

MedEdge-Conf: A Lightweight Confidence-Scoring Framework for Edge-Deployed LLMs in IoMT Health Monitoring

Fahd Koraiche^{1*}, Rachid Dehbi², Amine Dehbi³

Research Laboratory of Data Engineering and Intelligent Systems Laboratory (LIDSI),

Faculty of Science Ain Chock, Casablanca, Morocco^{1,2}

Research Laboratory of Mathematics and Computer Science (LTI), Faculty of Sciences Ben M'Sik, Casablanca, Morocco³

Abstract—Remote patient monitoring systems based on the Internet of Medical Things (IoMT) increasingly use large language models (LLMs) to convert vital signs into readable alerts, but lightweight edge-deployed LLMs may hallucinate, contradict abnormal vitals, or overstate confidence. This study proposes MedEdge-Conf, a lightweight confidence-scoring framework combining physiological coherence aligned with the National Early Warning Score 2 (NEWS2), token-level entropy, and semantic stability into a calibrated composite score driving a deliver/review/escalate decision. We evaluate MedEdge-Conf on 249 test scenarios from two open-access PhysioNet resources (CinC2019, MIT-BIH), using three edge-deployed models (Phi-2, TinyLlama, Gemma-2B, all Q5_K_M), with ground truth from two adjudicated human raters per model. Calibrated fusion reduces the false acceptance rate (FAR) of dangerous hallucinations from 100 percent under naive delivery to 26.4 percent for Phi-2, 17.3 percent for TinyLlama, and 32.1 percent for Gemma-2B, with AUROC improvement over naive scoring statistically significant in every case (DeLong's test, p less than 0.03); precision on delivered responses remains low (14.0-22.8 percent), so it filters dangerous delivery rather than guaranteeing reliability. An ablation study shows the signal driving this improvement is model-dependent: physiological coherence alone is strongest for Phi-2, token entropy helps TinyLlama, and semantic stability matters most for Gemma-2B. Against three literature baselines, MedEdge-Conf achieves the highest AUROC and lowest FAR for every model. Real Raspberry Pi 5 measurements across nine model-quantization configurations show TinyLlama is fastest and lightest, while Gemma-2B is markedly slower and heavier despite being the safest raw generator.

Keywords—Edge AI; internet of medical things; medical LLM; hallucination detection; confidence scoring

I. INTRODUCTION

Remote patient monitoring has become a central component of digital health systems, particularly for chronic disease follow-up and early detection of physiological deterioration. IoMT devices and wearables continuously collect heart rate, oxygen saturation, respiratory rate, temperature, ECG signals, and activity indicators, but these are useful only once translated into timely, clinically safe information. Recent progress in LLMs enables converting such measurements into natural-language explanations, but a standard cloud-centric architecture creates three limitations in health contexts: latency from

transmitting data to remote infrastructure before producing a response, privacy and governance risk from continuous transfer of raw physiological streams to external services, and dependence on connectivity that may be unreliable in home or low-resource settings.

A natural response is to move inference closer to the patient using lightweight quantized LLMs such as Phi-2, TinyLlama, and Gemma variants, executed locally via runtimes such as llama.cpp; quantization can reduce model size and energy requirements, making on-device LLM inference increasingly feasible for embedded applications [6]. Feasibility is not equivalent to safety, however: an edge LLM that labels a critical state as “normal” or “stable” may create direct patient risk, and compression and limited model capacity can worsen instability and hallucination when interpreting abnormal physiological values.

The research gap addressed here is therefore not how to run an LLM on a Raspberry Pi, but how an IoMT system can decide whether a local LLM response is trustworthy enough to show the patient. Existing hallucination-detection methods often rely on large verifier models, retrieval pipelines, or multi-agent debate, not always compatible with an edge gateway's resource budget; generic hallucination detection also does not necessarily capture physiological contradictions such as low SpO₂ described as normal, and even token-entropy-based uncertainty estimates can leave a model confidently wrong about a clinical fact [9] (Section II).

This study introduces MedEdge-Conf, a lightweight confidence-scoring framework for edge-deployed LLMs in IoMT health monitoring, scoring each generated alert on three signals (C_1 physiological coherence/NEWS2, C_2 token entropy, C_3 semantic stability) fused into a calibrated composite score that determines whether the alert is sent, reviewed, or rejected/escalated, so no response reaches the patient without reliability assessment.

The contributions of this work are fourfold:

- We propose MedEdge-Conf, among the first confidence-scoring frameworks for medical LLM outputs under edge-deployment constraints, integrating clinical validation, internal uncertainty, and generation stability within a single-board computer's memory and latency budget.

*Corresponding author.

- We align physiological coherence with the NEWS2 early warning system, so the safety layer respects established clinical thresholds rather than ad-hoc heuristics.
- We execute the complete benchmarking pipeline on a physical Raspberry Pi 5, reporting latency, throughput, memory, CPU, and SoC temperature across all nine model-quantization configurations (Section IV-E).
- We conduct ablation studies on human-annotated data from PhysioNet/CinC2019 and MIT-BIH, quantifying each signal's marginal contribution across all three models (Section V-A), and report a baseline comparison against three literature methods (Section IV-D).

The remainder of this study is organized as follows: Section II reviews related work; Section III details the MedEdge-Conf architecture, data sources, scoring components (C_1 – C_3), and statistical validation protocol; Section IV reports safety-filtering, ablation, baseline-comparison, and edge-performance results; Section V discusses results and limitations; Section VI concludes.

II. RELATED WORK

A. Physiological Monitoring Datasets and Clinical Scoring

PhysioNet provides open access to physiological databases and software resources for biomedical signal research [1]. MIMIC-III contains bedside physiological signals and vital-sign numerics from intensive care settings [3], with its successor MIMIC-IV extending to broader electronic health record data [15]; the PhysioNet/Computing in Cardiology 2019 Challenge targets early sepsis prediction [4]; and MIT-BIH is a standard annotated ECG benchmark [5].

NEWS2, developed by the Royal College of Physicians, aggregates respiration rate, oxygen saturation, systolic blood pressure, pulse, consciousness, temperature, and supplemental oxygen into a deterioration risk score [2], assigning 0–3 points per parameter and mapping the aggregate to low (0–4), medium (3–6 in any single parameter or aggregate 5–6), or high (≥ 7) risk. MedEdge-Conf uses NEWS2 as a constraint layer, penalizing responses that contradict the NEWS2-derived risk category, rather than replacing NEWS2 or applying arbitrary rules.

B. Edge LLM Deployment and Uncertainty Estimation in Medical LLMs

LLMPi investigates quantization-based optimization for Raspberry Pi [6]; llama.cpp supports local inference across hardware using GGUF quantized models [7]; and Nguyen et al. [11] benchmark LLM inference on single-board computers, informing our experimental protocol. These developments, like broader edge-AI/IoT monitoring architectures generally, establish technical feasibility — local inference, resource constraints, communication overhead — but not output trustworthiness in medical contexts, where uncertainty estimation remains difficult [8] and token-level probabilities estimate confidence without guaranteeing correctness, since a model may be confidently wrong [9]; MedEdge-Conf instead adds a reliability-aware decision layer that checks each response before delivery, rather than optimizing generation or transport

alone. Wang et al. [10] propose word-sequence entropy for medical QA, showing correlation between entropy and factual errors.

SelfCheckGPT compares stochastically sampled generations via n-gram overlap, NLI scoring, or LLM prompting to flag hallucinations without model-internal access [17], but is costly and fragile to paraphrase. Broader LLM uncertainty-quantification surveys categorize methods by cost and by reliance on internal signals versus external, consistency-based verification [22]. Guardrail toolkits such as NeMo Guardrails [23] enforce runtime constraints between an application and a general-purpose LLM; MedEdge-Conf is a domain-specific, resource-constrained instance of this paradigm, specialized for physiological alerting rather than open-domain dialogue safety.

