Physiologically Validated Synthetic ECG Generation via LLM-Guided Constraint Enforcement

Sujit Bhandari and Robin Chataut

Department of Computer Science, Texas Christian University,
Fort Worth, TX, USA
{sujit.bhandari, robin.chataut}@tcu.edu

Abstract. Training reliable arrhythmia detectors requires large and diverse electrocardiogram (ECG) datasets, but clinical ECG corpora are scarce, class-imbalanced, and tightly regulated. We present a physiologically validated synthetic ECG pipeline in which a large language model (LLM) is used solely as a clinical-knowledge encoder, producing JSON constraint envelopes over heart rate, RR interval, QRS duration, P-wave presence, and amplitude, with all constraints cross-verified against established cardiology guidelines. These envelopes govern rejection sampling across three architecturally distinct generators: interpolation, MLP WGAN-GP, and 1D-CNN-GAN. Under an inter-patient protocol on the MIT-BIH Arrhythmia dataset, even the temporally aware 1D-CNN-GAN produces 96.3% physiologically implausible Normal-class beats prior to filtering. TOST equivalence testing ($\delta = 0.5\sigma$) with Bonferroni correction establishes statistical equivalence on $6/8$ morphological features for supraventricular beats and $5/8$ for ventricular beats. Constrained augmentation preserves premature ventricular contraction detection (V-F1: 0.651 vs. 0.598 unconstrained) while improving supraventricular sensitivity by 35%. A direct comparison shows LLM-derived and manually curated rule-based envelopes reach statistically equivalent downstream performance (p=0.35): the LLM's contribution is reproducible, low-effort extensibility to new conditions, not accuracy. A preliminary qualitative audit further reveals that although 100% of Stage-3 validated beats satisfy the LLM-derived global envelopes, only 13% additionally pass a local morphological audit, indicating that global physiological constraints are necessary but not sufficient for clinically reliable synthesis. These findings support treating constraint-based physiological validation as a minimum requirement for synthetic ECG used in healthcare pipelines and motivate complementary local-smoothness criteria as an immediate direction for future work.

Keywords: Synthetic ECG · Clinical validation · Generative models · Healthcare AI safety · Arrhythmia detection · Privacy-preserving AI.

1 Introduction

Cardiovascular diseases (CVDs) remain the leading cause of death globally, responsible for approximately 17.9 million deaths annually according to the World Health Organization [1]. Electrocardiogram (ECG) monitoring through wearable devices, including consumer smartwatches and clinical-grade chest monitors, has created unprecedented opportunities for early detection of cardiac arrhythmias. Hannun et al. [2] demonstrated that deep neural networks can achieve cardiologist-level arrhythmia detection from single-lead ECG, establishing the clinical viability of automated analysis. However, robust automated detection faces two persistent challenges, compounded by an underappreciated safety risk in current synthetic-data practice: a downstream classifier trained on physiologically implausible synthetic beats risks encoding spurious decision boundaries. Consider a GAN-generated beat with 120 ms QRS duration, 5× nominal peak-to-peak amplitude, and inverted P-waves: while visually ECG-shaped, this beat corresponds to no real cardiac condition and cannot teach a detector anything valid about arrhythmia morphology. Such data does not augment; it corrupts the training distribution. In safety-critical healthcare applications, validating that synthetic data remains within the physiologically plausible parameter space is not optional, it is a prerequisite for responsible deployment.

First, annotated clinical ECG datasets are scarce and heavily class-imbalanced. The MIT-BIH Arrhythmia dataset, the most widely used benchmark, contains recordings from only 47 subjects with approximately 110,000 beats [3]. Under the AAMI five-class mapping, Normal (N) beats constitute 89.9% of training data, while clinically critical Ventricular ectopic (V) account for 7.4%, Supraventricular ectopic (S) for 1.9%, and Fusion (F) for <0.8%. This imbalance systematically degrades classifier performance on the minority classes that carry the greatest clinical importance. Each AAMI class has distinct clinical implications. N beats exhibit high inter-patient morphological variability, making them difficult to model. S beats (premature atrial contractions, PACs) are characterized by narrow QRS complexes (80–110 ms) and altered P-wave morphology, and are associated with increased risk of atrial fibrillation and stroke [18]. V beats (premature ventricular contractions, PVCs) exhibit wide QRS complexes (>120 ms) and represent the most clinically critical minority class, as frequent PVCs are linked to cardiomyopathy and sudden cardiac death [19,20]. Preserving V-class integrity during augmentation is therefore essential. F beats combine features of N and V beats, making them particularly difficult to synthesize. Second, regulatory constraints such as HIPAA and GDPR limit the sharing of patient-derived ECG data, creating barriers to multi-institutional model development.

Synthetic data generation has been proposed to address both challenges. Early augmentation strategies, including additive noise, time-warping, and amplitude scaling [13], are physiologically naive. Recent approaches employ GANs, VAEs, and diffusion models for ECG synthesis [5,6,7,8]. However, these methods lack explicit physiological validation. A recent scoping review [5] highlights the absence of standardized metrics for assessing synthetic ECG quality, with most studies relying on visual inspection, adapted FID scores, or downstream perfor-

mance, none of which directly verify physiological plausibility. A key limitation of existing approaches is the absence of external physiological constraints. Generative models optimize for distributional similarity but do not enforce clinical validity, such as QRS duration bounds or amplitude ranges. As a result, even temporally aware architectures can produce unrealistic signals. In our evaluation, a 1D-CNN-GAN generates 96.3% physiologically implausible Normal-class beats without external filtering.

