Private AI: Running an LLM Chatbot on Your Local Network

March 26th, 2026

Denis Rechkunov

elastic | The Search AI Company

Denis Rechkunov

Principal Software Engineer I,
Elastic Agent

Agenda

1. My requirements
2. Hardware choice
3. What components are involved
4. Model choice
5. Live Demos
6. Running services on macOS
7. Q & A

My Requirements

My Requirements

1. Open Source
2. Privacy: no telemetry, no spying, no profiling
3. Everything runs locally (except the web search)
4. Multilingual: at least English, German, Russian
5. Document processing: **O**ptical **C**haracter **R**ecognition, embedding
6. Using on multiple devices by multiple people on my **local network**
7. Reasonable speed and accuracy – purely subjective
8. *(Optional)* Conversational mode

Hardware Choice

Hardware Choice

When Apple is an **affordable** option

- Macs have solid GPUs nowadays
- **Unified memory**
- Multi-purpose computer

Mac Mini (512GB SSD)

Chip	Memory	Model Size	Price*
M4	16 GB (120 GB/s)	8 B	$799
M4	24 GB (120 GB/s)	14 B	$999
M4	32 GB (120 GB/s)	14 – 20 B	$1,199
M4 Pro	24 GB (273 GB/s)	14 – 30 B	$1,399
M4 Pro	48 GB (273 GB/s)	32 B	$1,799
M4 Pro	**64 GB (273 GB/s)**	**32 B**	**$1,999**

M4 Pro has optional +2 CPU cores and +4 GPU cores for $200
*prices are from apple.com for a 512 GB SSD variant, 2026-03-23

In Stock OFF
Sold by Newegg OFF
AI Ready ON
New OFF
Combo Deals Only OFF

Brands
- ASUS
- MSI
- ASRock
- GIGABYTE
- XFX
- PowerColor

SHOW MORE

Price

$ to $

- $400 - $500
- $500 - $750
- $750 - $1000
- $1000 - $1250
- $1250 - $1500
- $1500 - $2000

SHOW MORE

GPU
- GeForce RTX 50 Series
- GeForce RTX 40 Series
- Radeon RX 9000 Series

Search Within: GO

Sort By: Lowest Price

Useful Links: AI Ready ×

Page 1/1

View: 36

And It's only for a GPU!

AI Ready

(1)

ASRock

ASRock B60 Intel Arc Pro B60 B60 CT 24G 24GB 192-bit GDDR6 PCI Express 5.0 x8 Graphics

- Model #: B60 CT 24G

$659.99

FREE SHIPPING from United States

ADD TO CART

Compare

AI Ready

(2)

ARKN INTEL ARC PRO B60 24GB GDDR6 GRAPHICS CARD Quad DP 8357-00128 (Brown box)

Get free Team Group 256GB SSD w/ purchase, while supplies last

- Model #: 8357-00128

$659.99

FREE SHIPPING from United States

ADD TO CART

Compare

AI Ready

(7)

ASRock

ASRock Monster Hunter Wilds Radeon RX 9070 XT 16GB GDDR6 PCI Express 5.0 x16 Graphics Card RX9070XT MH 16G

Get CRIMSON DESERT w/ purchase, while supplies last

- Model #: RX9070XT MH 16G

$749.99

FREE SHIPPING from United States

ADD TO CART

Compare

AI Ready

PowerColor

PowerColor Hellhound Radeon RX 9070 XT 16GB GDDR6 PCI Express 5.0 x16 Graphics CARD Hellhound Spectral White AMD Radeon RX 9070 XT 16GB GDDR6

~~$900.99~~

$849.99

Save: $51.00 (5%)

FREE SHIPPING from United States

*prices are from www.newegg.com, 2026-03-18

elastic

(2)

NVIDIA

NVIDIA GeForce RTX 3090 Founders Edition Graphics Card 24GB GDDR6X - Titanium and Black

- Model #: 900-1G136-2510-000

$1,510.20

More options from $1,510.20 - $1,625.00

FREE SHIPPING from Hong Kong

Order will ship out in **4** business days

ADD TO CART

Compare

(6)

NVIDIA

NVIDIA GeForce RTX 4090 Founders Edition Graphics Card 24GB GDDR6X - Titanium and Black

