Red Hat Developer

t3chfest

Everything You Need to Know About **Running LLMs Locally**

T3chfest Madrid: 2025

Cedric Clyburn
Senior Developer Advocate
@cedricclyburn

Roberto Carratalá
Principal AI Architect
@rcarrata_

We’ve got **a lot** to cover today!

We've got **a lot** to cover today!

Wait, so you can run your own language models... completely local?

We've got **a lot** to cover today!

Wait, so you can run your own language models... completely local?

and there are plenty of open source tools to do so?!

vLLM

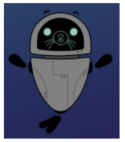
ramalama

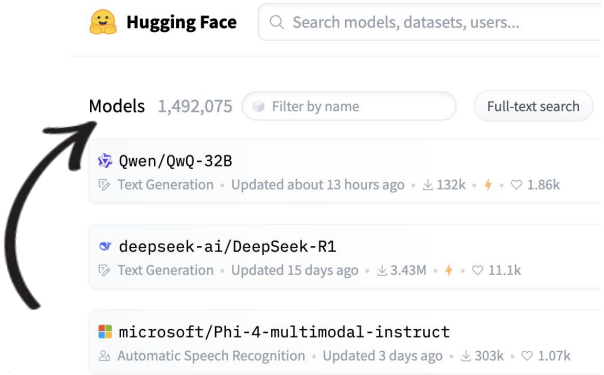
We've got **a lot** to cover today!

Wait, so you can run your own language models... completely local?

and there are plenty of open source tools to do so?!

vLLM ramalama





But there's over 1.5 million models, which to pick?

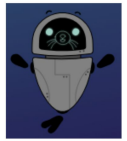
We've got **a lot** to cover today!

Wait, so you can run your own language models... completely local?

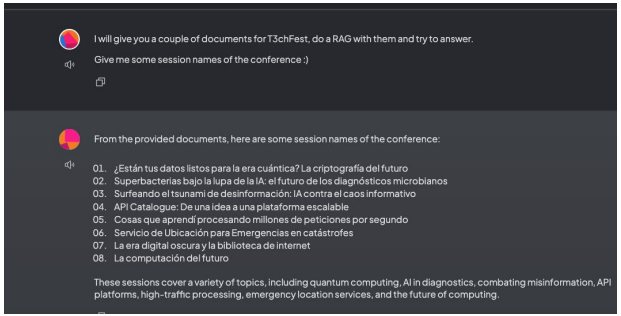
and there are plenty of open source tools to do so?!

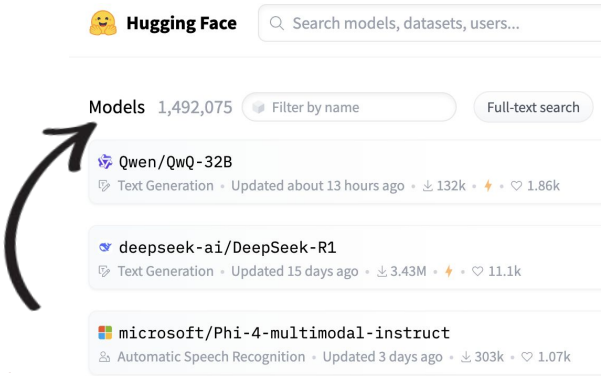
vLLM

ramalama



or how can i use my own PDF's or codebase or API's?





But there's over 1.5 million models, which to pick?

Today's Schedule

- Running **your own AI & LLMs**
- How to **choose the right model?**
- Integrating **your data & codebase!**

- **Demo #1:** Model serving & RAG
- **Demo #2:** Code assistance & agents
- **Demo #3:** Adding AI features to apps

Session Slides

red.ht/local-llm

Why’s everyone running their **own** AI models?

Why run a model locally?

Take advantage of total AI customization and control

For Developers

Convenience & Simplicity

Familiarity with the Development Environment and adherence of the developers to their "local developer experience" in particular for testing and debugging

Ease of Integration

Simplify the integration of the model with existing systems and applications that are already running locally.

Customization & Control

Easily train or fine-tune your own model, from the convenience of the developer's local machine.

Direct Access to Hardware

For Organizations

Data Privacy and Security

Data is the fuel for AI, and a differentiator factor (quality, quantity, qualification). Keeping data on-premises ensures sensitive information doesn't leave the local environment → crucial for privacy-sensitive applications

