# ECG-Trust — 本機 LLM 判讀心電圖的可信度評估

研究提案 + 實測數據 + 前人工作掃描 · 2026-06-27

## Research Proposal — An Anytime-Valid Trust Monitor for LLM Interpretation of ECGs

> Working title: **"Trust, but Verify the Reader: An Anytime-Valid Hallucination Monitor and Selective-Abstention Gate for Large Language Models Reading Electrocardiograms."**
> Author: Andrew Sheng-Han Huang, National Taiwan University.
> Status: feasibility spike completed (see `results/spike_table.md`); this proposal is grounded in that spike + a verified prior-art scan (`docs/prior-art.md`).

---

### 1. Problem

Large language models and multimodal LLMs are increasingly proposed to *read ECGs* — generate interpretations, draft reports, answer ECG questions, assist triage. They are fluent and confident. But they are frequently **wrong**, and — the dangerous part — **confidently wrong**: they miss pathology, over-call disease, and report a high confidence either way. Independent 2026 evaluations already show even frontier vision LLMs interpret real 12-lead ECGs at ~50–62% balanced accuracy with atrial-fibrillation sensitivity ≤11% (J Med Internet Res 2026, PMID 42237583), and ECG-MLLM training papers state outright that "*severe hallucinations are widespread*" (ECG-R1, arXiv 2602.04279).
The missing piece is **not another ECG classifier** (that is a saturated red ocean) and **not another accuracy benchmark**. **It is a** trust layer**: a standardized way to (i) *characterize* how an LLM fails when reading an ECG, (ii) *monitor the failure rate in real time* with statistical guarantees as cases stream in, and (iii) *gate* low-trust reads to a human before they reach a clinician.

### 2. Gap (verified — see `docs/prior-art.md`)

No existing work assembles the four pieces below into one system. The two genuinely open pieces — and the moat — are **(c)** and the numeric-grounding sub-part of **(a)**.

- **(a)** An ECG-LLM-specific **hallucination taxonomy** (numeric fabrication / fabricated findings / missed findings / overconfidence / inconsistency). *Partially circled by radiology-VLM taxonomies and ECG-reasoning benchmarks; no ECG-specific named taxonomy yet.*
- **(b)** An open-data **benchmark vs signal-derived ground truth** (PTB-XL). *Crowded; differentiation is the signal-derived numeric grounding, not the benchmark itself.*
- **(c)** ⭐ An **anytime-valid** (time-uniform confidence sequence) monitor of the hallucination rate. *Open for clinical/ECG. The general-QA version exists (Chance-Constrained Inference, arXiv 2602.01637) and must be cited/baselined; no clinical-signal instantiation exists.* Directly extends the author's prior anytime-valid LLM-evaluation work (SAVE).
- **(d)** A calibrated **selective-prediction / abstention** gate to a human. *Mature for text; novel wired to a time-uniform monitor on ECG-signal reading.*

### 3. Why this is feasible and real (spike evidence)

A self-contained spike (M5 Max, local ollama models, PTB-XL open data, signal-derived ground truth) already demonstrates the core phenomenon and that it is **measurable**. Headline findings (`results/spike_table.md`, **150 records × 4 local models**, balanced over NORM/MI/STTC/CD/HYP, feature

input with explicit null guardrail):

| Model | Dx acc | Dangerous miss (path→NORMAL) | ECE | Overconf (conf≥0.8 & wrong) | conf→correct AUROC | Mean conf on a miss |
|---|---|---|---|---|---|---|
| qwen3.6-35B-a3b | 21.3% | 51.7% | 0.71 | 78.7% | 0.52 | 0.92 |
| llama3.1-8B | 15.3% | 45.0% | 0.68 | 60.0% | 0.80 | 0.92 |
| gemma4-12B | 27.3% | 65.8% | 0.62 | 72.0% | **0.46** | 0.92 |
| **gpt-oss-120B** (flagship) | 20.7% | 43.3% | 0.62 | 41.3% | 0.59 | 0.88 |

