Keeping Your Intelligence Local (and Cheap)

by Nicholas Coats

% whoami

Name: Nicholas (Nick) Coats

Career: Sr. Consultant @ Nationwide

Specialties: Cloud Computing, Software Development and Architecture, DevSecOps

Hobbies: Science Fiction, Non-Fiction Science, Music Making, Social Gathering, History, Politics, Video Games, AI

Relevant Experience: Not much, honestly

All Presentation Materials on Github

https://github.com/coatsnmore/keeping-your-intelligence-local

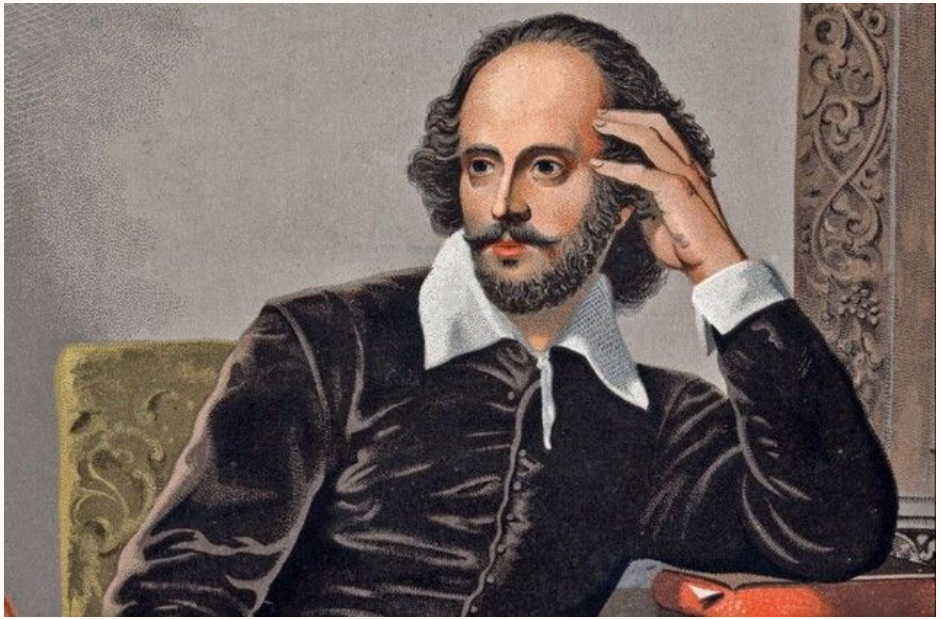
A Few Quick Definitions

- Gen AI: Generative Artificial Intelligence
- AGI: Artificial General Intelligence (roughly human)
- ASI: Artificial Super Intelligence (better than human)
- LLM: Large Language Model
- Token: Roughly a “word” but not limited to
- Embeddings: Numerical representations of text
- Context: Information surrounding the token that produces meaning
- Agent: AI system with perception; often characterized by independent action
- CAG: Context-Augmented Generation
- RAG: CAG with remote data
- Modalities: The type of data the model processes (text|vision|audio)
- Quantization: Reducing precision on model weights to create smaller models
- LoRA: Low-rank Adaptation - fine-tuning “low rank” matrices for better performance
- Model weights: matrices of number representing the relationship between tokens
- CoT (Chain of Thought): coercing the LLM to think “step by step”

Generative AI

“To be or” ~=> “not to be”

(inference)

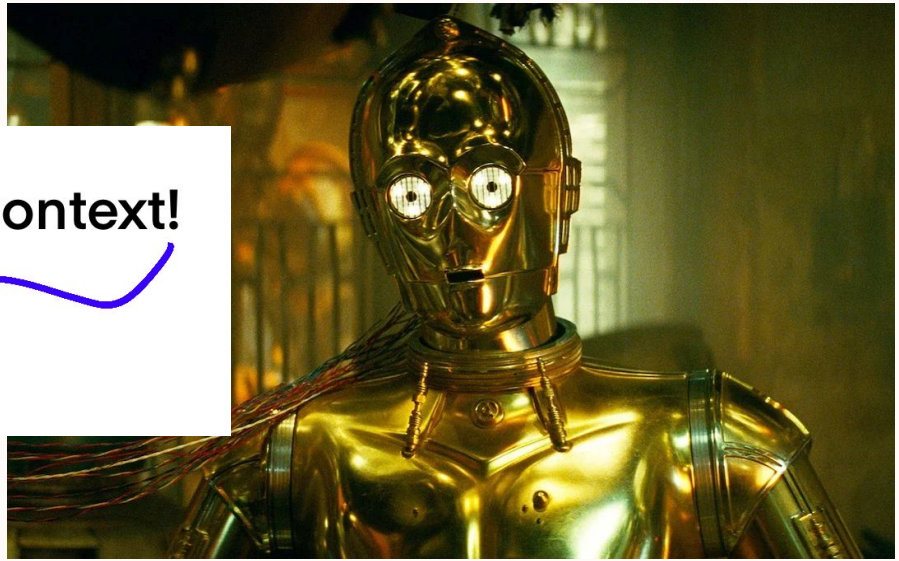


“Generative AI, a mere dispenser of tokens it be; ne'er shall it rival the wit and craft of me.”
- Shakesbeer

