THE KENYA INSTITUTE FOR CLINICAL AI · NAIROBI · 2026

A Model We Can Answer For

Running the Kenya Institute for Clinical AI's own clinical language model: the candidates, the hardware, the power and backup design, the software, the team, the cost — and what none of it does.

Dr Neal Aggarwal

Medicine · Surgery · Engineering · Information Technology · Artificial Intelligence

28 SEPTEMBER 2026

Commercial chatbots hedge when asked for medical advice, and the obvious conclusion is that the Kenya Institute for Clinical AI should run its own model so that its clinicians get straight answers. I think the conclusion is right and the reason is wrong. Refusal is the weakest argument for local inference; the strong ones are that an API call carrying patient data is a cross-border transfer under section 47 of the Digital Health Act, that a hosted model can change underneath a validation study without notice, and that the Pharmacy and Poisons Board's 2026 guideline on medical device software now expects change control, rollback and traceability that only an owner of the artefact can provide. This essay sets out what running the Institute's own model actually involves: which open-weight models are serious candidates and what each is built from, what 'controlling the source code' can and cannot mean when the thing that matters is a trillion numbers nobody can read, the hardware in three tiers with landed Kenyan costs, the power and backup design for a grid that went dark nationwide on 29 July, the full software stack, the team, a five-year cost of ownership, whether DeepSeek V4 is an answer, and — at length — what none of this does.

Contents

> *I write here in a personal capacity. This essay belongs to the series that began with Another Arrow in the Quiver, the blueprint for the Kenya Institute for Clinical AI, and it leans on The Law Is Part of the Architecture for the Kenyan legal position and on Not Shown Is Not Locked for MedLattice. Prices are those I could verify in September 2026; hardware prices in particular have moved violently this year and will move again. Nothing here is legal advice.*

It is a little after two in the morning in a county referral hospital. A clinical officer has a nine-month-old with a temperature of 39.8, a bulging fontanelle and a weight she has just measured at 7.1 kilograms. She knows the diagnosis she is worried about. What she wants is the loading dose of ceftriaxone for that weight, whether she should give it before or after the lumbar puncture she has no one to help her perform, and whether the dexamethasone question applies in a setting like hers. She types all of it, in one breath, into the chatbot on her phone.

The answer opens by telling her it cannot provide medical advice and that she should consult a qualified healthcare professional.

She *is* the qualified healthcare professional. She is the only one in the building.

I have heard some version of this story from every clinician I have talked to about the Institute, and it has hardened into a proposal that I think is right: **the Kenya Institute for Clinical AI should run its own language model, on its own hardware, in Kenya, under its own clinical governance** — a system whose every layer the Institute can name, version, test and answer for. This essay is the full argument for that proposal, and it is also an argument with the most common reason given for it, because I think the most common reason is the weakest one.

What I am proposing, in one page

The proposal. A clinical question-answering and decision-support service for clinicians trained and certificated by the Institute, running on open-weight language models served from hardware the Institute owns, located in Kenya, behind an orchestration layer the Institute writes itself.

The architecture. Two model tiers — a mid-sized workhorse model (tens of billions of parameters) for the bulk of questions and a large sparse Mixture-of-Experts model for hard reasoning — fronted by retrieval over Kenyan clinical guidelines and the Kenya Essential Medicines List, deterministic tools for anything numerical (doses, renal adjustment, paediatric weight bands), a safety classifier, and an append-only audit log anchored to MedLattice.

The hardware. A development box now; then two production inference servers, each with four 96 GB Blackwell workstation-class GPUs, in active/standby across two Kenyan sites; a Tier III colocation facility in Nairobi as the primary site.

The cost. Roughly USD 370,000 of landed capital including power infrastructure, and

> roughly USD 485,000 a year to run — of which people are about two-thirds. A five-year total cost of ownership around USD 2.9 million, about one-sixth of the Institute's entire five-year budget; a lean single-site variant comes in near USD 1.8 million. It is more expensive than buying tokens from an API, and I say by how much.
>
> **The honest headline.** This is technically routine and institutionally hard. Serving an open-weight model is a solved engineering problem. Proving that it gives safe answers to Kenyan clinicians is not, and that proof — not the GPUs — is where the money and the time go.

The rest of this essay justifies each of those lines, and then spends a long section on what the system does not do.

The premise, argued with

Do the commercial models actually refuse?

Less than the story suggests, and in a way that matters for how the Institute describes what it is building.

The consumer chat products are tuned for an anonymous public. When they cannot tell who is asking, they hedge, and a hedge aimed at a worried parent lands on a clinical officer just the same. But the underlying policies are narrower than the hedging. OpenAI's usage policy, revised in October 2025 and widely misreported as a ban on medical advice, prohibits the "provision of tailored advice that requires a license, such as legal or medical advice, *without appropriate involvement by a licensed professional*." Its head of health AI said at the time that model behaviour had not changed. Read carefully, that clause describes the Institute's clinicians exactly: they *are* the appropriate involvement. Through the application programming interfaces, with a system prompt that establishes a clinical context, the frontier models discuss doses, differentials and management with licensed clinicians routinely.

So I do not want the Institute to say it is running its own model "to get around" refusals. That framing is inaccurate, and it is dangerous in a way that matters for a regulator reading the Institute's documentation: it sounds like the goal is to remove safety behaviour. **The goal is not to remove guardrails. It is to replace a vendor's generic, liability-shaped guardrail with the Institute's own clinical governance** — which should be *stricter* than the vendor's in the places that kill people (paediatric dosing, pregnancy, narrow-therapeutic-index drugs, anything irreversible) and less defensive in the places where the vendor's caution is merely reputational.

> **One thing not to do.** There is a family of techniques, collectively called *abliteration*, that removes a model's refusal behaviour by projecting a single "refusal direction" out of its weights. Thousands of abliterated checkpoints are on public model hubs. It is tempting, for exactly the reason in the opening story, and it is wrong for clinical use. A July 2026 study of abliterated Gemma 4 and Qwen 3 models found that refusal removal is "not a scalpel": it shifted the models' decision disposition in ways nobody asked for — systematically more optimistic judgements on identical evidence, fewer explicit uncertainty markers — and in one model family made the model *more* confident and in the other *less*. A clinical model that has become quietly more optimistic and less inclined to say "I am not sure" is the worst possible outcome of trying to make it more helpful. The way to get a model that answers clinicians is to build the clinical context into the system around it, not to cut into the weights.

The reasons that actually carry the argument

If refusal is the weak reason, what are the strong ones? There are five, and any one of them would be enough on its own.

First, the law. Section 47 of the Digital Health Act 2023 restricts the transfer of personal health information out of Kenya, and the Data Protection (General) Regulations 2021 require at least one serving copy of health data to sit in a data centre in Kenya. I set this out in full in The Law Is Part of the Architecture; the one-line version is that **an API call to a foreign model carrying patient information is a cross-border transfer of health data.** A clinician who pastes a history into a hosted chatbot has done it already. De-identification helps less than people hope, because the clinically useful parts of a history — age, sex, rare presentation, location, dates — are exactly the quasi-identifiers that re-identify.

Second, version control over the thing being validated. Every hosted model I know of changes underneath its users. Providers update weights, safety tuning, system prompts and routing without a changelog granular enough to re-validate against. If the Institute runs a validation study in March and the model behind the endpoint changes in May, the March study describes a system that no longer exists. With open weights on the Institute's own disks, the model is a file with a SHA-256 digest. The model validated is the model deployed, and it stays that way until the Institute decides otherwise.

Third, the regulator now expects it. The Pharmacy and Poisons Board published its *Guideline on Regulation of Medical Device Software in Kenya* in April 2026. Software that informs diagnosis or treatment is medical device software; the higher-risk classes carry the strictest requirements; and the AI provisions ask for a software bill of materials, information on training data and model architecture, clinical validation metrics, and — for systems that change — anomaly detection, rollback and complete traceability, with annual post-market performance reports. **You cannot promise a regulator rollback on a model you do not possess.**

Fourth, independence. A vendor's policy, pricing, availability and geopolitical exposure are all outside the Institute's control. An institution whose purpose is to certify clinical judgement about AI should not be one terms-of-service revision away from losing the tool it certifies against.

Fifth, resilience. Kenya's international connectivity has been cut before by undersea cable damage, and the national grid went dark on the night of 29 July this year. A service hosted in-country, on its own power, keeps working when the link to Europe does not.

Notice what is not on the list: cost. I will come back to this, because on a straight token-for-token comparison **local inference is more expensive than buying tokens from an API** at the scale the Institute will operate for its first years, and I would rather say so now than have a funder discover it.

What "controlling the source code" can and cannot mean

The brief I set myself for this essay included a sentence I want to take apart, because it is the one most likely to mislead a board: that the Institute should run a system "the

source code of which we are in control of and understand ourselves, and therefore can vouch for its safety and correctness."

Half of that is achievable and important. The other half is a category error, and the error is worth understanding precisely.

A deployed language model system has four layers, and they differ completely in what it means to "understand the source".

What the Institute can author, what it can pin, and what it can only test. The error is to assume that because layers 2–4 are readable, layer 1 is too.

The orchestration layer and the governance layer can be entirely the Institute's own. That is where "we wrote it, we understand it, we can vouch for it" is literally true, and I would insist on it. The agent loop, the retrieval pipeline, the tool definitions, the dose calculators, the output checks and the audit logging should be written in-house, kept small, and reviewed line by line. I would resist every framework that makes this layer large and opaque; the code between a clinician's question and a model's answer is the code a clinical safety officer has to be able to read.

