// ARTICLE · PUBLISHED 25 OCTOBER 2026

# Local-First Medicine: The Box in the Clinic

Three October announcements make one question sharper: how much of a clinic's AI work has to leave the building at all

**Jason Newell** · MAX Research Collective · Article ART-HLC-016 · Published 25 October 2026
Licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/
Original: https://www.maxresearchcollective.com/articles/local-first-medicine.html
Cite as: Newell, J. (2026). Local-First Medicine: The Box in the Clinic. MAX Research Collective, ART-HLC-016. https://www.maxresearchcollective.com/articles/local-first-medicine.html

## How much has to leave

Most of a clinic's routine AI work could run on a machine in the clinic, with only the hard cases leaving. Nobody has measured that on real clinic workloads yet. Three announcements from October 2026 still turn it into a practical question. A European company now keeps patient data and model processing on a European-owned cloud. Anthropic released a small, cheap model built for routine work. And Google's open medical models were published in Nature Medicine.

Each answers part of the question and none answers all of it. Put together they sketch a clinic where routine work stays on a box under the desk. Harder work goes up through a gate, and a person still decides. That's the pattern MARA, the Medical Agentic Resident Assistant this project is designing, inherits from the AI resident she's built on.

## Why location matters

Two things travel when a clinic uses an AI tool: the patient's information and the work done on it. Most clinical AI today sends both to a vendor's servers, and that can be perfectly lawful. In the United States a vendor handling protected health

information signs a business associate agreement (BAA), the contract HIPAA requires; in Europe the General Data Protection Regulation (GDPR) sets the terms. A contract governs what a vendor may do with the data. It doesn't change where the data sits, or who could be compelled to hand it over.

The work can happen in three broad places.