Token

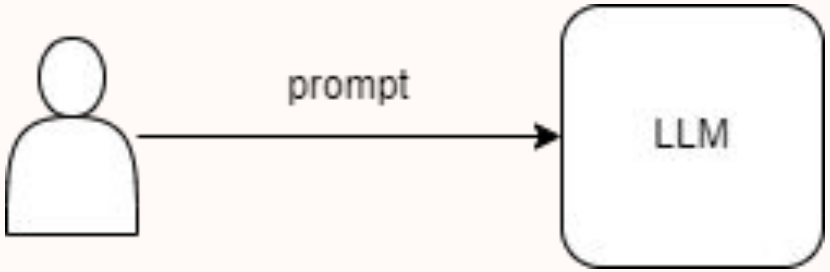
We're *gonna* need a bigger context!

context token context

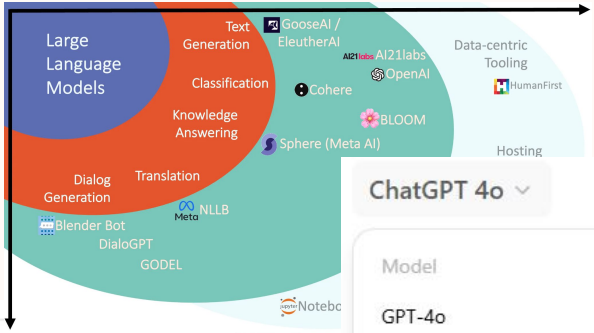


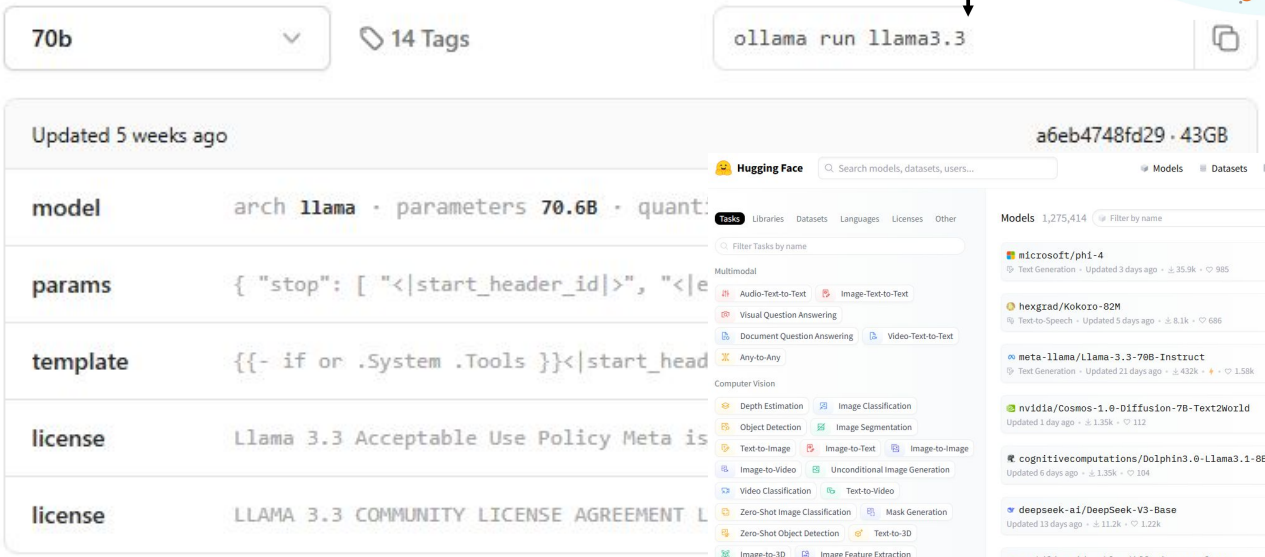
A Basic LLM

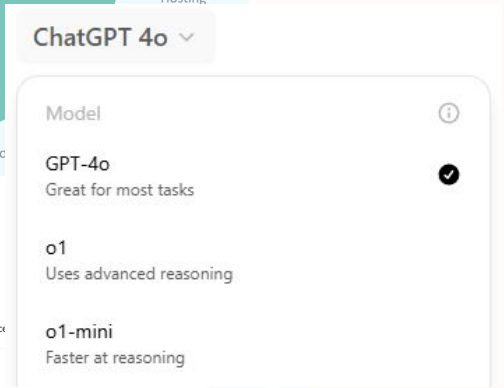
- LLM: Large Language Model

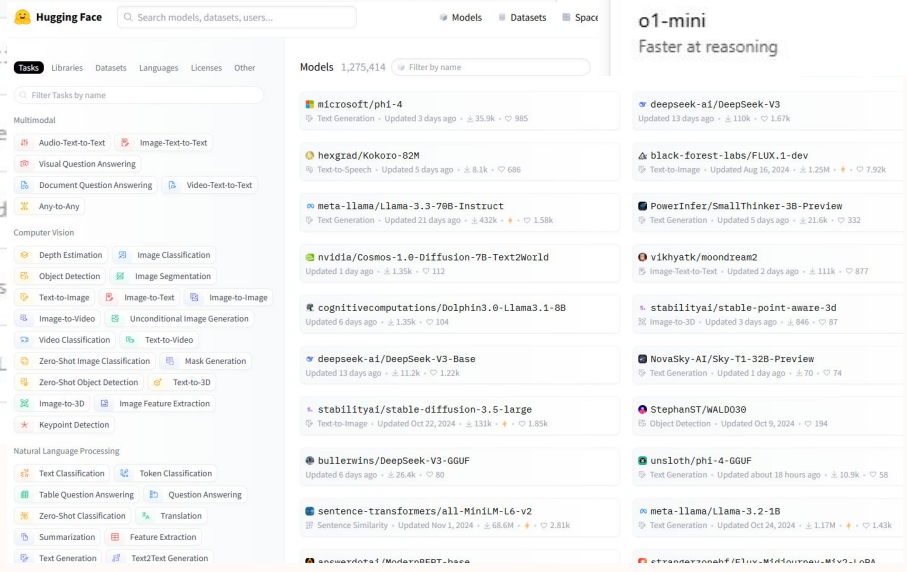


Choosing a Model



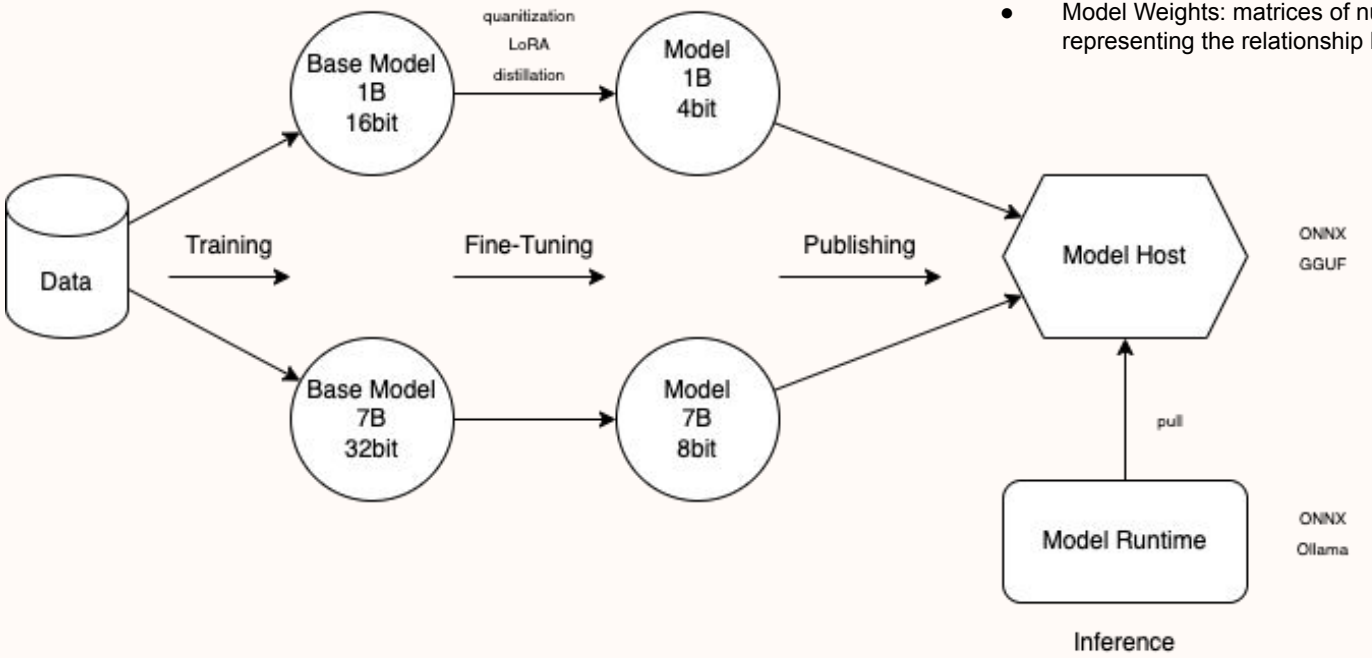