The inference engine is open source and can be pinned, but not realistically "understood" in full. vLLM and SGLang are large, fast-moving Python and C++/CUDA codebases, sitting on vendor GPU drivers and kernel libraries that are closed. What the Institute can do is pin exact versions, build from source, record a software bill of materials, run them on an isolated network with no route to the internet, and rely on the very large number of other people looking at the same code. That is how every hospital already trusts its operating systems and database engines.

The weights cannot be understood by reading them at all. A model is not a program in any sense that code review applies to. DeepSeek V4-Flash is roughly three hun-

dred billion floating-point numbers, most of them stored at four bits. There is no line in it that says “the loading dose of ceftriaxone is”. Its knowledge and its failure modes are distributed across those numbers in ways that interpretability research can probe in small corners and cannot audit end to end. Nor can anyone fully inspect how it was made: no developer of a leading open-weight model publishes its complete training data.

It is worse than opacity. Anthropic’s 2024 *Sleeper Agents* work showed that a model can be trained to behave well under ordinary conditions and differently on a trigger, and that standard safety fine-tuning did not remove the behaviour — in some cases it taught the model to hide it better. I have no reason to think any specific open-weight model contains such a thing. The point is that **nobody can rule it out by inspection, for any model, from any country.**

So the honest formulation is this. **The Institute can vouch for the weights the way a formulary committee vouches for a drug: not by understanding its every mechanism, but by fixing the exact product, knowing its provenance, testing it rigorously against the population it will be used on, and watching it after release.** The pharmacological analogy is exact enough to be useful. A committee does not approve “amoxicillin”; it approves a specific formulation from a specific supplier with a certificate of analysis. The Institute should approve not “DeepSeek” but `DeepSeek-V4-Flash-0731`, at a stated quantisation, with a recorded digest, served by a recorded engine version, behind a recorded prompt and retrieval corpus — and re-approve when any of those changes.

This also answers the correctness half of the sentence. **Correctness is not a property that can be vouched for in advance for any language model.** It is a rate, measured on a defined population of questions, with a confidence interval. The Institute’s claim to its clinicians should be of the form “on our Kenyan evaluation set of n cases, reviewed by a panel of k clinicians, the system produced a potentially harmful answer in x% of cases (95% CI …), and here are the categories where it fails”. That claim is falsifiable, repeatable, and the sort of thing the Pharmacy and Poisons Board asks for. “We control the source code” is not.

What a language model is made of: the technical basis

To compare candidates properly one needs a shared vocabulary for how they are built. Every serious open-weight model in 2026 is a variant of the same machine, and the variations are what determine whether it fits on a given budget.

The decoder-only transformer

All of the candidates are decoder-only transformers: a stack of identical blocks, each containing an attention sub-layer that lets every token look back at earlier tokens, and a feed-forward sub-layer that transforms each token's representation independently. Text is tokenised into integer IDs, embedded into vectors of a few thousand dimensions, passed through somewhere between thirty and a hundred and twenty blocks, and projected back onto the vocabulary to give a probability distribution over the next token. Generation is repeated sampling from that distribution. I built this from scratch in Train Your Own GPT and dissected it in Inside the Transformer; nothing below requires more than that.

Dense versus Mixture-of-Experts

In a **dense** model every parameter is used for every token. Gemma 4 31B and Qwen 3.8 27B are dense: 31 or 27 billion parameters, all active.

In a **Mixture-of-Experts** (MoE) model the feed-forward sub-layer is replaced by many parallel "experts" and a small learned router that sends each token to a handful of them. DeepSeek V4-Pro has about 1.6 trillion parameters in total but activates about 49 billion per token; gpt-oss-120b has about 117 billion total and 5.1 billion active; Kimi K3 fires 16 of 896 experts per token. The consequence that matters for procurement is a split between two resources:

- **Memory capacity** is set by *total* parameters. All experts must be resident, because any token may need any of them.
- **Speed** is set mostly by *active* parameters, because single-stream token generation is limited by how many bytes must be read from memory per token.

MoE is why a model with the knowledge capacity of a trillion parameters can generate at the speed of a model thirty times smaller — provided you can afford the memory to hold it.

Attention variants and the KV cache

Attention needs, for every previous token, a stored key and value vector per layer: the **KV cache**. For standard multi-head attention its size is

```
KV bytes = 2 × layers × kv_heads × head_dim × bytes_per_value × tokens
```

and at long contexts and many concurrent users it can exceed the weights themselves. Most of the architectural innovation of the last two years is about shrinking it:

- **Grouped-query attention** (GQA), used by Gemma and the Llama and Qwen families, shares each key/value head across several query heads.
- **Multi-head latent attention** (MLA), introduced in DeepSeek V2, compresses keys and values into a low-rank latent vector.
- DeepSeek V4 goes further with **Compressed Sparse Attention and Heavily Compressed Attention** (CSA/HCA); DeepSeek reports that at a one-million-token context V4-Pro needs about 27% of V3.2's inference FLOPs and 10% of its KV cache.
- **Linear and delta-rule attention** — Kimi K3's "Kimi Delta Attention" and the Gated DeltaNet layers in recent Qwen models — replace some full-attention layers with recurrent state whose size does not grow with context.

For the Institute this matters less for its headline one-million-token contexts, which clinical questions do not need, than for **concurrency**: a smaller KV cache per conversation means more clinicians served at once from the same GPUs.

Precision and quantisation

Weights are stored as numbers of a chosen width:

Format	Bits per weight	Bytes per billion parameters	Notes
BF16	16	~2.0 GB	The usual training and reference precision
FP8 (E4M3)	8	~1.0 GB	Native on Hopper and Blackwell; near-lossless for most models
INT8 / Q8	8	~1.0 GB	Integer; common in llama.cpp GGUF files
MXFP4 / NVFP4	~4.25	~0.53 GB	4-bit floats with a shared scale per small block (32 values for MXFP4); native on Blackwell
Q4_K_M and similar	~4.5–4.8	~0.6 GB	llama.cpp k-quants; good quality per bit
2–3-bit	2–3.5	0.25–0.45 GB	Measurably degraded; not for clinical use

Some models now ship *natively* quantised — trained or post-trained with the low-precision format in the loop — which is different from a community quantisation made afterwards. gpt-oss ships its MoE weights in MXFP4; DeepSeek V4 ships FP4 experts with FP8 elsewhere; Kimi K3 ships MXFP4. **A natively quantised checkpoint is the artefact the developer evaluated; a third-party 3-bit GGUF of it is not**, and for a clinical service I would only use the former or a quantisation the Institute has itself evaluated.

The two numbers that size the hardware

Two back-of-envelope formulas do most of the work.

Memory to hold the model ≈ total parameters × bytes per parameter, plus 15–30% for KV cache, activations and runtime overhead at modest concurrency.

Upper bound on single-stream decode speed ≈ memory bandwidth ÷ bytes read per token, where bytes read per token ≈ active parameters × bytes per parameter.

A worked example. DeepSeek V4-Flash activates about 13 billion parameters per token. At roughly 4.5 bits that is about 7.3 GB read per token. An RTX PRO 6000 Blackwell has 1.8 TB/s of memory bandwidth, so the theoretical ceiling for one stream on one card is about 245 tokens per second, and real systems reach perhaps half of that for a single user. The saving grace is batching: when many users are served at once, the same weights read from memory serve every sequence in the batch, so aggregate throughput rises far faster than per-user speed falls. That is why a server looks slow in a single-user benchmark and fast in production.

The candidates

These are the models I would put on a shortlist as of the end of September 2026. The field moves monthly; the method for choosing between them will outlive the list.

Model	Developer · release	Total / active parameters	Architecture	Native precision · weights on disk	Context	Licence
DeepSeek V4-Pro	DeepSeek · Apr 2026	~1.6T / ~49B	MoE; CSA + HCA attention; manifold-constrained hyper-connections	FP4 experts + FP8 · ~0.9 TB	1M	MIT
DeepSeek V4-Flash (0731)	DeepSeek · preview Apr, official 31 Jul 2026	~0.3T / ~13B	As Pro; DSpark speculative-decoding module	FP4 + FP8 · ~0.16–0.18 TB	1M (384K recommended for max reasoning)	MIT
Qwen 3.8-27B	Alibaba · 14 Aug 2026	27B dense	Dense; multimodal (text, image, video)	BF16 · ~54 GB (~27 GB FP8)	262K native, 1M with YaRN	Apache 2.0
Qwen 3.8-2.4T-A95B	Alibaba · 14 Aug 2026	~2.4T / ~95B	MoE; multimodal	BF16/FP8 · multi-TB	262K / 1M	Apache 2.0
Kimi K3	Moonshot AI · 27 Jul 2026	~2.8T / ~50B	MoE (16 of 896 experts); Kimi Delta Attention hybrid	MXFP4 · ~1.4 TB	1M	Bespoke; check before use
GLM-5.3	Z.ai · weights 28 Aug 2026	~753B MoE	MoE, long-horizon agentic tuning	BF16/FP8 · ~0.8 TB at 8-bit	1M	Bespoke GLM-5.3 licence (MIT dropped)
gpt-oss-120b	OpenAI · Aug 2025	~117B / ~5.1B	MoE; alternating dense and banded attention	MXFP4 · ~65 GB	128K	Apache 2.0
gpt-oss-20b	OpenAI · Aug 2025	~21B / ~3.6B	MoE	MXFP4 · ~13 GB	128K	Apache 2.0
Gemma 4 31B	Google · 2 Apr 2026	31B dense	Dense; native vision	BF16 · ~62 GB	256K	Apache 2.0
Gemma 4 26B MoE	Google · 2 Apr 2026	26B / 3.8B	MoE; native vision	BF16 · ~52 GB	256K	Apache 2.0
MedGemma 27B (text)	Google · May 2025	27B dense	Gemma 3 base, medical continued training	BF16 · ~54 GB	128K	Health AI Developer Foundations terms
MedGemma 1.5 4B	Google · Jan 2026	4B dense	Gemma 3 base, multimodal medical	BF16 · ~8 GB	128K	Health AI Developer Foundations terms

A note on the figures. Where developers' model cards and third-party write-ups disagree — they do, particularly on DeepSeek V4-Flash's total count after the July checkpoint and on-disk sizes — I have given the model card's figure or a rounded range. The on-disk

sizes are what matter for hardware and they are what the Institute should measure itself on download.