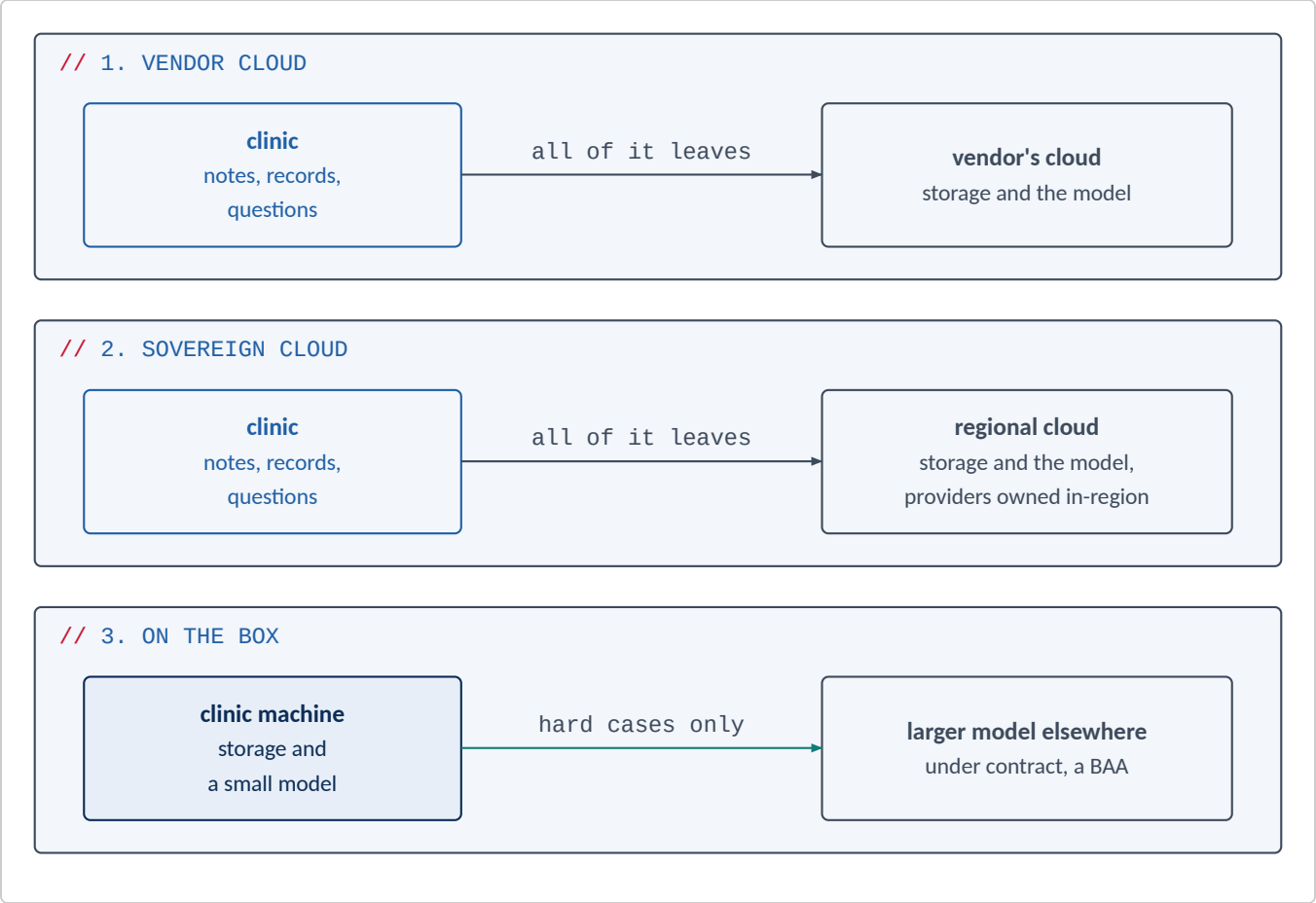



Figure 1. Read each row left to right: in the first two the whole record leaves the clinic, and in the third only the hard cases do.

The vendor's cloud is the default. A regional or "sovereign" cloud keeps both data and processing inside one jurisdiction, with providers owned there. The third place is the machine in the clinic itself, with a larger model elsewhere reached only for cases the local one can't handle — and only under a contract and a gate.

## The sovereign cloud: Tandem Health

Tandem Health sells a clinical assistant that writes notes and supports coding and decision support, and in September 2026 it raised a $100 million Series B, led by the Scaleup Europe Fund that EQT manages, on the strength of about 10,000 care organisations in 14 European markets. Its new option is called Sovereign Health. Patient data is stored and processed on Scaleway, a European-owned cloud, and Tandem holds the encryption keys. The models are open-weight ones Tandem hosts there (models whose trained parameters are published), with frontier models available on top. Enterprise customers get it on request.

The page also claims open-weight models are held to the same safety bar as frontier ones, and backs it with an evaluation of 47 models scored on 50 transcripts for note generation. Its table lists 12. It never says which of them run on Sovereign Health, so the claim can't be checked.

What Tandem does show is that keeping processing inside one jurisdiction sells at scale. The data still leaves the clinic.

## The small model: Claude Haiku 5.5

On 7 October Anthropic released Claude Haiku 5.5, which it calls "the cheapest, fastest, and most capable small model we've ever released." Anthropic pitches it at "high-volume, cost-sensitive tasks": summaries, classification, database queries, live customer support and work as a helper to larger models. Input costs $0.10 per million tokens (a token is a short piece of text) for prompts up to 100,000 tokens. Anthropic says that's about 75% cheaper on average than Haiku 4.5, the model it replaces.

Haiku 5.5 is hosted, so a clinic can't run it on its own machine. Its relevance is the pattern it makes explicit. Most of the work in a workflow is routine. Routine work can go to a small model, and the expensive model is kept for the decisions that need it; that split doesn't care where the small model runs.

## The open model: MedGemma

Google's MedGemma paper appeared in Nature Medicine on 6 October 2026 (Sellergren and colleagues, volume 32, pages 3695 to 3704). MedGemma is a family of medical vision-language models, models that read images and text together, built on Google's open Gemma 3. On tasks outside its training data the abstract reports gains over the base models of 2.6 to 10% in medical image question answering, 15.5 to 18.1% in classifying chest X-ray findings and 10.8% in agentic evaluations. The authors also found that fine-tuning MedGemma beat fine-tuning the base model when training data was scarce.

The weights are on Hugging Face in at least two sizes, about 4 billion and about 27 billion parameters, behind an access gate and under Google's own licence terms. The smaller one fits on a single modern workstation graphics card, which is the kind of hardware a clinic could actually buy and keep under a desk, while the larger needs a big card. Neither has been run here.

Three cautions apply. Every gain is measured against the base model, never against a clinician or a cleared device. Benchmarks aren't trials. And the licence has to be read before any use. MAX's rules require every dependency outside the language model itself to be open source, and every model licence gets checked before adoption.

## What could run on the box

Taken together, the announcements suggest a division of labour.

