IOWA CODE CAMP

Your Own Private AI

Aaron Grothe

www.grothe.us

November 7, 2026

Introduction



Private Setup

Run local AI environments.



Data Security

Zero exfiltration risk.



Custom Data

Extend models safely.

Housekeeping

- Ask questions anytime.
- Share relevant resources.
- Interactive session.
- Slides available now.

https://www.grothe.us

Audience Poll

1. Using ”free” LLM tiers?
2. Read the Terms of Service?
3. Paying for AI tools?
4. Tested Local/Private AI?

Gemini ”Free” Tier ToS

Data Privacy

- ✓ Human reviewers monitor prompts.
- ✓ Used for model training.
- ✓ Retained indefinitely decoupled.

Output Ownership

- ✓ Google claims no ownership.
- ✓ Outputs generated probabilistically.
- ✓ Identical outputs possible.

DISCLAIMER: NOT A LAWYER

ToS Summary

Google retains the right to use your data indefinitely to improve their models.

Most other commercial LLM terms are similar.

Hardware Requirements

Core Needs

Multi-core CPU, dedicated GPU, or modern NPU.

Demo System

System 76 Pangolin
AMD 7840U / Radeon 780M

TESTED RIGS

- HP Z620 (4x Nvidia K2200)
- Macbook Air M1
- GMKtec (Intel 285H/Arc B390)
- Tiiny.ai Custom Machine
- System 76 Pangolin

Ideal Candidates

Apple Silicon

Unified memory architecture.

Modern APUs

Shared CPU/GPU memory.

Retired Data Center

Cheap tensor accelerators.

Older Hardware

Run smaller models locally.

Sizing Your LLM

llmfit

Analyze system capability instantly.

www.llmfit.org

```
$ llmfit system


 # Outputs hardware breakdown
```

Model Recommendations

```
$ llmfit recommend --user-case coding --limit 5
```

Generates targeted top 5 models tailored for your hardware and use case.

Terminal UI

Interactive Interface

```
$ llmfit
```

Execution Agents

Ollama
Classic CLI standard.

LM Studio
Popular desktop GUI.

Open WebUI
Browser interface.

llama.cpp
Compile it yourself.

LLMserver
Rust-based engine.

AGENT CHOICE

LM Studio

It is easy and it works.

LIVE DEMONSTRATION

LM Studio in Action

Preloaded Models

Dynamic CPU/GPU Sharing

Benchmark Questions

Who was David Bowie?

What are Sea Monkeys?

Closest star to Earth?

Current Mac Mini CPU?

Model Licensing

1

Muse Glimmer

2

Meta Llama 3

3

Research Licenses

LICENSE DEEP DIVE

Muse Glimmer

Apache 2.0 License

Few restrictions. Use it however you want.

Meta Llama 3 Community



Scale Limit

Capped at 700M users.



Competition

Cannot train competing models.



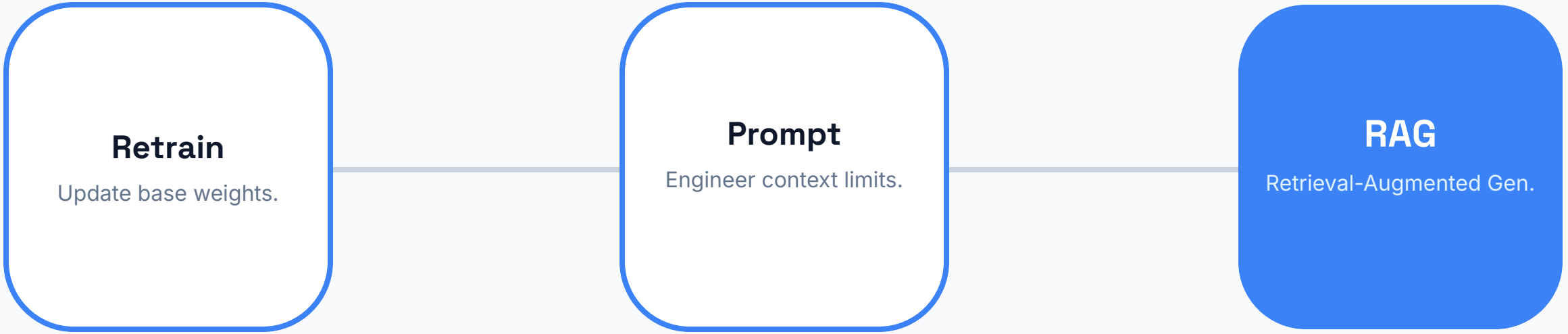
Trademarks

Strict naming restrictions.

Other Licenses

- Strict deployment restrictions.
- Limited to R&D purposes.
- Encourages innovation over competition.
- Review terms carefully.

Injecting Custom Data



Future Hardware Trends

Custom Devices

Tiiny.ai Architecture

- 12 ARM cores (v9.2)
- 80GB RAM
- 190 TFLOPS output

External GPUs

Modular Upgrades

- Separate compute unit.
- Independent system upgrades.

CHEAT GUIDE

The llamafile Paradigm

Combined Binary

LLM plus loader unified.

Single File

No complex dependencies.

Cross-Platform

Mac, Linux, Windows, BSD.

Windows Limit

Capped at 4GB models.

LIVE DEMONSTRATION

llamafile execution

LIVE DEMONSTRATION

Tiiny.ai Test

(If it cooperates)

Getting Started (1/2)

llamafile Entry

Easiest path to Private AI.

Open WebUI

Commercial-grade UX setup.

Mac Hardware

Used Mac Minis dominate.

Stay Flexible

Avoid tool attachment.

Getting Started (2/2)



Hallucinations

Never blindly trust outputs.



Thermals

Throttling can be an issue.



Power Issues

Laptop PSU driving system + GPU.



API Endpoints

OpenAI compatible by default.

Is Private AI Worth It?

Yes.

I use local models 80% of the time.

(Despite having a paid Gemini subscription)

Questions?

Enjoy the rest of the conference.

Slides available at

https://grothe.us

Reference Links

llmfit
github.com/AlexsJones/llmfit

Ollama
ollama.com

Open WebUI
openwebui.com

LM Studio
lmstudio.ai

llamafile
github.com/mozilla-ai/llamafile

llamacpp
github.com/ggml-org/llama.cpp

llmserve
github.com/AlexsJones/llmserve