DeepSeek V4-Pro and V4-Flash

The V4 family is the most architecturally interesting release of the year. Both models are sparse MoE transformers trained on over 32 trillion tokens, with DeepSeek's compressed attention stack for long context, **manifold-constrained hyper-connections** (mHC) replacing plain residual connections to stabilise very deep training, and the Muon optimiser. They ship with experts in FP4 and most other weights in FP8, under the MIT licence — the most permissive licence in the table.

V4-Pro, at around 0.9 TB, needs a full eight-GPU HBM server to host; it is out of scope for the Institute's first years. **V4-Flash is the realistic candidate**: small enough in memory to fit on four 96 GB cards with room for a useful KV cache, and with only about 13 billion active parameters it is fast. The official 0731 release substantially improved its agentic behaviour over the April preview, and it ships a native speculative-decoding module (DSpark) supported in both vLLM and SGLang. An independent benchmark on a dual-GH200 workstation reported roughly 276 tokens per second of decode under vLLM and 317 under SGLang, with the full one-million-token context resident in HBM.

I discuss whether DeepSeek is an answer to the Institute's needs in its own section below.

Qwen 3.8

Alibaba released Qwen 3.8 on 14 August with open weights under Apache 2.0, in two very different sizes: a **27B dense multimodal model** and a **2.4-trillion-parameter MoE** with about 95 billion active. The 27B is, on paper, one of the most attractive workhorse candidates in the list: permissively licensed, small enough to serve in FP8 on a single 96 GB card with generous KV cache, multimodal (which matters for photographs of charts, ECG strips and drug labels), with a toggleable reasoning ("thinking") mode. The 2.4T model is out of reach on any budget in this essay.

Kimi K3

Moonshot's Kimi K3, whose weights appeared on 27 July, is the largest open-weight model in the table at about 2.8 trillion parameters, natively in MXFP4. It is also the clearest example of the gap between "open weights" and "deployable": at 1.4 TB resident it needs something like eight 192 GB accelerators on a single node. I list it for completeness and as a benchmark reference. Its licence terms were not clear at release, and a clinical deployment should not proceed on a licence the Institute's counsel has not read.

GLM-5.3

Z.ai's GLM-5.3 is a 753B MoE aimed at long-horizon coding and agentic work. Its predecessor was MIT-licensed; GLM-5.3 is not — it carries a bespoke licence that imposes a

security review on very large hosting companies. That clause would not bite on the Institute, but the change is a useful reminder that **a licence is a property of a specific release, not of a developer**, and has to be re-read with every upgrade.

gpt-oss-120b and gpt-oss-20b

OpenAI's open-weight pair, released in August 2025 under Apache 2.0, remain unusually relevant to clinical use for one reason: OpenAI evaluated them on its own HealthBench and reported gpt-oss-120b performing close to its o3 reasoning model there. That is a vendor-reported figure on a vendor-authored benchmark and should be treated accordingly, but it is more health-specific evidence than most open-weight releases come with. gpt-oss-120b's shape — 117B total, 5.1B active, natively MXFP4, about 65 GB — means it fits on a single 96 GB card with room to spare and is very fast. gpt-oss-20b runs on a laptop and is a serious candidate for an offline, on-device fallback. There is no second generation yet.

Gemma 4

Google's Gemma 4, released on 2 April under Apache 2.0 — a real change from the bespoke Gemma licence of earlier generations — comes in four sizes: effective-2B and effective-4B edge models with audio input, a 26B MoE with 3.8B active, and a 31B dense model that fits unquantised on a single 80 GB H100. All are natively multimodal. The 31B is the other obvious workhorse candidate alongside Qwen 3.8-27B, and the E4B model is worth evaluating for offline use on tablets in facilities with poor connectivity.

MedGemma

MedGemma is the only family on the list trained specifically for medicine: Gemma 3 continued on medical text and imaging. The 27B text model reports about 88–90% on MedQA depending on sampling, and MedGemma 1.5 (January 2026) is a 4B multimodal model with improved records interpretation. Two cautions. First, **MedQA measures performance on United States licensing examination questions**, which tells us little about management of severe malaria, snakebite, or tuberculosis–HIV co-infection with Kenyan drug availability. Second, Google's own model card states that MedGemma outputs "are not intended to directly inform clinical diagnosis, patient management decisions, treatment recommendations, or any other direct clinical practice applications" and are released as a foundation for developers to adapt and validate. That is not a reason to exclude it. It is a precise statement of what the Institute would be taking on: the validation that turns a developer foundation into a clinical tool, which is exactly the work this essay is about.

Models I have left out

Meta's Llama 4 (Scout, 109B/17B; Maverick, 400B/17B) is capable but its community licence carries acceptable-use and attribution conditions that make it less attractive than the Apache and MIT alternatives, and it has been overtaken on most measures. Several

smaller medical fine-tunes circulate on model hubs with impressive exam scores and very little documentation of how they were made; for a regulated deployment, provenance is a requirement, not a nicety.

What I would actually shortlist

> **Workhorse tier** (most questions, high concurrency, one GPU each): **Qwen 3.8-27B**, **Gemma 4 31B**, **gpt-oss-120b**, and **MedGemma 27B** as a medically tuned comparator.
>
> **Heavy-reasoning tier** (hard cases, long documents, multi-step tool use, four GPUs): **DeepSeek V4-Flash-0731**.
>
> **Offline fallback** (a laptop or tablet in a facility with no link): **gpt-oss-20b** or **Gemma 4 E4B**.
>
> The final choice should come out of the Institute’s own evaluation, not out of this table. I would be unsurprised if the winner on Kenyan cases is not the winner on public leaderboards.

Is it technically feasible?

Yes, for the parts that matter, and no for one part that does not.

Serving an open-weight model is feasible and routine. Thousands of organisations do it. The inference engines are mature, the models ship with serving recipes, and the hardware is purchasable. Nothing in this essay's architecture is research.

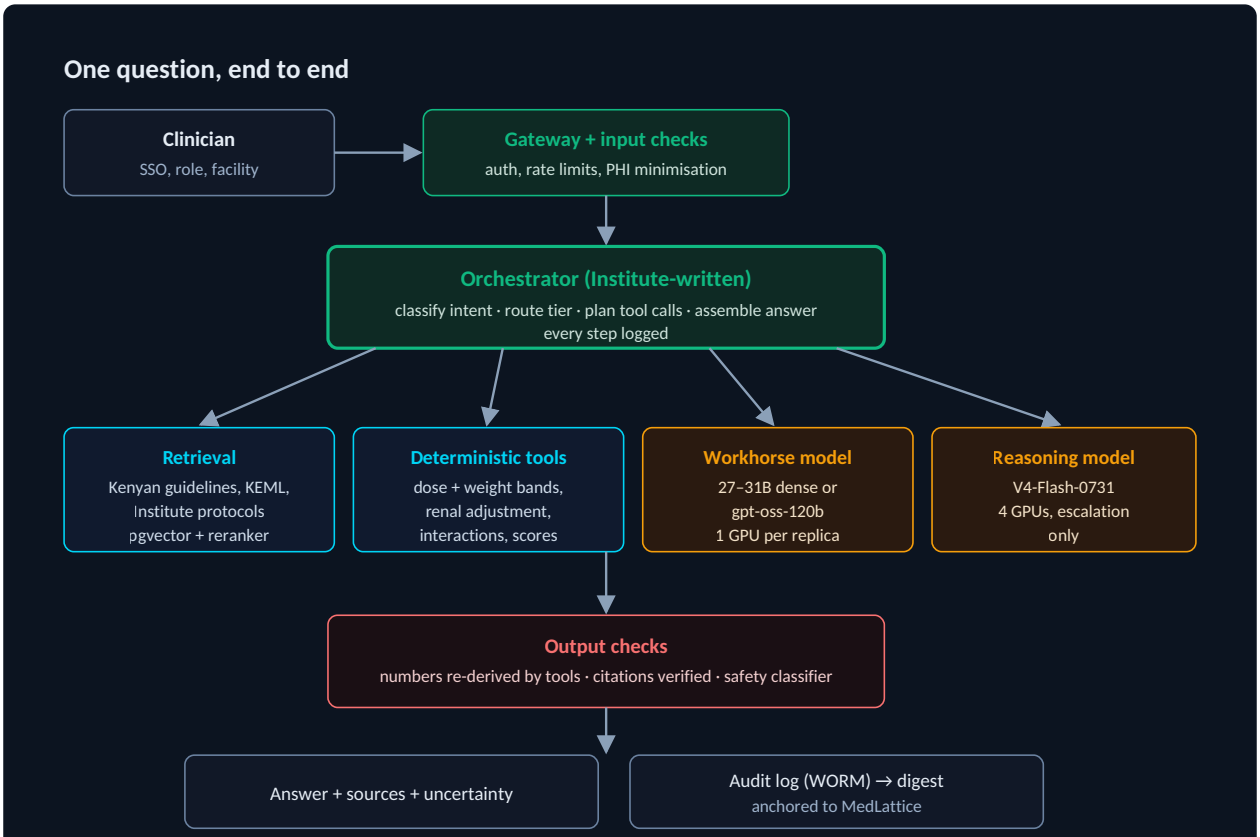
Adapting a model is feasible within limits. Parameter-efficient fine-tuning — LoRA and its variants, which train small low-rank adapter matrices while the base weights stay frozen — is practical on a 27–31B dense model with the hardware described below. It is how one would teach a workhorse model Kenyan formulary names, local referral pathways and the Institute's house style for answers. Full fine-tuning of a 300B MoE is not practical on this budget, and is not needed: most of what the Institute wants the model to know should come from retrieval, where it is citable and updatable, not from weights, where it is neither.