Lee et al. [12] highlight the gap between internal model confidence and clinical accuracy, and Rajpurkar et al. [13] emphasize the need for rigorous safety and reliability evaluation of AI in medicine. MedEdge-Conf addresses these gaps with a lightweight hybrid score combining uncertainty signals and physiological coherence for near-real-time, local execution on resource-constrained edge hardware, rather than cloud-based question answering or cloud-scale ensembles.

C. Clinical Safety Assessment, Linguistic Robustness, and Human Factors in Alert Systems

Asgari et al. [24] measure hallucination rates and clinical safety in LLM-generated medical summaries via structured clinician review categorizing errors by severity — an approach MedEdge-Conf shares through its four-way reliability taxonomy (Section III-B) — but target retrospective, cloud-hosted summarization rather than the latency and compute budget of an edge gateway.

NegEx, a curated list of negation trigger phrases determining whether a clinical finding is asserted or negated, remains widely used for its simplicity despite known failure modes on complex or multiply negated sentences [18]. C_1 's keyword-based contradiction detection (Section III-E) is conceptually related but has not been evaluated for robustness to negation or hedged language, a limitation revisited in Section V-C.

MedEdge-Conf's review tier routes uncertain cases to a human and so inherits the well-documented risk of alert fatigue from interruptive clinical alerts [25], discussed further in Section V-C.

III. METHODOLOGY

A. System Architecture

The proposed system targets an Edge-IoMT architecture for home or bedside monitoring: physiological measurements from sensors or wearables are transmitted to a local gateway (a Raspberry Pi 5, 8 GB RAM), which preprocesses, constructs a compact prompt, runs a quantized local LLM, evaluates the response via MedEdge-Conf, and routes the output by confidence level, with an optional cloud/hospital escalation layer. The pipeline comprises six stages: 1) sensor data acquisition and local validation; 2) prompt construction with structured vital signs; 3) LLM response generation with log-probability extraction; 4) confidence scoring via MedEdge-Conf; 5) decision routing based on calibrated thresholds; and

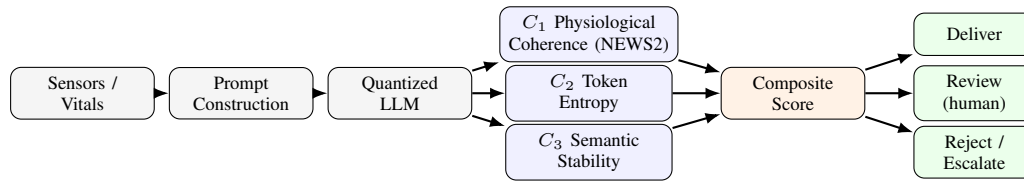


Fig. 1. MedEdge-Conf pipeline: Local LLM output is scored by $C_1/C_2/C_3$ (Sections III-E–III-G) and fused into a calibrated composite score (Section III-H) driving the delivery decision.

6) optional cloud escalation for rejected responses. Fig. 1 summarizes stages 3–6.

B. Data Sources and Scenario Construction

The experimental protocol uses two Open Access PhysioNet resources: the PhysioNet/Computing in Cardiology 2019 sepsis dataset (CC-BY 4.0, no credentialing) [4], providing real per-hour vital-sign records (heart rate, SpO₂, respiratory rate, systolic blood pressure, temperature, and fraction of inspired oxygen) for tens of thousands of ICU patients, and the MIT-BIH Arrhythmia Database (Open Data Commons Attribution License, no credentialing) [5], providing real annotated ECG recordings with beat-level rhythm labels. MIMIC-III’s waveform database [27] (distinct from the MIMIC-III clinical database [3] cited in Related Work as background) was not extracted, since it requires PhysioNet credentialing not pursued in this study.

We constructed 1,247 structured patient scenarios with vital signs and, where available, a trend descriptor, sourced per severity category in this preference order: 1) real CinC2019 patient-hours — heart rate, SpO₂, respiratory rate, systolic BP, and temperature are actual recorded values, supplemental-oxygen is derived from FiO₂, and trend descriptors come from within-patient sequential comparison against preceding hours, not templated text; 2) real MIT-BIH-derived scenarios — heart rate is computed from consecutive R-peak intervals (60/RR-interval-seconds at 360 Hz), the trend descriptor comes from the WFDB beat-annotation symbol, and the remaining ECG-only gaps (SpO₂, respiratory rate, systolic BP, temperature, consciousness) are filled from a documented synthetic range profile matched to beat severity; 3) fully-synthetic backfill from the same range profiles if the two real sources cannot supply enough eligible rows. No fully-synthetic backfill was needed: every scenario’s primary vital(s) — CinC2019’s full panel, or MIT-BIH’s ECG-derived heart rate — come from a real recorded source, though MIT-BIH scenarios’ non-ECG vitals remain profile-filled as noted above — stable: 423 (360 CinC2019, 63 MIT-BIH); borderline: 412 (401 CinC2019, 11 MIT-BIH); critical: 412 (350 CinC2019, 62 MIT-BIH) — with a minimum MIT-BIH share reserved per category. Consciousness (AVPU), unrecorded in either real source, defaults to “Alert.” Forcing every scenario’s AVPU to non-Alert instead shifts the NEWS2 risk category for 152/249 (61.0%) test scenarios (7 low→high, 77 low→medium, 68 medium→high) — not a minor simplification, though the true consciousness state is unrecorded in both sources and its concordance with “Alert” cannot be established from this data.

Each response was annotated into one of four categories, independently, by two raters (the first author and a second annotator):

- **Reliable:** clinically appropriate and consistent with vital signs.
- **Incomplete:** partially correct but missing critical information.
- **Contradictory:** contains statements that conflict with physiological evidence.
- **Unsafe:** provides false reassurance in the presence of abnormal vital signs.

Inter-rater agreement, computed with Cohen’s kappa [19] on the two raters’ independent labels before reconciliation, was $\kappa = 0.594$ (validation, $n = 249$) and $\kappa = 0.585$ (test, $n = 249$), both “moderate agreement” on the Landis and Koch scale; 69/249 validation items and 70/249 test items (27.7% and 28.1%) received different labels. Rather than discard these or default to one rater, each was resolved by joint review of the vitals, NEWS2 category, and response text to a single final label — the adjudicated label, not either rater’s original, is the ground truth used throughout Section IV. This two-stage protocol is standard for subjective clinical-text annotation and lets κ measure the task’s subjectivity directly (a limitation discussed in Section V-C).

Each scenario is also assigned one of three physiological categories — stable, borderline, or critical — a controlled grouping variable, not the reliability ground truth (which comes from the four-way annotation above). The dataset was split into training (60%, $n = 749$), validation (20%, $n = 249$), and test (20%, $n = 249$) sets, stratified by physiological category and data source.

C. Local LLM Configuration

The edge LLM layer uses lightweight quantized GGUF models — Microsoft Phi-2 (2.7B), TinyLlama-1.1B, and Google Gemma-2B — run through llama-cpp-python with a 512-token context window reflecting Raspberry Pi-class memory and latency constraints, decoding temperature 0.3 for stable responses, and output limited to 128 tokens for short alert generation.

We evaluate three quantization levels: Q4_0 (4-bit), Q5_K_M (5-bit), and Q8_0 (8-bit) — Q4_0 prioritizes speed and memory efficiency, Q8_0 accuracy, and Q5_K_M a middle ground. For each model-quantization pair, we measure generation latency, tokens per second, RAM consumption, and CPU utilization.

