

The End of Benchmarks. The Beginning of Certification.

How DeepSensi and the Gomola Framework move medical AI out of the leaderboard era and into the world of safety norms - where aviation and nuclear power have lived for half a century

Feature · DeepSensi PBC, Dover, Delaware · July 2026 · press@deepsensi.com

An aircraft engine, before it lifts its first passenger, must demonstrate a failure rate below one in a billion per flight hour. A nuclear reactor, before fuel is loaded, must satisfy the safety-integrity levels of IEC 61508. A drug, before it reaches a pharmacy, completes a Phase III trial. Clinical artificial intelligence - a technology whose errors are measured in human lives - is today required to demonstrate none of this. It can be deployed in a hospital with no quantitative reliability requirement, no certification framework, and no measurable bound on its most dangerous flaw: hallucination, the fluent, convincing untruth.

Tomasz Jan Gomola, the creator and architect of DeepSensi, set out to close that gap - not a product gap, but a civilizational one. His answer comes in two parts that must be judged together: the **Gomola Framework**, the first quantitative safety-certification standard for clinical AI, open and free to the entire industry including its competitors; and **DeepSensi Medical OS**, an operational system that became the first to submit itself to that standard's rigor.

Not a better model. An architecture above models.

The core idea is counterintuitive. DeepSensi does not compete in the race for ever-larger language models. Inside the system there is never one model at work - there are many, from different, independent vendors, and the architecture trusts none of them. Every diagnostic assertion must pass a cascade of twenty-three independent verification barriers, from deterministic checks of physiological coherence, through examination of the provenance and reliability of every cited source, to adversarial attacks by dedicated agents whose only job is to demolish a conclusion before a physician ever sees it. The mathematics of that cascade - derived by fault tree analysis, the same method that certifies power plants and avionics - yields a bound: under worst-case operating conditions, fewer than one undetected hallucination per roughly three hundred thousand assertions. That figure meets the numerical target of the highest safety-integrity level, SIL-4, with a thirty-one-fold margin.

From this construction follows something invaluable to investors and regulators, and unique in AI: **anti-fragility**. The question that kills most AI projects - "what happens when the next model is better?" - has here a one-sentence answer: it gets better. A better model in yields a better verified answer out. The model race, a threat to everyone else, is a free engine of progress for this architecture.

The first AI that says "I don't know"

The most human element of the system is a protocol called LIMBO. When the assembled evidence is insufficient for a safe conclusion, DeepSensi does not guess - it declares uncertainty in structured form: it names the core of the disagreement between its specialist perspectives, offers a working hypothesis, and specifies the exact tests that would resolve the

doubt. In medicine, where the most dangerous sentence is a confident mistake, an honest "I don't know" with a plan is often worth more than a brilliant-sounding answer.

That this is not a marketing line is shown by the numbers from a validation on 301 New England Journal of Medicine clinicopathological conference cases from 2014–2023 - the hardest public diagnostic examination in the world. After safety-first calibration the system reached 86% first-diagnosis accuracy and 93.7% within the top three, at a median deliberation time of fourteen seconds. But the study's most important number is a different one: **the rate of missed critical, life-threatening diagnoses fell from 4.7% to 0.0% - at a cost of 2.3 seconds of additional analysis.** If a single sentence captures the philosophy of the project, it is that one. Safety is not a tax on effectiveness - it costs two extra seconds, and it saves the patients that statistics call "the tail of the distribution."

The flawed yardstick

Along the way, Gomola's team made a discovery that may sting the whole industry. Testing the system against popular medical benchmarks, they turned its evidence-verification layers not on their own answers but on the test questions themselves - and found flaws. A companion paper, "The Flawed Yardstick," documents cases where automated scoring gave a system zero points for clinically exemplary behavior: for refusing to recommend physical therapy to a patient with signs of cauda equina syndrome; for verifying the quality of resuscitation before confirming adrenaline. The author calls it the Safe Triage Paradox: today's benchmarks reward the bold guess and punish the caution we teach medical students. The proposed alternative - the Automated Clinical Safety Audit - scores an honest uncertainty above a speculative certainty. If it is adopted, it will change how the world measures medical AI.

A system, not an app

What the newest technical papers reveal is that verification is only the foundation. On top of it runs a complete cognitive infrastructure. The **Hyper Consilium** treats the physician not as an approval gate but as a mathematically scored cognitive node deliberating alongside thirty-seven specialist agents - with a guaranteed floor ensuring a doctor's voice is never reduced to zero, and a formal economic model, a first of its kind, that pays physicians royalties for verified clinical knowledge, released only after patient outcomes confirm the contribution helped. The **AutoResearcher** engine autonomously generates and adversarially validates research hypotheses with a formal Bayesian convergence guarantee - anchored in deterministic biological-network validation so that no model can inflate a result with fabricated mechanism. And **Golden Horizon** turns compassionate access from a physician-initiated privilege into an actively advocated patient right: it scans the world's trial registries and deploys an autonomous agent that pursues providers on a dying patient's behalf - free by architectural mandate, a zero-cost constant hardcoded so that no future management can revoke it.

The standard, given away

Perhaps the boldest decision is the distribution model: the Gomola Framework and the DeepSensi Standard are open and free of licensing fees. Any vendor may certify against them, any hospital may write a required certification level into a tender, any regulator may reference the methodology, any insurer may price risk against a measurable error bound - because a quantified bound is the first step toward restoring AI's insurability, which the policy market is currently withdrawing. "A safety standard behind a paywall is not a standard - it is a product," Gomola writes. In the standard's documents one more sentence stands out, rare in any company's materials: the author classifies his own system more conservatively than the raw

mathematics would permit, on the ground that "a standard that flatters its own author is not a standard."

What comes next

The system is operational, running on edge nodes across four continents - including fully offline, from hardware costing tens of dollars to sovereign installations - and is entering the phase Gomola set for himself: hospital calibration, dialogue with the FDA through the Q-Submission pathway, and alignment with the EU AI Act ahead of its 2027 deadline. Six scientific papers are heading to medRxiv and to peer review, with the full benchmark package, a protocol open to auditors, and an invitation that, coming from the architect of an anti-hallucination system, is unusually credible: find, in a verified output, even one hallucination.

The history of technology knows this moment. It came to aviation when the question stopped being how far the best prototype flew and became what failure rate every unit guarantees. It came to pharmacy when enthusiasm was replaced by Phase III. If Gomola's proposal takes hold - and the openness of the standard, the hard mathematics, and the humility toward his own results give it a real chance - a decade from now no one will remember that medical AI could once be deployed without a certificate. That is what winning standards look like: they become invisible, because they become obvious.

DeepSensi PBC is a Public Benefit Corporation. Its three mission pillars - free global clinical-trial matching for patients, open research into autism and refractory epilepsy, and an open safety standard - are written into the company's charter. Technical documentation: www.deepsensi.com/papers · Press and auditor access to research protocols: press@deepsensi.com