related Work

2.1. ECG-Language Models

The automation of ECG analysis is advancing toward generative approaches with ELMs (Han et al., 2025). Previous ECG classification systems are limited to outputting probability scores for a fixed set of diagnostic labels (Qiu et al., 2023a; Hannun et al., 2019; Qiu et al., 2023b; Martin et al., 2021; Strodthoff et al., 2021; Liu et al., 2024; Na et al., 2024; Choi et al., 2023; Jin et al., 2025). In contrast, ELMs can process both text and ECG inputs and generate free-form textual responses conditioned on the given input. Thus, ELMs not only inherit the capabilities of classification systems but also extend them to more general tasks, such as conversational question answering and clinically relevant applications like ECG waveform analysis, patient context inference, and treatment planning. This broad potential motivates the transition from classification-based ECG analysis to generative approaches with ELMs.

Although promising, ELMs remain in their early stages, with most prior works still exploring different ECG representations, architectures, learning objectives, and training regimes (Lan et al., 2025; Zhao et al., 2024; Song et al., 2025). Following the architectural pattern of VLMs, many studies first train a strong ECG-specific encoder, then pair the pretrained encoder with a pretrained LLM for natural language generation (NLG) (Li et al., 2025). However, this approach requires substantial ECG data and computational resources to train an effective encoder. To address this limitation, recent works have proposed more efficient designs. MEIT (Wan et al., 2024), for instance, employs a lightweight ECG encoder composed of several 1-D convolutional layers with batch normalization (Ioffe and Szegedy, 2015), ReLU (Fukushima, 1969), and average pooling, trained end-to-end with the LLM. Another approach explores non-learning-based methods, such as applying Byte-Pair Encoding (BPE) (Gage, 1994) to tokenize the ECG signal and directly concatenate it with text query tokens (Han et al., 2024). We benchmark ELF against several recent ELM methods (Langer et al., 2025; Na et al., 2024; Liu et al., 2024; Han et al., 2025; Chen et al., 2026) and show that it consistently matches or outperforms them. This suggests strong performance on current benchmarks does not necessarily require complex encoders or elaborate training pipelines.

2.2. Encoder-Free Vision-Language Models

Fuyu-8B (Bavishi et al., 2023) is one of the first encoder-free VLMs developed. The authors argue that existing VLMs are overly complex, typically requiring separate image encoders, multi-stage training, and intricate connector modules. Fuyu-8B employs a simple decoder-only transformer where image patches are linearly projected into the first transformer layer. When benchmarked against state-of-the-art VLMs at the time, Fuyu-8B achieved competitive performance. We adopt the same motivation in developing ELF.

Since Fuyu-8B, there have been several advances in encoder-free VLMs (Diao et al., 2024; Team, 2025; Diao et al., 2025). However, we note that the "connector" between the image and the LLM has grown increasingly complex, resembling a vision encoder. This trend undermines the original motivation behind encoder-free VLMs: *simplicity*. To revisit this design principle in ECG-language modeling, we introduce a family of three encoder-free ELMs: Base ELF, Patch ELF, and Conv. ELF. Base ELF is the simplest variant, directly flattening a 12-lead, 10-second ECG and mapping it into the LLM embedding space with a single linear projection. Patch ELF extends this design by splitting the ECG into non-overlapping temporal patches and projecting each patch as a separate input token, enabling us to test whether patch-based tokenization improves performance. Conv. ELF further augments Patch ELF with lightweight temporal convolutions prior to projection, incorporating simple inductive biases to examine whether local feature extraction provides additional benefit. We provide a more comprehensive description of each ELF variant in Section 3.2. We deliberately choose these three designs to test whether controlled increases in architectural complexity yield meaningful gains in ELM performance.

3. Methods

In this section, we describe the datasets, ELMs, and learning objectives used in our study. We also detail the training and evaluation procedures to support reproducibility.

3.1. Datasets

We use preprocessed 12-lead, 10-second ECGs sampled at 250 Hz with 70:30 training/testing splits across two datasets comprising ECGs and text: ECG-Instruct 45K (Zhao et al., 2024) and ECG-QA-CoT (Oh et al., 2023; Langer et al., 2025). These datasets are derived from ECG recordings originating from MIMIC-IV-ECG (Johnson et al., 2023) and PTB-XL (Wagner et al., 2020). Lastly, we provide example instances from ECG-Instruct 45K and ECG-QA-CoT in Figures 12 and 11 of Appendix B.2.

ECG-Instruct 45k ECG-Instruct 45K (Zhao et al., 2024) is an instruction-following conversational dataset generated with GPT-4o (OpenAI et al., 2024). It consists of dialogues between a physician and a patient covering topics such as ECG interpretation, heart rate, waveform morphology, rhythm, and cardiac axis. The dataset is derived from MIMIC-IV-ECG (Johnson et al., 2023). In our study, we use ECG-Instruct 45K to train and evaluate the conversational capabilities of ELMs.

ECG-QA-CoT ECG-QA-CoT (Langer et al., 2025) extends ECG-QA (Oh et al., 2023) by generating Chain-of-Thought (CoT) (Wei et al., 2023) rationales with GPT-4o (OpenAI et al., 2024), conditioned on a plotted ECG image and its corresponding clinical context and question-answer pair. The clinical context consists of non-diagnostic PTB-XL metadata (Wagner et al., 2020), including patient demographics, basic recording information, signal quality notes, and additional technical observations such as extra beats or presence of pacemaker. The clinical context is also used during OpenTSLM's training and evaluation (Langer et al., 2025). In our study, we adopt the same train/test splits used in both OpenTSLM (Langer et al., 2025) and SLIP (Chen et al., 2026) for the ECG-QA-CoT dataset.

3.2. ELF

In this section, we describe each ELM architecture in the ELF family. A high-level visualization of each architecture is shown in Figure 1.

Base ELF Given an ECG signal $X \in \mathbb{R}^{C \times L}$, where C denotes the number of leads and L the signal length, Base ELF flattens the signal into a vector

$$x = \mathrm{vec}(X) \in \mathbb{R}^{CL},$$

and projects it into the LLM hidden space:

$$z = Wx + b, \qquad W \in \mathbb{R}^{d \times CL},\ b \in \mathbb{R}^{d}.$$

Let $E \in \mathbb{R}^{M \times d}$ denote the textual query embeddings, where M is the number of textual input tokens. We replace the embedding of a special placeholder token `<signal>` with z, yielding the combined input embedding sequence $\tilde{E}$. The model is then trained autoregressively to generate the response tokens $y_{1:T}$:

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

where θ denotes the model parameters. When computing $\mathcal{L}_{\mathrm{AR}}$, we mask all tokens except the response and end-of-sequence tokens.

Patch ELF Patch ELF extends Base ELF by partitioning the ECG signal $X \in \mathbb{R}^{C \times L}$ into $N = L/L_p$ non-overlapping patches of length L_p. We denote the patched signal by

$$\mathcal{X}^{(p)} = \{X_i\}_{i=1}^{N}, \qquad X_i \in \mathbb{R}^{C \times L_p}.$$

Each patch X_i is flattened into

$$x_i = \mathrm{vec}(X_i) \in \mathbb{R}^{CL_p},$$

and projected using a shared linear layer:

$$z_i = Wx_i + b, \qquad W \in \mathbb{R}^{d \times CL_p},\ b \in \mathbb{R}^{d}.$$

Collecting the patch embeddings gives

$$Z = [z_1, \ldots, z_N] \in \mathbb{R}^{N \times d}.$$

We then prepare N consecutive placeholder tokens `<signal`$_i$`>` for $i = 1, \ldots, N$, and replace the embeddings of the corresponding placeholder tokens in the textual query embedding sequence E with z_i. Unless otherwise specified, we use $N = 50$. Patch ELF investigates whether representing the ECG as a sequence of fixed-size patches, analogous to patching in VLMs, improves downstream performance.

Conv. ELF Conv. ELF extends Patch ELF by replacing the linear patch projection with a series of convolutional layers. Given the patched ECG signal $\mathcal{X}^{(p)} = \{X_i\}_{i=1}^{N}$, where each patch $X_i \in \mathbb{R}^{C \times L_p}$, Conv. ELF applies two 1D convolutions along the temporal dimension, treating the C leads as input channels. Let

$$h_i = G(X_i),$$

where G denotes a two-layer convolutional module, with a ReLU activation (Fukushima, 1969) applied after each convolution. The resulting feature map is globally average pooled and projected into the LLM hidden space:

$$z_i = W\,\mathrm{Pool}(h_i) + b, \qquad W \in \mathbb{R}^{d \times D_c},\ b \in \mathbb{R}^d,$$

where D_c denotes the channel dimension after convolution and pooling. Collecting the patch embeddings gives $Z = [z_1, \ldots, z_N] \in \mathbb{R}^{N \times d}$, where $N = 50$.