A typical prompt contains the patient’s vital signs and asks the model to produce a brief assessment and alert level, e.g., “Patient vital signs: Heart rate 118 bpm, SpO₂ 87%,

Temperature 39.2°C, Respiratory rate 28/min, Systolic BP 95 mmHg. Generate a short non-diagnostic alert using categories: normal, warning, or critical.” The generated response is never delivered directly; it is first evaluated by MedEdge-Conf.

D. MedEdge-Conf Composite Score

MedEdge-Conf computes a composite score from three components:

$$S = \alpha C_1 + \beta C_2 + \gamma C_3 \quad (1)$$

where, C_1 (physiological coherence: contradiction between text and NEWS2-aligned vitals; very low edge cost), C_2 (token entropy: uncertainty from output log-probabilities; low edge cost), and C_3 (semantic stability: sentence-embedding cosine similarity; moderate edge cost) are normalized to $[0, 1]$. In the initial heuristic configuration, $\alpha = 0.50$, $\beta = 0.25$, and $\gamma = 0.25$, fixed a priori per the design rationale detailed in the following subsections. These weights are subsequently calibrated using logistic regression on the validation set (yielding w_1, w_2, w_3 , Section III-H), with the three normalized components as input features to a calibrated probability of reliability, the final composite score driving the deliver/review/escalate decision.

E. C_1 : Physiological Coherence Score

C_1 verifies response consistency with the raw vital signs through a two-stage process.

Stage 1: NEWS2 Score Calculation. For each scenario, the NEWS2 aggregate score is computed from the vital signs (respiration rate, SpO2, systolic blood pressure, pulse, consciousness, temperature, supplemental oxygen) and mapped to the risk category defined in Section II-A.

Stage 2: Linguistic Contradiction Detection. We parse the LLM response for keywords indicating reassurance (“normal”, “stable”, “within range”, “satisfactory”, “no concern”, “fine”) or alarm (“critical”, “warning”, “abnormal”, “concerning”, “seek help”, “urgent”). If the NEWS2 category is medium or high but the response contains reassuring terms, C_1 applies a strong penalty; if NEWS2 is low but the response contains alarm terms without justification, a milder penalty avoids unnecessary anxiety. C_1 is bounded between 0 and 1, where 1 means no contradiction and 0 means a severe contradiction.

This component is deliberately interpretable — when C_1 lowers the score, the system reports the specific reason, e.g., “NEWS2=7 (high risk) but response states ‘stable’ ” or “SpO2 84% described as normal” — making it easier to audit than a purely neural verifier, and computationally cheap, requiring only dictionary lookups and arithmetic suitable for Raspberry Pi 5.

Stage 2a’s keyword classifier has known limitations: negated or hedged constructions (e.g., “no longer stable”) are not reliably handled and can invert the contradiction signal if left unaddressed [18]. Testing on annotated data exposed a second gap: Stage 2a compares keywords only to the *overall* risk category and misses checkable single-parameter claims, such as “systolic BP is above normal (126 mmHg)” when 126 mmHg is within the NEWS2 normal band (111–219 mmHg)

— an error that need not change the overall category or involve a keyword mismatch.

We therefore added Stage 2b: a per-parameter numeric contradiction check that scans the response (via parameter-name and directional-claim regular expressions, e.g. “heart rate ... above normal”) for an explicit directional claim (high/low/normal) on each of the five core vitals, compares it against that parameter’s true direction computed from the value and the NEWS2 Scale 1 normal band, and applies Stage 2a’s severe penalty on any mismatch (normal claimed abnormal, abnormal claimed normal, or wrong direction), independent of the overall NEWS2 category. Like Stage 2a, this regex/keyword-window approximation can miss indirect phrasing or unusual sentence structure.

On the annotated test set ($n = 249$), the combined C_1 detector (Stages 2a+2b) achieves standalone precision=74.4% and recall=79.2% against the ground-truth “unsafe or contradictory” label (adjudicated two-annotator ground truth, Section III-B), independent of the downstream composite score. A negation-aware layer (e.g., NegEx [18] or a lightweight learned classifier), and extending Stage 2b for co-occurring reassurance/alarm keywords that currently cancel out under Stage 2a’s logic, are future work (Section V-C).

F. C_2 : Token Entropy Score

C_2 estimates the uncertainty of the LLM response from token-level log-probabilities. For each generated token, the probability distribution over the vocabulary is converted into Shannon entropy:

$$H_t = - \sum_i p_i \log(p_i) \quad (2)$$

where, p_i is the probability of the i -th candidate token at position t . The mean entropy over the response is normalized into a confidence score using a sigmoid transform calibrated on the validation set:

$$C_2 = \frac{1}{1 + \exp(k(H_{mean} - H_0))} \quad (3)$$

where, H_0 is the median validation-set entropy and $k = 2.0$ is a steepness parameter. Higher entropy indicates less certainty about the next-token decision. C_2 reuses information already produced during inference, needing no additional model, but a response can be low-entropy yet medically wrong, so it is combined with C_1 and C_3 rather than used standalone.

G. C_3 : Semantic Stability Score

C_3 estimates response stability by comparing the delivered response (temperature=0.3) against one additional generation for the same prompt at temperature=0.7 (perturbed), using semantic similarity from lightweight sentence embeddings from the all-MiniLM-L6-v2 model (22M parameters), which runs efficiently on Raspberry Pi 5 via ONNX Runtime.

Semantic similarity is the cosine similarity between the two embedding vectors:

$$C_3 = \frac{e_1 \cdot e_2}{\|e_1\| \|e_2\|} \quad (4)$$

High similarity suggests stable generation; low similarity suggests sensitivity to decoding variation. Unlike Jaccard similarity over tokens, cosine similarity over embeddings captures semantic equivalence across different wording [16].

Where, the two responses differ in length by more than 50%, a length-normalization factor prevents trivial mismatches from dominating the score. The output is bounded to $[0, 1]$.

H. Score Calibration and Decision Policy

The three components are normalized, then combined via two fusion strategies: 1) heuristic fusion, fixed weights $\alpha = 0.50$, $\beta = 0.25$, $\gamma = 0.25$; and 2) calibrated fusion, logistic regression trained on the validation set to predict “reliable vs. unreliable” (unreliable includes contradictory and unsafe responses).

The logistic regression model learns optimal weights and a bias term:

$$P(\text{reliable}) = \sigma(w_1 C_1 + w_2 C_2 + w_3 C_3 + b) \quad (5)$$

On the validation set ($n = 249$, adjudicated ground truth, Section III-B), the calibrated weights converged to $w_1 = 2.393$, $w_2 = 0.319$, $w_3 = -0.129$, bias = -1.549 , confirming the dominance of physiological coherence but revealing an unexpected pattern: w_3 (semantic stability, C_3) is negative — higher stability between the two temperature samples was associated with *lower* reliability. A plausible hypothesis (Section V-A) is that a confidently wrong response (e.g., “vital signs are normal”) tends to repeat nearly verbatim across both samples, while a more nuanced response shows natural wording variation. 5-fold cross-validation confirms each weight’s sign is stable across folds ($w_1 = 2.345 \pm 0.161$, $w_2 = 0.292 \pm 0.121$, $w_3 = -0.120 \pm 0.191$, bias = -1.510 ± 0.180), though w_3 ’s magnitude is small relative to its fold-to-fold standard deviation.