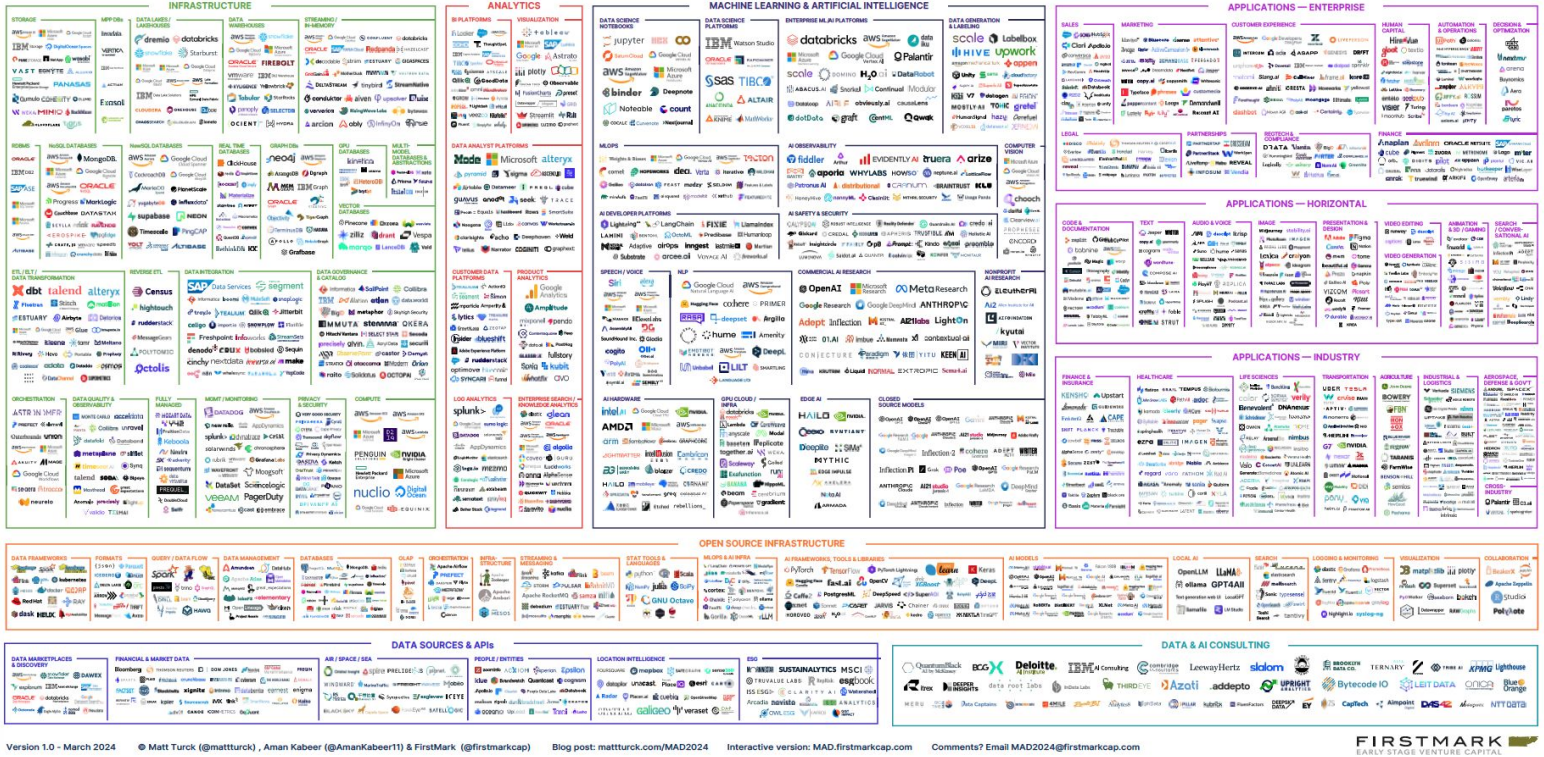
Cost Control

While there is an initial investment in hardware and setup, running locally can potentially reduce ongoing costs of cloud computing services and alleviate the vendor-locking played by Amazon, MSFT, Google

Regulatory Compliance

Some industries have strict regulations about where and how data is processed

But the stack can be **a bit** overwhelming!

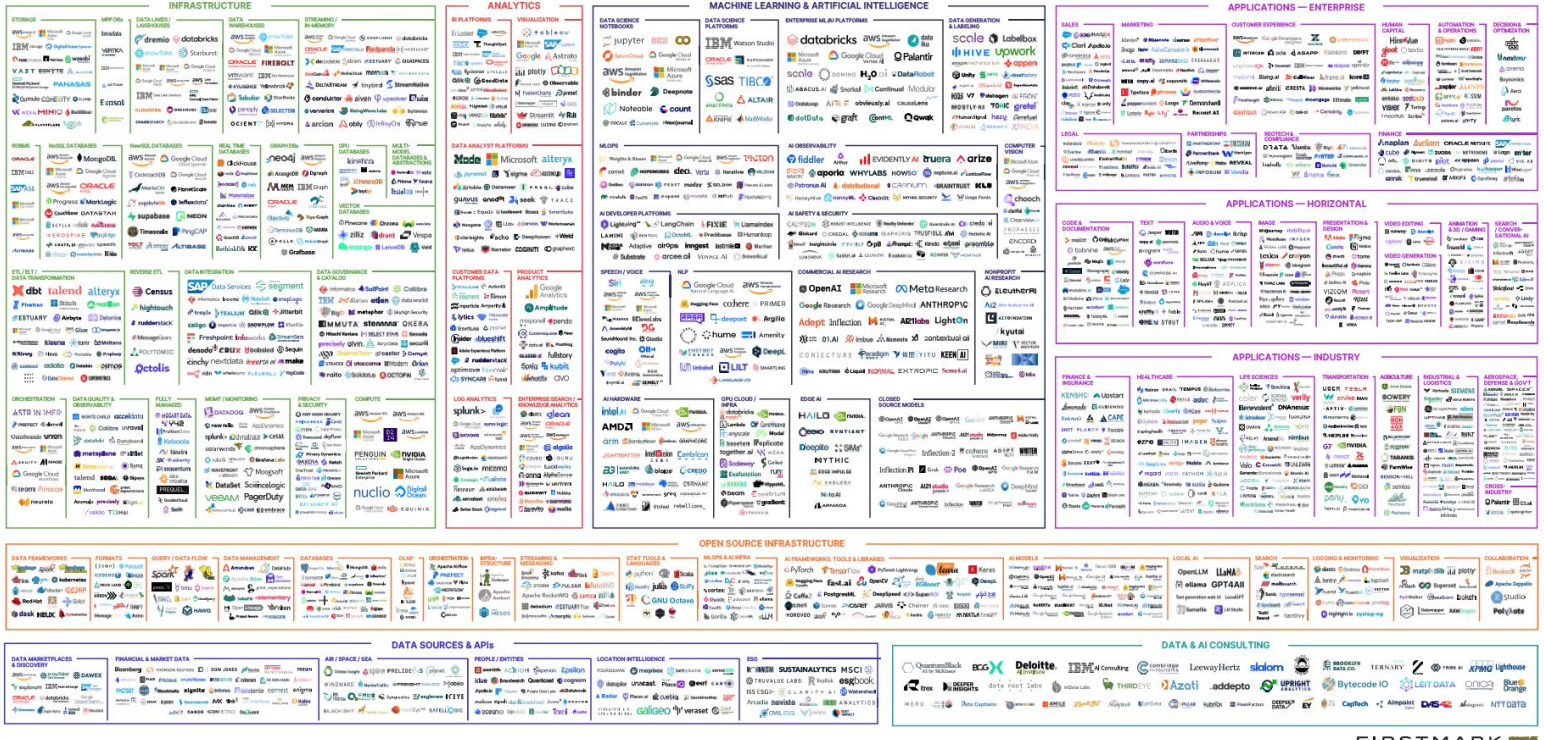


Version 1.0 - March 2024 Blog post: mattturck.com/MAD2024 Interactive version: MAD.firstmarkcap.com Comments? Email MAD2024@firstmarkcap.com

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

But the stack can be **a bit** overwhelming!