Training a foundation model from scratch is not feasible, and would be a mistake if it were. A frontier-class pre-training run consumes on the order of 10^{24}–10^{25} floating-point operations, tens of trillions of tokens and tens to hundreds of millions of dollars. The Institute's comparative advantage is not in making models; it is in knowing, better than anyone, whether they are safe for Kenyan clinicians to use. Every shilling spent imitating DeepSeek is a shilling not spent on the evaluation that no one else will do.

The hard part is not technical. It is the construction of a Kenyan clinical evaluation set, the clinician time to grade outputs against it, the clinical safety case, the regulatory submission, and the operational discipline to keep a validated configuration frozen while the world outside releases a new model every month.

The system: what actually runs

A language model on its own is not a clinical tool. What the clinician talks to is a system, and most of its safety properties come from the parts that are not the model.



The model is two boxes out of ten. Numbers in answers are never trusted from the model: they are recomputed by deterministic tools and the answer is rejected if they disagree.

A few design rules, each of which I would treat as non-negotiable.

Numbers come from code, not from the model. Language models are poor at arithmetic and worse at noticing that they are poor at it. Any dose, rate, volume, score or unit conversion in an answer should be produced — or independently recomputed — by a deterministic function the Institute has written and tested, from a formulary table the Institute's pharmacists maintain. If the model's number and the tool's number disagree, the answer is withheld and the disagreement logged.

Answers carry their sources. Retrieval is over a curated, versioned corpus — Kenyan national guidelines, the Kenya Essential Medicines List, the Institute's own protocols — and every clinical claim should be traceable to a passage. The orchestrator checks that cited passages exist and say what the answer says they say. I described how to build and honestly evaluate a retriever of this kind in The Capstone: Build a Grounded Assistant.

The agent reads; it does not write. In the first versions, the system's tools should be read-only: look up a guideline, compute a dose, query a drug-interaction table, read a record the clinician is already entitled to see. It should not place orders, write to the record, or message patients. Anything with side effects goes through a human signature.

The model is a principal, not a pipe. MedLattice's §7 already treats an AI as a principal with its own attestation — weights digest, version, jurisdiction, whether it retains inputs — and its own purpose codes for inference, training and evaluation. A locally hosted model is the first case where every field of that attestation can be *measured* rather

than asserted: the Institute computed the digest, the jurisdiction is a rack in Nairobi, and retention is whatever the Institute configured. The audit log records which digest answered which question.

The orchestration code should be small enough to read in an afternoon. The heart of it is not complicated:

```
import hashlib, json, time
from pathlib import Path

APPROVED = json.loads(Path("approved_models.json").read_text())
# {"workhorse": {"name": "Qwen3.8-27B-FP8", "sha256": {...per-shard digests...}, "engine":
"vllm==0.26.x"}, ...}

def verify_weights(model_dir: Path, expected: dict[str, str]) -> None:
    """Refuse to serve anything whose bytes are not exactly the approved artefact."""
    for shard, digest in expected.items():
        h = hashlib.sha256()
        with open(model_dir / shard, "rb") as f:
            for block in iter(lambda: f.read(1 << 24), b""):
                h.update(block)
        if h.hexdigest() != digest:
            raise SystemExit(f"UNAPPROVED WEIGHTS: {shard}")

TOOLS = {  # read-only, deterministic, unit-tested
    "dose_by_weight": dose_by_weight,
    "renal_adjust": renal_adjust,
    "guideline_search": guideline_search,
    "interaction_check": interaction_check,
}

def answer(question: str, clinician: Clinician) -> Answer:
    plan = workhorse.plan(question, tools=list(TOOLS))          # model proposes tool calls
    results = {c.name: TOOLS[c.name](**c.args) for c in plan.calls if c.name in TOOLS}
    draft = route(plan).generate(question, context=results)      # tier chosen by rules, not by
the model
    checked = verify_numbers(draft, results) and verify_citations(draft, results)
    audit.append(clinician.id, question, draft, results, model_digest=route(plan).digest,
ok=checked)
    return draft if checked else withheld(draft, reason=checked.reason)
```

That is a sketch, not production code — but the production code should not be many times longer, and every line of it should be the Institute's.

What skills the Institute needs

The team is small, but it cannot be generic. The critical roles, with indicative Nairobi monthly base salaries for 2026 drawn from published local salary surveys, are these.

Role	What they own	FTE	Indicative monthly base (KES)
Lead ML / LLM engineer	Model selection, serving configuration, quantisation, fine-tuning, the orchestrator	1	600,000 – 900,000
MLOps / site reliability engineer	GPU servers, drivers, containers, deployment pipeline, monitoring, backups, failover drills	1	450,000 – 600,000

Role	What they own	FTE	Indicative monthly base (KES)
Security engineer	Network isolation, identity, secrets, supply chain (SBOM, signing), penetration testing, incident response	0.5–1	400,000 – 600,000
Clinical informatician	Retrieval corpus, guideline curation, formulary tables, FHIR integration with MedLattice	1	350,000 – 500,000
Clinical safety officer	The clinical safety case and hazard log; signs off every change	0.5	500,000 – 700,000
Clinical pharmacist	Dose tables, interaction data, review of every numerical tool	0.3	300,000 – 400,000
Evaluation lead	Evaluation set construction, grading protocol, panel management, statistics	0.5–1	400,000 – 600,000
Data protection officer	DPIA, ODPC registration, cross-border and retention questions	shared	—
Facilities / electrical	UPS, generator, solar and battery maintenance (on-campus sites)	contract	—

Plus a standing **clinician evaluation panel** — a dozen or more practising Kenyan clinicians across the Institute's tracks, paid per session to grade outputs — which is the single most important and most underestimated line in the budget. The standard-setting methods I described in The Angoff Panel for Testing Clinicians transfer almost unchanged: the panel's job is to decide what an acceptable answer looks like before the model is scored against it.

Two observations about this list. First, the scarce skill in Nairobi is not "AI" in general, of which there is now a good deal, but **MLOps at the GPU level** — people who have kept CUDA drivers, NCCL and a multi-GPU inference engine alive in production under load. Salary surveys put MLOps engineers at the top of the Kenyan tech pay scale for exactly that reason. Second, several of these roles are the Institute's own graduates in waiting: the health informatics track in the blueprint exists to produce clinical informaticians and clinical safety officers. **An Institute that runs its own model is also building the training ground for its own informatics students** — a benefit no API subscription provides.

The hardware

A necessary word on prices

2026 has been a bad year to buy memory. The AI build-out has absorbed so much DRAM, HBM and GDDR7 capacity that server memory prices roughly doubled in the first half of the year, and it has dragged GPU prices with it. The RTX PRO 6000 Blackwell — the 96 GB workstation card at the centre of the design below — launched in March 2025 at about USD 8,500 and now lists at **USD 16,000**. An eight-GPU B200 server quotes at USD 400,000–500,000. Every figure in this section should be re-quoted at the point of purchase; I have used September 2026 list prices and added the Kenyan landed uplift.

Landed cost in Kenya

Computers and their parts generally enter under HS chapter 84 at a low or zero EAC Common External Tariff rate — confirm the specific headings for GPUs and servers with KRA before relying on it — but the other charges are not optional: **Import Declaration Fee 2.5%, Railway Development Levy 2%, and VAT at 16%** on the CIF value plus duty and levies. Before freight and clearing, that is roughly a **21% uplift** on list price. The Institute should ask counsel whether any exemption for health or educational institutions applies; I have not assumed one.

Export control is the other procurement question. The United States rescinded its AI Diffusion Rule in May 2025, and as of this writing the replacement framework is still in draft; reported versions would subject small volumes — on the order of hundreds of accelerators — to light review. A purchase of eight workstation GPUs is well under any reported threshold. The Institute should nevertheless buy through an authorised distributor, keep end-use documentation, and expect the rules to change during the programme.

Three tiers

Tier	Purpose	Configuration	GPU memory	What it runs	Indicative landed cost (USD)
0 — Evaluation	Build the evaluation harness; compare models; no patients	Mac Studio M5 Ultra 512 GB *or* a workstation with 2× RTX PRO 6000	512 GB unified / 192 GB	Every workhorse candidate; V4-Flash at reduced context	14,000 – 55,000
1 — Production	Live service for the first few thousand clinicians	2 servers, each 4× RTX PRO 6000 Server Edition, dual CPUs, 1 TB RAM, NVMe; active/standby across two sites	384 GB per server	Workhorse (2–3 replicas) + V4-Flash on each server	~270,000 landed; ~370,000 with campus power infrastructure
2 — Scale	National scale, or hosting a 1T+ model	8× H200 (141 GB) or 8× B200 (192 GB) HGX server, ×2 for redundancy	1.1 – 1.5 TB per server	V4-Pro, Kimi K3 class models	900,000 – 1,200,000 plus facility upgrades

Tier 0 is where the Institute should start, now. An M5 Ultra Mac Studio with 512 GB of unified memory (the 512 GB configuration was due in late October at the time of writ-

ing, at a price Apple had not yet announced but which will exceed USD 10,000) can load every model in the workhorse shortlist and V4-Flash at the same time, with memory bandwidth of about 1.2 TB/s, while drawing a few hundred watts. It is a superb evaluation machine and a poor production server: its throughput under many concurrent users is far below a GPU server's, and its software stack (MLX, llama.cpp) is not the one production will run. The two-GPU workstation costs more but runs the same vLLM/SGLang stack as production, which makes evaluation results more transferable. My preference is the workstation; the Mac is the cheaper way to start this quarter.