- Model #: 900-1G136-2530-000

$3,765.00

More options from $2,999.99 - $4,699.00

FREE SHIPPING from United States

ADD TO CART

(1)

NVIDIA

NVIDIA Founder Edition GeForce RTX 5090 32GB GDDR7 PCI Express 5.0 x16 ATX Graphics Card 900-1G144-2530-000

- Model #: RTX 5090

$4,498.99

More options from $3,599.99 - $5,599.00

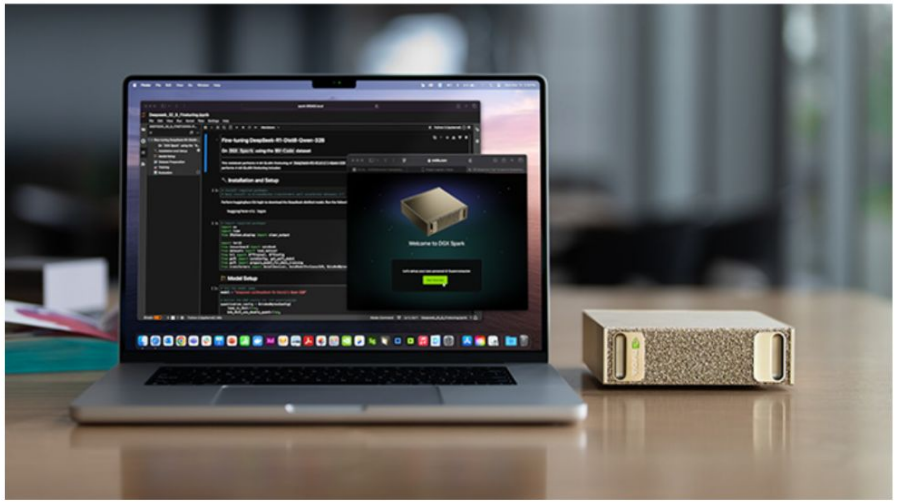
FREE SHIPPING from United States

ADD TO CART

Compare

*prices are from www.newegg.com, 2026-03-18

And this is what NVIDIA thinks you should buy...



NVIDIA DGX Spark

A Grace Blackwell AI Supercomputer on your desk.

> NVIDIA GB10 Grace Blackwell superchip

> 1 PFLOPS of FP4 AI performance

> 128GB of coherent, unified system memory

> ConnectX-7 Smart NIC

> 4TB NVME.M2 with self-encryption

> 150mm L x 150mm W x 50.5mm H

> Free 90-day NVIDIA AI Enterprise - DGX Spark License

Special Bonus: Enjoy a free, DLI hands-on AI course included with your NVIDIA DGX Spark purchase ($90 value).

$4,699.00 Add to Cart

Components

Chatbot Components

To run a private AI chatbot we would need:

Often combined

- Model manager (search, download, delete)
- Engine(s)
- Web Chat Client
- Web Search
- *(Optional)* Nginx + Let's Encrypt

Chatbot Components

oMLX includes:

- Model manager
- Engines:
  - Regular text models: **mlx-lm**
  - Models with vision: **mlx-vlm**
  - Document processing: **mlx-embeddings**
  - **T**ext-**T**o-**S**peech, **S**peech-**T**o-**T**ext: **mlx-audio** (Coming soon)
- Paged SSD KV caching

Compatible with any Anthropic / OpenAI-compatible client

(Claude Code, OpenClaw, Cursor, etc.)

Chatbot Components

Origin	**GGUF** (llama.cpp)	**MLX** (Apple)
Target Hardware	**Universal**: CPU, NVIDIA, AMD, Apple.	**Exclusive**: Apple Silicon (M1–M5) only.
Memory Strategy	**mmap + Page Faults**: Swaps model parts to disk if > RAM.	**Unified Memory (UMA)**: CPU/GPU shared memory; zero-copy, no explicit swap.
Execution Model	**Pre-compiled Kernels**: Fixed C++ kernels; no graph compilation.	**Lazy Graph + JIT**: Builds Metal kernel graph on-the-fly; auto-fusion.
Performance	**Scalable**: Handles massive models (>70B) via disk swapping.	**Fast (Small)**: Beats GGUF on models <22B; hits RAM wall on huge models.
Quantization	**Granular**: Explicit control (Q4_K_M, Q5_K_XL).	**Dynamic**: Automatic mixed, some layers in 4-bit, some in 8-bit.
Best Use Case	**Cross-platform, massive models, or strict memory constraints.**	**Native Mac apps, training/finetuning, and small-to-mid models.**