2024 MAD (Machine learning, Artificial Intelligence & Data) Landscape

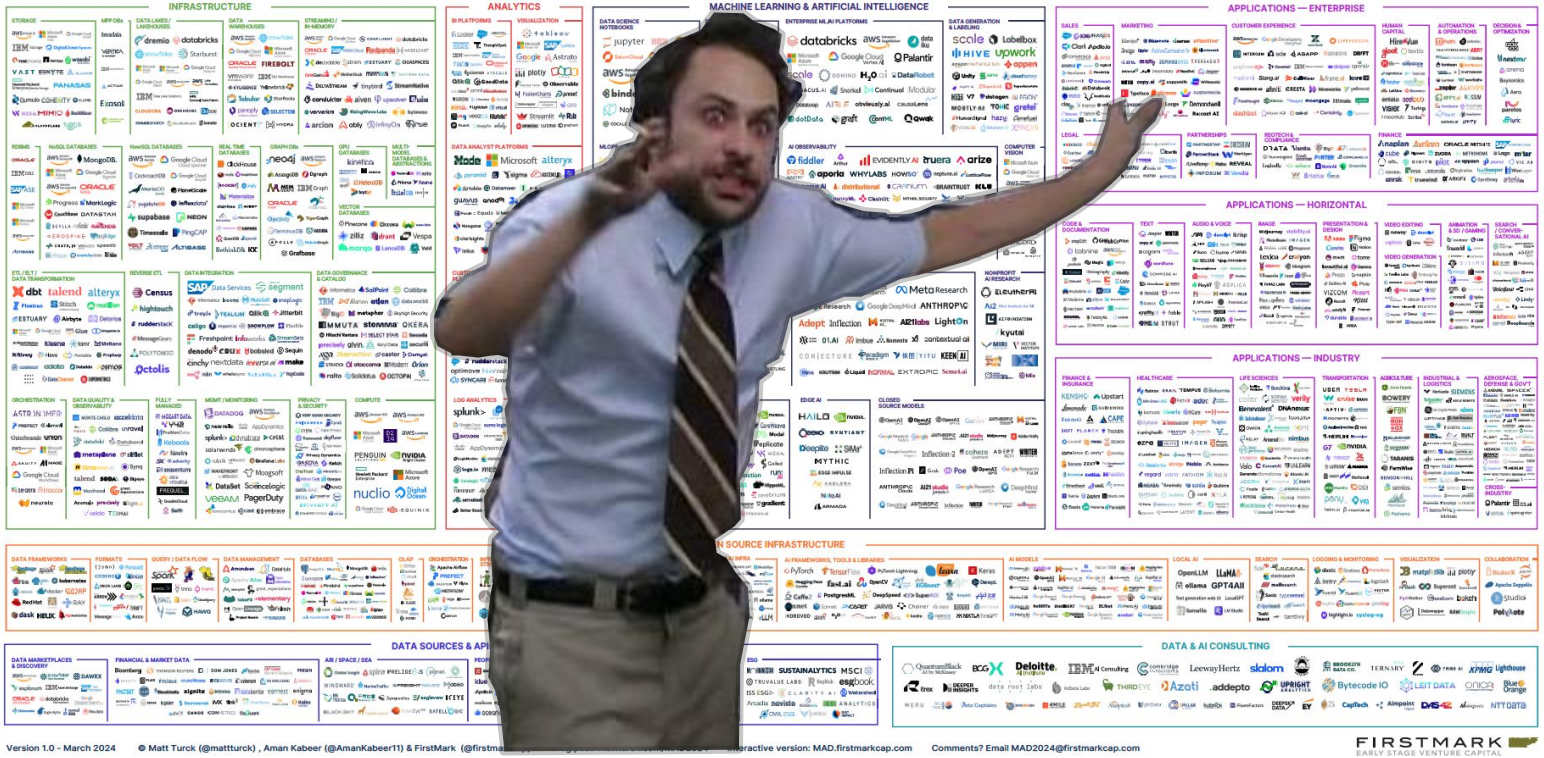


Version 1.0 - March 2024 Blog post: mattturck.com/MAD2024 Interactive version: MAD.firstmarkcap.com Comments? Email MAD2024@firstmarkcap.com

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

But the stack can be **a bit** overwhelming!

2024 MAD (Machine learning, Artificial Intelligence & Data) Landscape



Average developer trying to download & manage models, configure serving runtimes, quantize and compress LLMs, ensure correct prompt templates... (Colorized, 2023)

So, what **open source** tech can help us run AI?

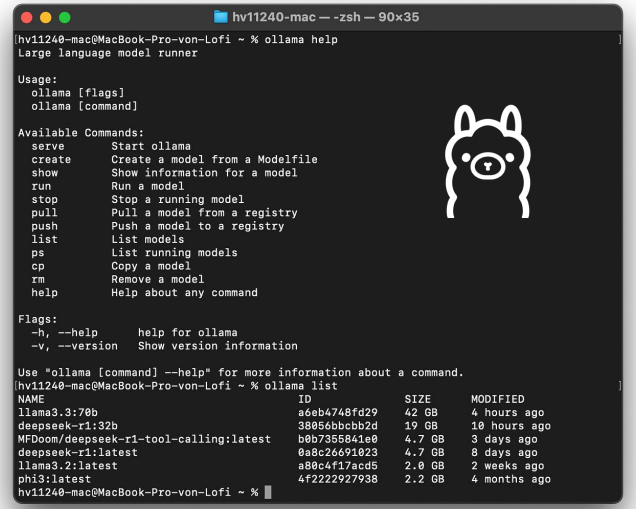
Fortunately, there's a lot... for **every use case**!

Fortunately, there's a lot... for **every use case**!