The calibrated confidence score controls a three-tier routing policy: deliver (score ≥ 0.43), review ($0.40 \leq \text{score} < 0.43$), and reject/escalate (score < 0.40). These thresholds were selected by grid search over deliver/reject candidates in $[0.05, 0.95]$ at 0.01 resolution (the same grid used for every threshold-fitting step in this study) to maximize F2-score on the validation set, prioritizing recall of unsafe/contradictory responses over precision, subject to a minimum review-band width of 0.02 enforced for all three models so this band cannot collapse to near-zero width regardless of the unconstrained F2 optimum (motivated by the narrow-review-band limitation discussed in Section V-C); on the test set, the review band now contains 1 of 249 scenarios (Section IV-F, Case 1). A parallel threshold pair was fit for the heuristic-fusion score by the same procedure (deliver ≥ 0.54 , reject < 0.52), since the two scores are on different scales.

Beyond discrimination metrics (AUROC, AUPRC), we assess whether $P(\text{reliable})$ is well-calibrated statistically — i.e., whether a predicted probability of roughly 0.7 corresponds

empirically to reliability about 70% of the time — using Expected Calibration Error and Brier score [20], reported in Section IV-B.

I. Statistical Validation Protocol

All point estimates in Section IV (FAR, AUROC, AUPRC, precision, recall, ECE, Brier score) are computed on the $n = 249$ test set (Section III-B), with 95% confidence intervals via nonparametric bootstrap resampling (2,000 resamples) and DeLong’s test [21] for AUROC-difference significance between scoring configurations. Bootstrap CIs and DeLong’s test cover all pairwise comparisons discussed in the text: Table I (Section IV-B), Table II (Section IV-C), and Table III (Section IV-D). Because DeLong’s test compares AUROC rather than the thresholded delivery decision FAR is computed from, paired McNemar’s exact tests supplement it for the FAR comparisons in Section IV-B. Since each table reports many pairwise tests, we applied Benjamini–Hochberg FDR correction ($q = 0.05$) within each table’s DeLong-test family (Tables I–III) and within Table I’s heuristic-vs-calibrated McNemar tests; every significance conclusion in Section IV is unchanged after correction. 5-fold cross-validation of the calibrated weights (w_1 , w_2 , w_3) was completed for all three models (Sections III-H, IV-G, IV-H).

IV. EXPERIMENTAL RESULTS

A. Dataset and Annotation Statistics

The test set comprises 249 scenarios: 84 stable, 82 borderline, and 83 critical, identical across all three evaluated models (Phi-2, TinyLlama, and Gemma-2B, all Q5_K_M). Sections IV-B–IV-D report results for all three models together; Sections IV-G–IV-H give each replicated model’s stats and weights. The Phi-2 LLM (a base model — Microsoft has not released an instruction-tuned Phi-2, discussed as a limitation in Section V-C) generated responses distributed, under the final adjudicated ground truth (Section III-B), as follows: 21 reliable, 103 incomplete, 84 contradictory, and 41 unsafe. Only 8.4% of responses were judged reliable — low enough that a safety layer is a precondition, not optional, for any patient-facing use.

B. Safety Filtering Performance

Table I presents the primary safety metrics for all three evaluated models, comparing MedEdge-Conf’s heuristic and calibrated fusion on the human-annotated test set.

The false acceptance rate (FAR) measures the proportion of contradictory or unsafe responses incorrectly classified as reliable and delivered. Calibrated fusion reduces FAR to 26.4% for Phi-2 (73.6% relative reduction), 17.3% for TinyLlama (82.7%), and 32.1% for Gemma-2B (67.9%), with AUROC improvement over naive scoring statistically significant for every model (DeLong’s test: Phi-2 $z = 2.219$, $p = 0.027$; TinyLlama $z = 3.839$, $p = 0.0001$; Gemma-2B $z = 4.299$, $p = 1.7 \times 10^{-5}$); heuristic fusion’s AUROC gain over naive is also significant for every model ($p \leq 0.016$). Precision is low in absolute terms across all models (10–23%) because “reliable” is a rare outcome (8.4%/8.0%/14.9% prevalence for Phi-2/TinyLlama/Gemma-2B): even a well-calibrated filter cannot reach high precision against a class this rare without

TABLE I. SAFETY FILTERING PERFORMANCE ON THE TEST SET, ALL THREE MODELS ($n = 249$)

Metric	Heuristic			Calibrated		
	Phi-2	TinyLlama	Gemma-2B	Phi-2	TinyLlama	Gemma-2B
FAR (% , 95% CI)	24.8 (17.2–32.5)	60.4 (52.2–68.6)	32.1 (23.4–41.2)	26.4 (18.9–34.2)	17.3 (11.1–23.9)	32.1 (23.7–41.3)
Unsafe delivered	11	11	0	11	2	1
Contradictory delivered	20	73	34	22	22	33
Precision (reliable)	13.3%	10.6%	21.8%	14.8%	14.0%	22.8%
Recall (reliable)	81.0%	100.0%	73.0%	95.2%	75.0%	83.8%
F1-score	22.8%	19.1%	33.5%	25.6%	23.6%	35.8%
AUROC (95% CI)	0.610 (0.512–0.698)	0.729 (0.613–0.831)	0.666 (0.578–0.753)	0.596 (0.509–0.676)	0.664 (0.577–0.747)	0.674 (0.593–0.753)

Naive delivery (all responses delivered) has FAR=100.0% (100.0–100.0) and AUROC=0.500 (0.500–0.500) for every model by construction; naive precision equals each model's reliable-response prevalence (Phi-2 8.4%, TinyLlama 8.0%, Gemma-2B 14.9%, Sections IV-A, IV-G, IV-H), recall is 100.0%, F1 is 15.6%/14.9%/25.9%. AUPRC (heuristic/calibrated): 0.104/0.098 (Phi-2), 0.198/0.112 (TinyLlama), 0.229/0.223 (Gemma-2B); naive AUPRC equals each model's reliable-response prevalence. ECE (heuristic/calibrated): 0.411/0.398, 0.604/0.383, 0.322/0.390; naive ECE is 0.916/0.920/0.851. Brier score (heuristic/calibrated): 0.287/0.270, 0.455/0.239, 0.260/0.289; naive Brier is 0.916/0.920/0.851. Calibrated-fusion review-tier counts (of 249, min. 0.02-wide band, Section III-H): Phi-2 1; TinyLlama 25; Gemma-2B 8.

being extremely conservative, at the cost of recall. (A paired McNemar's exact test likewise confirms naive vs. calibrated differs significantly for every model, $p < 10^{-20}$, trivially so since naive delivers every dangerous response by construction.)

Whether calibration improves on the fixed heuristic weights, however, is model-dependent. For Phi-2, it does *not* clearly help: heuristic fusion has a numerically lower FAR (24.8% vs. 26.4%) and higher AUROC (0.610 vs. 0.596), though the difference is not significant by AUROC (DeLong's test, $z = 0.357$, $p = 0.721$) or by McNemar's exact test on the delivery decision ($p = 0.50$). For TinyLlama, calibration is a large, unambiguous improvement, cutting FAR from 60.4% to 17.3%, confirmed by McNemar's exact test ($p = 2.7 \times 10^{-17}$). For Gemma-2B, heuristic and calibrated fusion land on an *identical* FAR (32.1%) because their deliver thresholds happen to bind on the same decision boundary (McNemar's exact test, $p = 1.00$); calibration's only measurable advantage there is in recall (83.8% vs. 73.0%) and F1 (35.8% vs. 33.5%). The practical value of data-driven calibration therefore varies by base model, discussed further in Section V-A.