Large language models (LLMs) have demonstrated substantial cardiological knowledge in ECG interpretation tasks. ECG-LM [10], ECG-Chat [11], and retrieval-augmented approaches [12] leverage LLMs for diagnostic reasoning. However, prior work uses LLMs for interpretation rather than for enforcing physiological constraints during data generation. In this work, we propose a three-stage pipeline in which an LLM is used strictly as a clinical-knowledge encoder. The LLM generates physiological constraint envelopes for each arrhythmia class, which are cross-validated against established cardiology guidelines [15,16,17]. These constraints are applied through rejection sampling to filter outputs from three generator architectures, an intentionally generator-agnostic post-hoc design that wraps any candidate model without retraining and yields an auditable per-beat rejection record. A multi-metric evaluation framework, including TOST equivalence testing and Bonferroni-corrected ablation, quantifies both statistical and clinical fidelity.

Contributions. (1) *A clinically grounded synthetic ECG pipeline* that enforces literature-derived physiological constraints (RR, QRS, P-wave, amplitude) on every generated beat. (2) *A validation framework bridging machine learning and clinical plausibility*, combining TRTS, TSTR, MMD, DTW, and TOST equivalence testing for per-feature validation. (3) *Empirical evidence of failure in unconstrained generators*, with rejection rates of 36.1–97.8% across architectures. (4) *A direct LLM-vs-rule-based comparison* showing statistically equivalent downstream performance (p=0.35); the LLM's value is reproducibility and low-effort extensibility, not classification accuracy. (5) *A privacy-preserving pathway for healthcare AI*, supported by preliminary membership inference analysis indicating low memorization risk. (6) *A qualitative audit showing global envelopes are necessary but not sufficient*: 100% of Stage-3 validated beats satisfy the LLM constraint envelopes, yet only 13% additionally pass a local-morphology audit, motivating complementary local-smoothness criteria as immediate future work.

Clinical Relevance. This work addresses three practical challenges in healthcare AI. First, it enables physiologically valid augmentation for underrepresented arrhythmia classes, improving robustness in safety-critical settings. Second, it supports cross-institutional model development by reducing reliance on patient data. Third, it mitigates the risk of training models on physiologically invalid signals, a failure mode not addressed in existing synthetic ECG pipelines.

2 Related Work

Generative models have been widely explored for synthetic ECG generation. Delaney et al. [7] used GANs to generate normal-rhythm ECG for privacy-preserving data sharing, evaluating only at the signal level. Adib et al. [6] compared five GAN architectures (vanilla GAN, DCGAN, ACGAN, BiGAN, and WaveGAN) for single-lead ECG, showing that augmentation improves classification accuracy without physiological validation of individual beats. Golany and Radinsky [9] proposed personalized GANs conditioned on patient-specific templates. More recently, Alcaraz and Strodthoff [8] introduced SSSD-ECG, combining diffusion models with structured state-space sequences and reporting state-of-the-art results on PTB-XL with a cardiologist Turing test. Despite these advances, a recent scoping review by Zanchi et al. [5] confirms that no standardized metrics yet exist for evaluating physiological validity of synthetic ECG; this gap motivates our framework. Diffusion-based generators are not included in our triple-generator evaluation because training per-class diffusion models on small minority subsets (e.g., S-class with 944 beats) risks overfitting, and our central claim is about the constraint layer rather than any single generator; the constraint layer is generator-agnostic and applies equally to diffusion outputs. We are also not aware of a published synthetic-ECG method that performs head-to-head physiology-aware constraint enforcement at the per-beat level, which precludes a direct numerical comparison against an alternative constraint-guided method. Table 1 summarizes representative prior work in comparison to our approach.

Table 1. Comparison with representative prior synthetic ECG work. PV = physiological validation; MM = multi-metric evaluation; EQ = equivalence testing (TOST); TG = temporally-aware generator.

Study	Generator	Metrics	PV	MM	EQ	TG
Adib et al. [6]	5 GANs	Acc, FID	–	–	–	–
Delaney et al. [7]	GAN	Sig. sim.	–	–	–	✓
Golany & Radinsky [9]	Seq2Seq GAN	Acc	–	–	–	✓
Alcaraz & Strodthoff [8]	Diffusion + SSM	FID, class.	–	–	–	✓
Ours	3 generators	4 metrics + TOST	✓	✓	✓	✓

Large language models have also been applied to ECG analysis. ECG-LM [10] couples an ECG signal encoder with a biomedical LLM for cross-modal interpretation, while ECG-Chat [11] uses contrastive learning to align ECG waveforms with cardiology reports. Yu et al. [12] propose a retrieval-augmented framework leveraging cardiology textbook knowledge. These approaches use LLMs for interpretation; in contrast, our work uses the LLM solely as a clinical-knowledge encoder that generates numerical constraint envelopes. The idea of incorporating external domain knowledge has precedents in adjacent fields such as physics-

informed neural networks for time series [23]. Unlike prior approaches, constraints in our framework are derived from published clinical guidelines and applied after generation via rejection sampling, making the method generator-agnostic. Traditional ECG data augmentation methods, including additive Gaussian noise, time-warping, amplitude scaling, and baseline-wander simulation [13], are physiologically naive; for example, time-warping may produce ventricular beats with unrealistic QRS durations that do not correspond to any valid cardiac pathology. In contrast, our framework constrains augmented beats to remain within class-specific, literature-verified physiological envelopes, ensuring clinical plausibility.