Tool #1: Ollama

For simple model downloading & serving

- **Simple CLI:** “Docker” style tool for running LLMs locally, offline, and privately
- **Extensible**: Basic model customization (Modelfile) and importing of fine-tuned LLMs
- **Lightweight**: Efficient and resource-friendly.
- **Easy API**: API for both inferencing and Ollama itself (ex. download models)

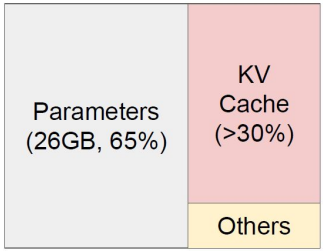


Tool #2: vLLM

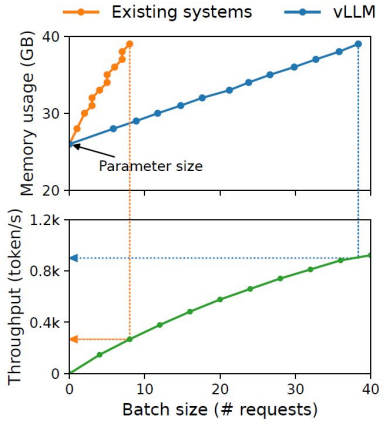
For scaling things up in production environments

- **Research-Based:** UC Berkeley project to improve model speeds and GPU consumption
- **Standardized:** Works with Hugging Face & OpenAI API.
- **Versatile:** Supports NVIDIA, AMD, Intel, TPUs & more.
- **Scalable:** Manages multiple requests efficiently, ex. with Kubernetes as an LLM runtime

vLLM



NVIDIA A100 40GB



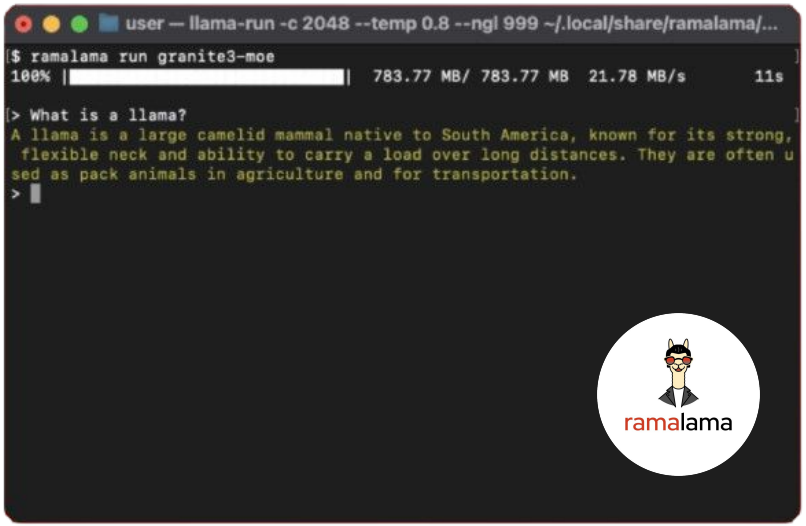
Red Hat Developer

Red Hat

Tool #3: Ramalama

To make AI boring by using containers

- **AI in Containers:** Run models with Podman/Docker with no config needed.
- **Registry Agnostic:** Freedom to pull models from Hugging Face, Ollama, or OCI registries.
- **GPU Optimized:** Auto-detect & accelerate performance.
- **Flexible:** Supports llama.cpp, vLLM, whisper.cpp & more.



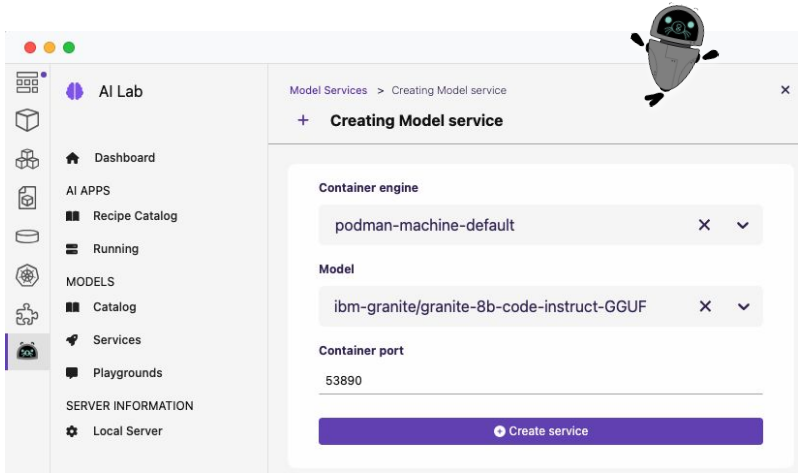
Red Hat Developer

Red Hat

Tool #4: Podman AI Lab

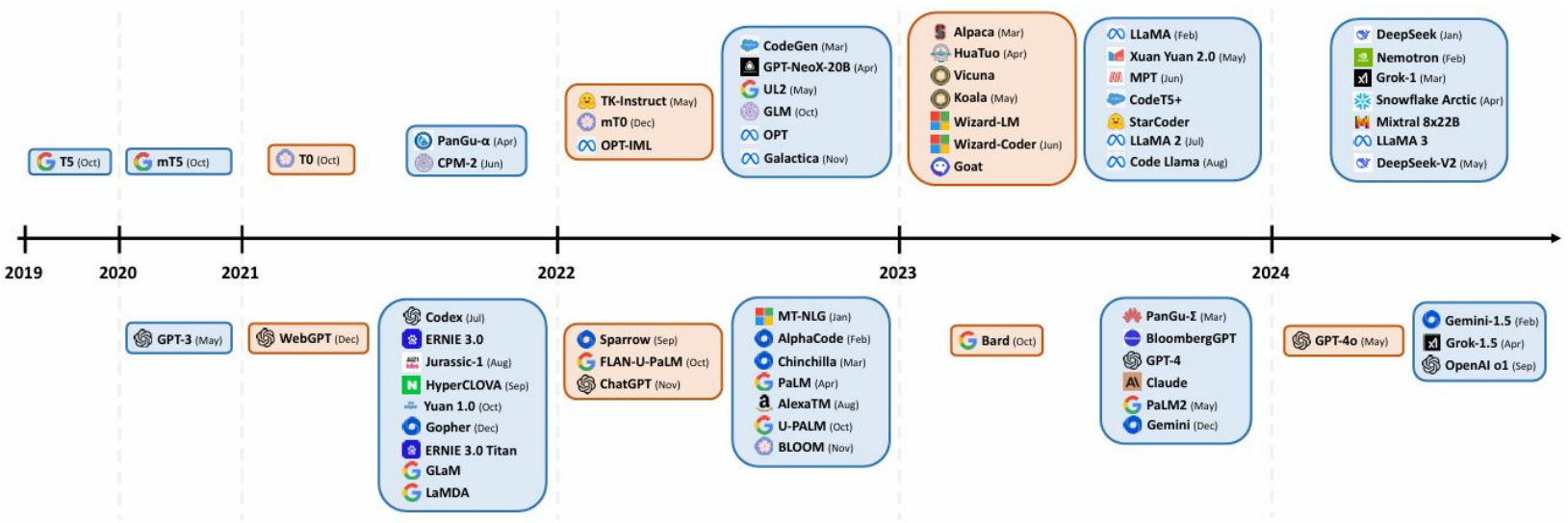
For developers looking to build AI features