The patch embeddings are then inserted into the textual query embedding sequence E following the same procedure as in Patch ELF. Conv. ELF investigates whether combining patching with lightweight temporal convolutions, and thus introducing simple inductive biases, improves ECG feature extraction and downstream performance.

3.3. Baselines

We provide descriptions of the baselines we compare against ELF. We utilize four strong baselines in this study: OpenTSLM (Langer et al., 2025), SLIP (Chen et al., 2026), ST-MEM (Na et al., 2024), and MERL (Liu et al., 2024).

OpenTSLM OpenTSLM (Langer et al., 2025) is a general, time-series language model (TSLM) with two variants: OpenTSLM SoftPrompt (OpenTSLM SP) and OpenTSLM Flamingo (OpenTSLM FL). OpenTSLM SP converts the signal into soft prompt tokens interleaved throughout the input sequence, whereas OpenTSLM FL encodes the signal separately and integrates it through gated cross-attention. Both variants are trained in two stages with a mixture of data: the first learns simple question-answering from paired time-series input, and the second introduces CoT rationales, which includes the ECG-QA-CoT dataset.

SLIP SLIP (Chen et al., 2026) is a general, sensor-language model designed to learn transferable, language-aligned representations from multivariate time-series data. It extends the CoCa framework (Yu et al., 2022) to the sensor domain with a Transformer-based sensor encoder, attention pooling module, text encoder, and multimodal decoder. The model is pretrained with a CLIP-style contrastive loss and an autoregressive captioning loss on over 600K paired sensor-text examples from diverse domains, then finetuned for question answering on ECG-QA-CoT using the same train/test splits as OpenTSLM (Langer et al., 2025), denoted as $\mathrm{SLIP}_{\mathrm{SFT}}$ (Chen et al., 2026). In our study, we exclusively compare with $\mathrm{SLIP}_{\mathrm{SFT}}$ and treat it as a strong finetuned sensor baseline.

ST-MEM and MERL as ELMs We convert both ST-MEM (Na et al., 2024) and MERL (Liu et al., 2024) into ELMs through a LLaVA-like architecture (Liu et al., 2023). ST-MEM is a masked autoencoder (MAE) (He et al., 2021) built on a Vision Transformer (ViT)

backbone (Dosovitskiy et al., 2021), while MERL is a multimodal ECG-text representation model trained to align ECGs and reports through contrastive learning. For MERL, we use its strongest backbone, based on the ResNet101 architecture (He et al., 2015). The output of each encoder is projected into the LLM embedding space through a linear layer and inserted at the placeholder tokens `<signal$_i$>` for $i = 1, \ldots, N$. The resulting model is then trained with the same autoregressive objective $\mathcal{L}_{\text{AR}}$. During this stage, the encoder is frozen, and only the projection layer and LLM backbone are trained. For both models, a 1D adaptive average pooling layer controls the number of encoder tokens N, where $N = 50$ unless otherwise noted. We provide a detailed table of the hyperparameters for training ST-MEM and MERL encoders in Table 1 of Appendix A. Lastly, we refer to these resulting ELMs as ST-MEM and MERL throughout the rest of the study.

3.4. Large Language Model Backbones

For ELF, we experiment with the Llama-3.2-1B-Instruct[1], Qwen2.5-1.5B-Instruct[2], and gemma-2-2b-it[3] checkpoints (Grattafiori et al., 2024; Qwen et al., 2025; Gemma Team et al., 2024) accessed via the HuggingFace API (Wolf et al., 2020). ST-MEM and MERL ELM baselines use only the Llama-3.2-1B-Instruct LLM backbone. All LLMs are initialized with their default hyperparameters.

We use the Muon (Liu et al., 2025) optimizer with learning rate 1e-3, weight decay 5e-2, $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 1e-8$. We employ LoRA (Hu et al., 2021) with rank 16, $\alpha_{LoRA} = 32$, and dropout 0.05. We use a batch size of 4 with 10 second 12-lead ECGs, max input sequence length of 2048, and warm-up for 10% of total steps across 10 epochs. We provide a detailed table of all hyperparameters used during ELM training in Table 1 of Appendix A.

3.5. Evaluation

We evaluate all models over three random seeds, reporting performance on the following text generation metrics: BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and BERTScore F1 (Zhang et al., 2020). We also report F1 and accuracy. We define accuracy as the proportion of generated responses that exactly match the ground truth (i.e., exact string matching). We define F1 using token-level overlap after lowercasing, removing punctuation, and normalizing whitespace. For each ground truth–prediction pair, we compute precision and recall from the overlapping tokens and take their harmonic mean. The final F1 score is the average of these per-sample token-level F1 values across the dataset. All metrics are in percentage form.

4. Results

In this section, we present results on the ECG-QA-CoT and ECG-Instruct 45K datasets and report several ablation studies.

1. `meta-llama/Llama-3.2-1B-Instruct`
2. `Qwen/Qwen2.5-1.5B-Instruct`
3. `google/gemma-2-2b-it`

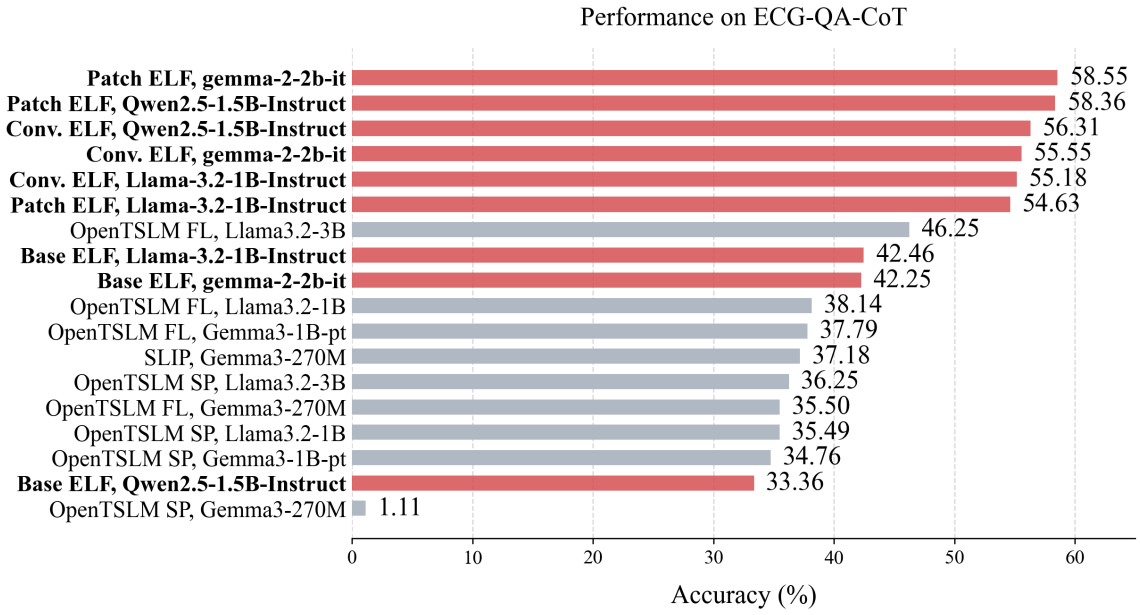

Figure 2: Comparing performance on ECG-QA-CoT between various baselines (e.g., OpenTSLM (Langer et al., 2025) and SLIP (Chen et al., 2026)) and ELF. ELMs from the ELF family are colored in red. A tabular summary of these results, including additional metrics and standard deviations, is provided in Table 3 of Appendix B.