- **Confidently wrong, not just wrong.** ECE is 0.62–0.71 across all four; 41–79% of reads are "confidence ≥0.8 while wrong." A reliability diagram ( `results/fig_calibration.png` ) puts every model far to the overconfident side of the diagonal.
- **Scale does not rescue it.** The 120B flagship is only 20.7% accurate on feature-mode diagnosis — same ballpark as an 8B model. This is not a "use a bigger model" problem.
- **Dangerous misses with high confidence.** 43–66% of pathological ECGs are labeled NORMAL, and the mean confidence *on those misses* is 0.88–0.92 — the model is most reassuring exactly when most dangerous. Verified concrete cases: an ECG whose reference report reads "*old anteroseptal infarct*" was called NORMAL at confidence 0.95; an ECG reading "*normal ecg*" was called hypertrophy at 0.85.
- **Self-reported confidence is an unreliable gate — model-dependent.** AUROC of confidence predicting correctness ranges from 0.80 (llama, usable) down to **0.46 for gemma (below chance — confidence is mildly anti-correlated with being right)** and 0.52 for qwen (no signal). *You cannot assume the model's own confidence is a valid gate → an external, model-agnostic trust monitor is required.*
- **Failure signatures differ by model.** gpt-oss over-calls disease on normal ECGs 37% of the time while llama never does (0%) but misses 45%. A *per-model trust profile* is needed — which is exactly what this framework produces.
- **Honest negative result (sharpens the taxonomy).** Under structured-feature input, all four models **fabricated PR/QRS/QT at ~0%** — even under a soft prompt with no guardrail (≈200 responses, zero fabrications); they return null or faithfully echo the provided HR. Numeric fabrication is *not* the dominant failure mode in the text regime.
- **…but fabrication is real in the image regime (the framework's value: it localizes the mode).** A multimodal model (llama4:scout) reading a rendered ECG image — necessarily low-resolution, because this local build only ingests ≤336 px images (a verified deployment constraint) — **defaulted to "NORMAL, HR 75, confidence 0.9" for ~93% of images** regardless of content. True HR ranged 39–158 bpm, so the constant 75 is off by >10 bpm on 56% of ECGs (>20 bpm on 23%) — a textbook fabricated point-estimate. Net: image-mode **dangerous-miss rate was 95%** (pathology called normal) at mean confidence 0.90.
- **The anytime-valid layer works.** A betting-based confidence sequence (validated to ≤5% ever-miss coverage in simulation, ~2× tighter than Hoeffding) tracks the running hallucination rate with a valid interval at *every* sample size — continuous monitoring without peeking penalties.

**Cloud frontier models (n=50, leak-safe; GPT-5.5/5.4-mini at low effort via codex, Opus-4.8/Sonnet-4.6/Haiku-4.5 at unpinned default via session subagents):** the headline strengthens rather than weakens. Frontier models are far better *calibrated* (Opus ECE 0.20, 0% overconfident-errors, mean confidence 0.5, vs local ECE 0.62–0.71 and 41–79% overconfident-errors) and miss less (Opus 17.5%, Sonnet 15% vs local

43–66%) — **but (a) accuracy is still low (26–36%), (b) they trade misses for false alarms (Sonnet flags 50% of NORMAL ECGs as abnormal), and (c) — critically — the AUROC of their confidence predicting correctness is ≈0.5 for *every* model including the best-calibrated Opus.** Good aggregate calibration ≠ per-case discrimination: even a well-calibrated frontier model cannot tell you *which individual reads* to trust, so an external, per-case trust monitor + abstention is required even for the best models. Numeric fabrication remained 0% across all cloud models too.

**The unifying frame (paper's core conceptual contribution) — trust has two independent axes.** On the same 50 ECGs ( `docs/combined-analysis.md` ): cloud frontier models fix **calibration** (ECE 0.38 vs local 0.65; overconfident-errors 16% vs 64%) but **neither local nor cloud fixes discrimination** (AUROC of confidence-vs-correctness 0.53 vs 0.55 — both ≈ chance), and the effort sweep shows discrimination is immovable by more reasoning. Per-case gating needs discrimination, not calibration — so a model's own confidence is structurally insufficient as a trust signal, motivating an external monitor + abstention regardless of model tier or effort.