3 Methodology

The proposed framework operates in three sequential stages (Fig. 1): clinical-knowledge encoding via an LLM, constrained triple-generator synthesis, and multi-metric validation with equivalence testing.

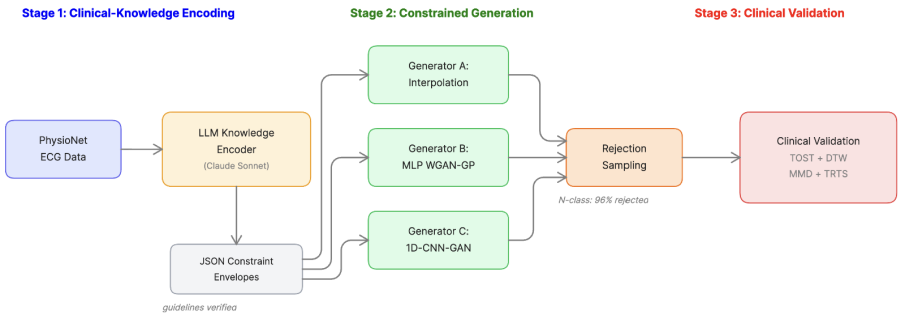


Fig. 1. Clinically validated synthetic ECG generation pipeline. An LLM encodes cardiology-guideline knowledge into physiological constraint envelopes. Three generators produce candidate ECG beats, which are filtered via rejection sampling and evaluated against real data. The "N-class: 96% rejected" annotation highlights the central finding that even temporally-aware generators produce many physiologically implausible beats without external validation.

3.1 Stage 1: LLM as a Clinical-Knowledge Encoder

For four AAMI classes (N, S, V, F), Claude Sonnet is prompted to emit JSON parameter envelopes specifying ranges for RR interval, heart rate, QRS duration, P-wave presence, rhythm regularity, SDNN, RMSSD, and amplitude scaling. Crucially, the LLM never touches an ECG waveform; its role is strictly to translate textbook cardiology knowledge into a machine-readable specification. All parameters are cross-validated against AHA/ACC [15], Chou's [16],

and ESC/NASPE [17] guidelines, and pass without correction. The LLM correctly captures pathological distinctions: wide QRS (120–200 ms) for ventricular ectopy versus narrow QRS (80–110 ms) for supraventricular ectopy, and absent P-wave association in ventricular beats. Prompt sensitivity analysis shows 100% clinical validity at ±20% perturbation and 87.5% at ±25%.

LLM versus rule-based constraints. A reasonable alternative is to hand-curate the envelopes directly from textbooks; we evaluate both. A rule-based baseline uses identical parameters manually specified from the same cardiology references. The LLM produces slightly tighter bounds (e.g., F-class QRS 100–140 ms vs. rule-based 100–160 ms; F-class amplitude scaling 0.6–1.2× vs. 0.5–1.5×). On the downstream arrhythmia task the two routes are statistically equivalent (macro F1 0.324 vs. 0.312, V-F1 0.669 vs. 0.628, p=0.35 across 10 seeds; Section 5). We explicitly do not claim the LLM produces superior envelopes. Its practical advantage lies elsewhere: a single prompt produces internally consistent envelopes in seconds, the same prompt template can be re-used to extend the framework to new cardiac conditions or pediatric/athletic sub-populations without re-curating a textbook, and the JSON output is auditable. The LLM *automates and standardizes* the translation of published guidelines into machine-readable constraints; the rule-based variant remains a valid, slightly more conservative drop-in for deployments that prefer fully manual provenance. Cardiologist-blinded review of the resulting synthetic beats is planned as a separate study.

3.2 Stage 2: Triple-Generator Synthesis

Generator A (Interpolation). For each target class c with beat library $\mathcal{L}_c$: two beats $\mathbf{x}_1, \mathbf{x}_2$ are sampled uniformly without replacement; mixing $\alpha \sim \text{Beta}(0.5, 0.5)$ produces $\mathbf{x}_{\text{syn}} = \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2$; amplitude scaling $s \sim \mathcal{U}(s^c_{\min}, s^c_{\max})$ within LLM bounds is applied, followed by Gaussian noise with $\sigma \sim \mathcal{U}(0.005, 0.03)$ and normalization to zero mean and unit variance.

Generator B (MLP WGAN-GP). Per-class Wasserstein GAN with gradient penalty [14]. The generator is a four-layer MLP mapping $\mathbf{z} \sim \mathcal{N}(0, I_{64})$ through hidden dimensions 128, 256, 512 to output dimension 250, with LeakyReLU (α=0.2), batch normalization, and Tanh output. The critic mirrors this architecture with dropout (0.3). This generator processes the full 250-sample vector without temporal structure.

