成功大學
National Cheng Kung University
WEARABLE TECHNOLOGY AND MOBILE HEALTHCARE (WTMH) LAB

Evidence-Grounded Offline ECG Interpretation on an 8GB Edge Device — Quantifying the Fidelity–Completeness Trade-off

生物醫學工程學系　學生：劉品妍　指導教授：林哲偉

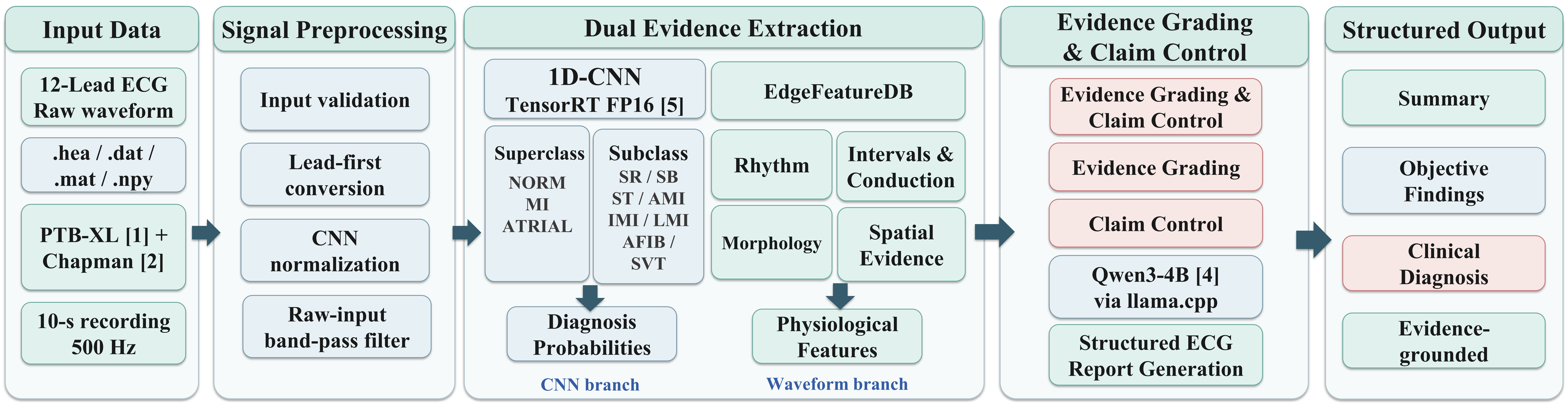
Introduction

Automated ECG interpretation may support rapid screening, but generated clinical reports must remain faithful to measurable ECG evidence. However, conventional LLM-based report generation provides limited control over whether generated findings are explicitly supported by upstream physiological evidence [3], [6]. We therefore developed a fully offline ECG-AI prototype on a Jetson Orin Nano 8GB, combining ECG classification, physiological feature extraction, evidence-to-claim control, and on-device LLM report generation.

Methods

Evidence-Grounded ECG Reporting Pipeline

Jetson Orin Nano 8GB



Results

Evidence Grounding — Full Evaluation (Mixed-source cohort: PTB-XL 80 + Chapman 60, n = 140)

FULL EVALUATION

755 / 755 waveform claims grounded

0 unsupported waveform claims

100% lead-level grounding on evaluable waveform claims—not "100% diagnostic accuracy."

Where Does Information Loss Occur?

OVERALL	CNN 0.786	→	REPORT 0.661
MI LOCALIZATION	CNN 0.733	→	REPORT 0.388

MAJOR BOTTLENECK
evidence-to-report transfer

MI localization was identified as the major information bottleneck during conversion from classifier evidence to natural-language reporting.

Edge Deployment Performance

CNN Inference	TIME TO FIRST TOKEN	GENERATION SPEED	AVERAGE POWER
~22 ms	1.77 s	17.6 tok/s	18.7 W

8GB EDGE DEVICE / TENSORRT FP16 / Qwen3-4B Q4_K_M / FULLY OFFLINE

Peak system RAM usage: 5.64 GB of 7.62 GB usable memory (20-case stable run). A separate 140-case sustained-load run revealed cumulative memory pressure, with recovery required in 3/140 cases.

Why Does Claim Control Matter?

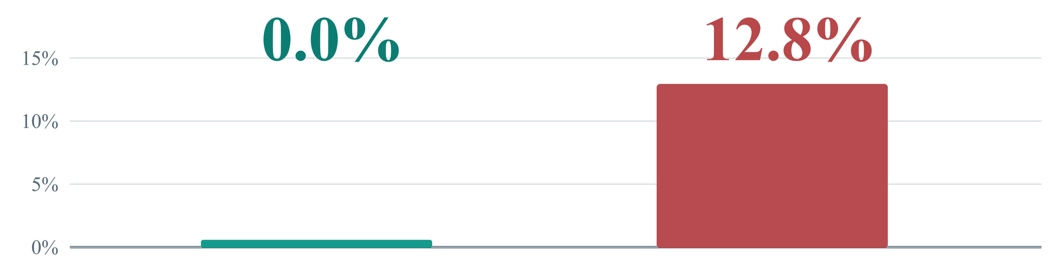
Evidence-to-Claim Constraints Reduce Unsupported Waveform Claims

EXPANDED 140-CASE ABLATION · PTB-XL subset
Independent of the mixed-source full-evaluation cohort

Full Constraint-Guided System
CNN + ECG Features
↓
Evidence / Claim Control
↓
Qwen3-4B
↓
Audited Report

VS.