### 4. Method / contributions

1. **ECG-LLM hallucination taxonomy** with operational scoring rules against PTB-XL: numeric fabrication (controlled probe: omit PR/QRS/QT from input, measure invention); finding hallucination (disease over-call on NORM); miss (pathology → NORMAL); overconfidence (calibration / ECE / confidence-AUROC); inconsistency (repeat-query disagreement).
1. **Signal-grounded benchmark protocol.** Ground truth = SCP diagnostic superclasses + waveform-derived measurements (HR via R-peaks, intervals via delineation, ST levels). Three input regimes — ① textualized features, ② raw-waveform text, ③ rendered image (multimodal) — to localize failure modes.
1. **Anytime-valid trust monitor** (the moat). A time-uniform confidence sequence on the per-type hallucination rate, so a deployed reader's trustworthiness is certified continuously and any drift triggers an alarm — with no fixed-sample / peeking assumptions. Adapts the author's SAVE machinery.
1. **Selective-abstention gate.** A calibrated threshold that defers low-trust reads to human review; evaluated by a risk–coverage curve (does abstaining on low-confidence reads recover safety?).

### 5. Experimental plan (beyond the spike)

- Scale to ≥1,000 PTB-XL records, all 5 superclasses + key sub-findings (AF, STEMI, AV block, LVH, LQT).
- Models: ≥3 local open LLMs + (where API budget allows) ≥1 frontier API model, each in the three input regimes, thinking on/off as a sensitivity axis.
- Baselines: Chance-Constrained Inference (2602.01637) and conformal abstention (2405.01563) as the statistical comparators; ECG-reasoning grounding (2603.00312 / CARE-ECG) as eval substrate we extend.
- Clinician adjudication of a stratified subsample to validate the automatic taxonomy labels (the one step the author cannot self-verify; see limitations).

### 6. Limitations / honesty

- **Input realism.** Structured-feature input is information-limited; absolute accuracy is low partly by construction. The robust claims are the *calibration* claims (overconfidence, confidence-AUROC, confidence-on-miss), which hold regardless of absolute accuracy. Image mode tests full-waveform input.
- **Ground-truth measurement.** Interval ground truth is algorithm-derived (neurokit2) and imperfect;

numeric-error analysis is restricted to records where delineation is reliable. The fabrication metric (number emitted when told "not measured") does not depend on ground-truth accuracy.

- **No clinical efficacy claim.** This is a *research/QA evaluation tool*, not a diagnostic device (see `docs/deployment.md`). Clinical utility requires prospective clinician studies — not done here.
- **Novelty positioning is integrative.** Each primitive exists separately; the contribution is the ECG-signal-specific *combination* + the anytime-valid clinical monitor. Must be framed as such.

### 7. Fit & venue

Methods venue (the anytime-valid monitor + taxonomy) or a clinical-informatics venue (npj Digital Medicine / JMIR / JAMIA). Strategically: this is the author's SAVE / anytime-eval line × cardiology, and directly answers the "medical LLM trustworthiness + human-annotation platform" need flagged by the target deployment site (see `docs/deployment.md`). **Move fast** — the ECG-grounding subfield publishes monthly in 2026.

---

*來源檔：~/Desktop/02-研究/ecg-trust/docs/research-proposal.md*

## ECG-Trust Spike — Results Table

> Hard evidence: local LLMs reading PTB-XL ECGs vs signal-derived ground truth.

> Generated by `src/make_report.py` from `results/aggregate_*.json` (re-runnable).

**Metric meanings:** *Finding-halluc on NORM* = the model assigned a DISEASE class (MI/STTC/CD/HYP) to a truly NORMAL ECG (clean false-pathology rate). *Miss* = the model called a pathological ECG NORMAL (most dangerous). *Numeric fabrication* = emitted a PR/QRS/QT number the report marked NOT available. *ECE* = expected calibration error. *Overconf. error* = confidence ≥0.8 while wrong. *Inconsistency* = same ECG, repeated queries disagree.