Generator C (1D-CNN-GAN). Per-class WGAN-GP with a 1D transposed-convolutional generator: $\mathbf{z} \in \mathbb{R}^{64} \rightarrow$ FC+Reshape (128×16) $\rightarrow$ four ConvT1d blocks with kernel 4, stride 2, and channel progression 128→64→32→16→1, interleaved with BN and LeakyReLU, producing length-256 output trimmed to 250. Output activation is Tanh. The critic mirrors this with four strided 1D convolutions, LeakyReLU, and dropout (0.25). Temporal convolutions learn localized patterns such as QRS complexes and P-waves. Training: n_{critic}=5, λ_{GP}=10, Adam (lr=10^{-4}, β_1=0.0, β_2=0.9), 300 epochs (MLP) or 500 epochs (1D-CNN)

per class on an NVIDIA T4 GPU. All three generators undergo identical LLM-constraint rejection sampling; unconstrained variants serve as ablation baselines.

Rejection Sampling. Algorithm 1 formalizes constraint enforcement. For each candidate $\hat{\mathbf{x}}$, morphological features $(A_{\text{ptp}}, w_{\text{QRS}}, A_R)$ are extracted and checked against the LLM envelope $\mathcal{C}_c$ and empirical 2–98$^{\text{th}}$ percentile ranges $\mathcal{R}_c$ (20% tolerance) from real beats.

Algorithm 1 Clinically governed rejection sampling.

Require: Generator G, LLM constraint envelope $\mathcal{C}_c$, target count N, empirical range $\mathcal{R}_c$
Ensure: Validated synthetic set $\mathcal{S}_c$ with $|\mathcal{S}_c| = N$

$\mathcal{S}_c \leftarrow \emptyset,\ n_{\text{rej}} \leftarrow 0$
while $|\mathcal{S}_c| < N$ **do**
  $\hat{\mathbf{x}} \leftarrow G(\mathbf{z}),\ \mathbf{z} \sim \mathcal{N}(0, I_d)$
  $(A_{\text{ptp}}, w_{\text{QRS}}, A_R) \leftarrow \textsc{Extract}(\hat{\mathbf{x}})$
  `valid` $\leftarrow$ `True`
  if $A_{\text{ptp}} \notin \mathcal{C}_c$[amp] **then** `valid` $\leftarrow$ `False`
  else if $w_{\text{QRS}} \notin \mathcal{C}_c$[qrs] **then** `valid` $\leftarrow$ `False`
  else if $A_{\text{ptp}} \notin [0.8\,\mathcal{R}_c^{2\%},\ 1.2\,\mathcal{R}_c^{98\%}]$ **then** `valid` $\leftarrow$ `False`
  end if
  if `valid` **then** $\mathcal{S}_c \leftarrow \mathcal{S}_c \cup \{\hat{\mathbf{x}}\}$
  else $n_{\text{rej}} \leftarrow n_{\text{rej}} + 1$
  end if
end while
return $\mathcal{S}_c$, rejection rate $= n_{\text{rej}}/(|\mathcal{S}_c| + n_{\text{rej}})$

3.3 Stage 3: Clinical Validation

We evaluate synthetic beats with four complementary metrics: TRTS (train on real, test on synthetic; >85% target), TSTR (train on synthetic, test on real), MMD (per-class RBF-kernel distributional test), and DTW (per-class morphological distance ratio). Distributional claims use Two One-Sided Tests [21] rather than NHST, with $\delta = 0.5 \cdot \sigma_{\text{pooled}}$ per morphological feature. This corresponds to a small-to-medium effect size (Cohen's d=0.5) with a concrete clinical interpretation: a difference below $0.5\,\sigma$ in normalized amplitude, QRS width, or RR interval falls within the range of intra-patient beat-to-beat variability in ambulatory recordings [15]. Equivalence is concluded when both one-sided tests are rejected at $\alpha = 0.05$. With k simultaneous comparisons, $\alpha_{\text{corr}} = 0.05/k$; for the arrhythmia ablation (k=7), $\alpha_{\text{corr}} = 0.0071$.

4 Experimental Setup

MIT-BIH Arrhythmia [3]. 48 half-hour two-channel recordings from 47 subjects at 360 Hz, approximately 110,000 beats annotated by two or more cardiologists.

We follow the inter-patient DS1/DS2 split of de Chazal et al. [4]: 22 training records (51,001 beats: N=45,847, S=944, V=3,788, F=414, Q=8) and 22 test records (49,694 beats) with no patient overlap. Signals are bandpass filtered 0.5–40 Hz (4^{th}-order Butterworth, zero-phase). R-peaks are detected with NeuroKit2. Beats are segmented into 250-sample R-centered windows and normalized to zero mean and unit variance.

Classifier. 1D CNN-BiLSTM (32/64/128 filters, kernels 5/5/3, two BiLSTM layers of 64 units/direction, 0.5 dropout; 143,557 parameters), focal loss (γ=2, class-weight cap 10×), Adam (lr=10^{-3}), cosine annealing, patience 15, 150 epochs. All experiments use 10 seeds (5 for 1D-CNN-GAN ablation) and paired t-tests with Bonferroni correction.

5 Results

We present quantitative and qualitative results evaluating the clinical validity, statistical fidelity, and downstream impact of the proposed framework. The central finding is that unconstrained generators produce a large fraction of physiologically implausible signals, while constraint-based validation substantially improves clinical reliability without degrading performance on critical arrhythmia classes.

