PHYSIOSKEPTIC: SKEPTICAL REASONING GROUNDED IN SIGNAL QUALITY FOR MULTIMODAL RHYTHM ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can reason over physiological summaries, but standard waveform-to-text interfaces often discard beat-indexed evidence and signal quality, limiting performance under noise and cross-modal conflict. We introduce PhysioSkeptic for retrospective ECG–PPG rhythm classification. Its PhysioPatch encoder aligns waveform patches to cardiac landmarks and emits heartbeat-level tokens and a Patch Report containing rhythm cues, confidence estimates, and signal quality. This interface exposes beat-indexed physiological evidence together with modality-level quality for traceable conflict checks and quality-conditioned belief revision. On MIMIC-III-Ext-PPG, PhysioSkeptic reaches 0.886 Macro-F1, 6.2 percentage points above generic Debate and 4.1 points above an encoder-matched control; it has the lowest observed ECE among GPT-5.2 methods and reduces anchor-trap retention from 56.6% to 16.8%. Removing SQI anchoring produces the largest observed reduction among reasoning controls. In a zero-shot SQI proxy-transfer audit on MC-MED, learned SQI remains informative without retraining, while an auxiliary weak-label AF stress test improves AUROC from 0.913 to 0.934. A distilled Llama-1B student reaches 0.847 Macro-F1 with local inference. These results show that exposing beat-indexed evidence together with signal quality can improve the traceability of multimodal rhythm reasoning under degraded measurements.

1 INTRODUCTION

Continuous physiological monitoring in intensive care and emergency settings produces high-volume ECG–PPG streams. These signals support rhythm analysis, but motion, electrode displacement, perfusion changes, and ambiguous class boundaries complicate automated interpretation (Wankhede et al., 2025; Khalili et al., 2024). Failure to identify atrial fibrillation, rate abnormalities, or paced-rhythm patterns may delay clinical assessment, particularly when they are associated with hemodynamic instability (Lubitz et al., 2010; Jin et al., 2024).

Large language models (LLMs) are increasingly used for clinical reasoning and waveform-informed diagnosis (Wang et al., 2023; Liu et al., 2025a; Christof & Armoundas, 2025; Guo et al., 2025). A common pipeline compresses raw waveforms into a textual summary before prompting the LLM (Liu et al., 2025b; Zhang et al., 2024c). This step removes beat-indexed reliability from the reasoning chain (Qian et al., 2025; You et al., 2025): a model can produce fluent explanations without retaining which observations were noisy or conflicting. Reliable rhythm analysis must instead propagate waveform quality to each reasoning step and flag cross-signal conflict before assigning a confident label.

A fixed encoder may learn morphology without exposing beat-indexed reliability; a text-only LLM can reason over a report but cannot audit its signal quality. We study two failure modes that follow from this missing interface. First, *anchoring bias*: autoregressive reasoning can preserve an early, incorrect rhythm hypothesis even after later evidence contradicts it (Lou & Sun, 2026; Zhang et al., 2024b; Turpin et al., 2023; Zhou et al., 2024). Second, *conflicting-data integration failure*: when PPG-derived rhythm features and simultaneous ECG features disagree, CoT and self-consistency can privilege one modality without explicitly auditing the conflict, producing overconfident errors.

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

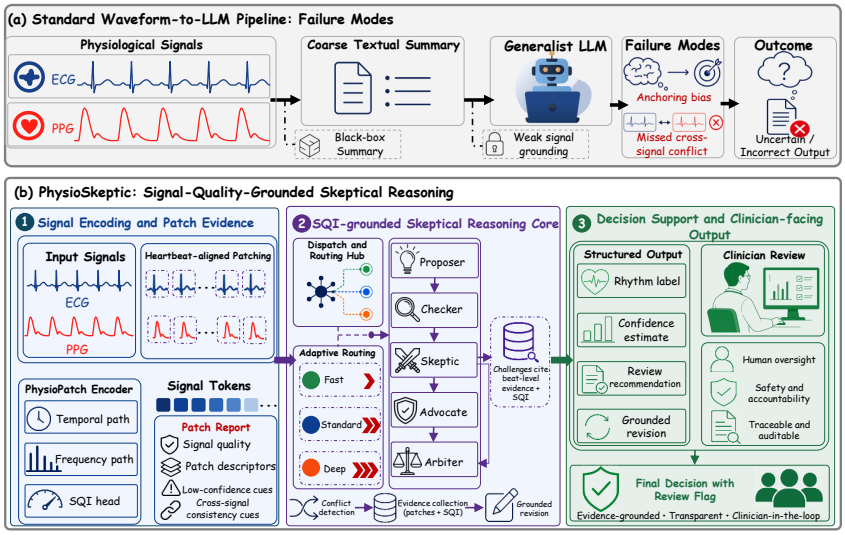

Figure 1: PhysioSkeptic overview with signal-quality-grounded evidence and review.

Multi-agent debate introduces explicit disagreement (Du et al., 2024; Zhuge et al., 2025), but standard critiques are usually linguistic rather than evidence-grounded (Liang et al., 2024; Du et al., 2024). A dissenting agent may object without citing the waveform segment or quality estimate that supports it. This is hard to audit clinically, especially when signal quality varies over time and modality (Banerji et al., 2023).

We propose **PhysioSkeptic**, a framework that turns beat-indexed physiological evidence and modality-level signal quality into traceable inputs for structured skeptical reasoning. Rather than using the Signal Quality Index (SQI) (Elgendi, 2016) as a rejection gate, its **PhysioPatch** encoder aligns ECG R-peaks and PPG systolic onsets (Davies et al., 2024a), emits a numerical Patch Report, and exposes quality to a five-role workflow comprising Proposer, Checker, Skeptic, Advocate, and Arbiter. The Skeptic cites beat indices, time windows, and quality values, whereas the Arbiter returns class probabilities and an explicit review flag (Fig. 1). The empirical policy uses a deterministic, quality-conditioned score update rather than a fitted Bayesian posterior. The resulting reasoning framework for ECG–PPG rhythm classification makes each belief revision traceable to an external physiological anchor. Our contributions are:

- We introduce a structured Patch Report that exposes beat-indexed confidence, cardiac landmarks, time windows, modality-level quality, and rhythm cues as numerical evidence for downstream reasoning. Each challenge can therefore be traced to an explicit observation rather than to an unsupported textual objection.
- PhysioPatch aligns ECG and PPG patches to cardiac landmarks and combines temporal beat tokens with a frequency-path summary. The resulting interface supports both cross-modal checks and quality-conditioned routing.
- The reasoning procedure couples a five-role debate with an SQI-dependent challenge budget and a deterministic score update. This makes the boundary between quality-triggered intervention and the executable policy explicit.
- Matched experiments quantify the consequences of this interface through Macro-F1, calibration, low-SQI performance, anchor-trap recovery, corruption robustness, backbone transfer, and local Student inference.

2 Related Work

Physiological signal models. CNNs and Transformers define strong supervised baselines for arrhythmia detection, extending earlier handcrafted-feature pipelines with fixed patches and temporal reshaping (Rasti, 2024; Kovalchuk et al., 2025; Zhou & Wang, 2024; Nie et al., 2023; Akan et al., 2023; Chen et al., 2024). SSMs provide linear-time recurrence and are effective on long biosignal sequences (Qiang et al., 2024; Jiang et al., 2025; Hu et al., 2025a), but their dense hidden states

are not structurally aligned to individual heartbeats, making beat-specific evidence citation difficult. SQI-aware ECG pre-training has also been studied (Weimann & Conrad, 2025), typically using quality as an inference filter or representation target rather than as a variable that changes the reasoning policy. In contrast, PhysioSkeptic uses heartbeat-aligned tokens and exposes quality estimates to downstream routing, challenge pressure, and review flags. Related signal-to-text approaches show that learned PPG prefixes can coexist with LLM contexts (Davies et al., 2024b); our setting further requires the resulting reasoning steps to remain tied to traceable cardiac cycles rather than unstructured summaries.

Clinical LLM reasoning. Instruction-tuned LLMs have been evaluated on medical licensing, clinical text, and multistep diagnosis benchmarks (Kweon et al., 2024; Wu et al., 2024; Christophe et al., 2024; Guo et al., 2025). These benchmarks have also exposed anchoring and conflicting-evidence failures under noisy inputs (Schmidgall et al., 2024). However, they are mostly text-centered: they do not test whether a model can propagate waveform reliability that changes across beats, time windows, and modalities. This distinction matters for physiological monitoring, where the same rhythm label may be supported by high-quality ECG evidence and contradicted by unreliable PPG evidence. It also separates our setting from clinical-agent systems that coordinate textual case reasoning but do not expose an external signal quality variable to each intermediate challenge.

CoT, self-consistency, and debate. CoT, self-consistency, self-critique, and tree/chain search improve reasoning by increasing the amount or diversity of inference-time computation (Wei et al., 2022; Wang et al., 2023; Snell et al., 2025; Huang et al., 2024; Zhang et al., 2024d; Yao et al., 2023). Their critiques remain internal to the model unless tied to external observations. Multi-agent debate and Society-of-Mind frameworks introduce explicit disagreement (Du et al., 2024; Hu et al., 2025b; Zhang et al., 2024a), but the disagreement is usually linguistic: an agent can object without grounding the objection in a verifiable signal segment. Divergent-thinking variants broaden the space of candidate rationales (Liang et al., 2024), but they still optimize language-level disagreement rather than evidence-level admissibility. PhysioSkeptic instead requires each challenge to trace back to a beat index, time window, and SQI value, and then measures whether these challenges trigger belief revision. The SQI-dependent challenge pressure and routing depth are the main structural differences from prior debate methods.

3 METHODOLOGY

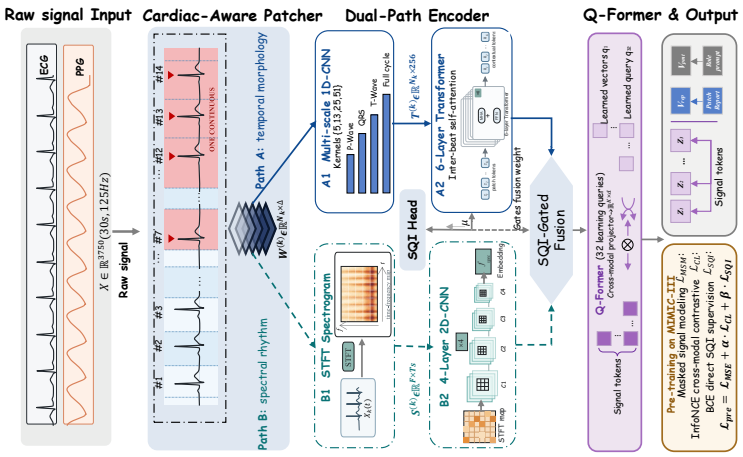

Figure 2: PhysioPatch dual-path encoder: landmark patching, dual-path features, and SQI-gated fusion.

3.1 PROBLEM FORMULATION

We formalize cardiac rhythm classification as inference of a latent rhythm state from partial, noisy ECG–PPG observations. Let $\mathcal{R} = \{\text{SR}, \text{STACH}, \text{SBRAD}, \text{AF_FAMILY}, \text{PACE}\}$ denote the five target rhythm classes. Each sample is represented as a synchronized 30-second ECG–PPG pair $\boldsymbol{X} = \left(\boldsymbol{x}_{\text{ECG}}, \boldsymbol{x}_{\text{PPG}}\right)$. The objective is to predict a normalized class distribution $\hat{\boldsymbol{p}} \in \Delta^{|\mathcal{R}|-1}$, its

label $\hat{r} = \arg\max_r \hat{p}_r$, a reported confidence $c \in [0, 1]$, and a structured natural-language rationale $\mathcal{C}$ for inspection.