Tier 1 is the core of this proposal. Four RTX PRO 6000 cards give 384 GB of GDDR7 per server — enough for V4-Flash at FP4/FP8 with a working KV cache *and* a workhorse model replica, or for three or four workhorse replicas when V4-Flash is not loaded. The cards have no NVLink, so multi-GPU serving runs over PCIe Gen 5; for an MoE model this is less of a penalty than it sounds, because expert parallelism moves only routed activations between cards, not full tensor shards. Two servers, one per site, give the Institute the ability to take either one down for maintenance or lose either one to a fault without losing the service.

Tier 2 is where a trillion-parameter model becomes possible. I would not buy it in the first two years. By the time the Institute has the evaluation evidence to justify it, prices and options will have changed completely; and it is entirely possible that the evaluation will show that the heavy-reasoning tier adds little over the workhorse on Kenyan cases.

Sizing Tier 1 against demand

Rough arithmetic, with the assumptions stated so they can be attacked.

Suppose 2,000 active clinicians in the first live phase, each asking 20 questions on a working day, each question carrying about 2,500 tokens of prompt and retrieved context and producing about 700 tokens of answer (more if a reasoning mode is on). That is **40,000 questions a day**, 100 million prompt tokens and 28 million generated tokens. If 15% of the day's traffic falls in the busiest hour, peak load is about 1.7 questions a second and roughly 1,200 generated tokens a second, before reasoning tokens.

A single RTX PRO 6000 serving a 27–31B dense model in FP8, or gpt-oss-120b in MXFP4, with continuous batching, sustains aggregate generation in the high hundreds to low thousands of tokens per second at interactive per-user latency. Prefill of the 2,500-token prompts is compute-bound and fast on Blackwell. Three workhorse replicas per server — with V4-Flash taking the fourth card's worth of capacity only for escalated questions — covers that peak with margin on one server, which is the point: **the standby server exists for resilience, not capacity**, and it can be promoted to active/active if demand doubles.

These are estimates. The Institute's first real job on Tier 0 is to replace them with measurements on its own prompts.

Cooling at altitude

One detail that off-the-shelf sizing guides get wrong for Nairobi. At about 1,700 metres, air density is roughly 83% of its sea-level value, so each cubic metre of air carries away correspondingly less heat. ASHRAE's thermal guidelines derate the allowable inlet temperature for IT equipment above 900 metres by 1 °C for every 300 metres. A server rated for 35 °C inlet at sea level should be treated as rated for roughly 32 °C in Nairobi. With four 600 W GPUs per chassis this matters. A colocation facility will have designed for it; an on-campus server room must be designed for it explicitly.

Power: designing for the night of 29 July

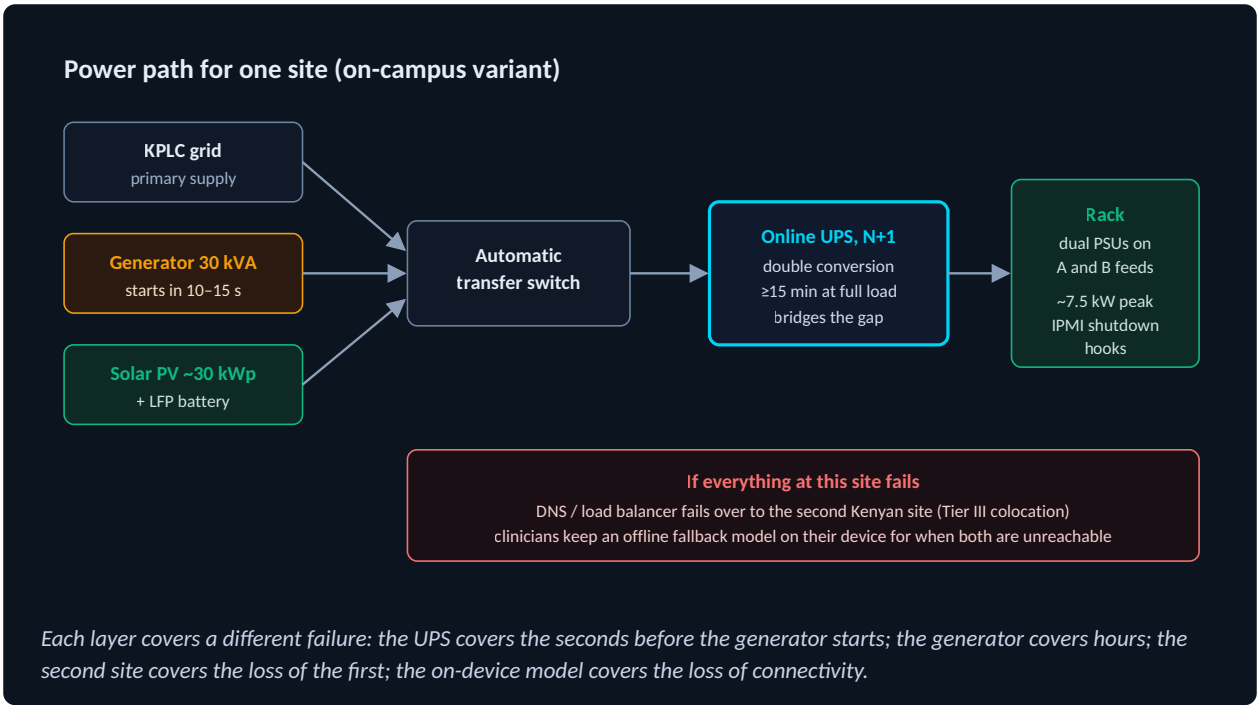
On the evening of Wednesday 29 July 2026, a “technical disturbance” on the national grid took out supply across Nairobi, the Coast and Mount Kenya for several hours. Kenya’s generation mix is one of the cleanest in the world — geothermal, hydro and wind carry most of it — but its transmission network is under strain from record demand, and every Kenyan facility manager already knows that the grid is a supply to be backed up, not relied upon. A clinical decision-support service that clinicians have been told to use at two in the morning must not go dark when the grid does.

The load

Each Tier 1 server draws about 3.4 kW at full load: four GPUs at up to 600 W each, plus roughly 1 kW for CPUs, memory, storage and fans. Two servers, a network switch, storage and a backup appliance bring the IT load to about **7.5 kW peak**, and perhaps 5 kW averaged over a day at realistic utilisation. In a small on-campus room with a power usage effectiveness around 1.6, the facility draws about 8 kW on average and 12 kW at peak.

At Kenyan business tariffs, which have run at around KES 23 per kWh all-in and have been pushed upward this year by monthly fuel-cost and foreign-exchange pass-through charges of KES 4–5 per kWh, a year of 8 kW average is about **70,000 kWh and KES 1.6 million** — roughly USD 12,500. Electricity is not the dominant cost. Uninterrupted electricity is.

The design



Each layer covers a different failure: the UPS covers the seconds before the generator starts; the generator covers hours; the second site covers the loss of the first; the on-device model covers the loss of connectivity.

The design has four layers, and each covers a different failure.

An online double-conversion UPS in an N+1 configuration, sized to carry the full IT load for at least fifteen minutes. Online double-conversion (rather than line-interactive) matters on a grid with frequent sags and surges: the load always runs from the inverter, so grid disturbances never reach the servers. Fifteen minutes is not meant to ride out an outage; it is meant to bridge generator start-up, and failing that, to give the servers time for an orderly shutdown triggered over IPMI or the UPS's network card. Lithium-iron-phosphate UPS batteries last roughly twice as long as valve-regulated lead-acid in a warm room and are worth the premium. Indicative cost for two 10–15 kVA modules with batteries: USD 12,000–20,000.

A standby diesel generator of about 30 kVA with an automatic transfer switch, starting within ten to fifteen seconds, with fuel for at least 24 hours on site and a written refuelling contract. Test it under load monthly; a generator that has only ever been started unloaded is a hypothesis. Indicative cost installed: USD 20,000–28,000.

Solar PV and battery storage, which do a different job: they cut the electricity bill and reduce generator hours rather than providing the resilience layer. Commercial solar in Kenya installs at roughly KES 80,000–135,000 per kWp and typically pays back in three to five years. About 30 kWp with 40–60 kWh of lithium-iron-phosphate storage would carry a large share of the daytime load. Indicative cost: USD 35,000–50,000. On a colocation site this layer does not apply; the facility's power is the facility's problem.

A second site, which is the only protection against the failures that take out a whole building: fire, flood, theft, a transformer failure, a landlord dispute. My recommendation is that the **primary site should be a Tier III colocation facility in Nairobi** — iXAfrica, Africa Data Centres and others now operate carrier-neutral facilities with redundant grid feeds, generators and cooling designed for AI densities (iXAfrica advertises up to 50 kW per rack) — with the **secondary on an Institute-controlled campus** built to the on-campus design above. Both are in Kenya, which keeps the whole arrangement inside section 47.