Models at a High Level

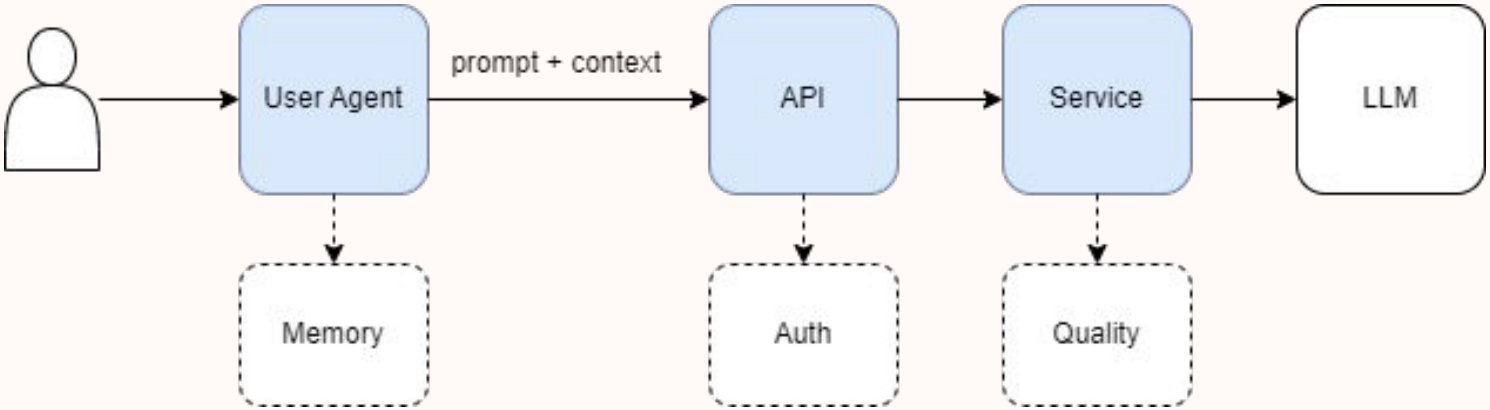
- Modalities: The type of data the model processes (text|vision|audio)
- Quantization: Reducing precision on model weights to create smaller models
- LoRA: Low-rank Adaptation - fine-tuning “low rank” matrices for better performance
- Model Weights: matrices of number representing the relationship between tokens