4.1. Baselines on ECG-QA-CoT and ECG-Instruct 45K

ECG-QA-CoT We first compare the accuracies of all methods on ECG-QA-CoT in Figure 2. While the figure highlights accuracy only, the full results, including BLEU-4, ROUGE-L, METEOR, and BERTScore F1, are provided in Table 3 in Appendix B. Overall, the ELF family is highly competitive, with Base, Patch, and Conv. ELF occupying eight of the top nine positions against prior ELM baselines such as OpenTSLM (Langer et al., 2025) and SLIP (Chen et al., 2026). Patch ELF with gemma-2-2b-it is the strongest model, achieving 58.55% accuracy. It substantially outperforms the best non-ELF baseline, OpenTSLM FL with Llama3.2-3B, despite the latter using a more complex architecture and training strategy (i.e., curriculum learning), and a larger pretrained backbone. Patch ELF with the Qwen2.5-1.5B-Instruct backbone follows closely at 58.36%. In contrast, Base ELF with Qwen2.5-1.5B-Instruct performs worst among the ELF variants at 33.36% accuracy. Notably, despite being the simplest variant of ELF, Base ELF with the Llama-3.2-1B-Instruct and gemma-2-2b-it backbones still outperforms SLIP and all OpenTSLM variants except OpenTSLM FL with Llama3.2-3B. Nonetheless, Patch ELF and Conv. ELF consistently outperform Base ELF, suggesting that patching and lightweight convolutional layers are beneficial on ECG-QA-CoT. Taken together, these results show that simple encoder-free designs can remain highly competitive with, and often outperform, more complex prior ELM pipelines.

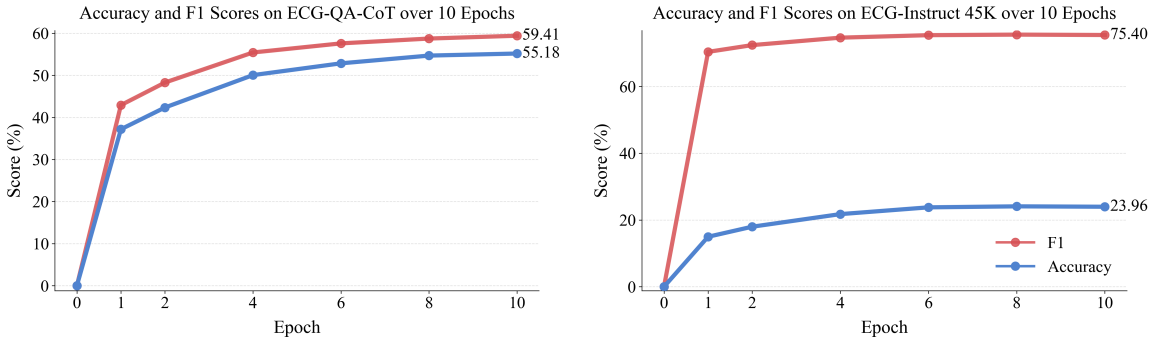

Figure 3: We plot the accuracy (Blue) and F1 (Red) scores on the test set of ECG-QA-CoT (**Left**) and ECG-Instruct 45K (**Right**) over 10 training epochs for Conv. ELF.

Findings 1 ELF occupies 8 of the top 9 positions on ECG-QA-CoT despite having simpler architectures and training procedures than baselines.

ECG-Instruct 45K Figure 4 and Table 4 (see Appendix B) compare ELF variants against ST-MEM and MERL on ECG-Instruct 45K. All ELM use the Llama-3.2-1B-Instruct LLM backbone. MERL achieves the highest accuracy (26.53%), with ST-MEM close behind. Conv. ELF narrows this gap considerably, reaching 23.96% accuracy, while Base ELF and Patch ELF trail at roughly 20% accuracy. Interestingly, while Base ELF consistently underperforms Patch ELF and Conv. ELF on ECG-QA-CoT, it outperforms Patch ELF on ECG-Instruct 45K, suggesting that the effectiveness of ELM design choices is dataset- and task-dependent. We note that accuracy is uniformly low on this dataset because ECG-Instruct 45K is conversational and substantially more free-form than ECG-QA-CoT. We therefore emphasize the F1 score in Figure 3, together with the additional metrics (i.e., BLEU-4, ROUGE-L, METEOR, BERTScore F1) reported in Table 4, as a more complete view of model performance.

Effect of Training Epochs We evaluate Conv. ELF with the Llama-3.2-1B-Instruct LLM backbone across selected epochs on the test sets of ECG-QA-CoT and ECG-Instruct 45K, shown in the left and right plots of Figure 3, respectively. On both datasets, performance improves rapidly in the early epochs before beginning to plateau, indicating that Conv. ELF should be trained close to convergence for a fair assessment. On ECG-QA-CoT, both F1 and accuracy increase from epoch 1 to epoch 10, starting at 37.16% accuracy after epoch 1 to 55.18% accuracy at epoch 10. On ECG-Instruct 45K, performance also improves early, with smaller changes after epoch 4. The highest scores are observed at epoch 8, with 75.51% F1 and 24.08% accuracy, followed by a slight decrease at epoch 10.

Training Efficiency Figure 4 reveals that this performance gap comes at a steep efficiency cost. In terms of total training time, all three ELF variants train in under 4 hours, whereas ST-MEM and MERL require 10-12 hours. This 3-7× increase in total training time is attributed largely to the time it takes to pretrain the ST-MEM (Na et al., 2024)

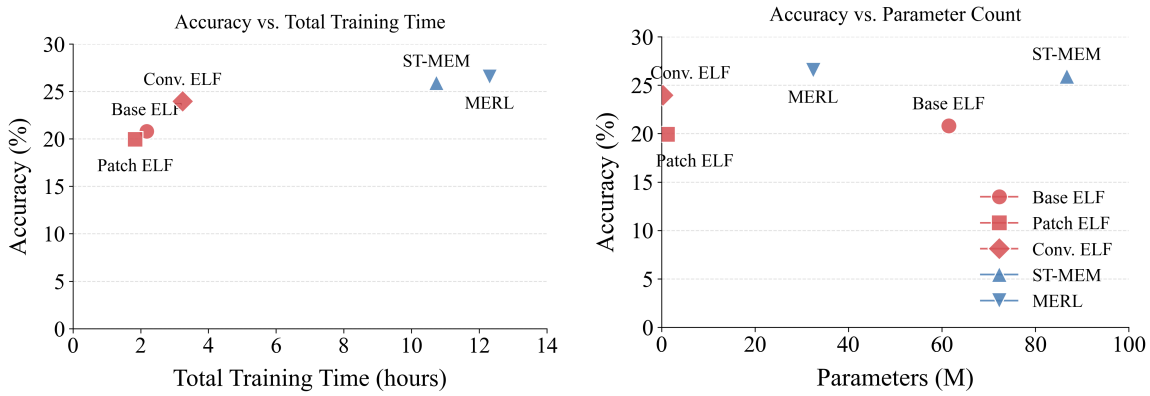

Figure 4: Accuracy vs. total training time (**Left**) and accuracy vs. parameter count (**Right**) for ELF variants (Red) and baseline ELMs (Blue) on ECG-Instruct 45K. Parameter counts only include non-LLM components.

and MERL (Liu et al., 2024) encoders. To view the total training times broken down for each ELM, please see Table 2 in the Appendix. In terms of parameter count, the Patch and Conv. ELF convolution and projection layers contain fewer than 1M trainable parameters, compared to roughly 30M for MERL and 85M for ST-MEM. Conv. ELF remains well below 5M parameters while recovering the majority of the accuracy gap with the pretrained encoders. These results suggest that pretrained ECG encoders offer diminishing returns relative to their computational overhead, and that lightweight projections can achieve competitive performance at a fraction of the compute.

Findings 2 On ECG-Instruct 45K, ELMs with pretrained ECG encoders yield only a modest accuracy gain over ELF while incurring 3-7× higher training cost and generally larger model footprints.

4.2. Ablation Study

Number of Tokens N In the left plot of Figure 5, we train and evaluate Patch ELF and Conv. ELF with the Llama-3.2-1B-Instruct backbone on ECG-QA-CoT using $N = 100$, and compare accuracy to the default $N = 50$ setting. Increasing the number of tokens does not improve performance, despite offering greater representational capacity. For additional metrics, please view Table 9 in the Appendix.

Findings 3 On ECG-QA-CoT, increasing the number of encoder tokens from $N = 50$ to $N = 100$ does not improve performance for either Patch ELF or Conv. ELF. In this setting, a larger token budget provides no measurable benefit over $N = 50$.