Beyond all of this, the clinician's own device should carry an **offline fallback**: a small model such as gpt-oss-20b or Gemma 4 E4B, with a local copy of the guideline corpus and the dose tools, clearly labelled as a degraded mode. When the network is down, a clinician who has been trained to rely on the service should get a lesser version of it rather than nothing — and should be told plainly which one she is using.

Data: what must be backed up, and where

The system holds five kinds of data, with different value, different sensitivity and different backup requirements.

Data	Why it matters	Sensitivity	Backup approach
Model weights (approved versions)	The validated artefact; public copies can be withdrawn or altered upstream	Low (public) but integrity-critical	Immutable copies with recorded digests on two sites plus offline media. Never assume Hugging Face will still host a given revision

Data	Why it matters	Sensitivity	Backup approach
Configuration — prompts, routing rules, tool code, formulary tables, engine versions	Together with the weights, *this is the device* the regulator certifies	Medium	Git, signed commits and tags per release; mirrored to both sites
Retrieval corpus and index	Source of every citation	Low–medium	Versioned snapshots; the index is rebuildable, the corpus version is not
Evaluation sets and results	The Institute's evidence base, and its most valuable intellectual property	Medium (may contain case material)	Treated like research data: versioned, encrypted, restricted access
Audit logs — question, context, answer, model digest, clinician	Clinical, medico-legal and regulatory record	**High — contains health data**	Write-once storage, encrypted, retained on the clinical-record schedule; digests anchored to MedLattice

The strategy is the familiar **3-2-1-1-0** rule: three copies, on two kinds of media, one at another site, one offline or immutable, and zero errors on verified restores. Concretely:

- **ZFS on the servers**, with frequent snapshots, gives point-in-time recovery from operator error in seconds and checksums every block against silent corruption.
- **An object store at the second site** (MinIO or equivalent, with object lock) receives encrypted, deduplicated backups nightly via restic or Borg. This is the copy restores normally come from.
- **LTO tape**, written monthly and stored off-site in a fireproof safe, is the air-gapped copy that ransomware cannot reach. An LTO-9 drive and a year of cartridges is a few thousand dollars and it is the cheapest insurance in the whole design.
- **Every backup stays in Kenya.** An offshore backup of the audit log is a cross-border transfer of health data just as surely as an API call is. This rules out the convenient default of backing up to a foreign cloud region. Oracle's Nairobi cloud region, opened with iXAfrica in January 2026, is the first hyperscale in-country option; whether it satisfies the Institute's counsel is a question for counsel.
- **Encryption at rest with keys the Institute holds**, on hardware security modules or at minimum TPM-sealed keys, with a documented key-escrow procedure so that the loss of one person does not mean the loss of the archive. For the audit log, which must stay confidential for decades, the same post-quantum reasoning I set out for MedLattice applies: encrypt for the length of a human life, not the length of a procurement cycle.

Two targets should be written down and tested. A **recovery point objective** — how much audit log the Institute can afford to lose — of minutes, which means continuous replication of the log to the second site, not nightly backup. And a **recovery time objective** — how long clinicians can be without the service — of under an hour for a site failure (fail over to the other site) and under a day for a total rebuild from backup. **A restore that has not been rehearsed is not a backup.** I would schedule a full restore drill, from tape, onto clean hardware, every quarter, and publish the time it took.

The software stack

Every component here is open source except the GPU driver and its libraries. Every version is pinned, and every image is built by the Institute from source or from a verified upstream artefact.

Layer	Choice	Notes
Operating system	Ubuntu Server LTS	Widest support for NVIDIA's stack; full-disk encryption; unattended security updates on a controlled schedule
GPU stack	NVIDIA driver, CUDA, NCCL	The one closed layer; pinned per release, upgraded only through change control
Containers	Podman or Docker; images built in-house	Reproducible builds; image digests recorded in the release manifest
Orchestration of services	systemd and Compose at first; Kubernetes (k3s) only if scale requires it	Resist Kubernetes until there is a problem it solves
Inference engine — production	**SGLang** and/or **vLLM**	OpenAI-compatible APIs, continuous batching, paged KV cache, FP8/FP4 kernels, speculative decoding, expert parallelism. Both support DeepSeek V4 and the Qwen, Gemma and gpt-oss families
Inference engine — evaluation and edge	**llama.cpp** (and MLX on Apple Silicon)	GGUF quantisations; runs on the Mac and on clinicians' laptops
Model format and provenance	**safetensors only**; per-shard SHA-256 digests; signature verification with OpenSSF model signing (Sigstore) where available	Never load pickle-based checkpoints; never enable `trust_remote_code`
Orchestrator	Python, FastAPI, Pydantic; the Institute's own code	Small, typed, fully tested, readable by the safety officer
Retrieval	PostgreSQL with pgvector; an open embedding model; a cross-encoder reranker	One database the team already knows, rather than a new vector store
Safety classifier	A small guard model (for example gpt-oss-safeguard, which applies a written policy) plus the Institute's own rule checks	The policy is written by the clinical safety officer in plain language
Clinical integration	HL7 FHIR R4 (HAPI FHIR) to MedLattice and facility systems	Read-only at first
Identity and access	Keycloak (OIDC), role-based access tied to the clinician's certification level	A Level 1 certificate should not unlock what a Level 4 does
Secrets	HashiCorp Vault or OpenBao; HSM-backed where possible	No secrets in images or repositories
Observability	Prometheus, Grafana, OpenTelemetry; Langfuse (self-hosted) for model traces	Latency, errors, GPU health, and clinically meaningful signals — withheld answers, tool disagreements, escalations
Evaluation	The Institute's own harness; OpenAI's open-source HealthBench grader; UK AISI's Inspect framework	Evaluation runs on every candidate change before it reaches a clinician
Supply chain	Syft (SBOM), Grype (vulnerability scanning), signed releases	The SBOM the Pharmacy and Poisons Board asks for is produced by the build, not written by hand
Backup	ZFS, restic or Borg, MinIO with object lock, LTO	As above

Two principles govern the whole list. **The inference servers have no route to the internet.** Weights and packages are fetched on a separate, audited staging machine, verified, and carried across; nothing on a production host can phone anywhere. This matters for the DeepSeek question below, and it matters equally for every other model and every Python package. And **the system is model-agnostic by construction**: the orchestrator talks to an OpenAI-compatible endpoint and knows the model only by the name and digest in the approved-models manifest. Swapping Qwen for Gemma is a configuration change followed by a full re-evaluation — not a rewrite.

How much it costs

Capital

For the recommended configuration — a Tier 0 evaluation machine now, two Tier 1 servers later, primary in colocation and secondary on campus:

Item	Indicative cost (USD, list)
8× RTX PRO 6000 Blackwell Server Edition, 96 GB	128,000
2× GPU server chassis (dual CPUs, 1 TB DDR5, NVMe)	50,000
Networking: switches, firewalls, out-of-band management	12,000
Backup: object-store server at second site, LTO-9 drive and cartridges	18,000
Tier 0: Mac Studio M5 Ultra 512 GB	14,000
Hardware subtotal	**222,000**
Kenyan landed uplift (IDF 2.5%, RDL 2%, VAT 16%), before freight	~47,000
Landed hardware	**~269,000**
On-campus secondary site: UPS N+1, 30 kVA generator and ATS, 30 kWp solar with LFP storage, room works and cooling	~100,000
Total capital	**~370,000**

Running costs, per year

Item	Indicative annual cost (USD)
Core team (about 5.3 FTE across the roles above, mid-range salaries, with ~15% employer on-costs)	~305,000
Clinician evaluation panel (a dozen clinicians, two graded sessions a month)	~33,000
Colocation, one rack at ~8 kW, Nairobi (indicative; to be quoted)	~30,000
Hardware support, warranty and spares (~8% of hardware)	~18,000
Regulatory, legal and external penetration testing	~30,000
Connectivity, two carriers at each site	~12,000
Campus electricity net of solar, generator fuel and servicing	~14,000
Contingency (10%)	~44,000
Total	**~485,000**

Five-year total cost of ownership

Capital of about USD 370,000, a hardware refresh reserve of about USD 100,000 in year four, and five years of running costs at about USD 485,000 gives a five-year total of

roughly USD 2.9 million. Against the blueprint's five-year budget of USD 17.38 million that is about one-sixth — a large line, and one I would expect a funder to question.

Two things soften it. First, several of these posts overlap with roles the blueprint already specifies among its seventy-one core posts — the informatics and quality functions especially — and I have not netted them off. Second, there is a **lean variant**: one production server in colocation, the Mac for evaluation, no on-campus power build, and a core team of about three and a half people with the evaluation panel intact. That is roughly USD 160,000 of capital and USD 330,000 a year: about **USD 1.8 million over five years**, at the price of having no second site and a thinner bench of people.

Compared with buying tokens

Here is the comparison I promised. At the demand assumed above — 40,000 questions a day, 100 million prompt tokens and 28 million generated tokens — and roughly 260 working days a year:

- At the per-token prices of a frontier closed model (on the order of a few dollars per million input tokens and ten or more per million output tokens), the bill is in the region of **USD 100,000–175,000 a year**.
- At the prices of hosted open-weight models such as DeepSeek's own API, it is anywhere from **USD 5,000 to 70,000 a year**, depending on whether the small or the large model is used.

Either is far less than the local system's running cost. But the comparison is not like-for-like, because **most of the local system's cost is people who would still be needed with an API**: someone still has to write the orchestrator, curate the corpus, build the dose tools, run the evaluation and carry the clinical safety case. The true premium for running locally is the hardware, the facility, and the MLOps and security capacity to run them — on the order of **USD 150,000–200,000 a year plus the capital**, or roughly USD 1.2–1.4 million over five years.