Chatbot Components – Web Chat

Open Web UI supports:

- User Management
- Document Processing (knowledge collections, attachments)
- Web Search (**R**etrieval-**A**ugmented **G**eneration or agentic)
- Skills, tools, code execution, MCP, etc.
- Image Processing / Generation
- Conversational Mode (STT+TTS)
- Supports many vector databases (Chroma by default)

Chatbot Components – Web Search

SearxNG – local meta search engine hosted by you:

- Searching the same query across a list of configured search engines
- Every search engine can have a configured weight
- Native support by Open Web UI
- Unfortunately **often gets blocked** by most of the search engines and web-sites

Chatbot Components – Web Search

Brave Search API – online service:

- Monthly: $5 per 1,000 search queries (first 1,000 is free)
- “Own, built-from-scratch index”
- “LLM-Optimized”
- “Does not profile you”
- They store your search queries for 90 days
- Native support by Open Web UI

Model Choice

Model Choice

You'll need multiple models:

- For the **R**etrieval-**A**ugmented **G**eneration Pipeline:
  - Content Extraction Engine
  - Embedding
  - Reranker
- Main Model
- Task Model (lightweight and fast): title, follow ups, search queries, etc.
- *(Optional)* STT / TTS – conversational mode

Search on huggingface.co.

Model Choice

It’s always a balancing act:

- Parameter count ↑ = Smarter ↑ = Memory ↑
  - How vast is the knowledge
- Quantization ↓ = Accuracy: ↓ = Generation Speed ↑
  - Like image compression, the lower the number, the lower the quality
- Capabilities:
  - Instruct or thinking mode?
  - What built-in tools does it support?
    - Vision
    - Web Search
    - Knowledge, Memory, etc.

Model Choice – RAG

- Content Extraction: Docling (**must be hosted as a separate daemon**)
  - Role: Reading a document, maintain the correct structure
  - **O**ptical **C**haracter **R**ecognition
- Embedding: jina-embeddings-v5-text-small-retrieval-mlx
  - Role: Converting documents into vectors, retrieving on query
  - Multilingual
- Reranker: jina-reranker-v3-mlx (will be supported in oMLX 0.2.20)
  - Role: Scoring and re-ordering candidate documents
  - Multilingual, good tests results

Model Choice – Main Model

- Originally settled on **Qwen3-14B-8bit**:
  - Multilingual, thinking **text** model
  - Decent accuracy
  - ~15 tokens/s
  - 15 GB of memory
- Currently evaluating **Qwen3.5-35B-A3B-8bit**:
  - Multilingual, thinking **multimodal** model
  - More competent
  - **Mixture-of-Experts** (fast but less accurate)
  - **~56 tokens/s (fast enough to also be a task model)**
  - **37 GB of memory**

Model Choice – MoE

Imagine a room with 256 specialists:

- In a **regular model**, all 256 specialists look at every word – slow and expensive
- A **MoE** model does something smarter.

  For each token it processes, it only wakes up 9 of those 256 specialists:
  - 8 "routed" experts – chosen dynamically based on what the token needs
  - 1 "shared" expert – always active, no matter what – generalist
  - The rest of the 247 specialists just sit idle for that token – saving compute

As of now, there are 194 MoE models available on Hugging Face.

Model Choice – Configuration

Find an optimal configuration for your model! For example, [Qwen3.5-35B-A3B-8bit]:

- **Thinking mode for general tasks**:
  `temperature=1.0`, `top_p=0.95`, `top_k=20`, `min_p=0.0`, `presence_penalty=1.5`, `repetition_penalty=1.0`
- **Thinking mode for precise coding tasks (e.g., WebDev)**:
  `temperature=0.6`, `top_p=0.95`, `top_k=20`, `min_p=0.0`, `presence_penalty=0.0`, `repetition_penalty=1.0`
- **Instruct (or non-thinking) mode for general tasks**:
  `temperature=0.7`, `top_p=0.8`, `top_k=20`, `min_p=0.0`, `presence_penalty=1.5`, `repetition_penalty=1.0`
- **Instruct (or non-thinking) mode for reasoning tasks**:
  `temperature=1.0`, `top_p=1.0`, `top_k=40`, `min_p=0.0`, `presence_penalty=2.0`, `repetition_penalty=1.0`