Selectively Training Patch ELF Components We study the effect of selectively training Patch ELF components on the ECG-QA-CoT and ECG-Instruct 45K datasets in the

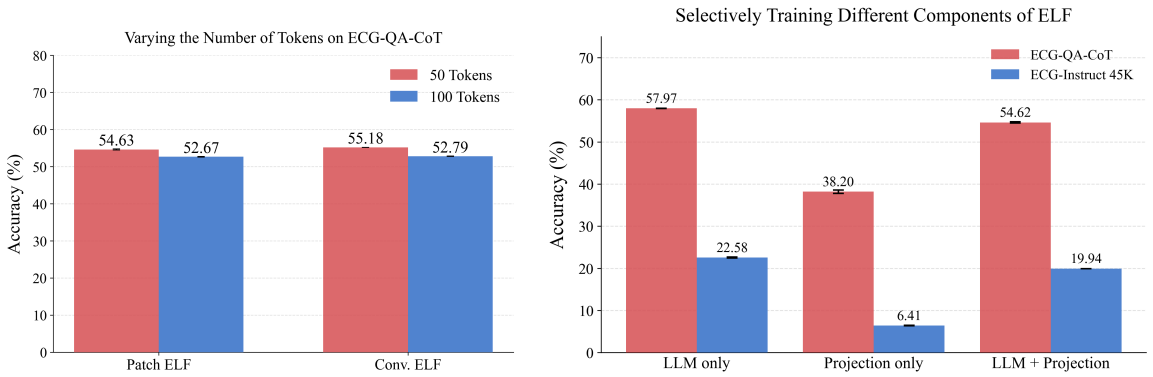

Figure 5: (**Left**) Observing the performances on ECG-QA-CoT when varying the number of tokens N for Patch and Conv. ELF. (**Right**) Selectively training Patch ELF's components on ECG-QA-CoT and ECG-Instruct 45K. LLM only, Projection only, and LLM + Projection indicate that the given component was trained while all others were frozen.

right plot of Figure 5. LLM only, Projection only, and LLM + Projection indicate that the given component was trained while all others were frozen. We use the Llama-3.2-1B-Instruct LLM backbone. Training only the LLM backbone while freezing the projection layer achieves the highest accuracy and consistently outperforms across most metrics (see Tables 7 and 8 in Appendix). Conversely, training only the projection layer while freezing the LLM yields the lowest accuracy. Training both components together slightly underperforms the LLM-only setting. We also repeat this experiment with the Gemma-2-2B-it backbone and observe similar trends, as shown in Table 7. These results suggest downstream performance in ELMs on both the ECG-QA-CoT and ECG-Instruct 45K datasets is predominantly driven by the LLM backbone.

Findings 4 Selective training results on ECG-QA-CoT and ECG-Instruct 45K show that Patch ELF performance is driven primarily by the pretrained LLM backbone, with the LLM-only setting achieving the strongest performance across two LLM backbones.

4.3. ELF's Performance Under Perturbations

We evaluate Patch ELF with the gemma-2-2b-it backbone under three inference-time perturbations on the ECG-QA-CoT dataset in Figure 6. The model is trained on normal ECG signals from ECG-QA-CoT, with perturbations applied only during inference. The three conditions are defined as follows:

Noise: A tensor of the same shape as the regular ECG signal filled with values sampled from a Gaussian distribution is used as input, alongside the textual query.

Zeros Tensor: A tensor of the same shape as the regular ECG signal filled with zeros is used as input, alongside the textual query.

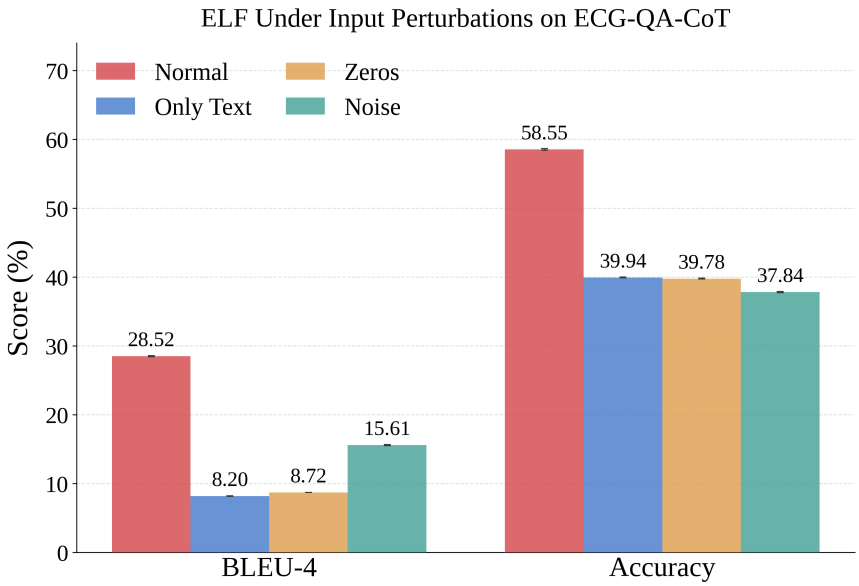

Figure 6: Patch ELF performance under three input perturbation settings on the ECG-QA-CoT dataset, using Gemma-2-2B-it as the LLM backbone.

Only Text: The ECG signal is completely omitted and only the textual query is used as input.
We note only the input is modified for each perturbation, and the architecture of ELF remains the same.

As shown in Figure 6, all perturbations degrade the performance of Patch ELF with the gemma-2-2b-it backbone on ECG-QA-CoT, with the largest drop observed for BLEU-4. While this suggests that ELF requires coherent ECG signal inputs to perform well on ECG-QA-CoT, these perturbation results alone do not establish ECG signal understanding and motivate further analyses of how ELF uses ECG information. Additional metrics are reported in Table 11 in the Appendix.

Findings 5 Patch ELF performance drops substantially under inference-time ECG perturbations on ECG-QA-CoT, suggesting that coherent ECG inputs are important for strong performance.

5. Discussion

As automated ECG analysis increasingly shifts from classification-based approaches to generative ones with ELMs, we introduce ELF, a family of three encoder-free ELM architectures in which each variant builds naturally on the previous one. We first introduce Base ELF, the simplest variant, which directly flattens the ECG signal and projects it with a linear layer into the LLM input space alongside the textual query. Despite its simplicity, Base ELF outperforms strong baselines such as OpenTSLM (Langer et al., 2025) and SLIP (Chen

et al., 2026) on ECG-QA-CoT. We then introduce Patch ELF and Conv. ELF. Patch ELF extends Base ELF by partitioning the ECG signal into non-overlapping patches and projecting each patch separately, while Conv. ELF further extends Patch ELF by adding temporal 1D convolutions before projection. Patch ELF and Conv. ELF also outperform all strong baselines on ECG-QA-CoT. Overall, ELF occupies eight of the top nine positions on this dataset despite using substantially simpler architectures and training procedures than prior baselines.

We also train and evaluate ELF on ECG-Instruct 45K alongside strong ELM baselines that use pretrained ECG encoders, namely ST-MEM (Na et al., 2024) and MERL (Liu et al., 2024). Although ST-MEM and MERL achieve higher performance than ELF on this dataset, they do so at a substantial computational cost, requiring 3–5× more training compute due to extensive pretraining on MIMIC-IV-ECG. In contrast, ELF requires no ECG encoder pretraining and only simple projection or lightweight convolutional layers, making it an attractive option in resource-constrained settings.

Across both datasets, we also find that added complexity within the ELF family does not consistently translate to better performance. On ECG-QA-CoT, Patch ELF, despite being architecturally simpler than Conv. ELF, often performs better. On ECG-Instruct 45K, Base ELF is able to outperform Patch ELF. Taken together, these results, along with ELF outperforming many strong baselines such as OpenTSLM and SLIP, suggest that simple encoder-free designs may already be sufficient to perform well on current ECG-language benchmarks.

We further examine the role of complexity by increasing the number of tokens N in Patch ELF and Conv. ELF from 50 to 100. On ECG-QA-CoT, this leads to a slight degradation in performance, indicating that a larger token budget does not necessarily improve results.

Lastly, we investigate what may be driving ELF's strong performance. By selectively training and freezing different components of Patch ELF on ECG-QA-CoT, we find that training only the LLM backbone while freezing the projection layer yields better performance than jointly training both the projection layer and the LLM, which is our default setting. This suggests that performance on ECG-QA-CoT may depend largely on the pretrained LLM backbone. We leave this direction for further study.

Overall, we hope this work contributes to the ELM community by introducing three new encoder-free ELM architectures and by providing further insight into what is, and may not be, necessary for strong performance on current ECG-language benchmarks.