Because the test set is CinC2019-majority (222/249 vs. 27/249 MIT-BIH), we checked whether calibrated-fusion performance differs by source. It does for Phi-2: AUROC drops from 0.616 to 0.340 (below chance) on MIT-BIH, significant by bootstrap on independent samples ($p = 0.013$), plausibly because MIT-BIH's non-ECG vitals are synthetic profile-fills C_1 may not handle as reliably as CinC2019's real panel. TinyLlama (0.682 vs. 0.600) and Gemma-2B (0.687 vs. 0.625) show the same direction but not significance at $n = 27$ ($p = 0.514$, $p = 0.800$).

C. Ablation Study

Table II presents the ablation study across all three models, comparing single-component and partial-combination configurations against the full MedEdge-Conf system on the same test set as Table I.

The three models show three distinct patterns, reported directly rather than reconciled into one story. For Phi-2, C_1 (physiological coherence) alone is the strongest configuration (AUROC=0.663) and *significantly outperforms* the full calibrated fusion (AUROC=0.596; DeLong's test, $z = 2.841$, $p = 0.0045$), though not heuristic fusion ($z = 1.627$, $p = 0.104$); $C_1 + C_3$ has the single highest point estimate

TABLE II. ABLATION STUDY AUROC, ALL THREE MODELS (TEST SET, $n = 249$)

Configuration	Phi-2	TinyLlama	Gemma-2B
C_1 only	0.663	0.607	0.635
C_2 only	0.315	0.625	0.588
C_3 only	0.553	0.367	0.639
$C_1 + C_2$	0.580	0.681	0.651
$C_1 + C_3$	0.675	0.505	0.675
$C_2 + C_3$	0.323	0.593	0.676
Heuristic (all 3)	0.610	0.729	0.666
Calibrated (all 3)	0.596	0.664	0.674

FAR values for the partial-component rows vary by model and configuration (Section III-H) rather than the discrimination shown here; only the full heuristic/calibrated rows are used for delivery decisions, and their FAR/precision/recall match Table I.

of any configuration for any model (AUROC=0.675), also significantly ahead of the calibrated fusion ($z = 3.008$, $p = 0.0026$) but not of C_1 alone ($z = 0.468$, $p = 0.640$). C_2 (token entropy) alone scores *below* chance for Phi-2 (AUROC=0.315). For TinyLlama, the opposite holds: the full heuristic fusion (AUROC=0.729) *significantly outperforms* C_1 alone (AUROC=0.607; $z = 2.253$, $p = 0.024$), and C_2 is a genuinely useful signal (AUROC=0.625) rather than a harmful one; the calibrated fusion (AUROC=0.664) is not significantly different from C_1 alone ($z = 1.391$, $p = 0.164$) or heuristic fusion ($z = 0.925$, $p = 0.355$). For Gemma-2B, a third pattern emerges: the full calibrated fusion (AUROC=0.674) again *significantly outperforms* C_1 alone (AUROC=0.635; $z = 2.716$, $p = 0.0066$), but here the improvement is driven by C_3 , not C_2 ($C_1 + C_3$ AUROC=0.675 is not significantly different from the full fusion, $z = 0.074$, $p = 0.941$, but is significantly better than C_1 alone, $z = 2.349$, $p = 0.019$); C_2 alone (AUROC=0.588) is neither clearly harmful nor clearly useful. C_3 is thus weak-to-harmful for Phi-2 and TinyLlama (AUROC 0.553 and 0.367) but is Gemma-2B's single strongest individual component (AUROC=0.639) — a third distinct role for the same signal across three models, interpreted further in Section V-A.

C_1 's standalone precision/recall against the adjudicated ground truth also varies sharply by model: 74.4%/79.2% for Phi-2, 98.0%/34.5% for TinyLlama (near-perfect precision but low recall, since TinyLlama's contradictions are disproportionately driven by fabricated non-vital details — invented medical history, medication, or diagnoses — a failure mode C_1 's

keyword/numeric-contradiction design was not built to catch), and 48.7%/72.6% for Gemma-2B (a more heterogeneous mix of the numeric-mismatch pattern C_1 targets and other failure types). Rows without C_1 are near or below chance for Phi-2 and TinyLlama but not Gemma-2B, showing physiological grounding is not equally load-bearing across models.

D. Baseline Comparison

We compare MedEdge-Conf against three baseline approaches from the literature: 1) Token-Entropy Only, a threshold on mean token entropy [9]; 2) SelfCheckGPT, sampling-based consistency with Jaccard similarity [17]; and 3) Semantic Entropy, entropy over semantic clusters of multiple generations [26], approximated here via embedding-based clustering rather than the original NLI-based bidirectional entailment clustering, since implementing a full NLI model was out of scope for an edge-resource-constrained comparison (documented in the implementation and revisited in Section V-C).

MedEdge-Conf has the highest point-estimate AUROC and lowest FAR of all four methods for every model, but the strength of its advantage is model-dependent. For Phi-2, MedEdge-Conf significantly outperforms only Token-Entropy Only ($p = 0.0003$) — the same signal Section IV-C already showed performs below chance in isolation (C_2 only, AUROC=0.315, since both reuse the identical entropy computation); its margin over the two sampling-based baselines does not reach significance at $n = 249$ ($p = 0.208$ and $p = 0.202$). For TinyLlama and Gemma-2B, MedEdge-Conf significantly outperforms all three baselines, including both sampling-based ones (TinyLlama: $p = 0.0062$ and $p = 0.0004$; Gemma-2B: $p = 0.00037$ and $p = 1.6 \times 10^{-6}$). Token-Entropy Only itself performs below chance for Phi-2 (AUROC=0.315) and Gemma-2B (AUROC=0.412) but is a reasonably strong baseline for TinyLlama (AUROC=0.625), consistent with Table II's finding that C_2 's usefulness varies by model. SelfCheckGPT and Semantic Entropy perform at or below chance for TinyLlama (0.427, 0.410) and Gemma-2B (0.481, 0.421) — the weakest baseline showing of the three models — strengthening the case for grounding the confidence score in NEWS2-aligned physiological rules over generic language-model uncertainty. Latency is the one advantage that holds unconditionally (Fig. 2): MedEdge-Conf's one extra generation is 2.6–2.9 $\times$ faster than the five-sample sampling-based baselines for every model, since its extra cost is a small fixed multiple of baseline generation while the five-sample baselines scale linearly with sample count.

E. Edge Performance and System Metrics

Table IV reports generation and scoring latency, throughput, RAM, CPU utilization, and SoC temperature across all nine model-quantization configurations, measured on the actual target device (a Raspberry Pi 5, 8 GB RAM, active cooling), using the same 249 test scenarios as Tables I–III. Each configuration was run as its own process to avoid allocator-fragmentation effects on later configurations' reported RAM.

