plainvue.ai

Physician-built lab-test interpretation

What Healthcare AI Really Costs

And who pays. The environmental, computational, and privacy costs of AI in healthcare.

A white paper on why Plainvue was built to be responsible, and why a frontier model in the cloud is not

September 2026 • Version 1.0

Prepared as a business and positioning brief

The moment

Healthcare is racing to adopt AI, and for most companies that means one thing: sending patient data to a large, general-purpose "frontier" model such as the ones behind ChatGPT, and getting back generated text. That approach carries three growing liabilities, each now the subject of active public concern, peer-reviewed research, and emerging regulation. It is environmentally costly. It puts private health data at risk. And it is clinically unreliable.

Plainvue was built deliberately to avoid all three. This paper explains what responsible AI means to us, and why our architecture, a lightweight, purpose-built, self-hosted system rather than a frontier model in the cloud, is responsible on each front.

Environmentally responsible

The environmental cost of generative AI is real, large, and rising, and it is driven specifically by big general-purpose models, not by computation itself.

Generative, multi-purpose AI models consume up to roughly 4,600 times more energy than task-specific models doing comparable work, even after accounting for model size.[1] The problem is a property of frontier, general-purpose models, not of AI in the abstract.

The ongoing cost is inference, not training. For large-scale deployments, LLM inference is estimated to consume about 25 times more computational resources annually than training, and one analysis puts inference's annual carbon footprint at up to 1,400 times that of training.[2] This matters because the recurring cost of a health chatbot is every question every patient asks.

Per query, it adds up fast. A single GPT-4o query is estimated at about 0.43 watt-hours, and the most intensive reasoning models exceed 33 watt-hours per long prompt, more than seventy times a lighter-weight model; at hundreds of millions of queries per day, this scales to electricity use comparable to tens of thousands of homes.[3] Water is part of the footprint too: large models were estimated to use roughly 500 milliliters of water per ten to fifty responses, and AI's 2025 water footprint has been projected in the hundreds of billions of liters.[4,5]

The macro trend is steep. AI was estimated to account for 15 to 20 percent of total data-center electricity demand by the end of 2024, and the International Energy Agency projects global data-center electricity consumption could double by 2026, driven largely by AI.[5,6] AI's 2025 carbon footprint alone has been estimated at 32.6 to 79.7 million tons of CO_2.[5]

The comparison that matters

Consider 100,000 lab-test interpretations. Running them through a frontier model, the approach behind tools like ChatGPT, means 100,000 calls to a massive general-purpose AI: each one, by the published estimates above, drawing meaningful energy and water, and sending patient data out of the covered environment. Running the same 100,000 through Plainvue means a lightweight, self-hosted, task-specific model, the category research shows uses a fraction of the energy of general-purpose AI, with no patient data ever leaving. The same volume of results, a categorically different footprint.

Plainvue does not run on a frontier model, and patient data is never sent to one. The platform is a lightweight, self-hosted system: the clinical interpretation is largely predetermined and governed rather than generated fresh on every request, and the only steps that use meaningful compute are reading an uploaded lab report and the patient chat. Both are minimal, both run on our own self-hosted model rather than any third-party frontier model, and the report-reading step largely disappears once lab partners feed results to us directly, leaving it relevant mainly when a patient adds older results from their

own history. In the language of the research above, Plainvue is a task-specific system, the category shown to use a fraction of the energy of general-purpose generative models. We do not claim a zero footprint. We claim, accurately, that Plainvue delivers AI-grade interpretation without the data-center energy and water profile of a frontier-model approach.

Privacy and security responsible

The dominant way to build health AI today, sending patient data to an external frontier model, is also the riskiest for privacy.

Health data is among the most sensitive and most regulated categories of personal information, and it is a frequent target. Healthcare has repeatedly ranked among the most breached sectors, and the cost of a healthcare data breach is consistently the highest of any industry.[7] Sending a patient's results to a third-party, general-purpose model means that data leaves the covered environment, which raises real questions under HIPAA and a growing patchwork of state privacy laws, and places protected health information in systems the provider does not control.

Plainvue is built the opposite way. Protected health information never leaves Plainvue's private environment and is never sent to an external or frontier model, for any purpose, including report reading and chat, which run on our own self-hosted model. There is no third-party AI in the data path. This is a deliberate architectural choice, and it is the same choice that keeps Plainvue lightweight: a self-hosted model rather than a call out to the cloud is what keeps patient data contained. Privacy and efficiency come from the same decision.

Clinically responsible

The third liability of the frontier-chatbot approach is clinical. General-purpose models can be confidently wrong, are not reproducible, and are not built to practice restraint.

General-purpose models hallucinate, producing fluent, authoritative text that can be factually wrong, a failure mode that is especially dangerous in medicine. They are not reproducible: the same question is expected to yield different answers on different runs or model versions, which is unacceptable when a patient's health decisions depend on the output. And they have no built-in clinical restraint: asked what to do about an abnormal result, a general model tends to suggest action rather than recognize when watchful waiting, or nothing at all, is the correct answer.

Plainvue's interpretation is governed. It is grounded in clinical guidelines and physician-reviewed rules, every recommendation is traceable to its source, and the same input produces the same output every time. It grades each abnormal result by severity and recommends follow-up only when a physician

would, which both improves care and avoids the over-testing that a flag-everything system produces. It is designed to support the clinician, not to replace the visit or diagnose the patient.

One idea, not three

These are not three separate features. They come from a single architectural decision. Choosing a lightweight, self-hosted, governed system over a frontier model in the cloud is simultaneously what makes Plainvue lighter on the environment, safer for patient data, and more reliable clinically. Responsibility is not a layer added on top of Plainvue. It is a property of how Plainvue is built.

Sources

1. Projective study of corporations' AI environmental impacts (generative versus task-specific energy). arXiv:2501.14334
2. Temperature-aware scheduling of large language model inference (inference versus training compute and carbon). arXiv:2603.07810
3. How Hungry is AI? Benchmarking energy, water, and carbon of LLM inference. arXiv:2505.09598
4. Comparing resource utilization of large language models and relational databases (water per response). arXiv:2404.08727
5. The carbon and water footprints of data centers and AI. ScienceDirect: S2666389925002788
6. Systematic review of electricity demand for large language models and data centers. ScienceDirect: S1364032125008329
7. IBM, Cost of a Data Breach Report (healthcare breach cost, annual).