Unconstrained Generators Produce Clinically Invalid Signals. Table 2 presents the central empirical finding of this work: the fraction of generated beats rejected by the clinical-validity filter. Two key patterns emerge. First, temporal inductive bias significantly improves generation for minority classes: the 1D-CNN-GAN reduces rejection from 35.8% to 15.8% (S), 70.3% to 19.1% (V), and 97.8% to 36.1% (F). Second, this inductive bias does not improve the N class; rejection instead increases from 94.7% to 96.3%. This is expected, as the N class exhibits the greatest morphological diversity across patients (45,847 beats from 22 patients). Clinically, this indicates that even a temporally aware generator cannot reliably produce valid normal-rhythm beats without external validation, motivating the need for a physiological constraint layer independent of architecture. Figure 2 presents a failure-mode gallery pairing rejected raw GAN outputs, annotated by their specific physiological failure, with validated Stage-3 outputs across all four AAMI classes.

Root cause analysis. 91.9% of 1D-CNN-GAN N-class rejections are due to peak-to-peak amplitude falling outside the real range [6.47, 7.23]. The GAN exhibits a PTP variance of 0.82, which is 3.3× higher than the real data variance (0.25), indicating failure to learn the narrow amplitude distribution of normalized ECG signals. The remaining 8.1% of N-class rejections correspond to QRS durations exceeding 120 ms, which would clinically be classified as ventricular. Pairwise cosine similarity among GAN outputs (0.34) is substantially lower than among real beats (0.98), indicating high diversity without mode collapse; however, this diversity is not clinically meaningful. For the F class, rejected beats typically

Table 2. LLM-constraint rejection rates (%) by generator architecture and AAMI class. Interpolation produces valid beats by construction, while both GAN variants generate a substantial fraction of clinically invalid signals.

Generator	**N**	**S**	**V**	**F**
Interpolation	0.0	0.0	0.0	0.0
MLP WGAN-GP	94.7	35.8	70.3	97.8
1D-CNN-GAN	96.3	15.8	19.1	36.1

violate amplitude scaling constraints (0.6–1.2×), reflecting the difficulty of synthesizing hybrid morphologies.

Rejection efficiency and practical cost. The high rejection rates imply non-trivial sampling overhead. From Table 2, one accepted 1D-CNN-GAN beat requires on average ∼27 draws for N, ∼1.57 for F, ∼1.24 for V, and ∼1.19 for S; the MLP WGAN-GP needs ∼19 (N) and ∼45 (F) draws per accepted beat; interpolation requires one. The N class is the dominant bottleneck. In absolute terms this remains tractable: the constraint check is a handful of arithmetic comparisons per beat, and a single T4 GPU produces the rejection-sampled training set used in our ablation (a few thousand beats per class) in under a minute. Two practical mitigations for higher target volumes are personalized N-class envelopes (tightening the distribution the generator must hit) and early-exit rejection on cheaper features, which we already exploit via the ordered checks in Algorithm 1. We view the current cost as an acceptable price for clinical auditability and the N-class bottleneck as a concrete signal for future generator design.

Multi-Metric Validation and TOST Equivalence. The constrained 1D-CNN-GAN achieves TRTS accuracy of 95.7% (target >85%), confirming morphological consistency. DTW ratios are 0.970 (V), 0.945 (S), and 0.876 (F), all well below the 2.0× threshold. TSTR reaches 82.1% (partial). MMD passes marginally when excluding the N class (p=0.057) but fails globally (p<0.001) due to the broad N-class distribution. Per-class TOST equivalence with δ=0.5σ shows equivalence on 6/8 features for S, 5/8 for V, and 1/8 for N, consistent with the high rejection rate in the N class. Table 3 reports per-feature TOST results. Timing (RR interval) satisfies equivalence universally, while amplitude and energy-related features remain challenging. The S class achieves the highest equivalence rate (6/8), aligning with its low rejection rate (15.8%). T-wave amplitude and signal energy fail across all classes, indicating key limitations in current generators.

The global MMD failure (p<0.001) is driven by the majority N class. When excluding N, the MMD statistic decreases by 97.8% (from 0.104 to 0.002), with p=0.057 approaching significance. The minority classes, which are the primary targets of augmentation, therefore achieve near-distributional equivalence. A direct comparison between LLM-derived and rule-based constraints using identical interpolation synthesis shows statistically equivalent downstream performance

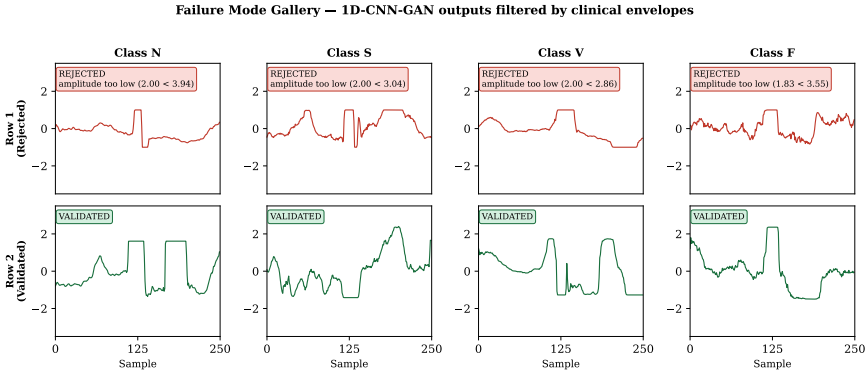


Fig. 2. Failure Mode Gallery: why constraint-based validation is necessary. For each AAMI class (N, S, V, F), the top row shows representative *rejected* 1D-CNN-GAN outputs annotated with the specific physiological failure mode (amplitude out of range), and the bottom row shows *validated* outputs that passed the Stage 3 envelope checks. Rejected beats are visually ECG-shaped but violate clinical envelopes that any cardiologist would recognize as non-physiological, motivating external constraint-based validation prior to any downstream training use.