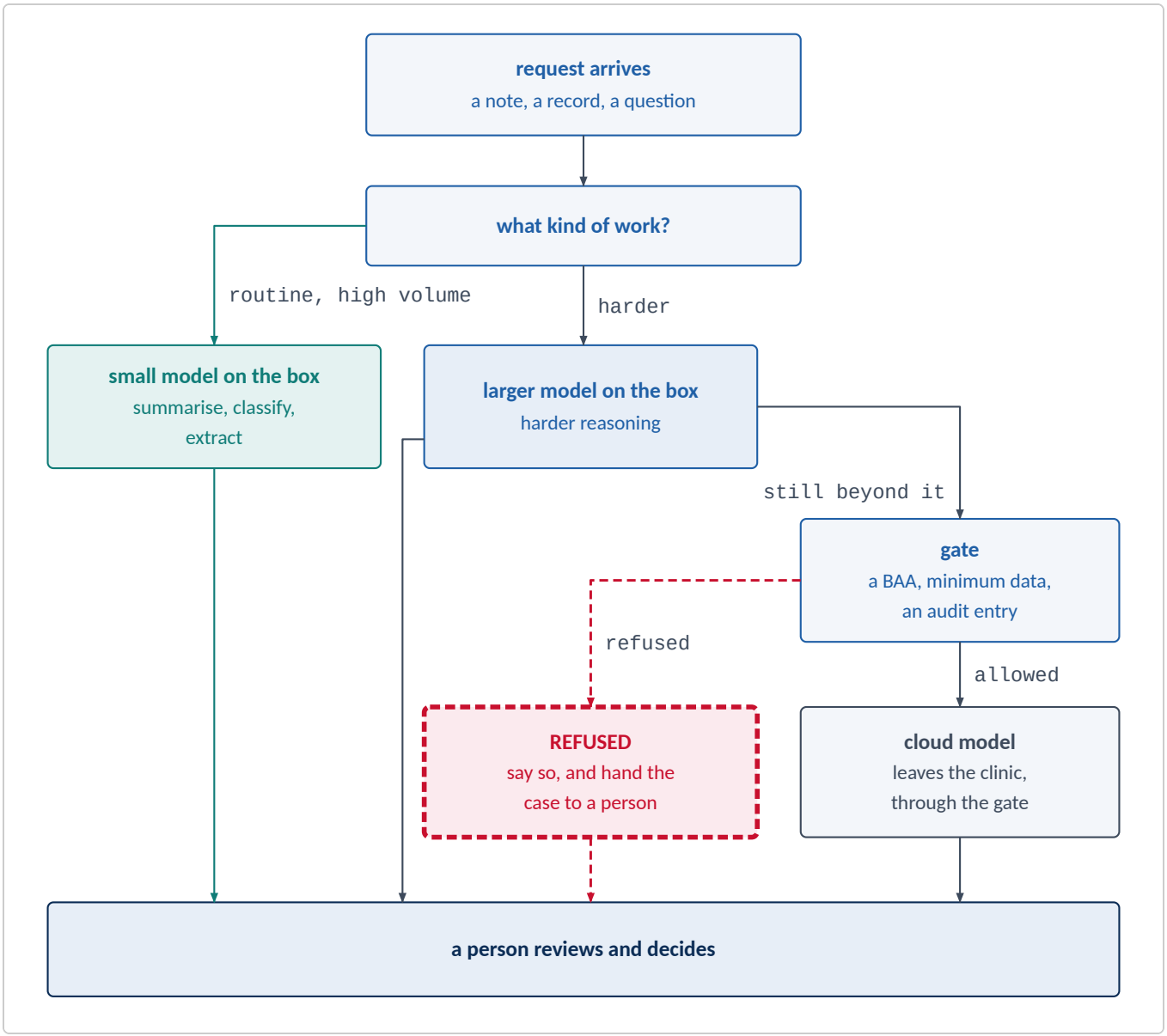



Figure 2. Only the right-hand path leaves the clinic, and it passes the gate first; every path ends at the same person.

Routine, high-volume work (summarising a record, classifying a document, pulling out a medication list) goes to a small model on the clinic's machine. Harder reasoning goes to a larger local model if the machine can hold one. Only what's beyond both leaves, and only through a gate: a signed BAA, the minimum data the task needs, and an audit entry. If the gate refuses, the system says so and hands the case to a person. On every path a person reviews the result and decides.

That's a design. No one has shown, on real clinic workloads, what share of the work a small local model handles well enough. That share decides everything. A high one

makes the pattern a privacy gain; a low one makes it a slower road to the cloud.

## From MAX's rule to MARA's case

MAX, the AI resident this collective runs on one PC, has ”local-first” as his first hard rule. Routine interactions must work without the internet, cloud models are ”an upgrade path for hard problems, not a default dependency,” and ”privacy-sensitive interactions must be possible without ever leaving the machine.” He was built that way for a household. MARA’s technical design carries the same rule into healthcare in one line: model inference on site where possible, and the cloud only under a BAA and a gate.