3.2 Reliability-Conditioned Belief Revision

The pipeline treats reliability as an explicit control variable during belief revision. The encoder emits a modality/window-level quality estimate $\widehat{q}^{(k)} \in [0, 1]$, which is collapsed to the routing score $\widehat{q}$; the Patch Report separately records a beat-indexed confidence vector $\boldsymbol{c}$. Thus, $\boldsymbol{c}$ identifies which physiological observations should be inspected, whereas $\widehat{q}$ controls challenge pressure and route depth. The two quantities are related but are not interchangeable SQI measurements; their implementation mapping is summarized in Supplementary Table S3.

At inference, the Proposer initializes a class distribution $b_0(r) = P_{\mathrm{LLM}}(r \mid \boldsymbol{o})$. The Checker tests that proposal against cross-modal and quality evidence, and the Skeptic selectively targets the Patch Report items most likely to invalidate it. The Advocate and Arbiter then revise and normalize the running scores using the deterministic rule in Eq. (6). This formulation makes the implemented reliability-to-action path explicit without claiming that the LLM transcript is a fitted Bayesian posterior or that the routing policy is Bayes-optimal.

3.3 PhysioPatch: Cardiac-Aware Dual-Path Encoder

PhysioPatch (Figure 2) transforms each raw 30-second, 125 Hz signal $\boldsymbol{x}_k \in \mathbb{R}^{3750}$ ($k \in \{\mathrm{ECG}, \mathrm{PPG}\}$) into a sequence of beat-aligned token embeddings $\boldsymbol{H}^{(k)}_{\mathrm{fused}}$, a global frequency feature vector $\boldsymbol{h}^{(k)}_{\mathrm{freq}}$, and a scalar quality estimate $\widehat{q}^{(k)} \in [0, 1]$. The lightweight encoder (≈15 M parameters) keeps its contribution separable from the debate pipeline.

Cardiac landmarks define each patch. Let $\{p_1, \ldots, p_{N_k}\}$ denote cardiac-landmark indices in modality k, detected as R-peaks by Pan–Tompkins (Pan & Tompkins, 1985) for ECG and as systolic onsets (Elgendi et al., 2013) for PPG. The i-th heartbeat patch is defined as

$$\boldsymbol{w}^{(k)}_i = \boldsymbol{x}_k\big[p_i - \lfloor \Delta/2 \rfloor \;:\; p_i + \lceil \Delta/2 \rceil\big], \tag{1}$$

where $\Delta = \min_j(p_{j+1} - p_j)$ is the shortest inter-landmark interval in the window. This defines a fixed beat-centered neighborhood while limiting overlap between adjacent patches. Shorter patches are right-padded to form $\boldsymbol{W}^{(k)} \in \mathbb{R}^{N_k \times \Delta}$.

The encoder has two complementary paths. Path A applies a multi-scale 1D-CNN followed by a 6-layer Transformer to produce temporal beat tokens $\boldsymbol{H}^{(k)}_{\mathrm{temp}} \in \mathbb{R}^{N_k \times d}$; its inter-beat attention matrix $\boldsymbol{A}^{(k)}$ is retained for the Student loss in Eq. (7). Path B applies an STFT over 0.5–30 Hz and a 2D-CNN to produce a global rhythm-regularity embedding $\boldsymbol{h}^{(k)}_{\mathrm{freq}} \in \mathbb{R}^{128}$. Exact kernel sizes, STFT settings, and token formulas are in Suppl. C.3. The frequency cues are descriptive rather than definitive AF-vs-flutter markers because ICU instrumentation, medication, and autonomic state can distort spectral features.

A dedicated SQI head applied to the mean temporal embedding produces

$$\widehat{q}^{(k)} \;=\; \sigma\big(\mathrm{MLP}(\bar{\boldsymbol{H}}^{(k)}_{\mathrm{temp}})\big) \;\in\; [0, 1],$$

and the fused per-beat representation is

$$\boldsymbol{H}^{(k)}_{\mathrm{fused}} \;=\; \widehat{q}^{(k)} \cdot \boldsymbol{H}^{(k)}_{\mathrm{temp}} \;+\; (1 - \widehat{q}^{(k)}) \cdot \mathbf{1}_{N_k} \big(\boldsymbol{h}^{(k)}_{\mathrm{freq}}\big)^{\top} \boldsymbol{W}_f, \tag{2}$$

where $\boldsymbol{W}_f \in \mathbb{R}^{128 \times d}$ projects the frequency embedding to the temporal token dimension and $\mathbf{1}_{N_k} \in \mathbb{R}^{N_k}$ broadcasts across beats. When $\widehat{q}^{(k)} \approx 1$ the fused token retains precise per-beat morphological detail; when $\widehat{q}^{(k)} \approx 0$ it falls back to the frequency-domain rhythm summary.

ECG and PPG embeddings are concatenated and passed through a Q-Former-style cross-modal projector (Li et al., 2023) with 32 learned queries. All evaluated API-hosted LLM backbones receive the same serialized Patch Report and role-specific prompt. Continuous PhysioPatch embeddings are used for encoder-side prediction and local Student distillation, not injected into API calls.

In addition to continuous embeddings, PhysioPatch outputs a JSON Patch Report enumerating: (a) the per-beat confidence vector $\boldsymbol{c} \in [0,1]^N$; (b) a modality/window-level scalar quality summary $\widehat{q}$ used for routing and challenge scheduling; (c) low-confidence beats with beat index, time window, and attention weight; (d) the ECG-derived P-wave detection ratio; and (e) frequency-domain features including a low-/high-band power ratio. This report is injected verbatim into the Skeptic's prompt, providing numerical anchors that make each emitted challenge amenable to rule-based traceability checks. For notation, we write the serialized report as

$$\mathcal{P}(\boldsymbol{X}) = \{\boldsymbol{c}, \widehat{q}, \mathcal{L}_{\text{low}}, \rho_P, \boldsymbol{h}_{\text{freq}}\}, \tag{3}$$

where $\mathcal{L}_{\text{low}}$ is the indexed low-confidence beat list and ρ_P is the ECG-derived P-wave detection ratio. The API roles consume these serialized fields; continuous embeddings remain local to the encoder and Student distillation.

PhysioPatch is pre-trained on the unlabeled MIMIC-III-Ext-PPG corpus using masked beat reconstruction, cross-modal contrastive alignment, and supervised SQI prediction; the full objective and training recipe are provided in Suppl. C.3. Pre-training uses only training and validation manifests; no test-patient waveform or SQI label enters pre-training.

3.4 Five-Role Adaptive Skeptical Debate Pipeline

The reasoning pipeline comprises five role-specific LLM modules with predefined JSON output schemas; each prompt prohibits evidence absent from the Patch Report. Suppl. Fig. S1 shows the full message flow, and Algorithm 1 gives the complete inference procedure.

Proposer. The Proposer emits the initial rhythm class $\hat{r}_0$, confidence c_0, and three supporting Patch Report items. This initial record fixes the evidence that the downstream roles are allowed to inspect.

Cross-modal checker. The Checker compares heart-rate, morphology, hemodynamic, and confidence–quality cues, then records conflict type and severity in a structured manifest. Thresholds are listed in Suppl. C.3.

Skeptic. Standard and Deep routing invoke the Skeptic for up to three evidence-grounded challenges, whereas Fast routing skips this module. The challenge budget is

$$n_{\text{chal}} = \begin{cases} 0, & \text{depth}(\boldsymbol{X}) = \text{FAST}, \\ 3 - \lfloor 2\widehat{q} \rfloor, & \text{depth}(\boldsymbol{X}) \in \{\text{STANDARD}, \text{DEEP}\}, \end{cases} \qquad n_{\text{chal}} \in \{0,1,2,3\}. \tag{4}$$

Each challenge targets one Proposer evidence item, cites a numerical Patch Report value, and proposes an alternative interpretation. Low-quality or conflicting evidence is examined first.

Advocate (response module). The Advocate tests each challenge against cross-signal evidence, rejects unsupported claims, and updates the running confidence before arbitration.

Arbiter. The Arbiter returns the final prediction, confidence, uncertainty annotation, and a short reasoning summary. An "expert review recommended" flag is raised when $c_{\text{final}} < \tau_{\text{rev}} = 0.70$; the threshold is fixed on the validation split and is not a clinically calibrated alert (Supplementary Table S16). The routing function maps signal characteristics to debate depth:

$$\text{depth}(\boldsymbol{X}) = \begin{cases} \text{FAST} & \text{if } c_0 > 0.9 \wedge \widehat{q} > 0.8 \\ & \quad \wedge\ \text{conflict} = \varnothing \\ \text{DEEP} & \text{if } \widehat{q} < 0.5 \vee c_0 < 0.6 \\ & \quad \vee\ \text{conflict} = \text{MAJOR} \\ \text{STANDARD} & \text{otherwise,} \end{cases} \tag{5}$$

Here $\widehat{q}$ denotes the scalar routing score obtained by collapsing the per-modality estimates $\widehat{q}^{(k)}$. The Checker assigns NONE, MINOR, or MAJOR; a MAJOR conflict triggers Deep routing regardless of $\widehat{q}$. Fast invokes Proposer, Checker, and Arbiter; Standard adds the Skeptic, and Deep invokes all five roles. For Standard and Deep routes, rule-audited challenge penalties are combined with Patch Report class scores and the Proposer distribution through the deterministic SQIWeightedScoreUpdate before arbitration. Fast routing passes the Proposer distribution directly to the Arbiter.

For clarity, the empirical update can be written as a normalized score rule:

$$b_1(r) = \frac{[s_r]_+}{\sum_{u\in\mathcal{R}}[s_u]_+}, \qquad s_r = p_{0,r}\left[\widehat{q}\,u_r(\text{report}) + \frac{1-\widehat{q}}{|\mathcal{R}|}\right] - \sum_{j:\,\hat{r}_j=r}\lambda_{\ell_j} + T_r(\text{challenges}), \quad (6)$$

where $p_{0,r}$ is the Proposer distribution, u_r is the rule-audited Patch-Report score, and T_r transfers challenge mass toward a cited alternative class. The penalty level λ_{ℓ_j} is determined by the challenge severity (0.03, 0.08, or 0.15); the update is deterministic and is not presented as a fitted Bayesian posterior.

Student distillation. To support local low-latency inference, a compact Llama-3.2-1B Student is trained offline from PhysioSkeptic's class distribution, confidence scores, and low-quality beat attention. Suppl. C.2 gives the full loss and training recipe.

4 Experiments

4.1 Datasets and Evaluation Protocol

The primary benchmark is MIMIC-III-Ext-PPG (Moulaeifard et al., 2026) (v1.0.0), whose rhythm codes are mapped to SR, STACH, SBRAD, AF_FAMILY, and PACE. AF_FAMILY follows the benchmark taxonomy and does not resolve the AF-versus-flutter treatment decision. Bench-100 retains 100 windows per eligible patient with patient-disjoint splits. The matched LLM test subset contains five fixed-seed windows for each of 791 patients (3,955 windows), sampled after class–SQI stratification so all methods receive identical cases. This enriched benchmark is not prevalence representative. MC-MED (Kansal et al., 2025) is reserved for an 8,340-window zero-shot SQI-transfer audit. Its auxiliary weak-label AF and pacemaker analyses are stress tests rather than external diagnostic validation; MC-MED is not used for MIMIC training or tuning.

Macro-F1 is the primary metric because it weights the five rhythm classes equally. AUROC, 10-bin ECE, Brier, and NLL are secondary metrics. We report five-run means and Macro-F1 standard deviations descriptively; the repeated windows are clustered within patients, so the runs are used as stability checks rather than patient-level inferential replicates. The matched benchmark, component ablations, controlled corruption, backbone transfer, and routing analyses evaluate both average performance and the low-SQI/conflict cases that motivate the method. This design keeps the primary comparison fixed while testing whether the proposed evidence interface changes behavior specifically when signal quality and physiological evidence disagree.