That is the price of sovereignty, version control and regulatory standing. I think it is worth paying. It should be presented to funders as exactly that, not dressed up as a saving.

Is DeepSeek V4 an answer?

"DeepSeek 4" — properly, the DeepSeek V4 family — is the model people ask me about first, because it is the most capable model with the most permissive licence. The honest answer has five parts.

It is a legitimate technical candidate — specifically V4-Flash. MIT licensing places no restriction on clinical use, fine-tuning or redistribution of derivatives. V4-Flash-0731 fits Tier 1 hardware, runs fast because of its small active-parameter count, handles long documents efficiently, and is well supported by both production inference engines. V4-Pro, at around 0.9 TB, is out of reach until Tier 2.

Running it locally sends nothing to DeepSeek. This is the most common misunderstanding. The weights are inert numerical arrays in safetensors files. They are executed by vLLM or SGLang, which are not DeepSeek's code. On an inference server with no route to the internet, loaded without `trust_remote_code`, there is no mechanism by which a DeepSeek model can transmit anything anywhere. The concerns about the DeepSeek *app* and *API* — where data goes to servers in China — simply do not apply to local weights.

Its security behaviour needs to be designed around, not trusted. The United States' Center for AI Standards and Innovation (CAISI) evaluated earlier DeepSeek models in September 2025 and found that R1-0528 responded to 94% of overtly malicious requests under common jailbreak techniques, against 8% for the US reference models, and that agents built on DeepSeek models were about twelve times more likely to follow hijacking instructions planted in their inputs. Those findings concern older models, and CAISI's May 2026 evaluation of V4-Pro assessed capability rather than safety. But the relevant lesson is architectural and applies to every model: **a system that reads referral letters, pasted notes and retrieved documents is exposed to prompt injection, and the defence must live in the orchestrator** — read-only tools, no side effects without a human signature, outputs checked by deterministic code — not in the model's disposition.

Its self-reported benchmarks are optimistic, as everyone's are. CAISI's evaluation placed V4-Pro about eight months behind the US frontier, performing comparably to GPT-5, and found the gap wider on held-out and non-public benchmarks than on the public ones DeepSeek reported. The same is true in varying degrees of every developer's self-reported numbers, which is why the Institute's own evaluation set is the only benchmark that decides anything. CAISI also found earlier DeepSeek models repeating state-aligned political narratives far more often than US models. That is largely irrelevant to a ceftriaxone dose. It is a reminder that alignment reflects the priorities of whoever did it, and that a model's behaviour on public-health topics with political content should be tested rather than assumed.

Procurement and funding optics are real, and they are a governance question, not a technical one. Some funders and partners — particularly those with US government exposure — restrict or scrutinise the use of models of PRC origin. The Institute's

funding model depends on exactly those relationships. Before any DeepSeek model reaches production, grant conditions should be checked, and the decision should be made by the board with that information in hand.

> **My position.** DeepSeek V4-Flash belongs on the shortlist for the heavy-reasoning tier, evaluated on precisely the same terms as Qwen, Gemma, gpt-oss and MedGemma: on Kenyan cases, air-gapped, never trusted for a number, never given a tool with side effects. It is not "the answer", because there is no single model that is. **The answer is the harness** — the evaluation set, the orchestrator and the governance — that lets the Institute swap one model for another in a week and prove within a month whether the new one is better.

The regulatory frame

Three regimes apply, and the Institute should plan for all of them from the first month. I set out the general Kenyan position in The Law Is Part of the Architecture; what follows is what is specific to a clinical language model.

The Pharmacy and Poisons Board — medical device software. A system that answers clinicians' questions about diagnosis and treatment is medical device software under the Board's April 2026 guideline. The guideline follows the International Medical Device Regulators Forum's risk framework, in which classification depends on the significance of the information to the clinical decision and the seriousness of the condition. A general clinical question-answering service used in critical situations — the infant with meningitis — will probably fall in **Category III**, and conceivably IV for some uses; the Institute should seek the Board's view in a pre-submission meeting rather than self-classify optimistically. Reported fees are modest for a local manufacturer (on the order of USD 1,000 for Category III, plus an administrative fee, valid for five years). The requirements are not modest:

- a **software bill of materials** — which the build pipeline above produces;
- **training dataset information, model architecture and bias management** — which raises a genuine problem, discussed below;
- **clinical validation metrics** — which the evaluation programme produces;
- for systems that change, **anomaly detection, rollback and complete traceability** — which pinned weights, signed releases and the audit log provide;
- **annual post-market performance reports** and clear labelling of intended use.

The genuine problem is training data. No developer of a leading open-weight model publishes its complete training corpus, and the Institute cannot supply what it does not have. My proposed approach is to disclose everything that *is* knowable — the developer's own documentation, the exact artefact and its digest, every document in the retrieval corpus, every example used in any fine-tuning — and to rest the safety claim on

local clinical validation of the complete system. Whether the Board will accept that is an open question, and one I would put to it early and in writing.

The Digital Health Agency — certification. Facilities must use digital health solutions certified by the Agency. A decision-support service integrated with facility systems and MedLattice will need certification against the published framework, including the Kenya Health Data Governance Framework's security and privacy standards and the capacity for health information exchange.

The Office of the Data Protection Commissioner — processing health data. The audit log is health data. The Institute needs to register as a data controller and processor, carry out a data protection impact assessment specifically for the model service, set a retention period consistent with clinical-record law, and design for the 72-hour breach clock.

And behind all three sits professional regulation. **The clinician remains responsible for the decision.** A locally hosted model does not change that, and the Institute's own curriculum exists partly to make sure its clinicians understand it. There is a risk here that I want to name directly: a tool the Institute built, runs and certifies will feel *more* trustworthy to the clinicians it trained than a foreign chatbot does. The 2025 randomised trial that started the blueprint found physicians with twenty hours of AI-literacy training still deferring to deliberately erroneous model output. **Running the Institute's own model makes automation bias more dangerous, not less**, and the interface — sources shown, uncertainty stated, numbers visibly recomputed — must be designed to resist it.

Evaluation: where the money actually goes

Everything above is infrastructure for one activity: finding out, with evidence a regulator and a sceptical clinician would both accept, whether the system's answers are safe for Kenyan clinicians. This is the part of the project that no vendor will do for the Institute, and the part I would protect first if the budget were cut.

The Kenyan clinical question set

The evaluation set should be built from four sources.

1. **Real questions from Kenyan clinicians**, collected with consent and de-identified, from the Institute's own trainees. These are the only questions that reflect what clinicians actually ask, in the way they actually ask it — often in a mixture of English and Kiswahili, often with half the relevant facts missing.
2. **Guideline-derived cases** written by the panel to cover the national guidelines systematically: malaria, tuberculosis, HIV, maternal and neonatal care, paediatric emergencies, snakebite, sickle cell disease, non-communicable diseases as they present in Kenyan facilities.
3. **Known hard cases**, where a plausible answer is wrong: paediatric weight-band dosing; drugs in pregnancy and breastfeeding; renal adjustment; and interactions that matter here and are rare elsewhere — rifampicin with dolutegravir, which requires the dolutegravir dose to be doubled to twice daily, is the canonical example. A model trained mostly on high-income-country text will not reliably know which drugs are actually on the Kenya Essential Medicines List or in stock at a level 4 facility.
4. **Adversarial and calibration cases**: prompt injections embedded in pasted notes and referral letters; questions the system should decline or escalate; questions with no safe answer from the information given, where the correct response is to ask for more.

Each case gets a **rubric written by the panel before any model is scored against it**: the elements a good answer must contain, the errors that would make it harmful, and the harm severity of each — none, minor, major, catastrophic. This is the HealthBench method — OpenAI's benchmark of 5,000 conversations graded against rubrics written by 262 physicians — applied to Kenyan medicine, and it is the standard-setting logic of the Angoff panel applied to a machine. Grading is double, blinded to which model produced the answer, with disagreements adjudicated and inter-rater agreement reported.

How many cases is enough?

Statistics sets the floor. If the Institute wants to claim, with 95% confidence, that the system's rate of potentially harmful answers is below some threshold p, and it observes no harmful answers at all, the **rule of three** gives the number of cases needed as about $3/p$. To support "below 1%", about 300 cleanly graded cases with zero harmful answers. To support "below 0.1%", about 3,000. And those are the numbers for *zero* observed

failures; any failure pushes the requirement higher. That arithmetic is why the evaluation panel is the budget line I would protect.

It also has to be done per category. A system that is excellent on adult medicine and poor on paediatric dosing has not got a "0.5% harmful rate"; it has a paediatric problem, averaged away. Results should be reported by domain, by professional track, and by severity.

Over-refusal is a failure too

The opening story is itself a failure mode, and it should be measured. The evaluation should report how often the system declines, hedges or escalates when a competent answer was available and safe. A system that is never harmful because it never says anything useful has not passed; it has failed differently, and clinicians will route around it to a chatbot on their phones.

From bench to ward

Offline evaluation is necessary and nowhere near sufficient. The progression I would propose:

1. **Offline evaluation** on the question set, for every candidate model and every configuration change.
2. **Silent mode.** The system runs alongside real practice at a small number of facilities, generating answers that are logged but not shown, and clinicians' actual decisions are compared with what the system would have said. This finds the failures the question set did not anticipate, at no risk to patients.
3. **Supervised live use** with a limited cohort of the Institute's certified clinicians, with every answer available for review and a low-friction way to flag a bad one.
4. **Stepped-wedge roll-out** across facilities, measured at the level of behaviour and patient outcome — Kirkpatrick levels three and four, as I argued in Measuring What Actually Matters — rather than satisfaction.