Limitations Our study has several limitations. First, we evaluate ELF on two datasets, ECG-Instruct 45K and ECG-QA-CoT, which are derived from MIMIC-IV-ECG and PTB-XL, respectively. Evaluating on additional ECG-language datasets from diverse clinical populations would strengthen the generalizability of our findings. Second, all experiments use 12-lead, 10-second ECGs sampled at 250 Hz. Performance on recordings with different lead configurations, durations, or sampling rates is not examined. Third, we rely on standard text generation metrics (BLEU-4, ROUGE-L, METEOR, BERTScore F1) alongside token-level F1 and exact-match accuracy. These metrics measure surface-level textual similarity and do not directly assess clinical correctness. Lastly, we do not conduct any clinical validation or evaluate ELF in a real-world deployment setting.

Acknowledgments

This work is done in collaboration with the Mario Lemieux Center for Heart Rhythm Care at Allegheny General Hospital.

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, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*, 2005.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, and Sagnak Tasirlar. Fuyu-8b: A multimodal architecture for ai agents, 10 2023. URL https://www.adept.ai/blog/fuyu-8b.

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.

Seokmin Choi, Sajad Mousavi, Phillip Si, Haben G. Yhdego, Fatemeh Khadem, and Fatemeh Afghah. Ecgbert: Understanding hidden language of ecgs with self-supervised representation learning, 2023.

Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models, 2024. URL https://arxiv.org/abs/2406.11832.

Haiwen Diao, Xiaotong Li, Yufeng Cui, Yueze Wang, Haoge Deng, Ting Pan, Wenxuan Wang, Huchuan Lu, and Xinlong Wang. Evev2: Improved baselines for encoder-free vision-language models, 2025. URL https://arxiv.org/abs/2502.06788.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Kunihiko Fukushima. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5:322–333, 1969. doi: 10.1109/tssc.1969.300225.

Philip Gage. A new algorithm for data compression. *The C Users Journal archive*, 12: 23–38, 1994. URL https://api.semanticscholar.org/CorpusID:59804030.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

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 Poe-

naru, 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, Chaojing Duan, Michael A. Rosenberg, Emerson Liu, and Ding Zhao. Ecg-byte: A tokenizer for end-to-end generative electrocardiogram language modeling, 2024. URL https://arxiv.org/abs/2412.14373.

William Han, Chaojing Duan, Zhepeng Cen, Yihang Yao, Xiaoyu Song, Atharva Mhaskar, Dylan Leong, Michael A. Rosenberg, Emerson Liu, and Ding Zhao. Signal, image, or symbolic: Exploring the best input representation for electrocardiogram-language models through a unified framework, 2025. URL https://arxiv.org/abs/2505.18847.

Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25:65–69, 01 2019. doi: 10.1038/s41591-018-0268-3. URL https://www.nature.com/articles/s41591-018-0268-3.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. URL https://arxiv.org/abs/1502.03167.

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.

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.

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, 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, 2025. URL https://arxiv.org/abs/2510.02410.

Haitao Li, Ziyu Li, Yiheng Mao, Ziyi Liu, Zhoujian Sun, and Zhengxing Huang. anyecg-chat: A generalist ecg-mllm for flexible ecg input and multi-task understanding, 2025. URL https://arxiv.org/abs/2506.00942.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*, 2004.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement, 2024. URL https://arxiv.org/abs/2403.06659.

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.

Harold Martin, Ulyana Morar, Walter Izquierdo, Mercedes Cabrerizo, Anastasio Cabrera, and Malek Adjouadi. Real-time frequency-independent single-lead and single-beat myocardial infarction detection. *Artificial intelligence in medicine*, 121:102179, 2021.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram, 2024. URL https://arxiv.org/abs/2402.09450.

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.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela

Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett,

Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

Jielin Qiu, William Han, Jiacheng Zhu, Mengdi Xu, Michael Rosenberg, Emerson Liu, Douglas Weber, and Ding Zhao. Transfer knowledge from natural language to electrocardiography: Can we detect cardiovascular disease through language models? In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 442–453, Dubrovnik, Croatia, May 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.33. URL https://aclanthology.org/2023.findings-eacl.33.

Jielin Qiu, Jiacheng Zhu, Shiqi Liu, William Han, Jingqi Zhang, Chaojing Duan, Michael A. Rosenberg, Emerson Liu, Douglas Weber, and Ding Zhao. Automated cardiovascular record retrieval by multimodal learning between electrocardiogram and clinical report. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh, editors, *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 480–497. PMLR, 10 Dec 2023b. URL https://proceedings.mlr.press/v225/qiu23a.html.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang

Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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.

Xiaoyu Song, William Han, Tony Chen, Chaojing Duan, Michael A. Rosenberg, Emerson Liu, and Ding Zhao. Retrieval-augmented generation for electrocardiogram-language models, 2025. URL https://arxiv.org/abs/2510.00261.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE Journal of Biomedical and Health Informatics*, 25:1519–1528, 2021.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL https://arxiv.org/abs/2405.09818.

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.

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, 2024. 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. URL https://arxiv.org/abs/2205.01917.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675, 2020.

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: Training hyperparameters used in this study. Left: ELF, ST-MEM, and MERL ELM training. Right: pretraining ST-MEM (Na et al., 2024) and MERL (Liu et al., 2024) encoders. All trainings were conducted on 4×H100 NVL NVIDIA GPUs.

ELM Training Hyperparameters

Hyperparameter	Value
Optimizer	Muon
Learning rate	1×10^{-3}
LR schedule	Cosine
Minimum LR ratio	0.1
Weight decay	0.05
β_1	0.9
β_2	0.95
ϵ	1×10^{-8}
Gradient clipping	1.0
Batch size	4
Reference global batch size	16
Gradient accumulation steps	1
Epochs	10
Patience	5
Patience delta	0.1
LoRA rank	16
LoRA alpha	32
LoRA dropout	0.05
LLM input length	2048

Encoder Pretraining Hyperparameters

Hyperparameter	Value
Optimizer	AdamW
Learning rate	3×10^{-4}
LR schedule	Cosine
Minimum LR ratio	0.1
Weight decay	0.01
β_1	0.9
β_2	0.95
ϵ	1×10^{-8}
Gradient clipping	1.0
Batch size	64
Reference global batch size	256
Gradient accumulation steps	1
Epochs	100
Patience	5
Patience delta	0.1

Table 2: Training time breakdowns for ELF, ST-MEM, and MERL. The MIMIC-IV-ECG dataset was used during the encoder training stage, while the ECG-Instruct 45K dataset was used for the ELM training stage.

Model	Stage	Training Time
ST-MEM	Encoder	6h 45m
	ELM	3h 59m 21s
MERL	Encoder	8h 35m
	ELM	3h 43m 29s
Base ELF	ELM	2h 10m 40s
Patch ELF	ELM	1h 50m 20s
Conv. ELF	ELM	3h 14m 28s

Appendix A. Hyperparameters

Table 1 shows the hyperparameters used for training the encoders ST-MEM (Na et al., 2024) and MERL (Liu et al., 2024), as well as the ELMs for ELF, ST-MEM, and MERL.

Table 3: Comparing performance on ECG-QA-CoT between various baselines and ELF.

ELM	LLM Backbone	BLEU-4	ROUGE-L	METEOR	BERTScore F1	F1	Accuracy
OpenTSLM SoftPrompt Langer et al. (2025)	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 Langer et al. (2025)	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 Chen et al. (2026)	Gemma3-270M	-	-	-	-	-	37.18
Base ELF	Llama-3.2-1B-Instruct	12.74 ± 0.06	47.06 ± 0.13	29.75 ± 0.12	91.94 ± 0.03	46.79 ± 0.13	42.46 ± 0.16
	gemma-2-2b-it	11.85 ± 0.02	46.58 ± 0.11	29.30 ± 0.09	91.90 ± 0.02	46.36 ± 0.10	42.25 ± 0.13
	Qwen2.5-1.5B-Instruct	15.75 ± 0.02	39.92 ± 0.02	29.00 ± 0.04	90.42 ± 0.01	39.87 ± 0.01	33.36 ± 0.08
Patch ELF	Llama-3.2-1B-Instruct	27.49 ± 0.12	59.21 ± 0.10	40.93 ± 0.11	93.93 ± 0.01	58.78 ± 0.10	54.63 ± 0.16
	gemma-2-2b-it	**28.52 ± 0.08**	**62.17 ± 0.04**	**42.83 ± 0.07**	**94.34 ± 0.01**	**61.83 ± 0.05**	**58.55 ± 0.11**
	Qwen2.5-1.5B-Instruct	27.07 ± 0.00	62.10 ± 0.01	42.65 ± 0.01	94.31 ± 0.00	61.75 ± 0.01	58.36 ± 0.02
Conv. ELF	Llama-3.2-1B-Instruct	27.95 ± 0.12	59.85 ± 0.04	41.50 ± 0.07	94.02 ± 0.01	59.41 ± 0.03	55.18 ± 0.02
	gemma-2-2b-it	26.06 ± 0.10	59.56 ± 0.03	40.92 ± 0.05	93.94 ± 0.01	59.22 ± 0.02	55.55 ± 0.01
	Qwen2.5-1.5B-Instruct	24.75 ± 0.03	60.28 ± 0.00	41.16 ± 0.01	94.00 ± 0.01	59.93 ± 0.00	56.31 ± 0.00

Table 4: Comparing performance of ELF on ECG-Instruct 45K against ELMs with strong encoders.

