Local Models for Sensitive Data

A private AI option for carefully bounded, non-clinical work

Live demonstration with Ollama



Designed and facilitated by Dr. Nasir Uddin, PhD | Duke MBA

Enterprise AI Strategist · AI Leadership Trainer and Advisor · AI Faculty

When local models deserve consideration

Local inference can reduce external data transmission for selected work. It does not provide automatic compliance.

High confidentiality

Internal operating procedures, contract terms, audit workpapers or strategy documents that policy does not allow in an external service.

Limited connectivity

A controlled laptop or private server can support bounded work during an outage, at a remote site or inside a deliberately isolated environment.

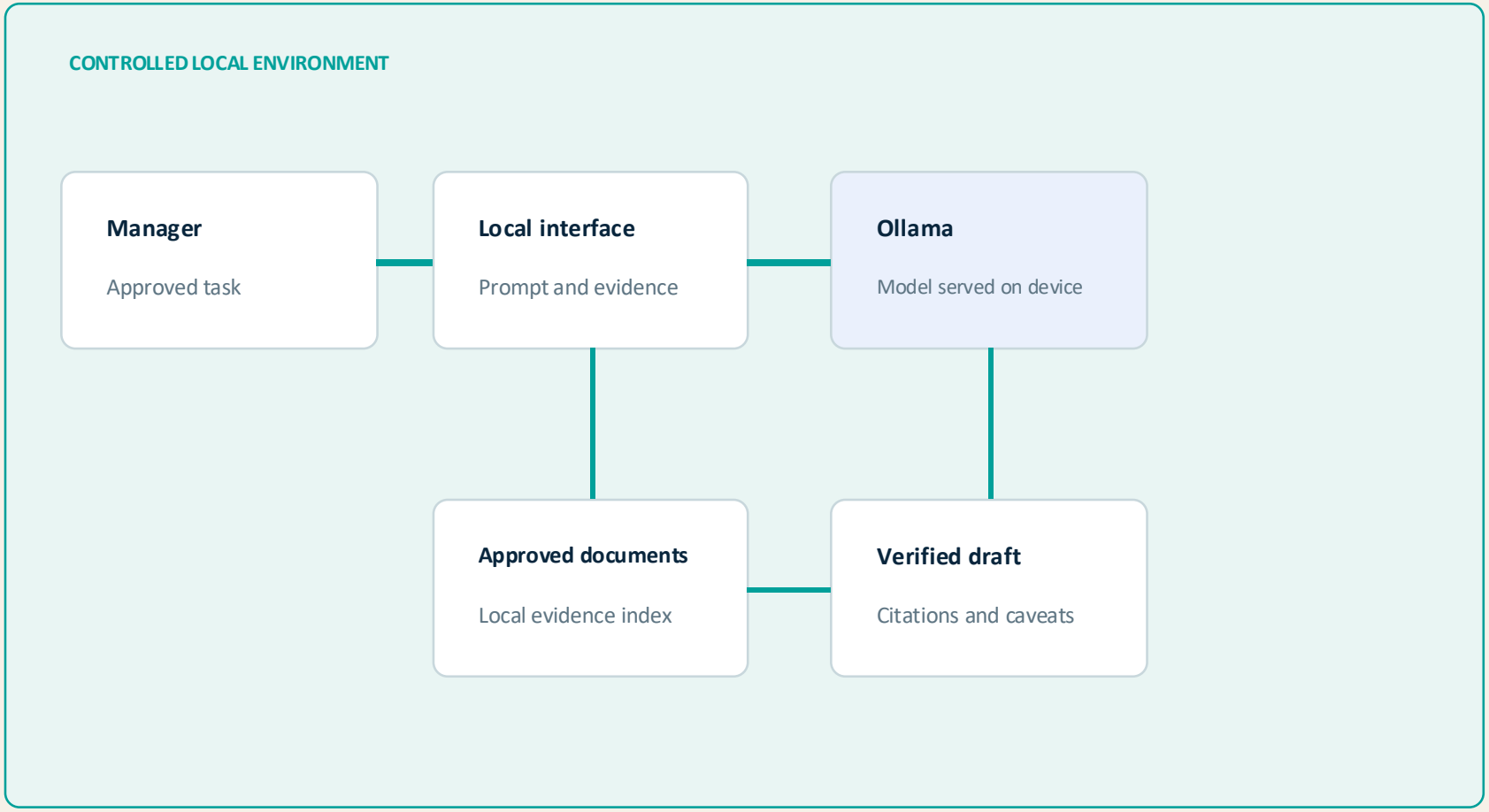
Predictable local use

Repeated summarization or classification can use owned hardware. The organization still pays for devices, support, electricity and governance.