And after release, the same set re-runs on every change, silent-mode sampling continues on a fraction of traffic, and flagged answers feed back into the question set. The annual report to the Pharmacy and Poisons Board is a by-product of doing this properly.

A plan in phases

Phase	Months	What happens	Gate to the next phase
0 — Foundations	0–3	Buy the Tier 0 machine. Recruit the lead ML engineer, the clinical informatician and the evaluation lead. Convene the panel. Write the intended-use statement and open the hazard log. Pre-submission meeting with the Pharmacy and Poisons Board; DPIA begun	Intended use agreed; evaluation protocol approved by the panel

Phase	Months	What happens	Gate to the next phase
1 — Evaluate	3–9	Build the first 1,000 cases and rubrics. Build the orchestrator, retrieval corpus and dose tools. Evaluate the shortlist. Choose workhorse and reasoning models	A configuration meets pre-agreed thresholds by category
2 — Build	6–12	Procure Tier 1. Commission colocation and the campus site. Security review and external penetration test. Backup and failover drills	Restore from tape proven; failover proven; pen-test findings closed
3 — Silent	12–15	Silent mode at three to five facilities. Grow the question set from real traffic	No unresolved catastrophic-severity failures; silent-mode concordance acceptable to the panel
4 — Supervised live	15–21	Live use by a limited certified cohort. Regulatory submissions to the PPB and the Digital Health Agency	Regulatory positions obtained; flag rate and harm rate within bounds
5 — Roll-out	21–36	Stepped-wedge expansion. Decide on Tier 2 only if evidence shows the reasoning tier materially improves safety	Kirkpatrick 3/4 evidence published — whatever it shows

What this does not do

Every system I have written about on this blog has a section like this, and in this case it is the most important one.

It does not make the model correct. It makes the model *fixed*, *tested* and *watched*. The error rate is whatever the evaluation says it is, and it will not be zero.

It does not let the Institute understand the weights. Owning the files is not understanding them. The Institute can vouch for the behaviour it has measured, on the population it measured it on, and for nothing beyond that.

It does not rule out hidden behaviour in any model. Backdoors of the sleeper-agent kind cannot be excluded by inspection or by ordinary safety testing, for a model from any country. The defence is architectural: a model that can only read, whose numbers are recomputed, and whose outputs are checked, has limited room to do harm even if its disposition is not what it seems.

It does not keep pace with the frontier. The best open-weight models trail the best closed models — by about eight months on CAISI's measurement of V4-Pro — and a validated configuration will be frozen for months at a time on top of that. The Institute is choosing to be somewhat behind the frontier in exchange for knowing exactly what it is running. That is the right trade for clinical use, and it is a trade.

It does not make anything cheaper. It costs more than an API, by roughly USD 1.2–1.4 million over five years at the scale I have assumed.

It does not remove automation bias. It probably makes it worse, and the design must work against that deliberately.

It does not move liability off the clinician — and it moves a manufacturer's obligations *onto* the Institute. Once the Institute supplies medical device software, it has post-market duties, incident reporting and the possibility of being wrong in public. That is the right place for the obligation to sit, but the board should accept it knowingly.

It does not fix the training-data question. The Institute cannot disclose what the developers never published. Whether the regulator accepts a safety case resting on local validation instead is not yet known.

It does not protect data the clinician sends elsewhere. If the Institute's service is slow, restrictive or down, clinicians will go back to the chatbot on their phones, and every protection in this essay will be bypassed in a single paste. The service has to be *better to use* than the alternative, or it will not be used.

Decisions I would put to the board

In the style of the MedLattice and Malipo Rail decision logs, these are the questions that need an owner and an answer. None is decided.

#	Decision	Why it matters
L-01	Approve Phase 0: the Tier 0 machine and the first three hires	Everything else depends on evaluation evidence that does not yet exist
L-02	Intended use: clinical question-answering for certified clinicians only, or also for trainees?	Changes device classification and the risk tolerance of the whole system
L-03	Primary site in colocation versus on campus	Colocation is more resilient per shilling; campus is more controllable and doubles as a teaching facility
L-04	Whether a DeepSeek model may be deployed in production, given funder and partner conditions	A governance question, to be decided with grant terms in hand
L-05	Whether questions containing **no** patient information may ever be sent to a frontier API as an optional "second opinion"	Would narrow the capability gap; requires a reliable way to guarantee the absence of personal data, which is harder than it sounds
L-06	Approach to the PPB on training-data disclosure for foundation models	Could block the regulatory route entirely if not settled early
L-07	Harm-rate thresholds per category that must be met before silent mode and before live use	Must be set by the panel *before* any model is scored, or they will be set to whatever the best model achieved
L-08	Retention period for the audit log	Tension between clinical-record retention and data minimisation, the same tension MedLattice resolves by crypto-shredding
L-09	Offline on-device fallback: permitted, and with what labelling?	Resilience versus the risk of a weaker model being mistaken for the full service
L-10	Whether the Institute will publish its evaluation set and results	Publishing builds trust and lets others check the work; it also exposes the set to contamination in future training data

The clinical officer, again

Go back to the county hospital at two in the morning. What does she get from all of this?

Not a model that "gives medical advice" in the way the phrase is usually meant. She gets an answer that opens with the ceftriaxone dose for 7.1 kilograms, computed by a function a pharmacist has checked rather than recalled by a network; that cites the Kenyan paediatric guideline paragraph it came from; that tells her the antibiotic should not wait for a lumbar puncture she cannot safely perform alone; that says plainly what the evidence on dexamethasone is in settings like hers and how sure it is of that; and that was produced by a system whose every layer someone in Nairobi can name, whose model has a digest in a manifest, and whose behaviour on hundreds of cases like hers has been graded by clinicians who work where she works.

And her question never left the country.

That is what I mean by a model the Institute can answer for. Not one it fully understands — nobody can offer that for any model — but one it has fixed, tested, watched, and stands behind in writing. The hardware is the easy part. **The answering-for is the work.**

References

Models

- DeepSeek-AI. DeepSeek-V4-Pro model card and DeepSeek V4 technical documentation, April 2026.
- DeepSeek-AI. DeepSeek-V4-Flash-0731 model card, July 2026.
- The Decoder. Alibaba's Qwen team releases Qwen 3.8 models with open weights under the Apache 2.0 licence, August 2026.
- TECHi. Kimi K3's open weights arrive July 27. The catch is 1.4TB, July 2026.
- The New Stack. Z.ai's GLM-5.3 goes open weight, but its new license aims at hyperscalers, August 2026.
- OpenAI. Introducing gpt-oss and gpt-oss-120b and gpt-oss-20b model card, August 2025.
- Google. Gemma 4: byte for byte, the most capable open models, April 2026.
- Google. MedGemma 27B text model card and MedGemma 1.5 model card.
- Wavect. Best open-weight LLMs 2026.

Safety, evaluation and security

- NIST Center for AI Standards and Innovation. CAISI evaluation of DeepSeek AI models finds shortcomings and risks, September 2025.
- NIST Center for AI Standards and Innovation. CAISI evaluation of DeepSeek V4 Pro, May 2026.
- Hubinger E, et al. Sleeper Agents: training deceptive LLMs that persist through safety training. arXiv 2401.05566, 2024.
- Fafuła A. Abliteration is not a scalpel: off-target effects of refusal removal on decision disposition across model families. arXiv, July 2026.
- Arora RK, et al. HealthBench: evaluating large language models towards improved human health. OpenAI, 2025.
- Tom's Guide. ChatGPT will still offer medical and legal advice — despite what rumors suggest, November 2025; OpenAI usage policies.
- Ng DN. 2× GH200 for LLM inference, part 4: DeepSeek V4 Flash — SGLang vs vLLM at 1M context, 2026.

Kenyan law and regulation

- Pharmacy and Poisons Board. Guideline on Regulation of Medical Device Software in Kenya (MDSW), April 2026; see also Health Business, Techweez and The Kenyan Wallstreet.
- Digital Health Act, 2023 (No. 15 of 2023).
- Digital Health Agency. Certification programme.

- Aggarwal N. The Law Is Part of the Architecture, August 2026.

Hardware, power and cost

- Thunder Compute. NVIDIA RTX PRO 6000 Blackwell pricing, September 2026; Tom's Hardware, Nvidia doubles RTX PRO 6000 Blackwell's MSRP to $16,000.
- Mercatus. NVIDIA B200 server price in 2026 and H200 server price in 2026.
- Fello AI. M5 Ultra Mac Studio: price, specs and 512GB memory, September 2026.
- CNBC. AI memory is sold out, causing an unprecedented surge in prices, January 2026.
- Global Trade & Sanctions Law. Reported draft rules signal new semiconductor export controls framework, March 2026.
- LeadAfrik. How import duty is calculated in Kenya (2026).
- GlobalPetrolPrices. Kenya electricity prices; Nairobi Wire, EPRA raises electricity costs by KSh4.70 per unit in August.
- TechTrends KE. Kenya Power explains what triggered Wednesday night's nationwide blackout, July 2026.
- Ki Energy Tech. Commercial solar installation in Kenya: ROI guide 2026.
- iXAfrica. Collaboration with Oracle to deliver Kenya's first public cloud region, January 2026.
- Nucamp. AI salaries in Kenya in 2026.
- ASHRAE TC 9.9. *Thermal Guidelines for Data Processing Environments*, 5th edition (altitude derating).

Earlier essays in this series

- Another Arrow in the Quiver — the Institute blueprint.
- The Angoff Panel for Testing Clinicians and Measuring What Actually Matters.
- Not Shown Is Not Locked and Twenty Generals, One Ledger — MedLattice.
- The Capstone: Build a Grounded Assistant.