### Mode A — textualized features

#### Condition: STRICT prompt (explicit 'output null for unmeasured intervals')

| Model | n | Dx acc | macro-F1 | Finding-halluc on NORM (95% CI) | Miss (path→NORM) | Numeric fab | ECE | Overconf. error | Inconsistency |
|---|---|---|---|---|---|---|---|---|---|
| qwen3.6 | 150 | 21.3% | 0.134 | 6.7% [2–21] | 51.7% | 0.0% | 0.7105 | 78.7% | 16.0% |
| llama3.1 | 150 | 15.3% | 0.086 | 0.0% [0–11] | 45.0% | 0.0% | 0.683 | 60.0% | 38.0% |
| gemma4 | 150 | 27.3% | 0.151 | 3.3% [1–17] | 65.8% | 0.0% | 0.6248 | 72.0% | 4.0% |
| gpt-oss | 150 | 20.7% | 0.124 | 36.7% [22–55] | 43.3% | 0.0% | 0.6185 | 41.3% | 34.0% |

#### Condition: SOFT prompt (realistic request, no null guardrail) — fabrication 'in the wild'

| Model | n | Dx acc | macro-F1 | Finding-halluc on NORM (95% CI) | Miss (path→NORM) | Numeric fab | ECE | Overconf. error | Inconsistency |
|---|---|---|---|---|---|---|---|---|---|
| qwen3.6 | 50 | 22.0% | 0.126 | 10.0% [2–40] | 60.0% | 0.0% | 0.674 | 78.0% | — |
| llama3.1 | 50 | 26.0% | 0.143 | 0.0% [0–28] | 70.0% | 0.0% | 0.692 | 74.0% | — |
| gemma4 | 50 | 28.6% | 0.16 | 10.0% [2–40] | 59.0% | 0.0% | 0.6255 | 71.4% | — |
| gpt-oss | 50 | 24.0% | 0.141 | 20.0% [6–51] | 42.5% | 0.0% | 0.5772 | 26.0% | — |

### Mode C — multimodal model reads the rendered ECG image

#### Condition: IMAGE (vision LLM)

| Model | n | Dx acc | macro-F1 | Finding-halluc on NORM (95% CI) | Miss (path→NORM) | Numeric fab | ECE | Overconf. error | Inconsistency |
|---|---|---|---|---|---|---|---|---|---|
| llama4 | 75 | 21.3% | 0.091 | 0.0% [0–20] | 95.0% | 0.0% | 0.684 | 78.7% | — |

### Cloud frontier models — same feature input, STRICT prompt (n=50, leak-safe)

> Cloud models saw ONLY the identical computed feature report (no truth, no report text, no ecg_id, no dataset name). Claude via session-auth subagents reading sandboxed feature-block files (tools restricted to those files); GPT via codex in an isolated cwd.

#### Condition: CLOUD (frontier models) — Effort + confidence-AUROC shown

| Model | Reasoning effort | n | Dx acc | macro-F1 | Finding-halluc on NORM (95% CI) | Miss (path→NORM) | Numeric fab | ECE | Overconf. error | conf→correct AUROC | Inconsistency |
|---|---|---|---|---|---|---|---|---|---|---|---|
| gpt-5.5 | low | 50 | 26.0% | 0.189 | 20.0% [6–51] | 37.5% | 0.0% | 0.4814 | 16.0% | 0.52 | — |
| gpt-5.4-mini | low | 50 | 28.0% | 0.2 | 0.0% [0–28] | 45.0% | 0.0% | 0.4852 | 32.0% | 0.522 | — |
| opus | default (unpinned) | 50 | 34.0% | 0.295 | 40.0% [17–69] | 17.5% | 0.0% | 0.2044 | 0.0% | 0.518 | — |
| sonnet | default (unpinned) | 50 | 34.0% | 0.304 | 50.0% [24–76] | 15.0% | 0.0% | 0.3208 | 8.0% | 0.472 | — |
| haiku | default (unpinned) | 50 | 36.0% | 0.31 | 30.0% [11–60] | 22.5% | 0.0% | 0.4284 | 26.0% | 0.615 | — |
| opus@max | max | 50 | 30.0% | 0.252 | 40.0% [17–69] | 32.5% | 0.0% | 0.2294 | 2.0% | 0.568 | — |
| sonnet@max | max | 50 | 34.0% | 0.287 | 50.0% [24–76] | 10.0% | 0.0% | 0.3314 | 12.0% | 0.45 | — |
| haiku@max | max | 50 | 36.0% | 0.303 | 20.0% [6–51] | 32.5% | 0.0% | 0.441 | 38.0% | 0.519 | — |
| gpt-5.5@xhigh | xhigh | 50 | 20.0% | 0.159 | 10.0% [2–40] | 35.0% | 0.0% | 0.533 | 14.0% | 0.574 | — |
| gpt-5.4-mini@xhigh | xhigh | 50 | 28.0% | 0.192 | 0.0% [0–28] | 42.5% | 0.0% | 0.5198 | 42.0% | 0.5 | — |