4.2 Baselines

Baselines are grouped by the capability they add. L0 includes waveform-only 1D-CNN (Hannun et al., 2019), PatchTST (Nie et al., 2023), and TimesNet (Wu et al., 2023). L1 uses GPT-5.2 in 0-shot and 5-shot settings with the same Patch Report. L2 adds single-agent reasoning through CoT, Self-Consistency ($\times$5), and Self-Refine (Madaan et al., 2023); L3 uses generic multi-agent Debate (Du et al., 2024). Two method controls isolate our design: *PhysioSkeptic w/o SQI anchoring* removes SQI-dependent grounding, and *PP+SP* routes all samples through only Proposer and Arbiter, separating encoder gains from skeptical-debate gains. Together, these tiers make the comparison interpretable: waveform-only rows measure the signal encoder, L1–L3 measure increasingly structured reasoning, and the two controls test whether the gains require SQI-grounded evidence rather than additional prompting alone.

Table 1: Matched-subject benchmark on MIMIC-III-Ext-PPG. Bold and underline mark the best and runner-up values.

Tier	Method	Macro F1↑	AUROC↑	ECE↓	Brier↓	NLL↓	F1 by SQI stratum	
							Low (<0.5)	Non-low (≥0.5)
L0	1D-CNN	0.779±.009	0.945	–	0.197	–	0.674	0.866
	PatchTST	0.794±.002	0.951	0.044	0.195	0.621	0.693	0.871
	TimesNet	0.800±.008	0.955	–	0.187	–	0.698	0.881
L1	GPT-5.2 0-shot	0.678±.003	0.913	0.041	0.251	0.807	0.561	0.752
	GPT-5.2 5-shot	0.754±.001	0.937	0.045	0.218	0.675	0.635	0.831
L2	GPT-5.2 + CoT	0.774±.010	0.942	0.128	0.204	1.097	0.671	0.860
	GPT-5.2 + SC×5	0.792±.009	0.952	0.112	0.194	1.005	0.692	0.876
	GPT-5.2 + Reflect	0.782±.010	0.945	0.121	0.201	1.051	0.674	0.869
L3	Debate (Du et al., 2024)	0.824±.001	0.962	0.041	0.183	0.512	0.718	0.897
Diag.	PhysioSkeptic w/o SQI anchoring	0.822±.008	0.973	0.085	0.174	0.894	0.726	0.907
	PP+SP	*0.845*±.001	*0.970*	*0.037*	*0.168*	*0.449*	*0.748*	*0.918*
Ours	**PhysioSkeptic (GPT-5.2)**	**0.886**±.003	0.981	0.033	**0.148**	**0.343**	**0.779**	**0.941**
	PhysioSkeptic (DeepSeek-V3.2)	0.866±.002	0.976	**0.032**	0.199	0.392	0.748	0.912
	PhysioSkeptic (Qwen3.5-27B)	0.851±.008	0.971	0.061	0.171	0.771	0.738	0.902
	Student (Llama-1B)	0.847±.008	**0.986**	0.065	0.164	0.811	0.767	0.919

Table 2: Per-class F1 on the patient-matched LLM test subset. Values are five-run means with standard deviations; class prevalence is shown in the header.

Method	SR (26.4%)	STACH (17.9%)	AF_FAM. (23.5%)	SBRAD (12.9%)	PACE (19.3%)
PatchTST	0.704±0.003	0.746±0.006	0.825±0.003	0.761±0.007	0.932±0.001
Debate (Du et al., 2024)	0.744±0.001	0.774±0.006	0.865±0.002	0.796±0.007	0.941±0.002
PP+Single-pass	*0.771±0.003*	*0.789±0.004*	*0.897±0.002*	*0.822±0.006*	*0.947±0.001*
PhysioSkeptic (GPT-5.2)	**0.829±0.006**	**0.844±0.005**	**0.925±0.003**	**0.872±0.008**	**0.960±0.001**
PhysioSkeptic (DeepSeek-V3.2)	0.799±0.004	0.814±0.006	0.913±0.002	0.850±0.008	0.954±0.002

4.3 Main Results

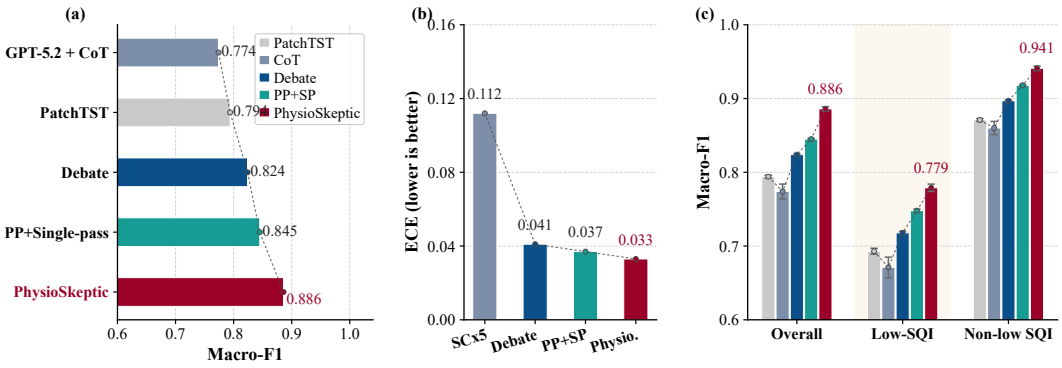

Figure 3: Benchmark Macro-F1, ECE, and SQI-stratified F1.

Table 1 and Fig. 3 use the same patient-matched test subset for every row. PhysioSkeptic reaches 0.886 Macro-F1, 6.2 pp above Debate, 4.1 pp above the encoder-matched PP+SP control, and 8.6 pp above TimesNet. The PP+SP row keeps the same beat-aligned evidence interface but uses a single Proposer–Arbiter pass, separating encoder gains from the effect of multi-role revision. The open-weight variants (0.866 for DeepSeek-V3.2 and 0.851 for Qwen3.5-27B) and the distilled Student (0.847) remain competitive, but the best GPT-5.2 result has the lowest observed ECE (0.033). Low-SQI F1 reaches 0.779 versus 0.718 for Debate and 0.748 for PP+SP, concentrating the gain in the quality-sensitive regime. The largest per-class gains occur for SR, STACH, and SBRAD; Table 2 gives the complete breakdown.

The comparison separates three sources of improvement. Signal-only rows test the encoder, L1–L3 add increasingly structured language-model reasoning, and PP+SP holds the beat-aligned interface fixed while removing the multi-role challenge process. The two diagnostic controls then test whether the gain depends on SQI-linked evidence rather than on additional prompt length or role count alone.

4.4 Ablation Studies

Table 3: Encoder and reasoning ablations on MIMIC-III-Ext-PPG.

Block	Variant	Macro-F1↑	Low-SQI↑	AUROC↑	ECE↓	ΔF1 (pp)	ΔLow (pp)
Encoder	PatchTST	0.794	0.693	0.951	0.044	–	–
	+ Cardiac align.	0.812	0.711	0.958	0.042	+1.8	+1.8
	+ Dual-path	0.830	0.731	0.964	0.040	+1.8	+2.0
	+ SQI gating	**0.838**	**0.741**	**0.968**	**0.038**	+0.8	+1.0
Reasoning	PP+SP	0.845	0.748	0.970	0.037	–	–
	w/o Checker	0.871	0.758	0.978	0.040	+2.6	+1.0
	w/o Skeptic	0.861	0.744	0.976	0.043	+1.6	-0.4
	w/o SQI anchoring	0.822	0.726	0.973	0.085	-2.3	-2.2
	w/o adapt. routing	0.863	0.757	0.978	0.034	+1.8	+0.9
	Full PhysioSkeptic	**0.886**	**0.779**	**0.981**	**0.033**	+4.1	+3.1

The Δ columns report percentage-point differences using the same measured rows: within the Encoder block each row is compared with the preceding configuration, whereas within the Reasoning block each row is compared with PP+SP. They are derived summaries, not additional runs.

The encoder rows are sequential: each row adds one component to the preceding configuration. Macro-F1 improves from 0.794 with PatchTST to 0.838 with SQI-gated PhysioPatch, followed by 0.845 with PP+SP; Low-SQI F1 and ECE move in the same direction. This pattern indicates that cardiac landmark alignment, dual-path encoding, and SQI gating contribute incrementally rather than acting as interchangeable prompt features.

The largest reasoning effect is quality-conditioned challenge selection: removing SQI anchoring lowers Macro-F1 by 2.3 pp and raises ECE from 0.037 to 0.085. Removing the Checker, Skeptic, or adaptive routing produces smaller changes. Random anchors and shuffled SQI values retain the prompt format but regress to approximately 0.83 Macro-F1 (Supplementary Table S12). These controls separate the evidence interface from the policy that decides when and where to spend additional reasoning. In particular, the role ablations retain the same input report and output space, so their differences are interpreted as changes in evidence checking rather than changes in the underlying classification target.

4.5 Robustness to Controlled Corruption

On initially high-quality windows ($\mathrm{SQI}>0.9$), PhysioSkeptic retains a positive margin over *PP+SP* across Gaussian noise, baseline wander, and motion artifact (Table 4; Supplementary Fig. S2). The three families probe diffuse high-frequency noise, low-frequency displacement, and localized morphology disruption. The corruptions are applied to initially high-quality windows ($\mathrm{SQI}>0.9$), which separates degradation caused by the perturbation from variation already present in the input. The margin persists as absolute Macro-F1 approaches the five-class chance range at the most severe levels. In the expanded table, the PhysioSkeptic scores are 0.791/0.640/0.252 for Gaussian noise, 0.820/0.793/0.493 for baseline wander, and 0.789/0.635/0.255 for motion artifact across their mild-to-severe levels; the corresponding margins over PP+SP remain positive throughout. The final two columns make the two views explicit: Δ reports the margin against the encoder-matched control, while "Drop vs. mild" reports the within-family loss relative to the mild or clean reference. Thus the table distinguishes relative ordering from absolute degradation under increasingly strong corruption.

4.6 Backbone Transfer and Diagnostic Analyses

Across the evaluated backbones, performance rises from Bare CoT through *PP+SP* to PhysioSkeptic (Fig. 4). The same stage-wise ordering holds for closed, open-weight, and distilled backbones: adding the beat-aligned encoder helps first, and quality-conditioned revision adds a further improvement. The encoder transition is larger than the debate transition for most backbones, while the Low-SQI gaps remain backbone dependent. This consistency supports portability of the evidence interface; absolute performance still depends on the underlying model and its local inference budget.

The class-wise gains are largest for SR, STACH, and SBRAD, where morphology, interval regularity, and quality cues must be combined. Relative to PP+SP, PhysioSkeptic improves these classes by 5.8, 5.5, and 5.0 pp, respectively; the gains for AF_FAMILY and PACE are 2.8 and 1.3 pp because irregular-RR and pacing-spike cues already provide stronger single-feature evidence.

Table 4: Controlled-corruption robustness across signal degradations.

Family	Level	PatchTST	GPT-5.2 + CoT	Debate	PP+SP	PhysioSkeptic	Δ vs. PP+SP (pp)	Drop vs. mild
Gaussian white noise	20 dB	0.723	0.738	0.739	0.763	0.791	+2.8	0.0%
Gaussian white noise	10 dB	0.581	0.591	0.620	0.613	0.640	+2.7	19.1%
Gaussian white noise	0 dB	0.200	0.210	0.203	0.228	0.252	+2.4	68.1%
Baseline wander	20 dB	0.739	0.748	0.755	0.783	0.820	+3.7	0.0%
Baseline wander	10 dB	0.707	0.716	0.729	0.751	0.793	+4.2	3.3%
Baseline wander	0 dB	0.416	0.422	0.417	0.448	0.493	+4.5	39.9%
Motion artifact	Mild	0.722	0.732	0.737	0.762	0.789	+2.7	0.0%
Motion artifact	Moderate	0.580	0.590	0.564	0.610	0.635	+2.5	19.5%
Motion artifact	Severe	0.197	0.202	0.204	0.226	0.255	+2.9	67.7%

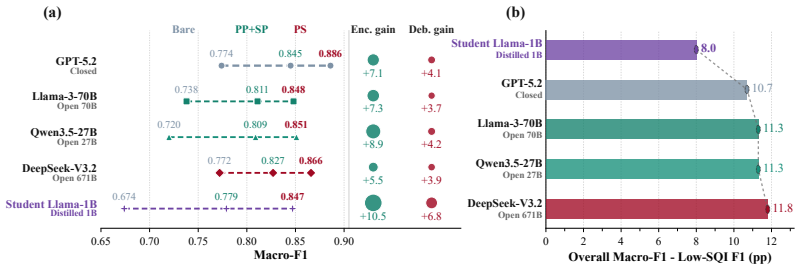

Figure 4: Backbone transfer across the Bare, PP+SP, and PhysioSkeptic stages.