- **For App Builders:** Choose from various recipes like RAG, Agentic, Summarizers
- **Curated Models:** Easily access Apache 2.0 open-source options.
- **Container Native:** Easy app integration and movement from local to production.
- **Interactive Playgrounds:** Test & optimize models with your custom prompts and data.



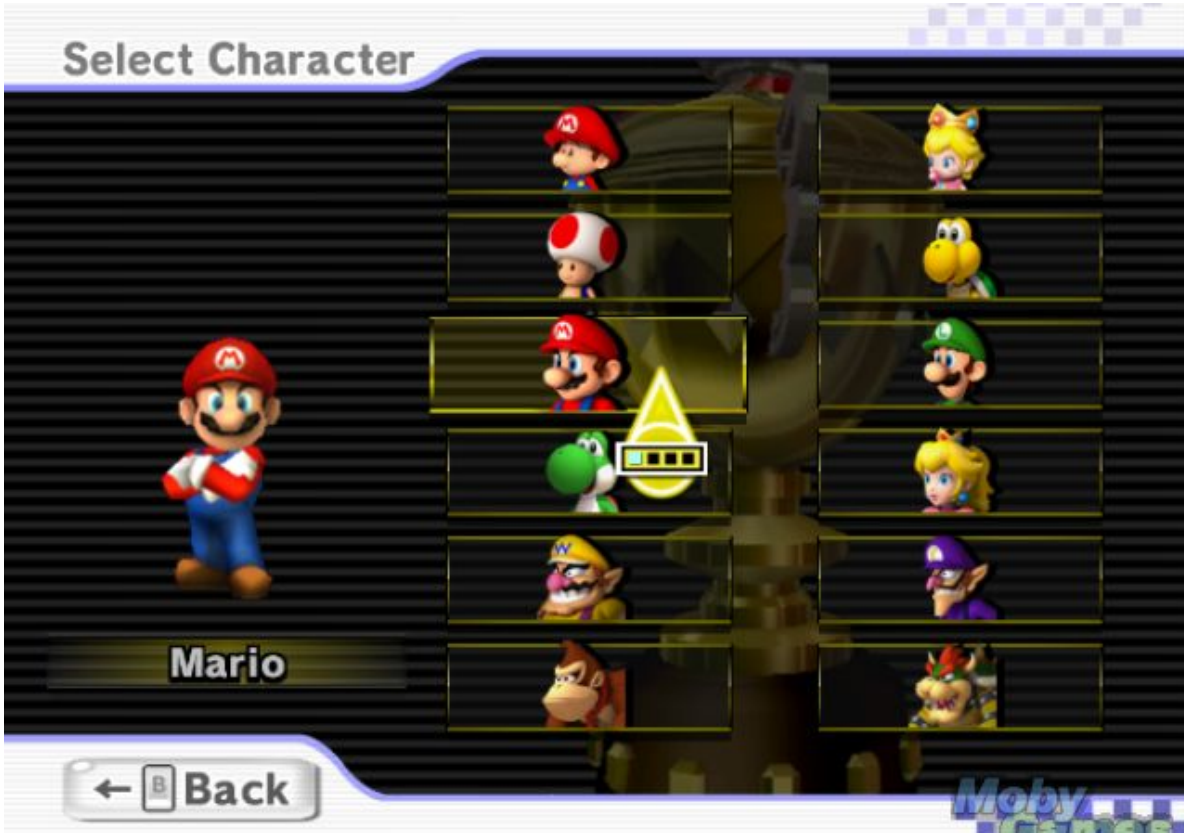
Cool! But, what specific **model** should I be using?

There are **plenty** of open, and closed model choices!



Darn, we’re back to here again!

But again, we'll use another video game analogy!

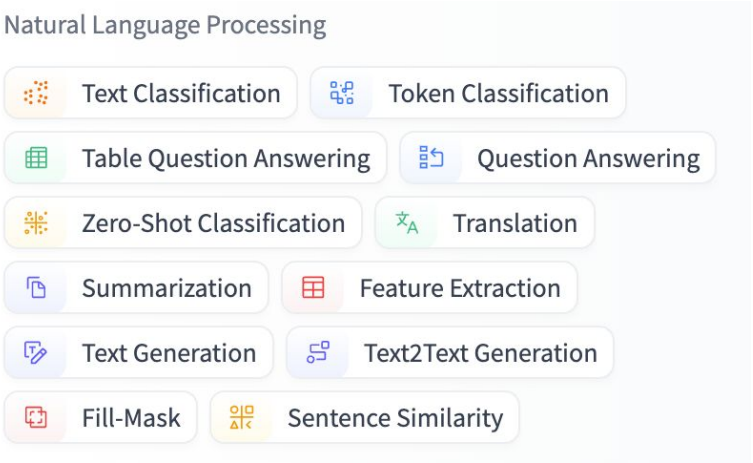


But again, we'll use another video game analogy!

So, which model should you select?