Free LLM Generation
CNN + ECG Features
reporting constraints removed
↓
Qwen3-4B
↓
Free Report



Full Constraint-Guided: 0 / 562 unsupported waveform claims
Free LLM Generation: 11 / 86 unsupported waveform claims
95% CI: Full 0.00–0.68%; Free 7.29–21.47%; Claim-level Fisher's exact test: $p < 0.001$

Verifiable Claims per Case
Full Constraint-Guided: 4.01
Free LLM Generation: 0.61

Lead accuracy: 100.0% → 89.6%
Lead accuracy evaluated on lead-specific claims only.

Constraint-guided generation improved both evidence grounding and verifiable waveform detail, while free generation produced fewer verifiable claims and a significantly higher unsupported-claim rate.

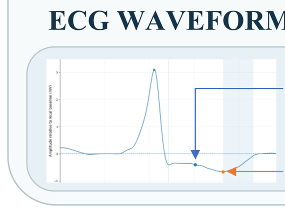
Fidelity–Completeness Trade-off

MACRO PRECISION	MACRO RECALL	AFIB PRECISION	LMI PRECISION
0.788	0.630	1.000	1.000

The system favors conservative omission over unsupported assertion.
Fidelity is prioritized; completeness remains the main improvement target.

ILLUSTRATIVE EXAMPLE

ECG WAVEFORM — Lead V5



This beat: ST-J60 = −0.400 mV
This beat: T-wave = −0.664 mV

AGGREGATED FEATURES — Lead V5

Median ST-J60 = −0.316 mV
Median T-wave amplitude = −0.451 mV
Reliable-beat T-inversion ratio = 1.00
ST-J60 and T-wave amplitudes are median-aggregated across their respective reliable beats; the waveform shows one representative complex.

GROUNDED CLAIMS MAPPED FROM FROZEN REPORT

ST-depression screening flag in V5 → negative/inverted T-wave finding in V5 → Final Report → ✓ grounded

Single-beat and median values differ in magnitude but agree in direction and threshold classification; grounding does not depend on beat selection.

Discussion & Key Findings

- **Constraint control improves both grounding and verifiable detail.** Free generation increased unsupported waveform claims to **12.8%**, reduced verifiable claim yield from **4.01 to 0.61 claims/case**, and decreased lead accuracy from **100.0% to 89.6%**.
- **The system is deliberately conservative.** Macro Precision (**0.788**) exceeded Macro Recall (**0.630**), reflecting a preference for omission over unsupported assertion.
- **The main remaining bottleneck is evidence retention.** MI localization declined from **0.733 to 0.388**, indicating loss during evidence-to-report translation.

Current Limitations

- **MI localization:** 0.733 → 0.388
- **Supporting Feature Recall:** 53.3%
- **P-wave detection F1:** 52.8%
- **Long-run deployment:** cumulative memory pressure still requires supervision

Conclusion

This work demonstrates that a clinically structured ECG report can be generated fully offline on an 8GB edge device while preserving strong evidence fidelity. Evidence-to-claim control significantly reduced unsupported waveform claims and promoted more verifiable, lead-specific reporting. Such an architecture is particularly suitable for privacy-sensitive or connectivity-limited clinical environments where cloud-based inference may be undesirable.

Future Work

- Supporting-feature retention optimization
- Proactive recycling for sustained inference
- Formal comparison between the explicit evidence-grounded pipeline and a latent multimodal ECG-LLM architecture

References

[1] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, Art. no. 154, 2020, doi: 10.1038/s41597-020-0495-6.

[2] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients," *Scientific Data*, vol. 7, Art. no. 48, 2020, doi: 10.1038/s41597-020-0386-x.

[3] Z. Wan, C. Liu, X. Wang, C. Tao, H. Shen, Z. Peng, J. Fu, R. Arcucci, H. Yao, and M. Zhang, "MEIT: Multi-Modal Electrocardiogram Instruction Tuning on Large Language Models for Report Generation," *arXiv preprint arXiv:2403.04945*, 2024.

[4] A. Yang *et al.*, "Qwen3 Technical Report," *arXiv preprint arXiv:2505.09388*, 2025.

[5] NVIDIA Corporation, "NVIDIA TensorRT Documentation," NVIDIA. [Online]. Available: https://docs.nvidia.com/deeplearning/tensorrt/latest/index.html. Accessed: Aug. 19, 2026.

[6] Y. Kim *et al.*, "Medical Hallucination in Foundation Models and Their Impact on Healthcare," *medRxiv*, 2025, doi: 10.1101/2025.02.28.25323115.