On the routing subset, Adaptive reaches 0.883 F1 at 1,430 ms, matching Standard while reducing latency relative to forced Deep routing at 3,510 ms; the default boundary misroutes 1.4% of difficult windows (Supplementary Table S17, Fig. S5). The distilled Student runs in 18±2 ms per CPU sample at 0.847 Macro-F1. SQI anchoring lowers anchor-trap retention from 56.6% to 16.8% and recovers 80.4% of anchored cases; on these cases, F1 exceeds CoT and PP+SP by 20.8 and 14.0 pp, respectively. The MC-MED transfer and age-stratified analyses (Supplementary Table S15) show agreement between learned and rule-based SQI, stronger review under weak-label shift, and lower performance in the class-sparse subgroup below 45 years ($n = 161$, Macro-F1$= 0.490$), defining a concrete generalization gap.

5 Discussion

The results support a simple interpretation: signal quality is useful when it changes which evidence is challenged, not merely when it is appended to a prompt. The ablations, controlled corruptions, backbone transfer, and routing audit point to the same mechanism—retain a fast path for clear cases and spend additional checking on low-quality or conflicting evidence. The resulting interface is therefore useful as an evidence-audit layer for retrospective rhythm analysis.

Several limitations bound the interpretation. The primary benchmark is retrospective and class–SQI enriched, with repeated windows clustered within patients; its calibration and subgroup behavior should not be read as population-level estimates. Synthetic corruptions probe controlled failure modes but do not reproduce every artifact in clinical acquisition. MC-MED is a zero-shot proxy-transfer audit with weak auxiliary labels, and the routing latency depends on the provider and hardware. The class-sparse age subgroup also calls for a larger, prospective, multicenter evaluation with independent readers and patient-level endpoints.

6 Conclusion

We introduced PhysioSkeptic, an SQI evidence interface for retrospective ECG–PPG rhythm classification. The Patch Report makes conflict checks and belief revision traceable to beat-indexed evidence and physiological anchors; under the matched protocol, the framework improves observed Macro-F1 and ECE over generic Debate and PP+SP. Results across open-weight backbones and a local Student suggest that the interface is not tied to GPT-5.2, while MC-MED remains a zero-shot proxy-transfer audit rather than external diagnostic validation. Prospective multicenter evaluation is still required.

AI USE STATEMENT

Generative AI tools assisted with code development. The authors verified the analyses and final manuscript.

ETHICS STATEMENT

This study uses de-identified data and an anonymous, minimal-risk interface without direct identifiers or sensitive information. Institutional ethics approval was not required.

REPRODUCIBILITY STATEMENT

The Supplementary Material documents the dataset mapping, splits, evaluation protocols, and compute details. An anonymized core-code archive and repository are available at `https://anonymous.4open.science/r/PhysioSkeptic-5B1B/`; the release includes the Patch Report, SQI-gated routing, role prompts, a deterministic mock backend, and tests, but no patient data or model weights.

REFERENCES

Taymaz Akan, Sait Alp, and Mohammad Alfrad Nobel Bhuiyan. Ecgformer: Leveraging transformer for ecg heartbeat arrhythmia classification. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 1412–1417. IEEE, 2023.

Christopher RS Banerji, Tapabrata Chakraborti, Chris Harbron, and Ben D MacArthur. Clinical ai tools must convey predictive uncertainty for each individual patient. *Nature medicine*, 29(12): 2996–2998, 2023. doi: 10.1038/s41591-023-02562-7.

Siyuan Chen, Hao Wang, Huijie Zhang, Cailiang Peng, Yang Li, and Bing Wang. A novel method of swin transformer with time-frequency characteristics for ecg-based arrhythmia detection. *Frontiers in Cardiovascular Medicine*, 11:1401143, 2024. doi: 10.3389/fcvm.2024.1401143.

Michael Christof and Antonis A Armoundas. Implications of integrating large language models into clinical decision making. *Communications Medicine*, 5(1):490, 2025. doi: 10.1038/s43856-025-01216-8.

Clément Christophe, Tathagata Raha, Svetlana Maslenkova, Muhammad Umar Salman, Praveenkumar Kanithi, Marco AF Pimentel, and Shadab Khan. Beyond fine-tuning: Unleashing the potential of continuous pretraining for clinical llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10549–10561, 2024.

Sayantan Dasgupta and Trevor Cohn. Improving language model distillation through hidden state matching. In *The thirteenth international conference on learning representations*, 2025.

Harry J Davies, Ghena Hammour, Marek Zylinski, Amir Nassibi, Ljubiša Stanković, and Danilo P Mandic. The deep-match framework: R-peak detection in ear-ecg. *IEEE Transactions on Biomedical Engineering*, 71(7):2014–2021, 2024a.

Harry J Davies, James Monsen, and Danilo P Mandic. Interpretable pre-trained transformers for heart time-series data. *arXiv preprint arXiv:2407.20775*, 2024b.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 11733–11763, 2024.

Mohamed Elgendi. Optimal signal quality index for photoplethysmogram signals. *Bioengineering*, 3(4):21, 2016. doi: 10.3390/bioengineering3040021.

Mohamed Elgendi, Ian Norton, Matt Brearley, Derek Abbott, and Dale Schuurmans. Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions. *PloS one*, 8(10):e76585, 2013.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019. doi: 10.1038/s41591-018-0268-3.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jiahao Hu, Muhammad Mahboob Ur Rahman, and Taous-Meriem Laleg-Kirati. Adaptive long-range modeling of eeg and ecg with mamba and dynamic graph learning. *Scientific Reports*, 15 (1):38762, 2025a. doi: 10.1038/s41598-025-22684-x.

Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Li Qing. Removal of hallucination on hallucination: Debate-augmented rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 15839–15853, 2025b.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

Huawei Jiang, Husna Mutahira, Shibo Wei, and Mannan Saeed Muhammad. ECG-Mamba: Cardiac abnormality classification with non-uniform-mix augmentation on 12-lead ECGs. *IEEE Journal of Translational Engineering in Health and Medicine*, 13:461–470, 2025. doi: 10.1109/JTEHM.2025.3613609.

Yanrui Jin, Zhiyuan Li, Mengxiao Wang, Jinlei Liu, Yuanyuan Tian, Yunqing Liu, Xiaoyang Wei, Liqun Zhao, and Chengliang Liu. Cardiologist-level interpretable knowledge-fused deep neural network for automatic arrhythmia diagnosis. *Communications medicine*, 4(1):31, 2024. doi: 10.1038/s43856-024-00464-4.

Aman Kansal, Emma Chen, Boyang Tom Jin, Pranav Rajpurkar, and David A Kim. Mc-med, multimodal clinical monitoring in the emergency department. *Scientific Data*, 12(1):1094, 2025. doi: 10.1038/s41597-025-05419-5.

Matin Khalili, Hamid GholamHosseini, Andrew Lowe, and Matthew MY Kuo. Motion artifacts in capacitive ecg monitoring systems: a review of existing models and reduction techniques. *Medical & Biological Engineering & Computing*, 62(12):3599–3622, 2024. doi: 10.1007/s11517-024-03165-1.

Oleksii Kovalchuk, Oleksandr Barmak, Pavlo Radiuk, Liliana Klymenko, and Iurii Krak. Towards transparent ai in medicine: Ecg-based arrhythmia detection with explainable deep learning. *Technologies*, 13(1):34, 2025. doi: 10.3390/technologies13010034.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. *Advances in Neural Information Processing Systems*, 37:124575–124611, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 17889–17904, 2024.

Fenglin Liu, Hongjian Zhou, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Yining Hua, Peilin Zhou, et al. Application of large language models in medicine. *Nature Reviews Bioengineering*, 3(6):445–464, 2025a. doi: 10.1038/s44222-025-00279-5.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, et al. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*, 2025b.

Jiaxu Lou and Yifan Sun. Anchoring bias in large language models: An experimental study. *Journal of Computational Social Science*, 9(1):11, 2026. doi: 10.1007/s42001-025-00435-2.

Steven A Lubitz, Emelia J Benjamin, Jeremy N Ruskin, Valentin Fuster, and Patrick T Ellinor. Challenges in the classification of atrial fibrillation. *Nature Reviews Cardiology*, 7(8):451–460, 2010. doi: 10.1038/nrcardio.2010.86.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Mohammad Moulaeifard, Peter H Charlton, and Nils Strodthoff. MIMIC-III-Ext-PPG: A PPG Benchmark Dataset for Cardiorespiratory Analysis. *PhysioNet*, February 2026. doi: 10.13026/nmwb-6h34. URL `https://doi.org/10.13026/nmwb-6h34`. Version 1.0.0.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Jiapu Pan and Willis J Tompkins. A real-time qrs detection algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, 1985. doi: 10.1109/TBME.1985.325532.

Wenhan Qian, Ziqu Shang, Detang Wen, and Tongran Fu. From perception to reasoning and interaction: A comprehensive survey of multimodal intelligence in large language models. *TechRxiv*, 2025. doi: 10.36227/techrxiv.176472911.13119857/v1.

Yupeng Qiang, Xunde Dong, Xiuling Liu, Yang Yang, Yihai Fang, and Jianhong Dou. Ecgmamba: Towards efficient ecg classification with bissm. *arXiv preprint arXiv:2406.10098*, 2024.

Tara Rasti. Ecg-based arrhythmia detection by a shallow cnn model. 2024.

Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. Evaluation and mitigation of cognitive biases in medical language models. *npj Digital Medicine*, 7(1):295, 2024.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

PR Wankhede, Devendra Bhuyar, Shrinivas Zanwar, Rohit Pawar, Mahendra R Jadhav, Nisarg Gandhewar, Madhusudan B Kulkarni, Manish Bhaiyya, and Hossam Haick. Artificial intelligence for noninvasive health diagnostics. *ACS Sensors*, 10(11):8217–8255, 2025. doi: 10.1021/acssensors.5c03171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Kuba Weimann and Tim OF Conrad. Self-supervised pre-training with joint-embedding predictive architecture boosts ecg classification performance. *Computers in Biology and Medicine*, 196: 110809, 2025. doi: 10.1016/j.compbiomed.2025.110809.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin Wang, Zhenxi Lin, et al. Medjourney: Benchmark and evaluation of large language models over patient clinical journey. *Advances in Neural Information Processing Systems*, 37:87621–87646, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Weiqiu You, Anton Xue, Shreya Havaldar, Delip Rao, Helen Jin, Chris Callison-Burch, and Eric Wong. Probabilistic soundness guarantees in LLM reasoning chains. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7506–7525. Association for Computational Linguistics, 2025.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 14544–14607. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.acl-long.782.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. How language model hallucinations can snowball. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 59670–59684, 2024b.

Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K Gupta, and Jingbo Shang. Large language models for time series: A survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 8335–8343, 2024c. doi: 10.24963/ijcai.2024/921.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37:333–356, 2024d.

Feiyan Zhou and Jiannan Wang. Heartbeat classification method combining multi-branch convolutional neural networks and transformer. *iScience*, 27(3):109307, 2024. doi: 10.1016/j.isci.2024.109307.

Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales? *Advances in Neural Information Processing Systems*, 37:123846–123910, 2024.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, et al. Mindstorms in natural language-based societies of mind. *Computational Visual Media*, 11(1):29–81, 2025. doi: 10.26599/CVM.2025.9450460.