ELM	LLM Backbone	BLEU-4	ROUGE-L	METEOR	BERTScore F1	F1	Accuracy
MERL Liu et al. (2024)	Llama-3.2-1B-Instruct	**39.17 ± 0.03**	**75.85 ± 0.02**	**73.16 ± 0.03**	**96.91 ± 0.00**	**78.06 ± 0.03**	**26.53 ± 0.09**
ST-MEM Na et al. (2024)	Llama-3.2-1B-Instruct	37.02 ± 0.01	74.92 ± 0.02	72.09 ± 0.03	96.74 ± 0.00	77.13 ± 0.02	25.91 ± 0.09
Base ELF	Llama-3.2-1B-Instruct	30.14 ± 0.07	70.97 ± 0.06	67.69 ± 0.08	96.09 ± 0.01	73.18 ± 0.06	20.80 ± 0.10
Patch ELF	Llama-3.2-1B-Instruct	30.50 ± 0.07	71.14 ± 0.08	67.72 ± 0.07	96.16 ± 0.01	73.33 ± 0.06	19.94 ± 0.01
Conv. ELF	Llama-3.2-1B-Instruct	33.60 ± 0.01	73.18 ± 0.01	70.20 ± 0.00	96.43 ± 0.00	75.40 ± 0.02	23.96 ± 0.07

Table 5: We evaluate Conv. ELF on the test set of ECG-QA-CoT on selected epochs.

Epoch	BLEU-4	ROUGE-L	METEOR	BERTScore F1	F1	Accuracy
1	14.91 ± 0.20	43.25 ± 0.01	28.88 ± 0.01	91.46 ± 0.01	42.92 ± 0.02	37.16 ± 0.00
2	20.42 ± 0.21	48.98 ± 0.05	33.62 ± 0.07	92.59 ± 0.02	48.28 ± 0.08	42.35 ± 0.03
4	27.27 ± 0.05	55.91 ± 0.22	39.20 ± 0.20	93.56 ± 0.03	55.43 ± 0.23	50.05 ± 0.23
6	**29.67 ± 0.28**	57.98 ± 0.40	41.08 ± 0.31	93.85 ± 0.05	57.59 ± 0.39	52.85 ± 0.45
8	27.77 ± 0.05	59.13 ± 0.04	40.75 ± 0.01	93.94 ± 0.00	58.74 ± 0.04	54.69 ± 0.07
10	27.95 ± 0.12	**59.85 ± 0.04**	**41.50 ± 0.07**	**94.02 ± 0.01**	**59.41 ± 0.03**	**55.18 ± 0.02**

Table 6: We evaluate Conv. ELF on the test set of ECG-Instruct 45K on selected epochs.

Epoch	BLEU-4	ROUGE-L	METEOR	BERTScore F1	F1	Accuracy
1	26.69 ± 0.01	68.11 ± 0.00	64.20 ± 0.01	95.71 ± 0.00	70.35 ± 0.01	14.97 ± 0.00
2	29.26 ± 0.03	70.14 ± 0.01	66.53 ± 0.01	95.98 ± 0.00	72.37 ± 0.00	17.98 ± 0.18
4	32.94 ± 0.11	72.41 ± 0.06	69.37 ± 0.08	96.38 ± 0.01	74.68 ± 0.07	21.75 ± 0.00
6	33.57 ± 0.03	73.11 ± 0.00	70.15 ± 0.03	96.44 ± 0.00	75.35 ± 0.01	23.78 ± 0.06
8	**33.83 ± 0.02**	**73.27 ± 0.00**	**70.37 ± 0.03**	**96.45 ± 0.01**	**75.51 ± 0.02**	**24.08 ± 0.14**
10	33.60 ± 0.01	73.18 ± 0.01	70.20 ± 0.00	96.43 ± 0.00	75.40 ± 0.02	23.96 ± 0.07

Table 7: We selectively train Patch ELF's components and report mean ± standard deviation over three seeds on ECG-QA-CoT across two LLM backbones. ✓indicates the corresponding component was set as trainable.

LLM Backbone	Components		BLEU-4	ROUGE-L	METEOR	F1	Accuracy
	LLM	Projection Layer W					
Llama-3.2-1B-Instruct	✓		26.81 ± 0.05	**61.83 ± 0.01**	**42.47 ± 0.03**	**61.49 ± 0.01**	**57.97 ± 0.02**
		✓	13.94 ± 0.05	43.99 ± 0.43	28.77 ± 0.22	43.54 ± 0.43	38.20 ± 0.41
	✓	✓	**27.49 ± 0.12**	59.21 ± 0.10	40.93 ± 0.11	58.78 ± 0.10	54.63 ± 0.16
gemma-2-2b-it	✓		**30.10 ± 0.03**	**62.98 ± 0.02**	**43.81 ± 0.05**	**62.64 ± 0.10**	**59.15 ± 0.07**
		✓	18.31 ± 0.06	49.00 ± 0.03	33.01 ± 0.08	48.60 ± 0.05	42.36 ± 0.05
	✓	✓	28.52 ± 0.08	62.17 ± 0.04	42.83 ± 0.07	61.83 ± 0.05	58.55 ± 0.11

Table 8: We selectively train Patch ELF's components and report mean ± standard deviation over three seeds on ECG-Instruct 45K using Llama-3.2-1B-Instruct. ✓indicates the corresponding component was set as trainable.

Components		BLEU-4	ROUGE-L	METEOR	F1	Accuracy
LLM	Projection Layer W					
✓		**31.76 ± 0.02**	**72.24 ± 0.01**	**68.85 ± 0.01**	**74.38 ± 0.01**	**22.58 ± 0.11**
	✓	19.53 ± 0.03	59.16 ± 0.05	54.95 ± 0.05	61.47 ± 0.04	6.41 ± 0.07
✓	✓	30.50 ± 0.07	71.14 ± 0.08	67.72 ± 0.07	73.33 ± 0.06	19.94 ± 0.01