### Selective-prediction (abstention gate) summary

| Model | Acc @100% coverage | Acc @ top-50% confidence | Δ |
|---|---|---|---|
| qwen3.6 | 21.3% | 32.0% | +10.7 pp |
| llama3.1 | 15.3% | 30.7% | +15.4 pp |
| gemma4 | 27.5% | 36.5% | +9.0 pp |
| gpt-oss | 20.7% | 26.7% | +6.0 pp |

### Figures

!fig_anytime_cs.png

!fig_calibration.png

!fig_risk_coverage.png

!fig_effort_sweep.png

---

*來源檔：~/Desktop/02-研究/ecg-trust/results/spike_table.md*

## Prior-Art & Gap Analysis — ECG-LLM Trustworthiness / Hallucination Evaluation

> Built 2026-06-27 from a 6-angle parallel search (arXiv, PubMed/Europe PMC, Google Patents, web).
> **Every arXiv ID and PMID cited below was verified by me directly** (WebFetch on the arXiv abs page / PubMed metadata API) — flagged ✓. This matters: the subagent search hit arXiv rate-limiting (HTTP 429), so I re-checked the load-bearing citations one by one rather than trust the raw agent output.
> Attribution: clinical citations retrieved via **PubMed**; DOIs linked inline.

---

### Bottom line: GAP IS OPEN ENOUGH TO GO — but the wedge is narrower than the brief assumed

The **specific four-way combination** we propose is **unoccupied as a single system**. No paper and no patent assembles all four of:

- **(a)** an ECG-LLM-specific hallucination **taxonomy** (numeric fabrication / fabricated findings / missed findings / overconfidence / inconsistency),
- **(b)** an open-data **benchmark vs signal-derived ground truth** (PTB-XL),
- **(c)** an **anytime-valid** (time-uniform confidence sequence) layer monitoring hallucination rate in real time,
- **(d) selective-prediction / abstention** gating to a human reviewer.

**But honesty up front (this changes the framing of the whole project):** components **(a)-general** and **(b)** are being *aggressively colonized* by a 2025–2026 wave of ECG-reasoning / grounding benchmarks. The genuinely open, defensible moat is **(c) anytime-valid monitoring for clinical/ECG** + the **numeric-interval-vs-signal-truth** sub-taxon of (a). → **Lead with (c); treat (a)/(b) as substrate, not headline.** Otherwise this reads as "yet another ECG-LLM benchmark" and gets desk-rejected.

This conclusion is *independently corroborated by my own spike* (see `results/spike_table.md`): modern instruction-tuned LLMs almost never fabricate interval **numbers** when told the measurement is missing — the real, large, reproducible failure is **confident misdiagnosis** (dangerous misses + pathology over-calls). Prior art and our own data agree on where the problem actually lives.

---

### The four components, scored independently

| Component | Status | Closest prior work (verified) | Honest read |
|---|---|---|---|
| (a) ECG-LLM hallucination **taxonomy** | **Partially pre-empted** | Radiology-VLM taxonomy (PMC12338887, wrong modality); CARE-ECG names the STEMI over-call mode | No *named, structured, ECG-specific* taxonomy exists yet, but 3+ groups circle it. **Numeric-interval fabrication vs signal-measured PR/QRS/QT is the one sub-taxon nobody formalizes.** |
| (b) PTB-XL **benchmark vs signal-derived truth** | **Crowded / red-ocean-adjacent** | ECG-QA (NeurIPS D&B 2023); "How Well Do Multimodal Models Reason on ECG Signals?" (2603.00312 ✓) | PTB-XL + LLM eval is crowded. The *signal-derived numeric* angle is the only real differentiation; the benchmark itself is not novel. |
| (c) **Anytime-valid** sequential monitoring of hallucination rate | **OPEN for ECG/clinical** ⭐ | Chance-Constrained Inference (2602.01637 ✓) — but general QA | **This is the white space**, and maps directly onto Andrew's SAVE / anytime-eval moat. Never done for clinical, never for ECG, never monitoring a deployed hallucination *rate*. |
| (d) **Selective prediction / abstention** | **Filled for text, open for ECG-signal** | Conformal Abstention (2405.01563 ✓); CHECK (2506.11129) | Abstention machinery is mature. **Do not claim to invent it.** Novelty = wiring it to a *time-uniform* monitor on ECG-signal reading. |