TinyLlama is the fastest and lightest configuration at every quantization level (24.5–33.7s total latency, under 2.5 GB RAM, comfortably inside the Pi 5's 8 GB budget), reinforcing

Discrimination vs. Latency: MedEdge-Conf vs. Baselines

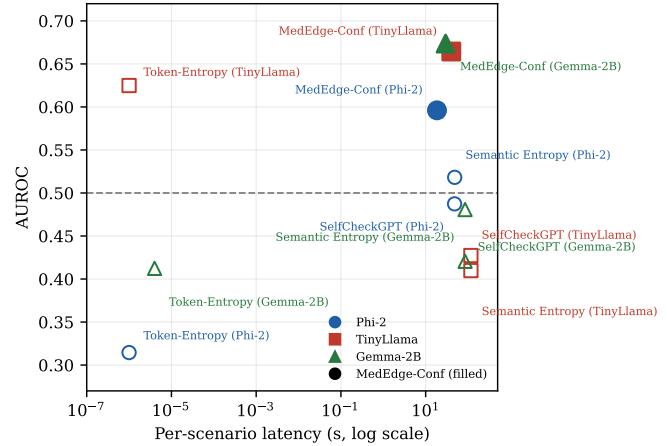


Fig. 2. AUROC vs. per-scenario latency (log scale), MedEdge-Conf vs. three baselines, all three models (Table III).

its deployment recommendation with both the strongest safety numbers (Sections IV-B, IV-D: lower FAR than Phi-2, 17.3% vs. 26.4%, and significant against all three literature baselines) and the fastest on-device performance of the three models. Gemma-2B is markedly slower (52.2–81.2s, 1.0–1.2 tokens/s) and heaviest (6.33 GB RAM at Q8_0, over three-quarters of the 8 GB budget; no OOM or swapping observed); the trade-off cuts both ways, since it is also the safest raw generator of the three by annotated label distribution (14.9% reliable, 0.8% unsafe vs. Phi-2's 8.4%/16.5% and TinyLlama's 8.0%/21.3%, Section IV-H).

A counter-intuitive pattern: Phi-2's Q4_0 (28.061s generation) is slower than its own Q5_K_M (15.583s) and Q8_0 (15.322s), plausibly because llama.cpp's legacy Q4_0 kernel is less well vectorized for this device's NEON pipeline than the newer K-quant paths — a known category of quantization-format-dependent variation not observed for TinyLlama or Gemma-2B, whose latency increases monotonically as expected. No thermal throttling occurred (SoC temperature 63.7–70.3°C across all nine configurations, vcgencmd get_throttled clean).

Table IV's figures are not directly comparable to Table III's baseline-comparison latencies (Section IV-D), which were measured on the development workstation, not the Pi 5, for relative-overhead comparison only; Phi-2 (Q5_K_M)'s real Pi 5 latency (34.574s) is roughly 1.9 $\times$ the workstation figure, illustrating why measuring the actual target device (Section III-A) matters for honestly reporting edge feasibility. Energy consumption (Wh) was not measured (no power meter available); this metric is omitted from Table IV, with a hook implemented for future use.

F. Qualitative Analysis

We present three test-set cases illustrating both successful operation and a current failure mode.

1) Case 1 (Review-tier catch): A borderline patient (NEWS2=6). LLM responds: "Patient vital signs: Heart rate

TABLE III. COMPARISON WITH BASELINE HALLUCINATION DETECTION METHODS, ALL THREE MODELS (TEST SET, $n = 249$)

Method	Phi-2			TinyLlama			Gemma-2B		
	AUROC	FAR	p vs. Ours	AUROC	FAR	p vs. Ours	AUROC	FAR	p vs. Ours
Token-Entropy Only	0.315	100.0%	0.0003	0.625	91.4%	0.035	0.412	100.0%	3.2e-5
SelfCheckGPT (Jaccard)	0.487	92.0%	0.208	0.427	97.8%	0.0062	0.481	100.0%	0.00037
Semantic Entropy (embed.)	0.518	30.4%	0.202	0.410	42.4%	0.0004	0.421	57.5%	1.6e-6
MedEdge-Conf (calibrated)	0.596	26.4%	—	0.664	17.3%	—	0.674	32.1%	—

“ p vs. Ours” is DeLong’s test for AUROC vs. MedEdge-Conf’s calibrated fusion for the same model. MedEdge-Conf’s per-scenario latency is 18.5s (Phi-2), 40.3s (TinyLlama), and 29.5s (Gemma-2B), vs. ~ 48.1 – 48.4 s / ~ 116.8 – 117.0 s / ~ 85.3 – 85.5 s for the sampling-based baselines (memory overhead 4.1–4.4 MB for MedEdge-Conf and Semantic Entropy, 0 MB otherwise). Pairwise DeLong comparisons among the three baselines themselves are reported in the released result files.

TABLE IV. RASPBERRY PI 5 HARDWARE BENCHMARK — ALL 9 MODEL $\times$ QUANTIZATION CONFIGURATIONS (REAL HARDWARE, $n = 249$ TEST SCENARIOS PER ROW)

Model	Quant.	Gen. Lat. (s)	Tok/s	Total Lat. (s)	RAM (GB)	CPU (%)	SoC Temp ($^{\circ}$ C)
Phi-2	Q4_0	28.061	3.73	52.925	3.54	42.3	63.7
Phi-2	Q5_K_M	15.583	2.61	34.574	4.43	50.7	70.3
Phi-2	Q8_0	15.322	2.05	34.346	5.97	54.0	67.5
TinyLlama	Q4_0	13.505	8.75	24.511	1.47	38.9	68.1
TinyLlama	Q5_K_M	16.828	7.08	30.132	1.75	41.8	70.3
TinyLlama	Q8_0	18.689	6.38	33.680	2.42	43.0	69.2
Gemma-2B	Q4_0	45.414	1.17	81.159	4.38	31.6	63.7
Gemma-2B	Q5_K_M	27.042	1.08	52.153	4.85	34.0	64.2
Gemma-2B	Q8_0	27.899	1.01	52.685	6.33	35.9	64.2

152 bpm, SpO2 96%, Temperature 36.7 $^{\circ}$ C, Respiratory rate 16/min, Systolic BP 87 mmHg. Category: Normal.” $C_1 = 0.40$ (Stage 2a fires only partially, checking the overall category, not the 152 bpm figure), $C_2 = 0.83$, $C_3 = 0.13$. Calibrated score=0.414. Both annotators labeled it unsafe. At the original unconstrained threshold (deliver ≥ 0.41), this item cleared delivery by a 0.004 margin, smaller than the score’s own bootstrap uncertainty, into a then-empty review tier. **Review**, under the minimum-width threshold now in use (Section III-H, deliver ≥ 0.43): the case that motivated the width constraint is now caught by it.

2) *Case 2 (Escalation)*: High-risk patient (HR 118 bpm, SpO2 100%, RR 34, SBP 88 mmHg, Temp 37.7 $^{\circ}$ C, on supplemental oxygen; NEWS2=10). LLM responds: “Patient’s heart rate is within normal limits.” $C_1 = 0.00$ (HR 118 bpm claimed normal, above the NEWS2 band of 51–90 bpm), $C_2 = 0.41$, $C_3 = 0.76$. Calibrated score=0.180. Rejected and escalated; the annotators’ disagreement (contradictory vs. unsafe) was resolved to unsafe.

3) *Case 3 (Failure mode)*: High-risk patient (HR 129 bpm, SpO2 78%, RR 28, SBP 82 mmHg, Temp 40.0 $^{\circ}$ C, ECG showing a premature ventricular contraction; NEWS2=13). LLM responds: “Warning. The patient has a premature ventricular contraction, which is a sign of arrhythmia. The patient’s vital signs are within normal limits, but the ECG indicates a potential cardiac problem. The patient should be monitored closely and receive further evaluation if the condition worsens or persists.” Both raters labeled this unsafe, but $C_1 = 1.00$: the response contains both an alarm keyword (“Warning”) and a reassurance keyword (“normal limits”), and Stage 2a only fires when reassurance appears without alarm, so the two cancel out. $C_2 = 0.69$, $C_3 = 0.01$. Calibrated score=0.743. **Delivered** (Sections III-E, V-C).