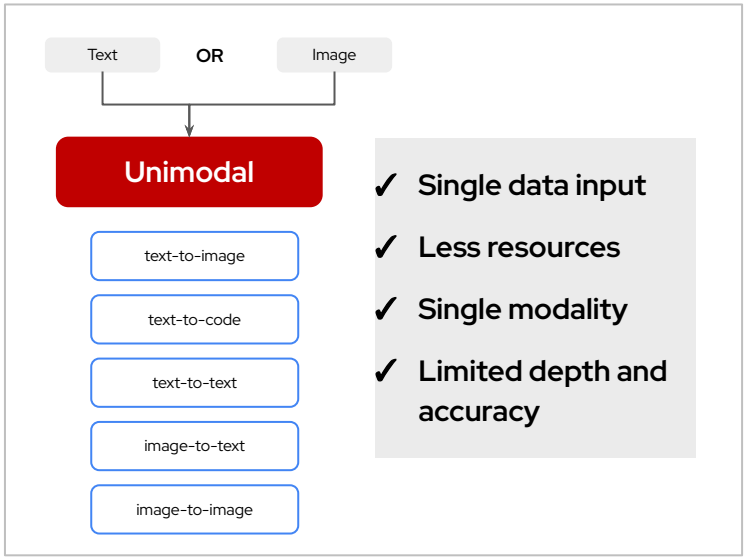
Well... it depends!

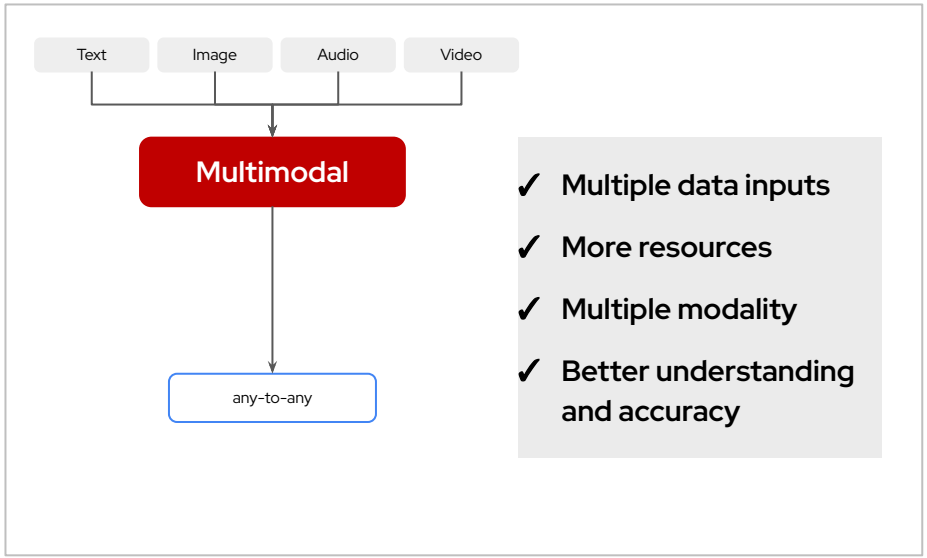
- It depends on **the use case** that you want to tackle.
- **DeepSeek** models excel in reasoning tasks and complex problem-solving.
- **Granite** SLM models perform well in various NLP tasks and multimodal applications.
- **Mistral** and **LLaMA** are particularly strong in summarization and sentiment analysis.



But not all models are the same!

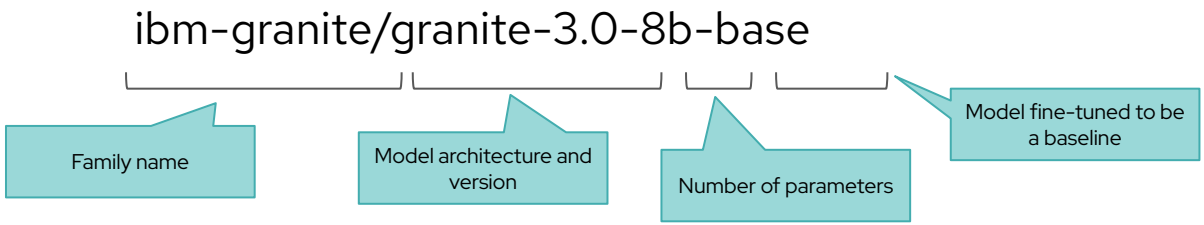
Our data isn't always in one format, it's text, image, audio, etc.

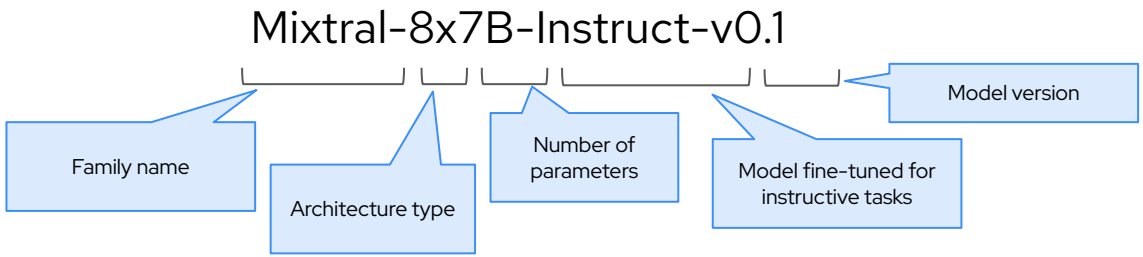




Also! There's a naming convention

Kind of like how our apps are compiled for various architectures!