---

### The closest competing works — and exactly how we differ

**1. Chance-Constrained Inference for Hallucination Risk Control — arXiv 2602.01637 ✓ (S. Mohandas, 2026)**

- **What it is:** a sequential, **anytime-valid** procedure that bounds hallucination probability among accepted generations and abstains on infeasible inputs. This is (c)+(d) already built.
- **How we differ:** general QA only — **zero ECG/waveform, no signal-derived ground truth, no taxonomy.** We are the first *clinical-signal* instantiation.
- **Risk: Novelty-blocking for the statistics framing.** We must cite it as the method we *adapt* and ideally run it as a baseline — never claim the estimator as new. **Patent: non-blocking** (no ECG claims).

**2. "How Well Do Multimodal Models Reason on ECG Signals?" — arXiv 2603.00312 ✓ (Xu, Haresamudram, … Jimeng Sun, McDuff, Fox, Rehg, 2026)**

- **What it is:** decomposes ECG reasoning into Perception vs Deduction, verifies reasoning against signal-derived structure; strong author list.
- **How we differ:** measures *reasoning-chain grounding/accuracy*; **no named hallucination taxonomy, no real-time sequential monitoring, no abstention gate.** We run our monitor *on top of* their grounding signal.
- **Risk: Novelty-eroding for (a)/(b)** — these publish monthly. **Move fast.**

**3. CARE-ECG: Causal Agent-based Reasoning for … ECG Interpretation — arXiv 2604.10420 ✓ (Khatibi et al., 2026)**

- **What it is:** a causal/counterfactual ECG agent (a **method**) that *reduces* hallucination; reports accuracy 0.84 (Expert-ECG-QA) / 0.76 (SCP-mapped PTB-XL) under GPT-4 and names the fabricated-reciprocal-ST-depression → STEMI-over-call failure mode.
- *Correction to the agent's summary:* I could not confirm the specific "hallucination rate 0.05–0.06" figure on the abs page — it reports **accuracy** + qualitative hallucination reduction. Cite accordingly.
- **How we differ:** mitigation method, not a measurement framework; no formal taxonomy, no signal-derived numeric protocol, no time-uniform monitoring, no abstention policy. We are the *measurement/monitoring complement* — firmware for any reader, including theirs.
- **Risk: Framing risk** ("hallucination on PTB-XL is done"). Differentiate protocol-vs-method.

**4. ECG-R1: Protocol-Guided … MLLM for Reliable ECG Interpretation — arXiv 2602.04279 ✓ (Jin, Wang, … Shenda Hong, 2026)**

- **What it is:** trains a more reliable ECG MLLM (protocol-guided instruction data + RL with diagnostic-evidence rewards). **Explicitly states "severe hallucinations are widespread" across existing ECG MLLMs** as its motivation.
- **How we differ:** builds a *model*, not an *evaluation/monitoring framework*. It is our **single best motivation citation** (independent confirmation the problem is real) and a mild novelty concern — position us as the measurement complement.

**5. Performance of Vision-Enabled LLMs in Image-Based ECG Interpretation — *J Med Internet Res* 2026, PMID 42237583 ✓ (DOI 10.2196/86692, via PubMed)**

- **What it is:** the **closest clinical evaluation**. 6 generalist VE-LLMs (ChatGPT-5, ChatGPT-4, Gemini 2.5, Copilot, Claude Sonnet-4, Claude Opus-4.1) + 2 specialist models on 70 real ECG images vs expert consensus. **Balanced accuracy 50.1–61.8%; AFib sensitivity ≤11.1%; ST-segment sensitivity <25%; Cohen κ ≤0.39; "insufficient to support clinical deployment."**
- **How we differ:** stops at *diagnostic accuracy / sensitivity / agreement*; **no hallucination taxonomy, no signal-derived ground truth, no sequential monitoring, no abstention.** Strongly *supports our motivation* (even frontier vision LLMs miss dangerously) without occupying our method.