Demos

Demos

- oMLX
- Open Web UI
  - Agentic web search via Brave Search API
  - Document Attachment (Docling + RAG)
  - Knowledge Collections (RAG)
  - Skills
  - Vision
  - Admin Panel

macOS as a Server

macOS as a Server

To do things right, you'd need:

- A **separate service account**:
  - No shell
  - No access to your personal files
  - Normal users don't have access to service account configs, settings, dbs.
- Property List (`*.plist`) files in `/Library/LaunchDaemons`:
  - Start services on boot without login
  - Logging locations
  - Watch & restart
- Use `pf` to forward ports (`/etc/pf.conf`). Only root can listen to port 443.
- TLS with Nginx (TLS is required for mic access and push notifications).

```
# Create a new group first
sudo dscl . -create /Groups/_services
sudo dscl . -create /Groups/_services PrimaryGroupID 450

# Create the user
sudo dscl . -create /Users/_svcuser
sudo dscl . -create /Users/_svcuser UserShell /usr/bin/false
sudo dscl . -create /Users/_svcuser RealName "Service Account"
sudo dscl . -create /Users/_svcuser UniqueID 451
sudo dscl . -create /Users/_svcuser PrimaryGroupID 450
sudo dscl . -create /Users/_svcuser NFSHomeDirectory /var/svcuser

# Set a password (required for launchd)
sudo passwd _svcuser

# Create home directory
sudo mkdir -p /var/svcuser
sudo chown -R _svcuser:_services /var/svcuser

# Hide from login screen
sudo dscl . -create /Users/_svcuser IsHidden 1

# Hide home folder from Finder
sudo chflags hidden /var/svcuser
```

/Library/LaunchDaemons/

com.omlx.serve.plist

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE plist PUBLIC "-//Apple//DTD PLIST 1.0//EN"
"http://www.apple.com/DTDs/PropertyList-1.0.dtd">
<plist version="1.0">
<dict>
    <key>Label</key>
    <string>com.omlx.serve</string>
    <key>UserName</key>
    <string>_svcuser</string>
    <key>GroupName</key>
    <string>_services</string>
    <key>ProgramArguments</key>
    <array>
        <string>/opt/homebrew/bin/omlx</string>
        <string>serve</string>
        <string>--model-dir</string>
        <string>/var/svcuser/.omlx/models</string>
    </array>
    <key>EnvironmentVariables</key>
    <dict>
        <key>HOME</key>
        <string>/var/svcuser</string>
    </dict>
    <key>KeepAlive</key>
    <true/>
    <key>RunAtLoad</key>
    <true/>
    <key>StandardOutPath</key>
    <string>/var/svcuser/logs/omlx/omlx.log</string>
    <key>StandardErrorPath</key>
    <string>/var/svcuser/logs/omlx/omlx.error.log</string>
</dict>
</plist>
```

Daemon Prioritization

```
<key>Nice</key>
<integer>-5</integer>
```

Lower nice values cause more favorable scheduling.

```
<key>ProcessType</key>
<string>Interactive</string>
```

If left unspecified, resource limits (CPU, I/O) are applied to the job:

- **Background** – Limits applied
- **Standard** – Standard jobs are equivalent to no ProcessType being set
- **Adaptive** – Move between the **Background** and **Interactive** based on activity
- **Interactive** – No limits, critical to maintaining a responsive user experience

More can be found in `man 5 launchd.plist`.

Conclusion

Is This Practical?

- I spent only a few evenings putting this together (with Claude's help)
- Now I use this chatbot daily for:
  - Research
  - Simple code generation
  - Document analysis
  - Language tools: translation, text-proofing, etc.
- It's almost as good as the online models
- I would have this Mac mini anyway, now it brings even more value
- The most privacy you can get

Her (2013)

Links

- https://omlx.ai
- https://openwebui.com
- Qwen3.5-35B-A3B-8bit (original, MLX)
- Document Processing & OCR https://www.docling.ai
- Jina Models:
  - jina-embeddings-v5-text-small (original, MLX)
  - jina-reranker-v3 (original, MLX)
- https://docs.searxng.org
- https://brave.com/search/api/

Q & A

Thank you!

Personal blog: rdner.de