Supplementary Material for PhysioSkeptic

A Definitions, Pipeline, and Theory

This section collects the notation, a concrete five-role transcript, and the idealized information-theoretic motivation used to interpret the empirical pipeline. The three views are complementary: the abbreviation table fixes the terminology, the illustration makes the evidence flow inspectable, and the analytic discussion states exactly which part is a design motivation rather than an implementation guarantee.

A.1 Abbreviations

The table fixes the abbreviations used in the main text, figures, and audit tables. Dataset-specific names are retained rather than silently expanded into clinical labels that would imply a different annotation scheme.

Table S1: Abbreviations used throughout the paper.

Abbrev.	Full Form
AF_FAMILY	Atrial Fibrillation / Flutter family
ABP	Arterial Blood Pressure
API	Application Programming Interface
AUROC	Area Under ROC Curve
BW	Baseline Wander
CNN	Convolutional Neural Network
Clinical5	Five-class rhythm label scheme
CoT	Chain-of-Thought
ECE	Expected Calibration Error
ECG	Electrocardiogram
ED	Emergency Department
FFT	Fast Fourier Transform
GWN	Gaussian White Noise
ICD-10	International Classification of Diseases, Tenth Revision
LLM	Large Language Model
MA	Motion Artifact
MC-MED	Multimodal Clinical Monitoring in the Emergency Department
MIMIC-III-Ext-PPG	Medical Information Mart for Intensive Care III – Extended PPG
MLP	Multilayer Perceptron
NLL	Negative Log-Likelihood
PACE	Paced Rhythm family
PP+SP	PhysioPatch with one Proposer–Arbiter pass
PPG	Photoplethysmogram
PS	PhysioSkeptic
PhysioSkeptic-D	PhysioSkeptic with DeepSeek-V3.2 backbone
RESP	Respiratory waveform
RR	R–R interval
SBRAD	Sinus Bradycardia
SARRH	Sinus Arrhythmia
SC	Self-Consistency
SNR	Signal-to-Noise Ratio
SQI	Signal Quality Index
SQIJK	Rule-based SQI baseline
SR	Sinus Rhythm
STACH	Sinus Tachycardia
STFT	Short-Time Fourier Transform

A.2 Five-Role Debate Illustration

The illustration follows one Patch Report through proposal, conflict checking, challenge generation, response, and arbitration. It is a concrete instance of the JSON contracts described in the main text; it does not introduce an additional model or training stage.

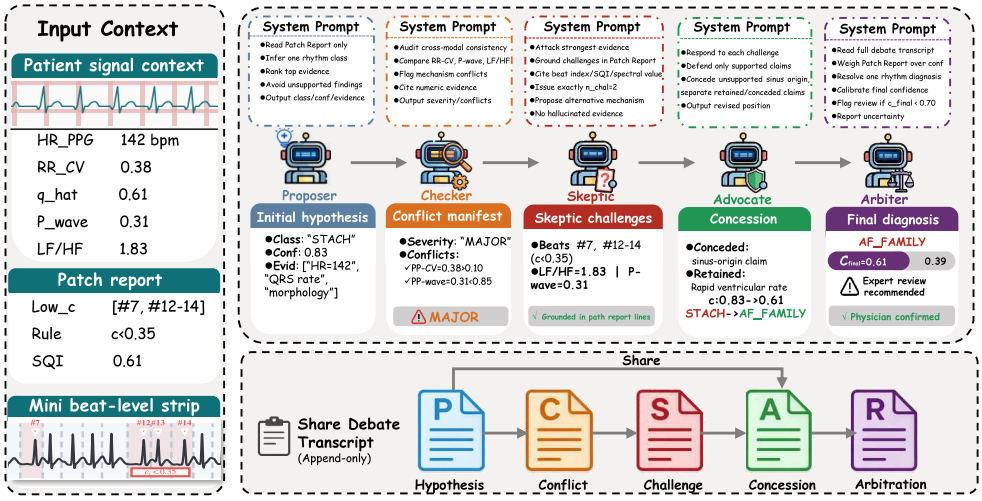

Figure S1: Five-role pipeline from Patch Report to confidence-scored arbitration.

A.3 INFORMATION-THEORETIC MOTIVATION

The following derivation gives the narrow motivation for using signal quality to control evidence checks. We separate this idealized argument from the empirical rule-score update so that the theoretical notation is not read as a claim that the LLM transcript is Bayesian inference.

Let $R \sim b_t$ denote the latent rhythm state and let O_{t+1} be sampled from a correctly specified idealized observation channel $P(\boldsymbol{o} \mid R; \mathrm{SQI})$, whose noise component is a valid distribution over the observation space. If $b_{t+1} = P(R \mid O_{t+1})$, then

$$\mathbb{E}_{O_{t+1}}[H(b_{t+1})] = H(R \mid O_{t+1}) \leq H(R) = H(b_t),$$

with equality when $I(R; O_{t+1}) = 0$. This conditional-entropy identity is a design motivation for using signal quality when selecting evidence checks; it is not a novel theoretical contribution or a convergence guarantee for the LLM transcript.

The implementation does not fit this observation channel or compute a Bayesian posterior. Its SQI-WEIGHTEDSCOREUPDATE function is a deterministic rule-score update: it combines class scores from the Patch Report with rule-audited challenge penalties and normalizes the resulting non-negative class scores. Thus, the analytic identity above applies only to the idealized model, while the empirical system is evaluated through the reported ablations and audits.

Increasing SQI shifts the idealized observation channel toward its signal-dependent component, but the resulting information gain depends on the full class-conditional distributions and need not scale linearly with SQI. Accordingly, we use SQI only as a monotone design signal for routing depth and challenge pressure. We make no general convergence-rate claim for either the analytic mixture model or the correlated LLM debate transcript.

B Data and Transfer Audits

This section documents the cohorts, implementation mapping, and external SQI audits used to interpret the reported results. The dataset table establishes the sample accounting; the mapping table connects the formal motivation to the implemented modules; the remaining analyses test transfer and weak-label stress behavior without treating either proxy as external diagnostic validation.

B.1 Dataset Statistics

Table S2 summarizes the primary benchmark and external SQI-transfer cohort used throughout the paper.

Table S2: Dataset statistics for the primary and external SQI-transfer cohorts.

Dataset	Scope	Pts./Visits	Segs.	SQI L/M/H	ECG/ABP/RESP avail.
MIMIC-III Ext-PPG	Source	6 131	4.92 M	1.5/24.4/74.2	98.4/45.7/89.2
	Clinical5	6 038	4.78 M	1.5/24.2/74.4	98.5/45.7/89.1
	Bench-100	4 532	453 K	2.2/26.2/71.7	98.1/29.5/81.8
	Train	3 029	303 K	2.1/26.0/71.9	98.3/29.1/81.6
	Valid	712	71 K	2.4/25.5/72.2	97.6/32.7/80.6
	Test	791	79 K	2.1/27.3/70.6	97.6/28.2/83.8
MC-MED*	Eval	14 221	118 K	24.1/41.2/34.7	71.0/38.0/52.0

Bench-100 retains patients with $\geq$100 usable windows and 100 windows/patient. All listed cohorts require synchronized ECG–PPG; the final column reports availability of ECG, ABP, and RESP beyond this requirement. Main experiments use ECG–PPG only. SQI L/M/H denotes < 0.5, 0.5–0.8, and ≥ 0.8; Table 1 instead reports Low (< 0.5) and Non-low (≥ 0.5). *MC-MED Eval is the signal-available visit pool; the SQI audit uses 8 340 quality-filtered ECG–PPG windows, and weak-label rhythm detection uses the stricter visit-level subset in Appendix B.4.

The five-class mapping is fixed before model development: SR and SARRH map to SR; STACH and SBRAD retain their labels; AF and AFLT map to AF_FAMILY; and APACE, VPACE, AVPACE, and WAPACE map to PACE.

B.2 THEORY-TO-IMPLEMENTATION MAPPING

This mapping is the bridge between the idealized equations and the executable pipeline: each theoretical object is paired with the concrete field, score, or routing decision that consumes it. It also makes clear which quantities are design abstractions and which are actually computed.

Table S3: Theory-to-implementation mapping.

Theoretical term	**Practical implementation**
Observation $\boldsymbol{o}_t$	Patch Report fields (per-beat c_i, low-conf. beat list, ECG-derived P-wave ratio, low-/high-band power ratio) plus cited role evidence.
SQI (SQI $\in [0, 1]$)	Learned $\widehat{q}^{(k)} \in [0, 1]$ from the SQI head (Eq. 2), supervised during pre-training.
Critique intensity	Fast routing skips the Skeptic; Standard and Deep routing request up to three challenges, prioritizing low-quality and conflicting evidence.
Belief $b_t(r)$	Class-confidence vector updated across roles (Proposer c_0, post-Advocate c_1, final c_{final}).
SQI-weighted score update	Rule-based class scores, challenge penalties, and normalization over $\mathcal{R}$ (implemented by SQIWEIGHTEDSCOREUPDATE in Algorithm 1).
Routing depth	Eq. (5): Fast/Standard/Deep gated by $(c_0, \widehat{q}, \mathrm{conflict})$.

The learned SQI estimate affects the empirical system through discrete routing and the SQI-weighted score update. The API-based deliberation transcript is not differentiated through.

B.3 SQI AGREEMENT ON MC-MED

Table S4 reports Spearman correlations between PhysioPatch's learned SQI and a rule-based SQI proxy on the 8,340-window MC-MED audit set. The proxy is not an adjudicated ground truth for signal quality.

Table S4: Agreement between learned and rule-based SQI on MC-MED.

Modality	**Spearman ρ**	**MAE**	**N**
PPG	0.806	0.075	8 340
ECG	0.779	0.087	8 340
ECG–PPG mean	*0.793*	*0.081*	*8 340*

Spearman correlations of $\rho = 0.779$–0.806 indicate agreement with the rule-based proxy without retraining. The MAE (0.075–0.087) indicates some domain shift; SQI routing thresholds in MC-MED may therefore be slightly misaligned without adjustment, consistent with the transfer boundary stated in the Conclusion.

B.4 MC-MED Stress Tests with Weak Labels

The two weak-label analyses below are reported as transfer stress tests. They probe whether the learned SQI remains useful when the target labels are proxy annotations, while keeping the primary MIMIC benchmark and its evaluation claims separate from these external checks.

Table S2 reports the broad MC-MED signal-available pool (14 221 visits; 118 K windows) used for cohort accounting and SQI-transfer screening. For weak-label rhythm detection, we apply a stricter visit-level filter to the full MC-MED corpus (118 385 visits; 70 545 patients), retaining visits with simultaneous ECG (Lead II) and PPG (Pleth) recordings of at least 30 s; this yields 12 705 visits from 10 969 patients. Labels are automatically derived by combining ICD-10 discharge codes, past medical history, and rule-based signal-feature scoring (RR coefficient of variation, P-wave ratio, SQI); 1 706 visits pass a confidence threshold of 0.6 and receive a non-empty rhythm label (SINUS_FAMILY 1 487; AF_FAMILY 169; PACE 50). For AF detection we pair all 169 ICD-10-confirmed AF/flutter visits with 169 quality-filtered sinus controls selected after SQI thresholding (controls are not propensity-matched on SQI; mean SQI_{PPG}: 0.955 for controls vs. 0.818 for AF). For the pacemaker-presence stress test we pair all 50 visits with ICD-10 Z95.0, which indicates pacemaker presence but does not independently confirm a paced waveform, with 50 quality-filtered sinus controls. No MC-MED signal is used for MIMIC training or hyperparameter tuning. Because weak-label construction uses signal features also present in the Patch Report and the controls are not SQI-matched, these experiments are auxiliary case–control stress tests rather than independent external diagnostic validation. Their ECE values characterize these constructed sets, not calibration under clinical prevalence.