Her compliance notes add the rest. Protected health information is encrypted at rest, access is by role and audited, and only the minimum necessary is used. Her graph holds de-identified patient data by default and identifiable data only where the law allows it.

The privacy case for MARA is a claim about architecture: for routine work, the data never leaves. It’s testable, and the first test runs in MAX4, the desktop generation of MAX from October 2026 on, which holds no patient data. There it can be measured how much work stays local, how often the gate is reached, and what happens when the network goes down.

## What isn't known

- Local quality is unmeasured. How well small local models handle real clinical documents, against larger hosted ones, hasn’t been shown; MedGemma’s published comparisons are against its own base model.
- Tandem’s safety claim is unsupported as published. Its page doesn’t show which models in its evaluation run on Sovereign Health.
- The hardware bill is unknown. A box in a clinic needs updates, backups and security, and that work moves from the vendor to the clinic.
- The regulator’s view is open. A model that supports clinical decisions may be a medical device wherever it runs.

## Why it matters

A business associate agreement is a promise about a vendor's conduct.

A box in the clinic is a fact about where the data is. Ask who could have seen a record. Under the first, the answer depends on a contract holding up; under the second, it depends on a cable. If the routine share of clinical AI work turns out to be large, most records will never need the promise at all.

## Sources


- Tandem Health. "Sovereign Health" product page, opened 21 October 2026. https://tandemhealth.ai/sovereign-health
- HLTH. "Tandem Health raises $100M to expand clinical AI across Europe." 15 September 2026. https://hlth.com/insights/news/tandem-health-raises-100m-to-expand-clinical-ai-across-europe
- Anthropic. "Claude Haiku 5.5." Released 7 October 2026, opened 21 October 2026. https://www.anthropic.com/claude-haiku-5-5
- Sellergren A, et al. "An open vision-language model for diverse medical applications." Nature Medicine 32(10), 3695–3704 (2026), abstract via PubMed, PMID 42839104. https://doi.org/10.1038/s41591-026-04626-w
- Google. MedGemma model pages on Hugging Face, google/medgemma-4b-it and google/medgemma-27b-it. https://huggingface.co/google/medgemma-4b-it
- MAX3 repository, STANDING_RULES.md hard rule 1; MARA technical design and compliance notes (mara/spec/MARA_TECHNICAL_DESIGN.md, mara/compliance/PHI.md).

# Glossary

| TERM | DEFINITION |
|---|---|
| Audit entry | a record of what was sent, by whom and under which permission. The gate writes one before any case leaves the clinic. |
| BAA (business associate agreement) | the US contract under which a vendor may handle protected health information for a provider. It's the first check at the gate. |
| De-identified data | patient data with the details that identify a person removed. MARA's graph holds it by default. |
| Fine-tuning | further training of an existing model on a smaller, specific data set. MedGemma's authors found it worked better from MedGemma than from the base model when data was scarce. |
| Gate | the check a hard case must pass before it leaves the clinic, covering the contract, the minimum data and the audit entry. If it refuses, a person takes the case. |
| GDPR (General Data Protection Regulation) | the European Union's data protection law. It sets the terms for vendors handling health data in Europe. |
| Local-first | a design rule that routine work runs on the user's own machine and the network is an upgrade. It's MAX's first hard rule and MARA's privacy case. |
| MARA (Medical Agentic Resident Assistant) | this project's design for a clinical assistant that recommends while a person decides. She inherits the local-first rule from MAX. |
| MAX4 | the desktop generation of the MAX AI resident, from October 2026 on. It's where the local-first pattern is measured before MARA uses it. |

Newell, J. (2026). MAX Research Collective, ART-HLC-016. CC BY 4.0

| TERM | DEFINITION |
|---|---|
| Open-weight model | a model whose trained parameters are published so others can run it, under licence terms that still apply. Tandem hosts such models, and MedGemma is one. |
| Protected health information | health data that identifies a person, which HIPAA protects. It's what the gate keeps from leaving without cause. |
| Sovereign cloud | cloud infrastructure owned and run within one jurisdiction, so data and processing stay under its law. Tandem's Sovereign Health is an example. |
| Token | a short piece of text a language model reads or writes. Model prices are quoted per million tokens. |
| Vision-language model | a model that reads images and text together. MedGemma is one, aimed at medical images and records. |

HEALTHCARE ART-HLC-016

// ABOUT THE AUTHOR

**Jason Newell** leads organizations and people in technology, and is a researcher, deep technologist and builder — from a CRAY to agentic AI. He founded MAX Research Collective to build AI for regulated industries that runs where its owner has control, with healthcare as its strongest focus. About Jason → Connect on LinkedIn →