Table 3. Per-feature TOST equivalence results ($\delta = 0.5\sigma$) across AAMI rhythm classes.

Feature	**N**	**S**	**V**	**F**
A_{ptp}	—	✓	✓	—
w_{QRS}	—	✓	✓	✓
A_R	—	✓	—	—
RR	✓	✓	✓	✓
A_P	—	✓	✓	—
A_T	—	—	—	—
ST	—	✓	✓	✓
E	—	—	—	—
Passed / 8	*1*	*6*	*5*	*3*

(macro F1 0.324 vs. 0.312, V-F1 0.669 vs. 0.628, p=0.35 across 10 seeds). The LLM's value is therefore reproducibility and low-effort extensibility rather than classification accuracy.

Under the inter-patient evaluation protocol with Bonferroni correction (k=7, α_{corr}=0.0071), no augmentation strategy significantly improves macro F1 over the real-only baseline (0.321 ± 0.018). This aligns with prior findings that inter-patient variability dominates generalization. Traditional augmentation achieves 0.340±0.018, while unconstrained and constrained 1D-CNN-GAN achieve 0.305± 0.048 and 0.308±0.024, respectively. A clinically relevant tradeoff is observed: unconstrained augmentation improves S-F1 (0.052 to 0.115, +121%) but degrades

V-F1 (0.679 to 0.598, −12%), whereas constrained augmentation preserves V-F1 (0.651, −4%) while still improving S-F1 (0.070, +35%). A preliminary privacy assessment using a shadow-model membership inference attack yields 52.4% accuracy on real data. Only 0.3% of synthetic beats are classified as training members, compared to 50.2% of real samples. The average membership probability is lower for synthetic data (0.395 vs. 0.502), suggesting reduced memorization; these results are preliminary and do not constitute formal privacy guarantees.

Preliminary Qualitative Assessment

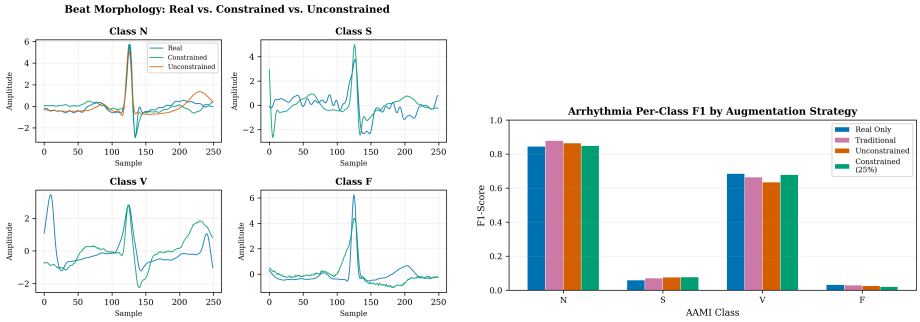
In lieu of a cardiologist-blinded review, which is planned as a separate study, we conducted a structured audit of 100 synthetic beats drawn deterministically from the Stage 3 validated output (50 Normal, 25 PVC, 25 SVB). Each beat was scored on four objective morphological-discontinuity criteria: (i) *baseline drift*, the difference in mean amplitude between the first and last 70 ms of the beat window; (ii) *high-frequency ringing*, measured as total variation; (iii) *discontinuous segments*, the largest single-sample amplitude jump outside the QRS region; and (iv) *flatline regions*, defined as any 30-sample window outside the QRS with near-zero total variation. A beat is reported as *indistinguishable from real* only if all four criteria pass at fixed thresholds calibrated against real-data statistics.

All 100 audited beats satisfy the LLM-derived global envelopes by construction (Stage 3 acceptance), but only 13 % (10/50 N, 0/25 V, 3/25 S) additionally pass all four local-morphology criteria. None of the audited beats exhibit high-frequency ringing, confirming that the rejection sampler effectively suppresses the noisy hallucinations characteristic of unconstrained GAN output. The dominant flagged criteria are discontinuous segments (79/100) and flatline regions (29/100), artifacts of the Tanh output activation saturating at the edges of its range. This audit reveals a finer-grained morphological failure mode that the amplitude- and QRS-based Stage 3 filter does not catch: the filter ensures synthetic beats fall within clinically plausible *global* parameter ranges, but does not enforce *local* waveform smoothness. We view this finding as supportive of the framework's central claim that physiological validation is *necessary but not sufficient* for safe synthetic ECG. Two complementary future directions follow: cardiologist-blinded review (to determine which local discontinuities are clinically meaningful versus benign) and extending the constraint envelope to include local-smoothness criteria (bounded total variation outside the QRS, single-sample derivative thresholds).

Finally, Fig. 3 illustrates the morphological impact of constraints. Interpolation-based beats exhibit lower RMSE relative to class templates (e.g., N: 0.450 vs. 0.655), indicating conservative but accurate synthesis. GAN-generated beats that pass constraint filtering are more diverse while maintaining clinical validity.

6 Discussion

The triple-generator evaluation surfaces a finding that neither architecture alone could have established: the need for clinical validation is *class-dependent*, not



(a) Real (blue), constrained (green), unconstrained (red) beats per class.

(b) Per-class F1 across augmentation strategies.