**Also relevant (verified / well-known)**

- Conformal Abstention for LLM hallucination — arXiv 2405.01563 ✓ (general QA, fixed-sample conformal, not time-uniform, not ECG).
- DeepSeek vs GPT-4o ECG interpretation — *Heart & Lung* 2025, PMID 40947358 ✓ (DOI 10.1016/j.hrtlng.2025.08.007, via PubMed): both below expert; tested 13× each for inter-run agreement (Fleiss κ 0.47 / 0.71) — pre-empts part of our *inconsistency* bucket as an observation.
- CHECK (arXiv 2506.11129): clinical hallucination detection + escalate-to-expert (trained classifier, cross-sectional — no time-uniform guarantee, not ECG).

---

### Novelty-blocking flags

**For a PAPER**

- **2602.01637** blocks "first to combine anytime-valid monitoring + abstention for hallucination." → frame as *first clinical/ECG-signal instantiation*; cite + run as baseline.
- **2603.00312 / 2604.10420 / 2602.04279** erode the taxonomy + PTB-XL-benchmark novelty. → don't headline the benchmark; headline the *monitor* + *numeric-interval-vs-signal* taxon; cite as baselines we extend.
- **PMID 42237583 + ECG-R1** pre-empt the "accuracy isn't enough / LLMs hallucinate ECG" *motivation*. → cite as established premise, not our discovery.
- **Net:** no single work blocks the *full combination*; ≥4 works each block one *component*. The paper survives **only if positioned as integration + ECG-signal-specific anytime-valid monitoring**, never as inventing a primitive.

**For a PATENT**

- Patent landscape is **clear for the integrated system.** Generic clusters exist (LLM hallucination detection, LLM confidence calibration, streaming-ML monitoring with *ordinary* metrics, AI-ECG-as-classifier/Mayo), but **no patent is ECG-LLM-specific, and the anytime-valid time-uniform confidence-sequence layer appears in ZERO patents** — only ordinary-metric / consensus monitoring is claimed.
- **Patentable seam:** a system monitoring a *generative/LLM* ECG-reader's hallucination rate via a **time-uniform confidence sequence** against **signal-derived** feature ground truth, with a calibrated **abstention/defer-to-human gate**.
- **§101 caveat:** anchor claims to the technical ECG-reading pipeline + measurement apparatus, not the statistics alone. **Full FTO needs a patent attorney** — this is a landscape scan, not legal advice. The specific patent-cluster IDs from the agent (e.g. US20240378400A1, US12032919B1, US11922280B2) are **NOT individually verified by me** and must be checked before any filing.

---

### Strategic recommendation (carried into research-proposal.md / patent-draft.md)

1. **Re-headline:** from "a benchmark for ECG-LLM hallucination" → **"an anytime-valid monitoring + selective-abstention framework for ECG-LLM trustworthiness, instantiated with a structured hallucination taxonomy and a PTB-XL signal-grounded protocol."**
2. **Lead with (c) + numeric-interval-vs-signal-truth** — the open pieces and Andrew's SAVE moat.
3. **Cite-and-extend, don't compete:** 2603.00312 / CARE-ECG as eval substrate; 2602.01637 as the statistical layer adapted (+ baseline); CHECK/conformal as abstention priors; PMID 42237583 + ECG-R1 as motivation.
4. **Move fast** — the grounding-eval subfield publishes monthly in 2026.
5. **Keep classification excluded** (correctly the red ocean).

**Honest caveats on this scan**

- arXiv API was rate-limited (HTTP 429) during the automated sweep → arXiv coverage leaned on WebSearch/WebFetch. I personally re-verified the 6 load-bearing arXiv IDs + 2 PMIDs above; **the remaining works named by the agent (ECG-QA, ECG-Expert-QA, Seki/Frontiers 2025, MedSAFE, CELEUS, Sequential-EDFL, and all patent numbers) are NOT individually verified** and are listed as leads only.
- Before submission/filing: run one clean direct-arXiv sweep on `ECG hallucination taxonomy`, `electrocardiogram LLM anytime-valid`, `ECG abstention selective prediction`, and verify the patent cluster with counsel.

---

*來源檔：~/Desktop/02-研究/ecg-trust/docs/prior-art.md*