Table S5: AF weak-label stress test using MC-MED.

Method	AUROC	F1	ECE	Rev. %
w/o SQI routing	0.897	0.872	0.124	–
PP+SP	0.913	0.889	0.097	–
PhysioSkeptic	**0.934**	**0.908**	**0.063**	26.9

On this constructed set, PhysioSkeptic has an AUROC 2.1 pp higher than PP+SP and a lower benchmark ECE (0.063 vs. 0.097). These descriptive differences do not establish independent external diagnostic validation.

Table S6: Pacemaker weak-label stress test using MC-MED.

Method	AUROC	Sensitivity	Specificity
PP+SP	0.836	0.720	0.880
PhysioSkeptic	**0.858**	**0.760**	**0.900**

On this constructed set, PhysioSkeptic has an AUROC 2.2 pp higher than PP+SP, with sensitivity 0.760 versus 0.720 and specificity 0.900 versus 0.880. The elevated review-flag rate across both tasks (26.9 % vs. 18.2 % on MIMIC at the default threshold; Table S16, Part B) is descriptive of these differently constructed cohorts and is not a calibrated clinical alert.

Algorithm 1 PhysioSkeptic Inference Pipeline

Input: signal $\boldsymbol{X}$, LLM $\mathcal{L}$, encoder ϕ.
$\boldsymbol{H}, \hat{\text{SQI}}, \text{report} \leftarrow \phi(\boldsymbol{X})$; ctx $\leftarrow$ SERIALIZE(report)
$(\hat{r}_0, c_0, \boldsymbol{p}_0, e_{1:3}) \leftarrow \mathcal{L}_{\text{Proposer}}(\text{ctx})$
check $\leftarrow \mathcal{L}_{\text{Checker}}(\text{report}, c_0)$; depth $\leftarrow$ ROUTE$(c_0, \hat{\text{SQI}}, \text{check})$
$b \leftarrow \boldsymbol{p}_0$; chal $\leftarrow \varnothing$; reb $\leftarrow \varnothing$
if depth $\neq$ FAST **then**
    chal $\leftarrow \mathcal{L}_{\text{Skeptic}}(\hat{r}_0, e_{1:3}, \text{report}, \text{check})$
    if depth $=$ DEEP **then**
        reb $\leftarrow \mathcal{L}_{\text{Advocate}}(\hat{r}_0, c_0, \text{chal}, \text{report})$
    end if
    $b \leftarrow$ SQIWEIGHTEDSCOREUPDATE$(\boldsymbol{p}_0, \text{report}, \hat{r}_0, \text{chal}, \hat{\text{SQI}})$
end if
return $\mathcal{L}_{\text{Arbiter}}(\text{report}, b, \text{depth}, \text{chal}, \text{reb})$

C IMPLEMENTATION AND EXPERIMENTAL DETAILS

This section provides the executable inference procedure, distillation recipe, and additional experimental details needed to reproduce the reported comparisons. The order follows the computation graph: inference first, then the local Student, then prompting, subset construction, and resource details.

C.1 ALGORITHM PSEUDOCODE

Operationally, SQIWEIGHTEDSCOREUPDATE scores each class from heart rate, RR variability, P-wave ratio, and spectral cues; it mixes these scores with a uniform class-score component according to the global SQI, multiplies the result by the Proposer's complete class-probability vector, and renormalizes. Weak, moderate, and strong challenges apply penalties of 0.03, 0.08, and 0.15 to the proposed class while transferring mass toward the cited alternative. This deterministic score update is a traceable implementation rule, not a fitted generative model.

C.2 SKEPTICISM-AWARE DISTILLATION DETAILS

Following knowledge distillation with soft targets and intermediate alignment (Hinton et al., 2015; Dasgupta & Cohn, 2025), we transfer the Teacher's class distribution, confidence score, and low-quality beat attention to a compact Student. The Teacher (PhysioSkeptic with GPT-5.2 backbone) annotates a matched training subset offline, with five windows per training patient (15,145 windows from 3,029 patients), producing for each sample a tuple $(\tilde{p}, c_{\mathrm{T}}, \bar{\boldsymbol{A}}^{*}_{\mathrm{low}})$, where $\tilde{p}$ is the soft class distribution, c_{T} is the Teacher confidence score, and $\bar{\boldsymbol{A}}^{*}_{\mathrm{low}}$ is the normalized Teacher attention distribution over beats identified as low confidence: $\bar{A}^{*}_{\mathrm{low},i} = (A^{*}_{\mathrm{low},i} + \epsilon)/\sum_j (A^{*}_{\mathrm{low},j} + \epsilon)$. The Student (Llama-3.2-1B with PhysioPatch encoder fine-tuned via LoRA (Hu et al., 2022)) minimizes

$$\begin{aligned}\mathcal{L}_{\mathrm{KD}} \;&= \alpha\, \mathcal{L}_{\mathrm{CE}}(\boldsymbol{s}, y) \;+\; \beta\, \mathrm{KL}(\tilde{p} \,\|\, \tilde{p}_S) \\ &\quad +\; \delta_1\, \mathrm{MSE}(c_T, c_S) \;+\; \delta_2\, \mathrm{KL}(\bar{\boldsymbol{A}}^{*}_{\mathrm{low}} \,\|\, \bar{\boldsymbol{A}}_S),\end{aligned} \tag{7}$$

with $(\alpha, \beta, \delta_1, \delta_2) = (1.0, 0.5, 0.3, 0.2)$ selected on the benchmark validation split. The four terms supervise, respectively, the ground-truth label, the Teacher's soft class distribution, confidence score, and Patch-attention alignment over low-quality beats. Both attention arguments in the KL term are normalized across beat indices after ϵ smoothing; the term therefore aligns distributions rather than binary masks. Here $\boldsymbol{s}$ denotes Student logits.

C.3 Additional Experimental Details

The temporal path computes

$$\boldsymbol{H}^{(k)}_{\text{temp}} = \text{TransformerEncoder}\big(\boldsymbol{T}^{(k)} + \boldsymbol{E}_{\text{pos}}\big) \in \mathbb{R}^{N_k \times d}, \tag{8}$$

where $\boldsymbol{E}_{\text{pos}}$ encodes each beat's index within the 30-second window. For all API-backed roles, the LLM-facing context is

$$\text{ctx} = \Big[\underbrace{\text{Serialize}(v_{\text{rep}})}_{\text{Patch Report}}\,;\quad \underbrace{v_{\text{pmt}}}_{\text{role prompt}}\Big]. \tag{9}$$

Continuous encoder embeddings remain inside PhysioPatch and the local Student; they are not transmitted to public LLM APIs.

The Cross-Modal Checker flags HR discrepancy when $|\text{HR}_{\text{PPG}} - \text{HR}_{\text{ECG}}| > 5$ bpm, rhythm-morphology conflict when irregular RR co-occurs with ECG-derived sinus P-waves, ECG–PPG rhythm conflict when AF-like rhythm co-occurs with regular PPG pulse intervals, and confidence-quality mismatch when $c > 0.8$ with $\widehat{q} < 0.5$.

For modality k, the frequency path computes

$$S^{(k)}_{f,t} = \left| \sum_{\tau=0}^{W-1} x_k[\tau + tH]\, w[\tau]\, e^{-2\pi i f \tau / N} \right|^2,$$

with window $W = 256$, hop $H = 64$, and $N = 512$ FFT points. PhysioPatch is pre-trained on the rhythm-unlabeled MIMIC-III-Ext-PPG corpus via three jointly optimized objectives:

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{MSM}} + \lambda_1\, \mathcal{L}_{\text{CL}} + \lambda_2\, \mathcal{L}_{\text{SQI}}, \tag{10}$$

where $\mathcal{L}_{\text{MSM}}$ reconstructs 30% randomly masked beats, $\mathcal{L}_{\text{CL}} = \text{InfoNCE}(\boldsymbol{h}_{\text{PPG}}, \boldsymbol{h}_{\text{ECG}})$ aligns co-acquired PPG and ECG representations, and $\mathcal{L}_{\text{SQI}} = \text{BCE}(\widehat{q}, q_{\text{label}})$ uses MIMIC-III-Ext-PPG SQI annotations as direct supervision. We set $(\lambda_1, \lambda_2) = (0.5, 0.1)$ on the benchmark validation split. Pre-training reads only the patient-disjoint training and validation manifests; no test-patient waveform or SQI label enters pre-training, and validation is used only for model selection.

Pre-training runs for 200 000 steps, batch size 512, on 8 NVIDIA A100 80 GB GPUs using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 10^{-4}). Learning rate follows a cosine schedule with 10 000 warmup steps and peak 3×10^{-4}. Contrastive loss uses temperature $\tau = 0.07$ and a queue of 4 096 negatives. Fine-tuning (PhysioPatch + linear head) runs 30 epochs at 5×10^{-5}.

LoRA (Hu et al., 2022) rank-16 adapters applied to all attention projections of Llama-3.2-1B-Instruct. 20 epochs, batch 64, learning rate 2×10^{-4}. Loss coefficients are as in Eq. (7).

On the matched test subset, the Student disagrees with the Teacher on 9.4% of samples and is correct in 21.7% of those disagreements, indicating that distillation does not simply copy Teacher errors. Patch-attention alignment over low-confidence beats (c_i<0.40) remains substantial: top-k overlap is 0.71 at k=3, versus 0.49 without the attention term in Eq. (7). Low-SQI ECE is 0.072 versus 0.118 for the ablated variant, suggesting partial transfer of quality-aware behavior.

GPT-5.2 and open-model results were collected through API inference during January–March 2026. The public `gpt-5.2` alias was not pinned to a model-specific snapshot, so alias resolution may vary after this access window. Decoding uses T=0.7, top-p=1.0, JSON output, an 8 192-token context, and run-specific seeds where the provider supports deterministic seeding; five-run averages are reported in Table S14. API request timestamps, response hashes, package versions, fixed input IDs, and recorded seeds are archived for each run.

Prompts use fixed JSON schemas; routing thresholds and all hyperparameters are fixed on the validation split before test evaluation. Matched validation/test subsets use fixed seeds and shared subject–window IDs across all LLM methods. Teacher annotations are retained where PhysioNet and API terms permit; regeneration scripts use the same fixed input IDs.

All LLM baselines receive identical structured Patch reports. Debate (Du et al.) uses three GPT-5.2 agents, two rounds. GPT-5.2 + Reflect follows Self-Refine (Madaan et al., 2023): one self-critique, one revision. The *PP+SP* baseline routes all samples through Proposer + Arbiter only (no Cross-Modal Checker, Skeptic, or Advocate; no SQI routing constraint). This differs from the adaptive Fast path, which applies SQI and confidence gating before bypassing the debate roles.

D Additional Analyses and Audits

The final section expands the evidence ladder beyond the main tables. It groups per-class results, mechanism controls, simpler role-count baselines, failure cases, and subgroup analyses so that each additional result is interpreted against the same matched protocol and its corresponding evidence boundary.

D.1 Additional Result Details and Audits

The next tables unpack the aggregate main-text score along three complementary axes: class composition, controlled corruption, and routing safety. They use the same matched evaluation protocol and should be read as diagnostic breakdowns, not as new training or tuning results.

Table S7: Per-class F1 across methods.

Method	SR (26.4%)	STACH (17.9%)	AF_FAM. (23.5%)	SBRAD (12.9%)	PACE (19.3%)
PatchTST	0.704±.003	0.746±.006	0.825±.003	0.761±.007	0.932±.001
Debate	0.744±.001	0.774±.006	0.865±.002	0.796±.007	0.941±.002
PP+SP	*0.771±.003*	*0.789±.004*	*0.897±.002*	*0.822±.006*	*0.947±.001*
PhysioSkeptic	**0.829±.006**	**0.844±.005**	**0.925±.003**	**0.872±.008**	**0.960±.001**
PhysioSkeptic-D	0.799±.004	0.814±.006	0.913±.002	0.850±.008	0.954±.002

PhysioSkeptic gives the highest F1 on every class in the matched test subset. The largest margins over *PP+SP* occur for SR, STACH, and SBRAD, where borderline interval and morphology cues are most sensitive to degraded or conflicting waveforms.