G. TinyLlama (Q5_K_M) Replication

We replicated the full generation, annotation, calibration, and evaluation pipeline (Sections III-B through III-I) on TinyLlama-1.1B (Q5_K_M), the deployment-recommended configuration, using the identical scenario set and two-annotator-plus-adjudication protocol as Phi-2. Two independent raters labeled all 498 TinyLlama responses; inter-rater agreement before adjudication was $\kappa = 0.568$ (validation) and $\kappa = 0.558$ (test), both moderate agreement, close to but slightly below Phi-2’s $\kappa = 0.594/0.585$ — consistent with TinyLlama’s more template-like output making four-way labeling marginally harder. 76/249 (validation) and 77/249 (test) items were resolved by adjudication. The test-set label distribution was 20 reliable (8.0%), 90 incomplete, 86 contradictory, and 53 unsafe (21.3%) — a higher unsafe rate and lower reliable rate than Phi-2 (16.5%, 8.4%), indicating TinyLlama is, if anything, a somewhat less safe raw generator despite its smaller footprint.

The calibrated weights (fit on TinyLlama’s own validation set) converged to $w_1 = 3.152$, $w_2 = 2.388$, $w_3 = -0.096$, bias = -4.029 , $h_0 = 0.486$ — w_1 dominant as for Phi-2, but w_2 (token entropy) large and positive rather than small, matching C_2 ’s genuine usefulness for this model (Section IV-C). 5-fold cross-validation confirms the sign of all three weights is stable across folds ($w_1 = 2.943 \pm 0.055$, $w_2 = 2.000 \pm 0.174$, $w_3 = -0.080 \pm 0.225$, bias = -3.657 ± 0.185). Results for TinyLlama are reported alongside Phi-2 and Gemma-2B throughout Tables I–III and Sections IV-B–IV-D.

H. Gemma-2B (Q5_K_M) Replication

We replicated the pipeline a third time on Gemma-2B (Q5_K_M), the largest and slowest of the three evaluated models (Table IV), using the identical protocol. Two independent raters labeled all 498 responses; inter-rater agreement before

adjudication was $\kappa = 0.631$ (validation) and $\kappa = 0.618$ (test) — the *highest* of the three models (Phi-2: 0.594/0.585; TinyLlama: 0.568/0.558), suggesting Gemma-2B’s responses were somewhat less ambiguous to categorize. 55/249 (validation) and 57/249 (test) items were resolved by adjudication — fewer than TinyLlama’s 76/77, closer to Phi-2’s 69/70. The test-set label distribution was 37 reliable (14.9%), 106 incomplete, 104 contradictory, and 2 unsafe (0.8%) — nearly double Phi-2/TinyLlama’s reliable rate and dramatically lower unsafe rate, making Gemma-2B a meaningfully safer raw generator, though still not reliable enough to deliver unfiltered (85.1% not judged fully reliable).

The calibrated weights converged to $w_1 = 1.754$, $w_2 = -0.979$, $w_3 = 0.200$, bias = -0.209 , $h_0 = 0.720$; w_2 ’s sign is negative here, the same direction as Phi-2’s small weight and the opposite of TinyLlama’s large positive w_2 — a third data point for the model-dependent-weighting conclusion of Section V-A. 5-fold cross-validation confirms the sign of all three weights is stable across folds ($w_1 = 1.716 \pm 0.084$, $w_2 = -0.832 \pm 0.112$, $w_3 = 0.190 \pm 0.129$, bias = -0.262 ± 0.126). Section III-H’s minimum-width threshold search leaves Gemma-2B’s review band unchanged, since its unconstrained F2 optimum already exceeded the 0.02 floor, but reshapes the other two: TinyLlama’s now receives the *most* review-tier traffic of the three models (Table I footnote, Section V-C), improving its calibrated FAR as a side effect (20.9% $\rightarrow$ 17.3%, Section IV-B); Phi-2’s goes from zero traffic to the Case 1 near-miss (Section IV-F). Results for Gemma-2B are reported alongside Phi-2 and TinyLlama throughout Tables I–III and Sections IV-B–IV-D.

V. DISCUSSION

A. Interpretation of Results

MedEdge-Conf addresses a clinically meaningful problem: local LLM inference alone is insufficient for safe patient-facing monitoring, as Section IV-A shows directly — only 8.4% of Phi-2 (Q5_K_M)’s responses were judged reliable. The core design choice is separating response generation from delivery: the LLM is a generator whose output must be checked, not an autonomous decision maker, and the safety-filtering results support this (calibrated fusion cuts FAR from 100.0% to 26.4%, $p = 0.027$). The framework is lightweight and thus well suited to edge-IoMT: C_1 is rule-based and explainable, C_2 reuses log-probabilities already produced during generation, and C_3 requires only one extra generation and a 22M-parameter embedding model — more suitable for Raspberry Pi 5 than large-verifier or cloud-scale approaches.

Section IV-C’s ablation study shows which components help is model-dependent, not fixed: C_1 alone is Phi-2’s strongest configuration, the full heuristic fusion is TinyLlama’s, and the full calibrated fusion (via C_3) is Gemma-2B’s. A plausible hypothesis for Phi-2’s pattern is that this base model is most confident — lowest entropy, most repeatable — precisely when confidently wrong (e.g., defaulting to “vital signs are within normal limits” regardless of the actual vitals), while more clinically engaged responses show more natural wording variation; if so, C_2 and C_3 as currently defined point in the wrong direction for some hallucination classes here. TinyLlama’s and Gemma-2B’s distinct dominant failure modes

(fabricated non-vital detail and a more heterogeneous contradiction mix, respectively) do not share this property, consistent with C_2/C_3 remaining informative for them. A fixed, model-agnostic weighting of $C_1/C_2/C_3$ is therefore unsupported; per-model (re)calibration (Section III-H) is doing necessary work.

B. Comparison with State of the Art

Section IV-D’s baseline comparison (Table III) supports grounding the confidence score in NEWS2-aligned physiological rules over generic language-model uncertainty: MedEdge-Conf significantly outperforms all three baselines (Token-Entropy Only [9], SelfCheckGPT [17], Semantic Entropy [26]) for TinyLlama and Gemma-2B, where the baselines are weak or near/below chance, but only Token-Entropy Only for Phi-2, and the case is weaker for TinyLlama, where Token-Entropy Only alone remains a viable signal. Latency favors MedEdge-Conf unconditionally across all three models.

C. Limitations and Future Work

Several limitations must be acknowledged:

First, several components remain simplified: NEWS2 uses a single-timepoint version rather than continuous trends; the sentence embedding model (all-MiniLM-L6-v2) is general-domain rather than medical-domain; and C_3 estimates stability from only two generations per prompt rather than the five-to-twenty typical of SelfCheckGPT [17]. C_3 ’s negative calibrated weight for two of three models (Sections III-H, V-A) remains only partly explained.

Second, C_1 ’s two-stage detector (Section III-E) improved from AUROC=0.517 (Stage 2a alone) to 0.663 with Stage 2b added (Table II), but two limitations remain: 1) negated or hedged constructions are not reliably handled by either stage, documented in the clinical NLP literature [18]; and 2) Stage 2a’s reassurance-vs-alarm logic cancels out when both keyword types co-occur in the same response, which Case 3 (Section IV-F) shows caused an unsafe response to be delivered despite an overall-correct composite score.

Third, several scope limitations bound generalization. Weights and thresholds were calibrated on retrospective scenarios from PhysioNet/CinC2019 and MIT-BIH (Section III-B), not a live deployment. MedEdge-Conf operates at the single-response level without patient history or temporal checks, and prospective clinical validation with real patient data and clinician feedback remains necessary before deployment. Relatedly, Phi-2 (Q5_K_M), the primary evaluated model, is a base model whose zero-shot behavior was, in early testing, dominated by drifting into Python code completion rather than alert text, mitigated by appending a plain completion cue (“Alert:”); a differently-prompted or differently-tuned Phi-2 could plausibly shift the reliability distribution reported in Section IV-A.