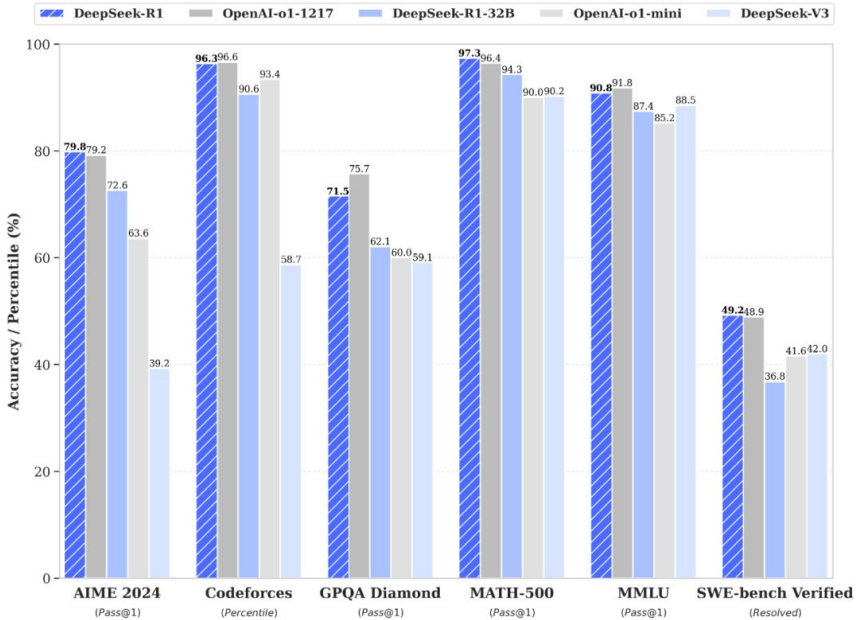




What about **model size**?

How to deploy a larger model?

For example, DeepSeek-R1 or Llama 3.3-405B



Let's say you want the best benchmarks with a frontier model

How to deploy a larger model?

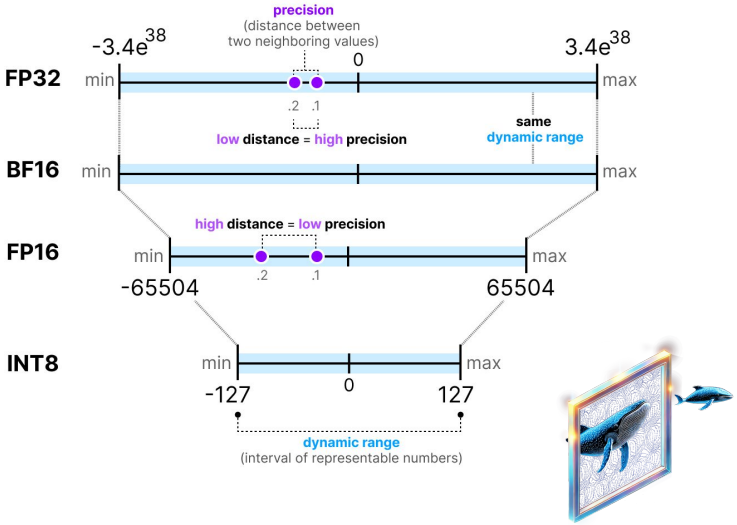
For example, DeepSeek-R1 or Llama 3.3-405B

Neither of these situations is ideal :)

Well, most models for local usage are quantized!

It's a way to compress models, think like a .zip or .tar

- **Quantization:** A technique to compress LLMs by reducing numerical precision.
- Converts high-precision weights (FP32) into lower-bit formats (FP16, INT8, INT4).
- **Reduces memory footprint,** making models easier to deploy.



Red Hat Developer

Red Hat

Well, most models for local usage are quantized!

It's a way to compress models, think like a .zip or .tar

- **The Benefit?** Run LLMs on "any" device, not just your local machine but IoT & Edge too
- Results in **faster and lighter models** that still maintain reasonable accuracy
  - Testing with Llama 3.1, for W4A16-INT resulted in **2.4x performance speedup** and **3.5x model size compression**
- Works on **GPUs & CPUs!**

Red Hat Developer

Source: https://neuralmagic.com/blog/we-ran-over-half-a-million-evaluations-on-quantized-llms-heres-what-we-found

Red Hat

& there's a open repository of Quantized Models

Check it out on Hugging Face, & save resources on LLM serving!

Broad Collection

Llama
Qwen
Gemma
Mistral
DeepSeek
Phi
Molmo
Granite
Nemotron

Comprehensive Validation

Open LLM Leaderboard evaluation scores

Benchmark	Meta-Llama-3.1-70B-Instruct	Meta-Llama-3.1-70B-Instruct-FP8(this model)	Recovery
MMLU (5-shot)	83.83	83.73	99.88%
MMLU-cot (0-shot)	86.01	85.44	99.34%
ARC Challenge (0-shot)	93.26	92.92	99.64%
GSM-8K-cot (8-shot, strict-match)	94.92	94.54	99.60%
Hellaswag (10-shot)	86.75	86.64	99.87%
Winogrande (5-shot)	85.32	85.95	100.7%
TruthfulQA (0-shot, mc2)	60.68	60.84	100.2%
Average	**84.40**	**84.29**	**99.88%**

Extensive Selection

Formats
- W4/8A16
- W8A8-INT8
- W8A8-FP8
- 2:4 sparse

Algorithms
- GPTQ / AWQ
- SmoothQuant
- SparseGPT
- RTN

Hardware

NVIDIA GPUs
Google TPUs
AMD Instinct
intel CPUs

Cut GPU costs in half ready-to-deploy inference-optimized checkpoints

Red Hat Developer

Red Hat

AI Engine? **Check** ✔

AI Model? **Check** ✔

What about your data?

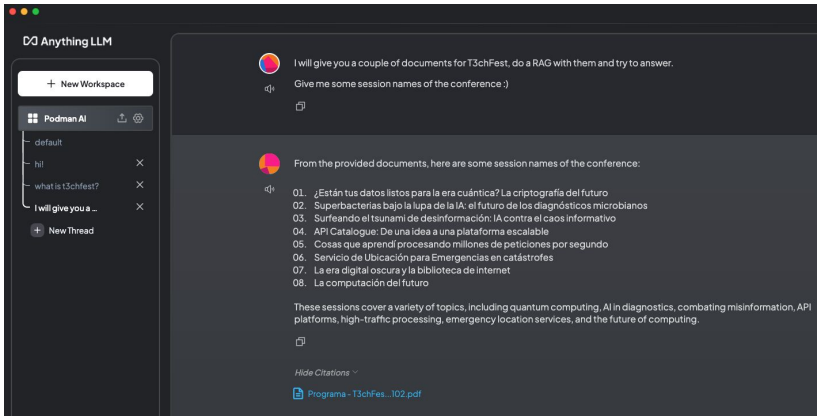
How can you integrate AI with your unique data?

Fortunately, many tools exist for this too!