Table S8: Noise robustness on initially high-SQI windows.

Gaussian White Noise (GWN)							
SNR	PatchTST	GPT-5.2+CoT	Debate	PP+SP	Physio Skeptic	Δ vs PP↑	Drop vs Ref.↓
20 dB	0.723±.004	0.738±.005	0.739±.007	0.763±.003	**0.791±.004**	**+2.8**	0.0%
10 dB	0.581±.012	0.591±.010	0.620±.010	0.613±.010	**0.640±.008**	**+2.7**	19.1%
0 dB	0.200±.003	0.210±.004	0.203±.004	0.228±.004	**0.252±.005**	**+2.4**	68.1%
Baseline Wander (BW)							
SNR	PatchTST	GPT-5.2+CoT	Debate	PP+SP	Physio Skeptic	Δ vs PP↑	Drop vs Ref.↓
20 dB	0.739±.005	0.748±.005	0.755±.006	0.783±.007	**0.820±.004**	**+3.7**	0.0%
10 dB	0.707±.005	0.716±.006	0.729±.006	0.751±.005	**0.793±.003**	**+4.2**	3.3%
0 dB	0.416±.008	0.422±.006	0.417±.009	0.448±.008	**0.493±.009**	**+4.5**	39.9%
Motion Artifact (MA)							
Severity	PatchTST	GPT-5.2+CoT	Debate	PP+SP	Physio Skeptic	Δ vs PP↑	Drop vs Ref.↓
Mild	0.722±.003	0.732±.004	0.737±.007	0.762±.006	**0.789±.002**	**+2.7**	0.0%
Moderate	0.580±.014	0.590±.012	0.564±.012	0.610±.010	**0.635±.008**	**+2.5**	19.5%
Severe	0.197±.004	0.202±.002	0.204±.004	0.226±.003	**0.255±.004**	**+2.9**	67.7%

The corruption table separates a method's baseline quality from its degradation under progressively stronger perturbations. The final two columns report the margin over PP+SP and the drop relative to the corresponding clean reference, so the absolute values and the robustness trend can be inspected together.

Figure S2 visualizes the same means and standard deviations as Table S8; it is a view of the table, not an additional result. The shaded ranges are five-run variability bands and the numeric labels are the reported means. Gaussian noise primarily tests broadband corruption, baseline wander tests slow displacement of the signal baseline, and motion artifact tests localized shape disruption. Reading the curves together with the table separates the absolute degradation from the retained advantage over PP+SP.

The ablation diagram identifies which encoder and reasoning components are removed in the corresponding controls. Its purpose is to document the scope of each ablation; the measured values remain in the main and supplementary tables.

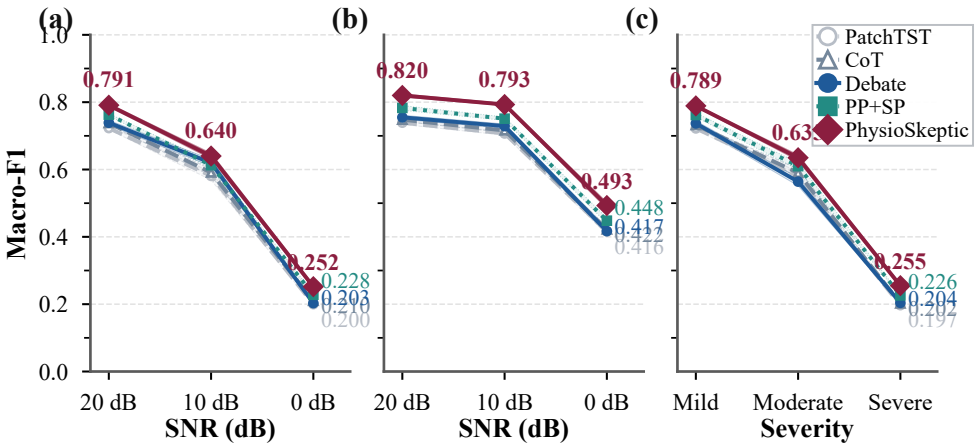

Figure S2: Macro-F1 under Gaussian noise, baseline wander, and motion-artifact corruption.

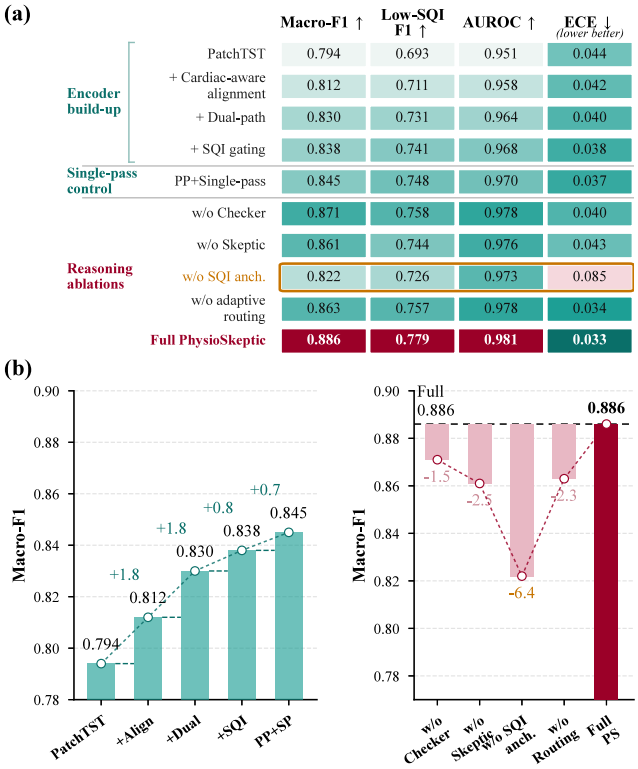

Figure S3: Ablation visualization of encoder and reasoning components.

The safety table compares fixed-depth policies with the adaptive policy on the same routing-analysis subset. “Misrouted difficult” counts difficult windows sent to an insufficiently deep route, while ECE measures calibration on this subset.

Table S9: Routing safety on the routing-analysis subset.

Policy	F1	ECE↓	Misrouted difficult ↓
Fixed Fast	0.821	0.067	100.0%
Fixed Standard	0.883	0.041	0.0%
Fixed Deep	0.880	0.038	0.0%
Adaptive	**0.883**	**0.039**	**1.4%**

This audit isolates cases where the initial Proposer class is wrong but later Patch Report evidence contradicts it. Anchor rate measures retained incorrect confidence, recovery measures corrected anchor cases, and F1 reports the final classification on the audited subset.

Table S10: Anchor-trap audit results.

Method	**Anchor Rate↓**	**Recov.↑**	**F1↑**
GPT-5.2 + CoT	56.6%	0%	0.569
GPT-5.2 + SC×5	45.9%	12.3%	0.609
Debate (Du et al., 2024)	35.2%	26.1%	0.663
PP+SP	*39.3%*	*0%*	*0.637*
PhysioSkeptic (GPT-5.2)	**16.8%**	**80.4%**	**0.777**

D.2 Mechanism Analysis: When Does Skepticism Help?

This section examines whether observed gains are associated with the challenge–response mechanism conditioned on SQI. Low-SQI and ECG–PPG conflict cases may trigger more Skeptic challenges, more response acknowledgments, more prediction revisions, and more net error correction. We therefore log, per sample, the emitted challenge count n_{chal}, the fraction of challenges conceded by the Advocate (*concession rate*), the fraction of samples whose final class differs from the Proposer's (*revision rate*), and net error correction (*revisions that fix a wrong Proposer minus revisions that flip a correct Proposer to a wrong class*).

Anchor-trap construction follows three rule-based conditions: the Proposer's initial prediction $\hat{r}_0$ is incorrect; at least one Patch Report item produced after the Proposer step contradicts $\hat{r}_0$ under a deterministic clinical rule (e.g., $\text{RR_CV} > 0.20$ or ECG-derived P-wave ratio < 0.40 for a sinus claim); and at least one beat with $c_i < 0.40$ falls within the cited evidence span. A sample counts as an anchoring error iff the final output retains the Proposer's incorrect class with $c_{\text{final}} \geq \tau_{\text{anc}}$=0.60. The rules are applied blind to model identity; two annotators verified a 200-case random subsample with Cohen's $\kappa = 0.86$.

Table S11: Mechanism analysis across SQI and conflict subsets.

Subset (n)	**Mean Chal. Count**	**Conces. rate**	**Rev. rate**	**Err.corr. rate**	**ΔF1** (vs PP+SP)
High SQI (n=2,119)	1.08	11.4%	5.9%	+3.7%	+2.3
Low SQI (n=918)	2.41	38.6%	28.4%	+20.1%	+3.1
Cross-modal conflict (n=1,242)	2.27	41.2%	33.5%	+24.7%	+6.9
Anchor-trap (n=500)	2.58	47.8%	58.1%	+52.6%	+14.0
Low SQI + Conflict (n=331)	2.74	49.1%	37.2%	+28.5%	+8.0

Mean challenge count, concession rate, and error-correction rate increase as signal quality degrades or modalities conflict: challenge count rises from 1.08 per sample on high-SQI cases to 2.74 on low-SQI + conflict cases. Revision and error-correction rates rise in the same order. The ΔF1 over the *PP+SP* control is small on high-SQI cases (+2.3 pp), where the Proposer is usually already correct, and largest on low-SQI + conflict (+8.0 pp) and anchor-trap cases (+14.0 pp). This pattern is consistent with the SQI-linked challenge–response–revision interpretation, although it does not isolate a unique causal mechanism. Sections D.3 and 4.4 provide complementary controls and component removals.

One alternative is that the Skeptic simply adds more prompt context and that any sufficiently rich text would produce similar gains. We probe this with two control conditions in Section D.3 (*Random beat anchors* and *Shuffled SQI*)—both retain the prompt format but break the SQI-to-challenge link, and both regress Macro-F1 to the SQI-unanchored range (0.83 / 0.83 vs. 0.886 for the full model). These controls suggest that the intended SQI-to-evidence link accounts for part of the observed revision behavior, without matching all routing or compute differences.

D.3 Simpler Baselines: Is the Five-Role Pipeline Necessary?

We next ask whether five roles are necessary or whether a simpler SQI-aware prompt is enough. Under matched evaluation and the same GPT-5.2 backbone, we compare five controls that vary role count and anchor content.

Table S12: Simpler role-count and grounding baselines.

Method	Macro F1↑	Low-SQI F1↑	ECE↓	Anch.↓
Single-agent + SQI prompt	0.831±.005	0.717	0.061	33.1%
SQI-aware Verifier (2-role)	0.847±.004	0.736	0.052	28.4%
3-role (Prop.+Skep.+Arb.)	0.864±.003	0.752	0.045	22.5%
Debate + random beat anchors	0.832±.006	0.722	0.082	31.7%
Debate + shuffled SQI values	0.829±.005	0.719	0.087	32.4%
PhysioSkeptic (full, 5-role)	**0.886**±.003	**0.779**	**0.033**	**16.8%**

More roles improve performance, but not linearly: the single-agent SQI prompt is competitive, the 3-role debate captures most of the gain, and the full 5-role pipeline adds $+2.2$ pp by separating conflict detection (Checker) from response control (Advocate). Perturbing the anchor content moves performance toward the SQI-free setting. Random beat anchors and shuffled SQI values preserve the prompt format but remove the intended evidence link, degrading F1, calibration, and anchoring. These controls suggest that correctly linking SQI to cited beat-indexed evidence accounts for a substantial part of the observed improvement.

D.4 Failure Case Analysis

The cases below summarize where debate conditioned on SQI still fails or appropriately escalates uncertainty, clarifying the boundaries of the mechanism.