Fig. 3. Clinical effect of physiological constraints. *(a)* Constrained beats preserve class-specific features (narrow QRS for N, wide complex for V) while unconstrained beats frequently violate envelope parameters. *(b)* Constrained augmentation preserves V-class detection at essentially baseline level, whereas unconstrained augmentation visibly degrades V-class F1 while boosting S-class F1, the core clinical tradeoff surfaced by the framework.

architecture-dependent. The 1D-CNN-GAN dramatically reduces rejection for minority classes (S: 35.8%→15.8%, V: 70.3%→19.1%, F: 97.8%→36.1%), yet N-class rejection increases slightly (94.7%→96.3%). For *every* AAMI class, at least one GAN architecture in our study produces a large fraction of clinically invalid beats. The constraint layer therefore plays two complementary roles: essential quality filtering for classes that no architecture handles well (N, F), and safety verification for classes where temporal generators perform better (S, V). A central implication is that unconstrained generative models should not be deployed directly in healthcare pipelines, and constraint-based physiological validation should be considered a minimum requirement for synthetic ECG used in safety-critical settings.

Why post-hoc filtering, not an in-loop constraint? A natural question is whether the constraint should be folded into the generator's training loss. We chose post-hoc rejection sampling deliberately. First, it is generator-agnostic: the same physiological filter wraps an interpolator, an MLP-WGAN, a 1D-CNN-GAN, and (in principle) any future diffusion or transformer model with no retraining. Second, the rejection record is interpretable—every discarded beat carries an explicit, named physiological reason, which is critical for clinical audit. Third, in-loop physiology losses risk vanishing-gradient pathologies when the constraint surface is non-differentiable (e.g., a hard 80–110 ms QRS bound), and soft surrogates dilute the very property we want to guarantee. The cost is generation efficiency, which we report transparently above; we treat this as a productive signal for future generator design rather than as a defect of the validation layer.

The qualitative assessment refines this picture. While global LLM-derived envelopes catch 36–97% of GAN failures at the amplitude/QRS level, the audit of validated beats shows that the remaining outputs, despite satisfying every global constraint, still contain local discontinuities and flatline artifacts in 87% of cases. This is the expected behavior of a global filter applied to outputs from a generator whose final Tanh activation saturates at the edges of its range. Physiological governance for synthetic ECG should therefore be conceived as a *layered* validation pipeline: global envelopes (clinically interpretable, easy to specify from guidelines) followed by local-smoothness criteria (mathematically well-defined, computationally cheap), with cardiologist review as the final arbiter for clinically meaningful local deviations.

The per-class ablation further reveals a safety-relevant tradeoff. Unconstrained 1D-CNN-GAN augmentation maximizes S-class detection (F1 = 0.115, +121% vs. baseline) but degrades V-class detection (F1 = 0.598, −12%). Constrained augmentation moderates the S-class gain (F1 = 0.070, +35%) while preserving V-class performance (F1 = 0.651, −4%). V-class beats carry substantially higher mortality risk than S-class beats: frequent PVCs are associated with PVC-induced cardiomyopathy and elevated risk of sudden cardiac death [19,20], whereas frequent PACs primarily increase the risk of atrial fibrillation [18]. Constrained augmentation therefore sacrifices some S-class sensitivity to preserve detection of the more clinically dangerous arrhythmia class, a tradeoff a safety-conscious deployment should explicitly favor. The persistent N-class rejection rate ($> 94\%$ across all GAN architectures) reflects a fundamental mismatch between per-class generative modeling and the patient-specific nature of normal sinus rhythm: 45,847 N-class beats originate from 22 patients with heterogeneous anatomy and electrode placement. We adopt TOST equivalence testing to avoid inferring distributional equivalence from failure to reject the null; Bonferroni correction (α=0.0071) removes all previously significant arrhythmia improvements, which we report explicitly. The primary contribution is therefore not improved classification under the inter-patient protocol but rather *safe augmentation*: privacy-preserving training without degrading performance on the clinically critical V class.

7 Conclusion

We presented a clinically grounded synthetic ECG pipeline in which an LLM serves strictly as a clinical-knowledge encoder, translating cardiology guidelines into JSON constraint envelopes that govern rejection sampling across three generators. Even a temporally aware 1D-CNN-GAN produces 96.3% physiologically implausible Normal-class beats and 19.1–36.1% invalid minority-class beats before external validation. Constrained augmentation preserves V-class detection (V-F1 0.651 vs. 0.598 unconstrained) while improving S-class detection by 35%, and TOST equivalence holds on 6/8 features for S and 5/8 for V. A direct LLM-vs-rule-based comparison shows statistically equivalent downstream performance (p=0.35); the LLM contributes reproducibility and low-effort extensibility, not

classification accuracy. A preliminary qualitative audit refines this picture: 100% of Stage-3 validated beats satisfy the LLM-derived global envelopes, but only 13% additionally pass a local-morphology audit, showing global physiological constraints are necessary but not sufficient. The practical implication is clear: *constraint-based physiological validation should be a minimum requirement for synthetic ECG in healthcare AI*, conceived as a layered pipeline combining global envelopes with local-smoothness criteria.