Fourth, the evaluation covers three open-weight models (Phi-2, TinyLlama, Gemma-2B) at one quantization tier each (Q5_K_M); results may not generalize to larger models, other architectures, or quantization levels, though Table IV covers three tiers per model. The ground truth rests on inter-rater agreement ($\kappa = 0.594/0.585$ for Phi-2, 0.568/0.558 for TinyLlama, 0.631/0.618 for Gemma-2B); the moderate-to-substantial agreement reflects the labeling task’s inherent subjectivity, not annotator carelessness.

Fifth, because MedEdge-Conf's review tier routes uncertain cases to a human reviewer, its practical value depends on how clinicians respond to alerts; clinical decision-support systems document high override rates for frequent or unspecific alerts [25]. Enforcing a minimum 0.02-wide review band (Section III-H), motivated by Case 1's original near-miss delivery (Section IV-F), gives every model non-zero review-tier traffic (Phi-2 1/249; TinyLlama 25/249, 10.0%; Gemma-2B 8/249, 3.2%; Table I footnote) and, for TinyLlama, incidentally improved calibrated FAR (20.9% $\rightarrow$ 17.3%); Phi-2's single review item remains a thin margin for a tier meant to catch borderline cases at scale. Human-in-the-loop evaluation of override behavior and alert burden remains important future work.

Finally, MedEdge-Conf is a safety layer, not a diagnostic tool; alerts remain non-diagnostic and informational, and the framework supports GDPR compliance by escalating only uncertain or high-risk cases to the cloud while raw physiological streams stay local [14]. Patient-facing deployment would still require clinician review, risk analysis, and regulatory consideration (e.g., EU MDR, FDA guidance on AI/ML-based SaMD).

VI. CONCLUSION

This study proposed MedEdge-Conf, a lightweight confidence-scoring framework for edge-deployed LLMs in IoMT health monitoring, combining physiological coherence aligned with NEWS2, token entropy, and semantic stability into a calibrated composite score that drives a send/review/escalate decision policy.

Experimental results on scenarios from PhysioNet/CinC2019 and MIT-BIH, using Phi-2, TinyLlama, and Gemma-2B (all Q5_K_M) and human annotation of all 1,494 generated responses (498 per model) by two independent raters with adjudicated disagreements (Cohen's $\kappa = 0.594/0.585$ for Phi-2, $0.568/0.558$ for TinyLlama, $0.631/0.618$ for Gemma-2B), demonstrate that MedEdge-Conf's calibrated fusion reduces the false acceptance rate of dangerous hallucinations from 100.0% to 26.4% for Phi-2, 17.3% for TinyLlama, and 32.1% for Gemma-2B, with AUROC improvement over naive scoring statistically significant in every case (DeLong's test, $p = 0.027$, $p = 0.0001$, $p = 1.7 \times 10^{-5}$). This required extending C_1 with a numeric contradiction check after the unextended version performed close to chance on annotated data. The ablation study surfaces a model-dependent pattern a single- or two-model evaluation would not have revealed: C_1 alone is Phi-2's strongest configuration, the full heuristic fusion is TinyLlama's (with token entropy a genuine contributor), and the full calibrated fusion is Gemma-2B's (driven by semantic stability rather than token entropy) — component weighting should therefore be calibrated per deployed model rather than fixed a priori. Against three literature baselines, MedEdge-Conf achieves the best AUROC and FAR for all three models, reaching significance over all three baselines for TinyLlama and Gemma-2B but only Token-Entropy Only for Phi-2 (Section IV-D). Edge feasibility measured on the physical Raspberry Pi 5 shows TinyLlama is the fastest and lightest of the three models, confirming the deployment recommendation, while Gemma-2B is markedly slower and

heavier despite being the safest raw generator by annotated label distribution.

The contribution is not merely deploying a quantized LLM on edge hardware, but adding a safety-oriented trust layer that prevents unreliable patient-facing communication, evaluated with a commitment to reporting what the data show, including where it complicates the original design's assumptions, on the same footing for all three models. Immediate next steps are resolving the keyword-cooccurrence gap in C_1 identified by Case 3 (Section IV-F) and investigating TinyLlama's fabricated-non-vital-detail failure mode (Section IV-C); the remaining directions are detailed in Section V-C.

REFERENCES

- [1] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [2] Royal College of Physicians, "National Early Warning Score (NEWS) 2: Standardising the Assessment of Acute-Illness Severity in the NHS," London, 2017.
- [3] A. E. W. Johnson et al., "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, 160035, 2016.
- [4] M. A. Reyna et al., "Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019," *Computing in Cardiology*, pp. 1–4, 2019.
- [5] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH Arrhythmia Database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.
- [6] M. Ardakani, J. Malekar, and R. Zand, "LLMPi: Optimizing LLMs for High-Throughput on Raspberry Pi," *CVPR Workshops*, 2025.
- [7] G. Gerganov et al., "llama.cpp: LLM inference in C/C++," open-source project, 2023–2026.
- [8] J. Wu, Y. Yu, and H.-Y. Zhou, "Uncertainty Estimation of Large Language Models in Medical Question Answering," arXiv:2407.08662, 2024.
- [9] A. V. Varshney et al., "Token Probabilities to Mitigate Large Language Models Overconfidence in Medical Question Answering," *Journal of the American Medical Informatics Association*, 2025.
- [10] Z. Wang et al., "Word-Sequence Entropy: Towards Uncertainty Estimation in Free-Form Medical Question Answering Applications and Beyond," arXiv:2402.14259, 2024.
- [11] T. N. Nguyen and collaborators, "An Evaluation of LLMs Inference on Popular Single-Board Computers," arXiv preprint, 2025.
- [12] H. Lee and collaborators, "Probabilistic Medical Predictions of Large Language Models," *npj Digital Medicine*, 2024.
- [13] P. Rajpurkar et al., "AI in Health and Medicine," *Nature Medicine*, vol. 28, pp. 31–38, 2022.
- [14] European Parliament and Council of the European Union, "Regulation (EU) 2016/679: General Data Protection Regulation," *Official Journal of the European Union*, 2016.
- [15] A. E. W. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, 2023.
- [16] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *EMNLP-IJCNLP*, 2019.
- [17] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," *Proc. 2023 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9004–9017, 2023.
- [18] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries," *Journal of Biomedical Informatics*, vol. 34, no. 5, pp. 301–310, 2001.
- [19] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

- [20] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," *Proc. 34th Int. Conf. on Machine Learning (ICML)*, pp. 1321–1330, 2017.
- [21] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.
- [22] X. Liu, T. Chen, L. Da, C. Chen, Z. Lin, and H. Wei, "Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey," *Proc. 31st ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD)*, pp. 6107–6117, 2025.
- [23] T. Rebedea, R. Dinu, M. Sreedhar, C. Parisien, and J. Cohen, "NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails," arXiv:2310.10501, 2023.
- [24] E. Asgari, N. Montaña-Brown, M. Dubois, S. Khalil, J. Balloch, J. Au Yeung, and D. Pimenta, "A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation," *npj Digital Medicine*, vol. 8, no. 1, 274, 2025.
- [25] H. van der Sijs, J. Aarts, A. Vulto, and M. Berg, "Overriding of drug safety alerts in computerized physician order entry," *Journal of the American Medical Informatics Association*, vol. 13, no. 2, pp. 138–147, 2006.
- [26] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, pp. 625–630, 2024.
- [27] G. B. Moody et al., "MIMIC-III Waveform Database," PhysioNet, 2020.