Case F1 (AF_FAMILY mistaken for STACH under noisy ECG). A sample with $\mathrm{HR}_{\mathrm{PPG}}$=138 bpm, $\mathrm{RR_CV}$=0.22, and an ECG-derived P-wave detection ratio of 0.44 (just above our 0.40 threshold). PhysioPatch reports $\widehat{q}$=0.74 and no low-confidence beats; the Skeptic emits only n_{chal}=1 challenge. The Advocate defends STACH citing the above-threshold P-wave ratio; the Arbiter retains STACH at c_{final}=0.79. Ground truth: AF_FAMILY (atrial fibrillation with rapid ventricular response). The boundary is misplaced: a P-wave ratio just above the rule threshold receives no challenge pressure even though $\mathrm{RR_CV}$ is borderline. This failure mode is consistent across $14/47$ AF_FAMILY-vs-STACH errors in the matched test subset. *Implication:* continuous (rather than thresholded) P-wave evidence in the Patch Report would likely yield a softer Skeptic activation here.

Case F2 (SBRAD vs. SR boundary at HR $\approx$ 58 bpm). A sample with $\mathrm{HR}_{\mathrm{PPG}}$=57.6 bpm, sinus P-waves present, $\widehat{q}$=0.91, no cross-modal conflict. The Proposer commits to SR (c_0=0.86), the Skeptic does not challenge the rate (because $\widehat{q}$ is high), and the Arbiter retains SR. Ground truth: SBRAD per the PhysioNet code. The error is a near-boundary labeling artifact rather than an SQI failure: the encoder *can* distinguish 58 from 65 bpm, but the rule basis the model has learned treats 58 ± 2 as ambiguous. *Implication:* the SQI conditioned mechanism does not address class-boundary calibration on inherently borderline rhythms; a separate threshold-uncertainty mechanism would be required.

Case F3 (PACE missed when pacing spikes are masked by motion artifact). A patient with severe motion artifact ($\widehat{q}$=0.31) and intermittent ventricular pacing. PhysioPatch reports six low-confidence beats; the Skeptic correctly issues n_{chal}=3 challenges flagging beats $\#3, \#9, \#15$. The Advocate concedes two of three, but because the visually salient pacing spike is itself *within* the low-confidence span, no role can confidently anchor on PACE; the Arbiter emits AF_FAMILY at c_{final}=0.52 and raises the expert-review flag. Ground truth: PACE. *Implication:* when the diagnostic feature itself coincides with the corrupted region, no amount of debate conditioned on SQI can recover the correct class; the system *does* surface uncertainty (review flag at $c<0.70$), but its substantive prediction is wrong.

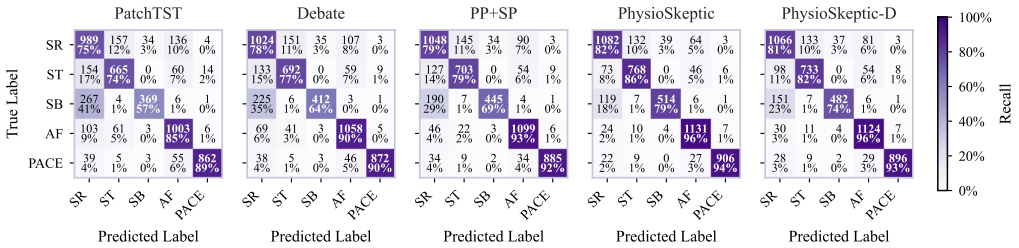

Figure S4: Confusion matrices for PatchTST, Debate, PP+SP, and PhysioSkeptic variants.

D.5 Additional Performance Analyses

This section provides diagnostic-difficulty stratification, multi-seed reliability statistics, and demographic subgroup audits.

These strata distinguish ordinary performance from the cases most likely to benefit from skeptical routing. In particular, the low-SQI-plus-conflict column combines the two evidence hazards rather than averaging them away.

Table S13 complements Section D.2 by reporting performance over the same diagnostic-difficulty strata.

Table S13: Macro-F1 by diagnostic difficulty.

Method	**Full** matched	**Low** SQI	**X-modal** conflict	**Low+** conflict	**High** SQI
PatchTST	0.794	0.693	0.729	0.587	0.871
GPT-5.2 + CoT	0.774	0.671	0.695	0.531	0.860
GPT-5.2 + SC×5	0.792	0.692	0.720	0.554	0.876
Debate (Du et al.)	0.824	0.718	0.741	0.579	0.897
PP+SP	*0.845*	*0.748*	*0.775*	*0.615*	*0.918*
PhysioSkeptic	**0.886**	**0.779**	**0.844**	**0.695**	**0.941**
Δ (PS vs PP+SP)	*+4.1*	*+3.1*	*+6.9*	*+8.0*	*+2.3*

The multi-seed table reports mean and standard deviation across the same fixed evaluation protocol used for the main comparisons. It is included to show the spread behind the aggregate Macro-F1 values rather than to introduce a new benchmark.

Table S7 reports the class-wise breakdown, while the main text contains the backbone-transfer figure. This appendix keeps only the aggregate multi-seed summary.

Table S14: Multi-seed Macro-F1 statistics.

Method	**Mean**	**Std**
PatchTST	0.794	0.002
GPT-5.2 + CoT	0.774	0.010
GPT-5.2 + SC×5	0.792	0.009
Debate (Du et al.)	0.824	0.001
PP+SP	*0.845*	*0.001*
PhysioSkeptic (GPT-5.2)	0.886	0.003
PhysioSkeptic (DeepSeek-V3.2)	0.866	0.002
Student (Llama-1B)	0.847	0.008

The confusion matrices expose the error redistribution behind the aggregate scores: they are useful for checking whether gains are concentrated in one class or spread across the five-class label space.

Table S15 reports PhysioSkeptic performance across age and sex subgroups on the validation proxy set (n=3,506 matched windows). This metadata-complete proxy is used only for subgroup auditing and is not directly comparable to the matched LLM test subset in Table 1. Race and BMI subgroups are omitted: race is missing for more than 60% of MIMIC-III-Ext-PPG records and BMI is not available in the synchronized waveform cohort. Rev. Flag% counts windows where $c_{\text{final}}<0.70$.

Table S15: Subgroup performance on the validation proxy set.

Subgroup	n	Macro-F1	ECE	Rev. Flag%
Age <45	161	0.490	0.051	25.5
Age 45–65	1,068	0.903	0.050	17.9
Age >65	2,277	0.869	0.041	14.9
Female	1,292	0.880	0.037	17.8
Male	2,214	0.887	0.036	15.4
Low SQI (<0.5)	182	0.682	0.038	22.5
Non-low SQI (≥ 0.5)	3,324	0.888	0.038	15.9

The final three-block table varies routing boundaries, the review threshold, and the source of SQI while holding the rest of the evaluation protocol fixed. The blocks separate compute-policy sensitivity from the quality-signal ablation.

Table S16: Sensitivity to routing thresholds and SQI sources.

Part A: routing boundary sensitivity

Setting	Fast/Std/Deep	Routing F1	ECE	Misrouted%
Conservative	20 / 52 / 28	0.882	0.038	0.3
Default	**31 / 49 / 20**	**0.883**	**0.039**	**1.4**
Aggressive Fast	42 / 43 / 15	0.876	0.043	3.1

Part B: review-flag threshold τ_{rev}

τ_{rev}	Rev. Flag%	Macro-F1	ECE
0.60	10.6	0.886	0.033
0.70 (default)	**18.2**	**0.886**	**0.033**
0.80	30.2	0.886	0.033

Part C: SQI source robustness

SQI source	Macro F1	Low-SQI F1	ECE	Anch. trap%
Learned SQI (full)	**0.886**	**0.779**	**0.033**	**16.8**
Rule-based (SQIJK)	0.868	0.748	0.042	22.4
Constant $= 0.5$	0.841	0.728	0.071	31.2
Shuffled SQI	0.829	0.719	0.087	32.4
No SQI anchoring	0.822	0.726	0.085	–

F1 remains stable across routing settings (0.876–0.883), while aggressive fast-routing raises difficult-case misrouting from 1.4% to 3.1%. Changing τ_{rev} adjusts review volume without changing the diagnostic output. Learned SQI yields the highest F1 and lowest ECE in this comparison; shuffled or constant SQI approaches the no-anchoring range. Rule-based SQI uses SQIJK (Elgendi, 2016); shuffled and no-anchoring rows reuse Tables S12 and 3.

These proxy-set subgroup results are descriptive and are not directly compared with Table 1 or its baselines. The low-SQI stratum (F1 0.682) should not replace the matched-test Low-SQI F1. The young-adult stratum (age<45, n=161) has markedly lower Macro-F1 (0.490); its small, class-sparse, rhythm-distribution-shifted sample precludes a precise demographic-fairness estimate, but identifies an important generalization limitation.

The routing analysis reported in the main text is expanded below. The figure keeps the former three views but composes the runtime and safety summaries as a single left block and the latency–accuracy trade-off as a right-hand view. The table reports the underlying route-level values and the threshold sensitivity used by the adaptive policy.

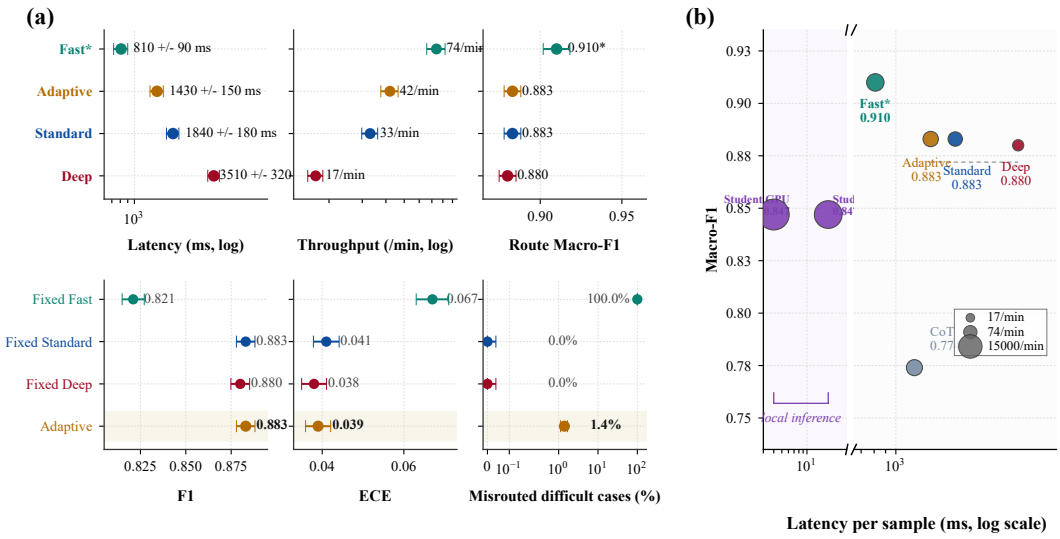

Figure S5: Routing latency, safety, and F1 trade-offs.

Table S17: Runtime, routing quality, and threshold sensitivity.

(a) Runtime and routing subsets

Config.	Latency (ms)	Thr.	F1↑	Roles
Fast path	810 ± 90	74	0.910	P+C+Arb
Standard path	$1{,}840 \pm 180$	33	0.883	P+C+S+Arb
Deep path	$3{,}510 \pm 320$	17	0.880	All roles
Adaptive	$\mathbf{1{,}430 \pm 150}$	**42**	**0.883**	SQI-gated mix
GPT-5.2 + CoT	$1{,}210 \pm 110$	50	0.774	Single
Student (CPU)	$\mathbf{18 \pm 2}$	**3,330**	0.847	Local
Student (A10G)	4 ± 1	15,000	0.847	Local

(b) Adaptive-routing boundaries

Setting	Fast/Std/Deep	Routing F1	ECE↓	Misrouted%↓
Conservative	20 / 52 / 28	0.882	0.038	0.3
Default	**31 / 49 / 20**	**0.883**	**0.039**	**1.4**
Aggressive Fast	42 / 43 / 15	0.876	0.043	3.1