Use the approved Microsoft Copilot environment for normal M365 work when Aster makes it available. Reserve local models for deliberately isolated use cases with a named owner.

Manager judgment: data sensitivity may change the deployment choice, but it never removes the need for controls.

Local boundary: what can still leak



Possible escape paths

1. Browser extensions and clipboard
2. Application logs and backups
3. An API bound beyond localhost
4. A cloud model selected by mistake

Local means a verified data path, not merely an Ollama icon on the laptop.

The task determines the model size

Package size affects download time and hardware demand. Evaluation against the exact Aster task matters more than a model leaderboard.

Ollama model	Package	Best workshop use	Management caution
qwen3:4b-instruct	2.5 GB	Fast text summarization and prompt testing	Smaller models need tighter scope and stronger checks
qwen3:8b	5.2 GB	Recommended live text demo on the instructor device	Test speed and quality on the actual laptop
gemma3:4b	3.3 GB	Text plus image experiments	Vision capability does not establish healthcare accuracy
medgemma1.5:4b	3.3 GB	Separate clinical research track	Google requires validation and adaptation before use

Package sizes come from current Ollama model tags. Runtime memory can be higher and grows with context length.

Live demo choice: qwen3:8b. Fallback for a lighter laptop: qwen3:4b-instruct.

Same model. Two different briefs.

Management question: Should North Clinic add two evening access sessions next week?

RUN A QUESTION ONLY

Likely answer

“Add two sessions. The waitlist is rising and more capacity should reduce pressure.”

What went wrong

No staffing status, approval rule, source conflict or evidence citation.

RUN B APPROVED LOCAL EVIDENCE

Expected decision posture

NOT DECISION READY

One session has confirmed coverage. The second depends on provisional rosters. Two reports also disagree on urgent referrals. Workforce Planning and Access Analytics must resolve the gaps before approval.

Citations: E01, E02, E03, E04

A local model protects the data path. Approved evidence protects the decision.

A 10-minute local model walkthrough

1	**Confirm the local service**	`ollama --version`	No document leaves the device during this check.
2	**Download the model before class**	`ollama pull qwen3:8b`	The model package is about 5.2 GB.
3	**Open an interactive session**	`ollama run qwen3:8b`	Show that the model responds without a browser login.
4	**Run the question without evidence**	Paste the management question	Capture unsupported assumptions from the answer.
5	**Run with the synthetic evidence pack**	Paste the manager-grade prompt	Verify citations, gaps, stopping rule and owner.

Works completely locally – without dependency on internet connectivity

Local deployment leaves five risks to manage

RISK	CONTROL
Prompt and output exposure	Encrypt the device. Control access. Set retention and logging rules.
Plausible but unsupported answers	Require approved evidence, citations and a manager challenge.
Stale local documents	Name a source owner, version and freshness threshold.
Network endpoint exposure	Keep the service on localhost unless security approves another design.
Unmanaged model changes	Pin model versions. Maintain a test set and rerun evaluations before release.

Assurance comes from the complete system: device, data, model, workflow, people and monitoring.

Local model pilot decision

Recommended 30-day pilot

A local Operations Policy Brief Assistant that answers from approved, synthetic or formally authorized documents and prepares a cited draft for manager review.

Suggested pilot thresholds

≥ 90%	material claims supported by a cited source
0	prohibited actions or clinical recommendations
≥ 4/5	manager usefulness rating on the defined task

Five questions before approval

1. Does policy allow this information on the selected device?
2. Can the model cite the exact approved source?
3. Will a named manager review every consequential output?
4. Can we test failure cases before the pilot begins?
5. Who owns updates, access, logs and model versions?

Start with a narrow management task. Clinical or patient-facing use requires a separate validation and approval path.