Data Interfaces

Pull in documents (PDF), web results, and agents together.

Tools: AnythingLLM, OpenWebUI, LM Studio



Ask a question to a PDF & receive citations!

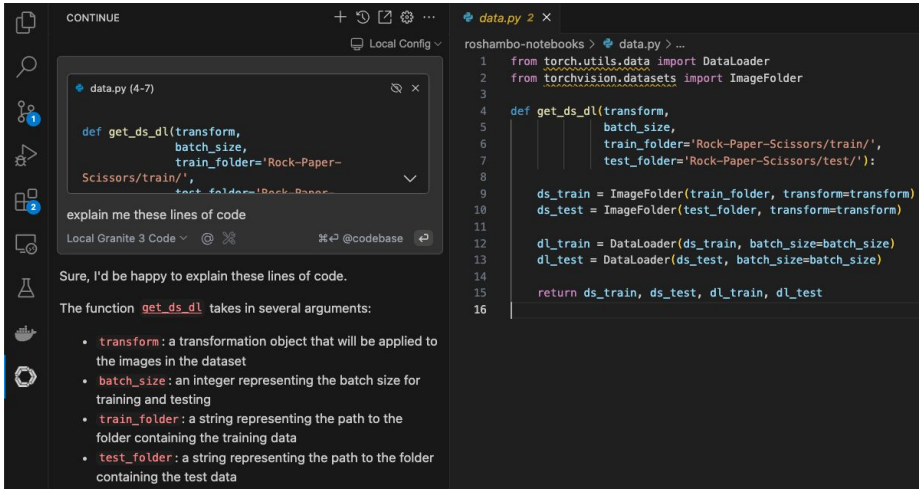
How can you integrate AI with your unique data?

Fortunately, many tools exist for this too!

Code Assistance

Use a model as a pair programmer, to generate and explain your codebase.

Tools: Continue, Cody, Cursor, Windsurf



No more copy/pasting, it's part of the IDE!

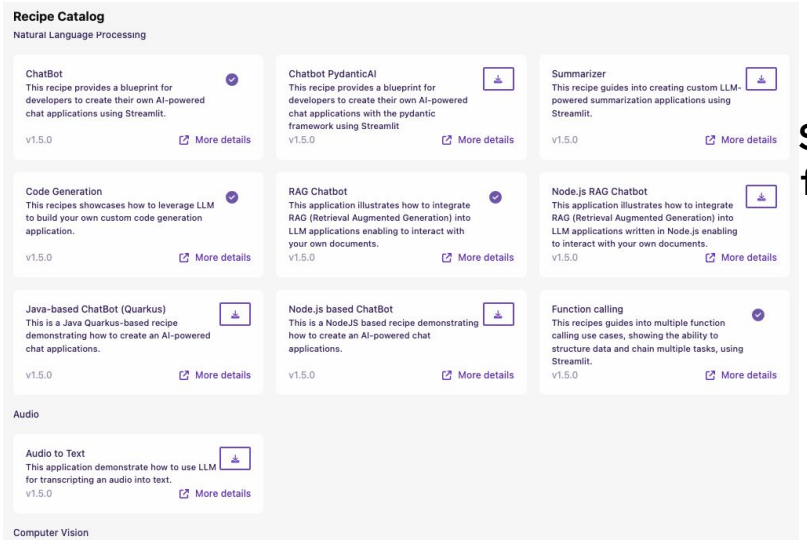
How can you integrate AI with your unique data?

Fortunately, many tools exist for this too!

Prompting & Building Apps

Experiment with data, build Proof of Concepts, and integrate AI into apps.

Tools: Podman AI Lab, Docker Gen AI Stack



Starting points for common AI apps

How can you integrate AI with your unique data?

Fortunately, many tools exist for this too!

Data Interfaces

Pull in documents (PDF), web results, and agents together.

Tools: AnythingLLM, OpenWebUI, LM Studio

Code Assistance

Use a model as a pair programmer, to generate and explain your codebase.

Tools: Continue, Cody, Cursor, Windsurf

Prompting & Building Apps

Experiment with data, build Proof of Concepts, and integrate AI into apps.

Tools: Podman AI Lab, Docker Gen AI Stack

Demo Time!

Model serving & RAG

Code Repository URL
red.ht/t3chfest-demo

Red Hat Developer

Demo Repository:
https://github.com/rh-aiservices-bu/t3chfest-demo

Red Hat

Demo Time!

Code assistants & agents

Code Repository URL
red.ht/t3chfest-demo

Red Hat Developer

Demo Repository:
https://github.com/rh-aiservices-bu/t3chfest-demo

Red Hat

Demo Time!

Adding AI features to apps

Code Repository URL
red.ht/t3chfest-demo

Red Hat Developer

Demo Repository:
https://github.com/rh-aiservices-bu/t3chfest-demo

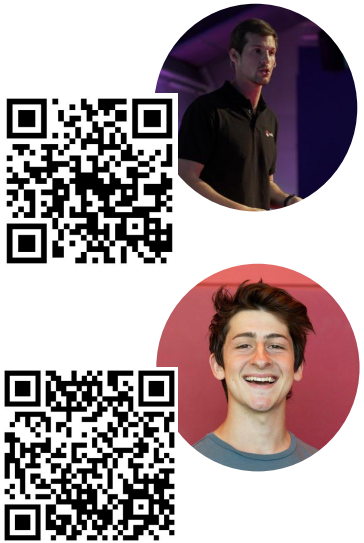
Red Hat

Thank you! You're awesome!

- Running **your own AI & LLMs**
- How to **choose the right model?**
- Integrating **your data & codebase!**

Session Slides

red.ht/local-llm



Connect on LinkedIn!

Red Hat Developer

Red Hat

Thank you

Join the DevNation

Red Hat Developer serves the builders. The problem solvers who create careers with code. Let's keep in touch!

- Join Red Hat Developer at **developers.redhat.com/register**
- Follow us on any of our social channels
- Visit **dn.dev/upcoming** for a schedule of our upcoming events

Red Hat Developer

Build here. Go anywhere.


linkedin.com/company/red-hat


youtube.com/user/RedHatVideos


facebook.com/redhatinc


twitter.com/RedHat

Red Hat