7.1 Limitations and Future Work

Five limitations frame these results. *First*, envelope validation is based on published guidelines; cardiologist-blinded review has not yet been conducted, and our four-criterion qualitative audit substitutes for but does not replace expert judgment. A cardiologist Turing-style evaluation is the immediate next study. *Second*, evaluation operates at the per-beat (250-sample) level and does not assess long-range temporal consistency—RR-interval sequences, beat-to-beat morphological drift, or rhythm-level patterns—which matters when synthetic beats are concatenated into multi-beat strips for rhythm classification. Extending the constraint envelope to rhythm-level features (HRV statistics, ectopic coupling patterns) is a natural extension the layered design supports without architectural change. *Third*, the membership inference attack achieves only 52.4% on real data, limiting interpretability; stronger attacks [22] and formal (ε, δ)-DP via DP-GAN integration are required before deployment-grade privacy claims. *Fourth*, the triple-generator evaluation excludes diffusion- and transformer-based generators [8]; we expect the generator-agnostic constraint layer to transfer, but empirical confirmation is required. We are also not aware of a published baseline that performs head-to-head per-beat physiology-aware constraint enforcement, precluding a direct numerical comparison—a gap noted by [5] that motivates rather than undermines the framework. *Fifth*, the inter-patient protocol inherently constrains observable augmentation gains, and downstream robustness was evaluated on a single 1D CNN-BiLSTM; cross-architecture and cross-dataset studies are needed.

Future work will address these through cardiologist-blinded evaluation; extending the envelope with local-smoothness and rhythm-level criteria; DP-GAN integration; diffusion- and transformer-based generators wrapped by the same constraint layer; multi-LLM constraint comparison; 12-lead synthesis; federated deployment; cross-architecture downstream-robustness studies; and personalized N-class envelopes from per-patient templates.

Acknowledgments. This work was supported in part by the Texas Christian University Research and Creative Activities Fund (RCAF) and the Louise Dilworth Davis College of Science & Engineering SERC Grant Program.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. World Health Organization: Cardiovascular diseases (CVDs) fact sheet (2021)
2. Hannun, A.Y., et al.: Cardiologist-level arrhythmia detection from single-lead ECG using deep neural networks. Nat. Med. **25**(1), 65–69 (2019)
3. Moody, G.B., Mark, R.G.: The impact of the MIT-BIH arrhythmia database. IEEE Eng. Med. Biol. Mag. **20**(3), 45–50 (2001)
4. de Chazal, P., O'Dwyer, M., Reilly, R.B.: Automatic heartbeat classification using ECG morphology and interval features. IEEE Trans. Biomed. Eng. **51**(7), 1196–1206 (2004)
5. Zanchi, B., et al.: Synthetic ECG signals generation: A scoping review. Comput. Biol. Med. **184**, 109453 (2025)
6. Adib, E., Afghah, F., Prevost, J.J.: Synthetic ECG signal generation using generative neural networks. PLoS ONE **20**(3), e0271270 (2025)
7. Delaney, A.M., Brophy, E., Ward, T.E.: Synthesis of realistic ECG using generative adversarial networks. arXiv:1909.09150 (2019)
8. Alcaraz, J.M.L., Strodthoff, N.: Diffusion-based conditional ECG generation with structured state space models. Comput. Biol. Med. **163**, 107115 (2023)
9. Golany, T., Radinsky, K.: PGANs: Personalized generative adversarial networks for ECG synthesis. In: AAAI **33**(1), 557–564 (2019)
10. Yang, K., et al.: ECG-LM: Understanding electrocardiogram with a large language model. Health Data Sci. (2024)
11. Zhao, Y., et al.: ECG-Chat: A large ECG-language model for cardiac disease diagnosis. arXiv:2408.08849 (2024)
12. Yu, H., Guo, P., Sano, A.: Zero-shot ECG diagnosis with large language models and retrieval-augmented generation. In: ML4H. PMLR **225**, 650–663 (2023)
13. Berkaya, S.K., et al.: A survey on ECG analysis. Biomed. Signal Process. Control **43**, 216–235 (2018)
14. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: NeurIPS, pp. 5767–5777 (2017)
15. Rautaharju, P.M., Surawicz, B., Gettes, L.S.: AHA/ACCF/HRS recommendations for ECG standardization: Part IV. Circulation **119**(10), e241–e250 (2009)
16. Surawicz, B., Knilans, T.K.: Chou's Electrocardiography in Clinical Practice, 6th edn. Saunders/Elsevier (2008)
17. Task Force of the ESC and NASPE: Heart rate variability: Standards of measurement, interpretation, clinical use. Circulation **93**(5), 1043–1065 (1996)
18. Binici, Z., et al.: Excessive supraventricular ectopic activity and increased risk of atrial fibrillation and stroke. Circulation **121**(17), 1904–1911 (2010)
19. Latchamsetty, R., Bogun, F.: Premature ventricular complex-induced cardiomyopathy. JACC Clin. Electrophysiol. **5**(5), 537–550 (2019)
20. Baman, T.S., et al.: Relationship between burden of premature ventricular complexes and left ventricular function. Heart Rhythm **7**(7), 865–869 (2010)
21. Schuirmann, D.J.: A comparison of two one-sided tests and the power approach for assessing average bioavailability. J. Pharmacokinet. Biopharm. **15**(6), 657–680 (1987)
22. Carlini, N., et al.: Membership inference attacks from first principles. In: IEEE Symp. Secur. Priv., pp. 1897–1914 (2022)
23. Raissi, M., Perdikaris, P., Karniadakis, G.E.: Physics-informed neural networks. J. Comput. Phys. **378**, 686–707 (2019)