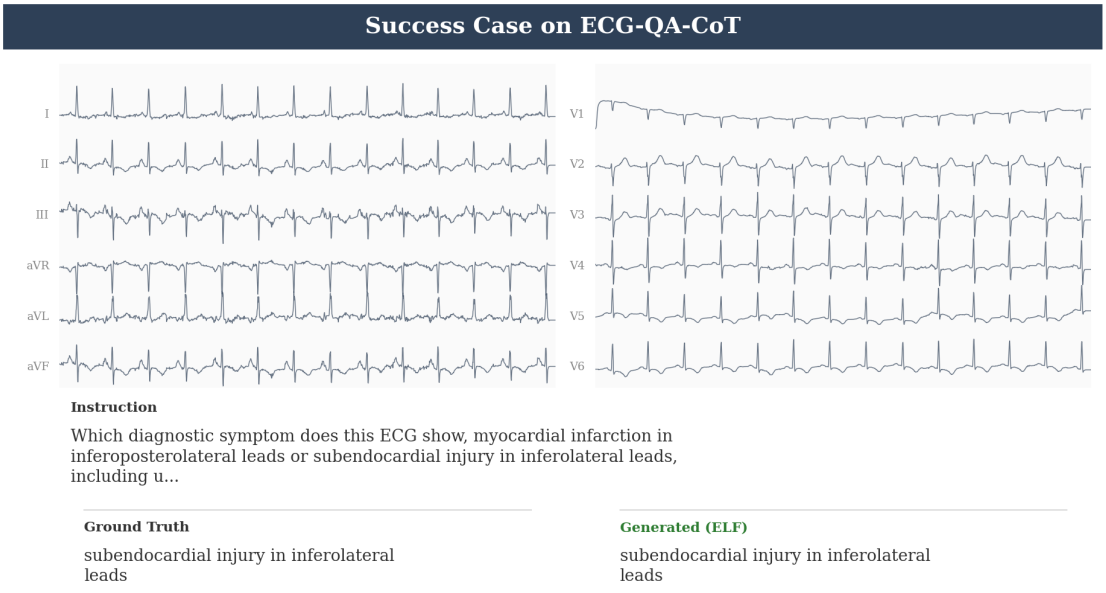

Figure 7: A successful generation with Conv. ELF on the ECG-QA-CoT dataset

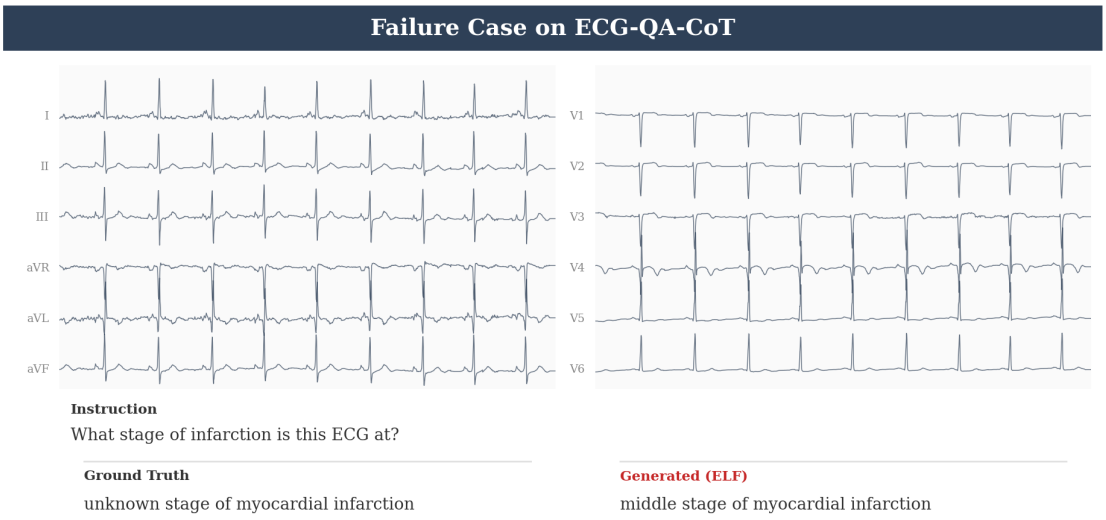

Figure 8: Failed generation with Conv. ELF on the ECG-QA-CoT dataset

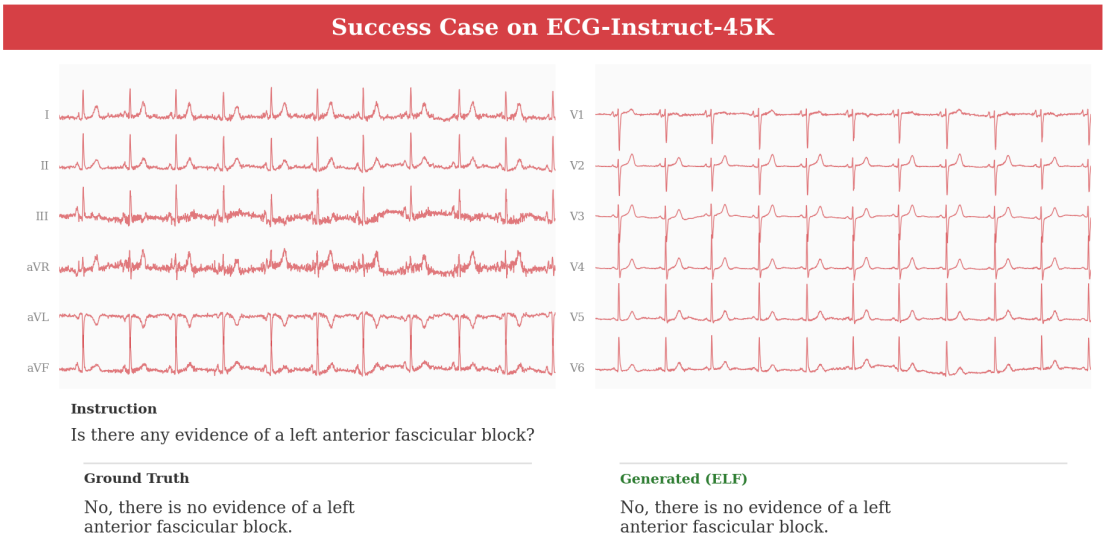

Figure 9: A successful generation with Conv. ELF on the ECG-Instruct 45K dataset

Table 9: ECG-QA-CoT results when varying the number of encoder tokens $N = \{50, 100\}$ for Patch ELF and Conv. ELF.

Model	Tokens	BLEU-4	ROUGE-L	METEOR	BERTScore	F1	Accuracy
Patch ELF	50	**27.49 ± 0.12**	**59.21 ± 0.10**	**40.93 ± 0.11**	**93.93 ± 0.01**	**58.78 ± 0.10**	**54.63 ± 0.16**
	100	25.88 ± 0.02	57.06 ± 0.07	39.04 ± 0.09	93.59 ± 0.01	56.66 ± 0.08	52.67 ± 0.09
Conv. ELF	50	**27.95 ± 0.12**	**59.85 ± 0.04**	**41.50 ± 0.07**	**94.02 ± 0.01**	**59.41 ± 0.03**	**55.18 ± 0.02**
	100	21.67 ± 0.07	56.93 ± 0.05	37.83 ± 0.04	93.47 ± 0.01	56.64 ± 0.06	52.79 ± 0.04

Table 10: Performance under different learning rates.

Learning rate	BLEU-4	ROUGE-L	METEOR	F1	Accuracy
1e-3	**28.52 ± 0.08**	**62.17 ± 0.04**	**42.83 ± 0.07**	**61.83 ± 0.05**	**58.55 ± 0.11**
1e-4	21.02 ± 0.05	45.96 ± 0.04	33.70 ± 0.10	45.89 ± 0.06	40.32 ± 0.09
1e-5	21.24 ± 0.07	46.84 ± 0.01	34.20 ± 0.09	46.42 ± 0.04	40.41 ± 0.10

Appendix B. Additional Results

We provide the tabular form with additional metrics for each figure in the main paper. Specifically, Table 3 corresponds to Figure 2, Table 4 corresponds to Figure 4, Tables 5 and 6 correspond to Figure 3, Tables 7, 8, and 9 correspond to Figure 5, and Table 11 corresponds to Figure 6.

B.1. Learning Rate

In Table 10, we present additional experiments varying the learning rate when using Patch ELF, gemma-2-2b-it on the ECG-QA-CoT dataset. We observe our original learning rate (1e-3) consistently outperforms 1e-4 and 1e-5 across all metrics.

B.2. Dataset Instance Visualizations

We visualize examples of instances from the ECG-QA-CoT and ECG-Instruct 45K datasets in Figure 11 and 12 respectively.

