

DropTicks AI Article

Scaling Notes: local LLM prototyping

Local AI | Mar 06, 2026

How the workflow changes as traffic, users, or data volume grows. how local models support demos, privacy-sensitive tests, and offline workflows

Article

Scaling Notes for local LLM prototyping

This article explains how local models support demos, privacy-sensitive tests, and offline workflows. It is written for teams that need production behavior, not only a convincing prototype.

Start by defining the workflow boundary. List the inputs, outputs, permissions, failure states, and human review points. Most software and AI failures become easier to manage when the boundary is explicit.

The implementation should separate product intent from infrastructure mechanics. Keep validation, storage, observability, retries, and deployment rules outside of prompts or UI copy. Those rules belong in code, configuration, and tests.

A strong delivery plan includes:

- â€¢ A small reference workflow that proves the approach.

- â€¢ Test data that represents real edge cases.

- â€¢ Logging and metrics that expose failures instead of hiding them.

- â€¢ A rollback path for bad releases.

- â€¢ Documentation that explains ownership and maintenance.

For DropTicks projects, the best solution is usually the one that makes the system easier to inspect. If a team cannot explain why a tool was selected, how it fails, and how it is measured, the system is not ready.

Related source links:

- â€¢ <https://github.com/ollama/ollama/tree/main/docs>

â€ https://github.com/ggerganov/llama.cpp

Web link

<https://www.dropticks.com/articles/scaling-notes-local-llm-prototyping>

Related links

<https://github.com/ollama/ollama/tree/main/docs>

<https://github.com/ggerganov/llama.cpp>