Table 11: ELF's performance under different input perturbation settings on the ECG-QA-CoT dataset

Perturbation	BLEU-4	ROUGE-L	METEOR	F1	Accuracy
Normal	**28.52 ± 0.08**	**62.17 ± 0.04**	**42.83 ± 0.07**	**61.83 ± 0.05**	**58.55 ± 0.11**
Only Text	8.20 ± 0.04	44.72 ± 0.04	27.28 ± 0.06	44.30 ± 0.05	39.94 ± 0.07
Zeros	8.72 ± 0.02	44.76 ± 0.05	27.70 ± 0.05	44.33 ± 0.06	39.78 ± 0.06
Noise	15.61 ± 0.07	43.55 ± 0.05	28.36 ± 0.07	43.36 ± 0.06	37.84 ± 0.08

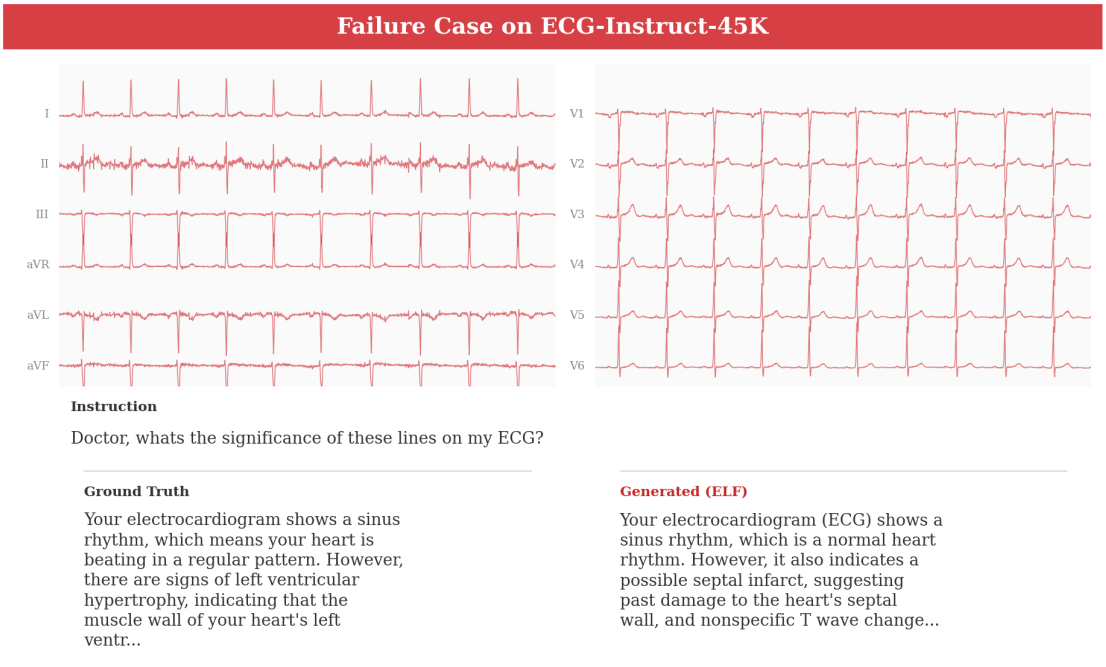

Figure 10: Failed generation with Conv. ELF on the ECG-Instruct 45K dataset

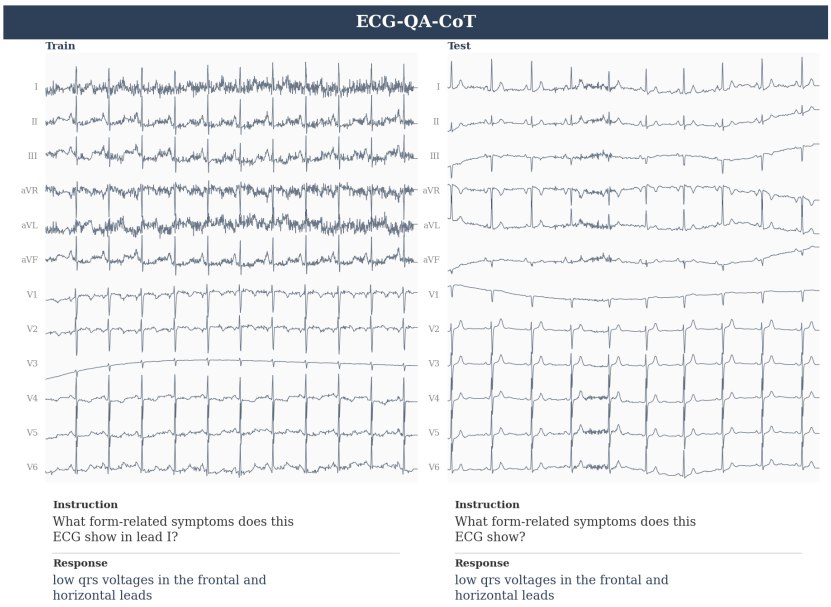

Figure 11: Instances from the training and testing set of ECG-QA-CoT.

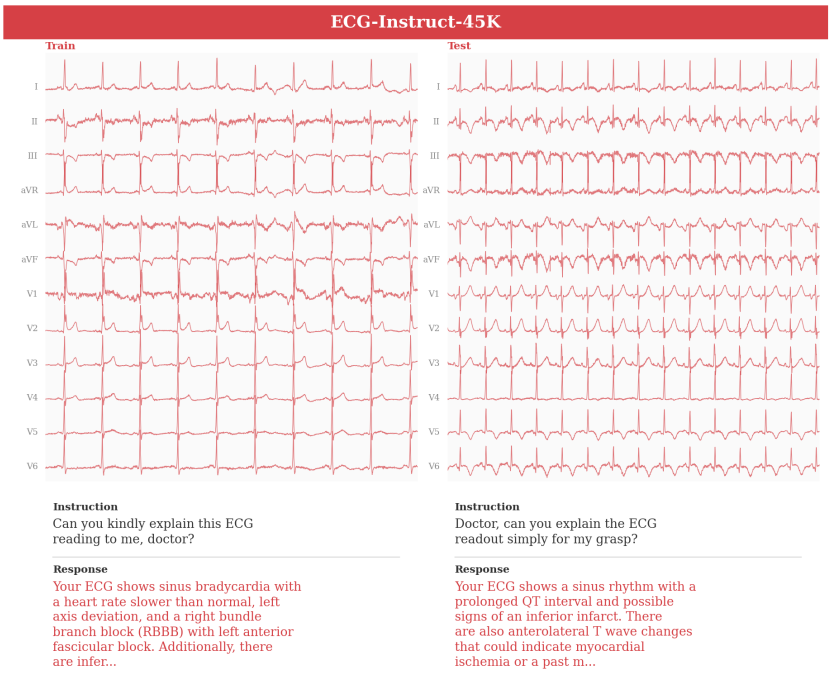

Figure 12: Instances from the training and testing set of ECG-Instruct 45K.

Appendix C. Templates

C.1. Large Language Model Conversation Templates

We provide the conversation templates for Llama-3.2-1B-Instruct, gemma-2-2b-it, and Qwen-2.5-1.5B-Instruct. q_{sys}, q_n, y_n refers to the system prompt, textual query and response respectively.

Llama-3.2-1B-Instruct Conversation Template

< |begin_of_text| >< |start_header_id| > system < |end_header_id| >

q_{sys} < |eot_id| >< |start_header_id| > user < |end_header_id| >

<signal>
q_1 < |eot_id| >< |start_header_id| > assistant < |end_header_id| >

y_1 < |eot_id| >< |start_header_id| > user < |end_header_id| >

q_2 < |eot_id| >< |start_header_id| > assistant < |end_header_id| >

y_2 < |eot_id| >
...

q_n < |eot_id| >< |start_header_id| > assistant < |end_header_id| >

y_n < |eot_id| >

gemma-2-2b-it Conversation Template

< bos >< start_of_turn > user
<signal>
q_1 < end_of_turn >
< start_of_turn > model
y_1 < end_of_turn >
< start_of_turn > user
q_2 < end_of_turn >
< start_of_turn > model
y_2 < end_of_turn >

...

< start_of_turn > user
q_n < end_of_turn >
< start_of_turn > model
y_n < end_of_turn >

Qwen-2.5-1.5B-Instruct Conversation Template

< |im_start| > system
q_{sys} < |im_end| >
< |im_start| > user
<signal>
q_1 < |im_end| >
< |im_start| > assistant
y_1 < |im_end| >
< |im_start| > user
q_2 < |im_end| >
< |im_start| > assistant
y_2 < |im_end| >

...

< |im_start| > user
q_n < |im_end| >
< |im_start| > assistant
y_n < |im_end| >

C.2. System Prompt

We provide the system prompts utilized throughout the study. Namely, we provide the ELM system prompt used for all experiments with ELF, ST-MEM Na et al. (2024), and MERL Liu et al. (2024).

ELM System Prompt

You are an expert multimodal assistant with advanced knowledge in **clinical cardiac electrophysiology**.

Input Identification
- Detect whether the input is **text**, **ECG signals (time-series data)**, or **both**.

ECG Signal Analysis
- Treat ECG as cardiac time-series data.
- Provide expert interpretation of **heart rate, rhythm, conduction, arrhythmias, and other electrophysiologic abnormalities**.

Multimodal Reasoning
- When both text and ECG are given, integrate them into a unified, **cardiac electrophysiologist-level assessment**.

Response Style
- Deliver responses that are **precise, concise, and clinically authoritative**, grounded in electrophysiology and natural language reasoning.
- For general, non-ECG questions, respond as a capable **general assistant**.