Hardware Requirements

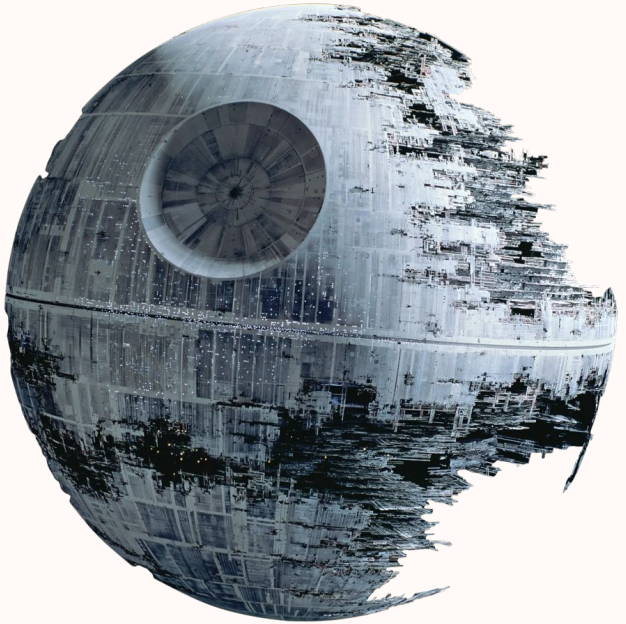
Model	VRAM Requirement (Full Precision)	Quantized VRAM	Compute Recommendations	Notes
OpenAI GPT-4	~350 GB+	N/A	A100/H100 clusters	Accessed via API.
LLaMA 3	14–160 GB	7–40 GB	RTX 4090, A100, H100	Optimized for inference.
TinyLlama	<2 GB	N/A	Consumer GPUs (GTX 1660, RTX 3050)	Lightweight and CPU-compatible.
GPT-NeoX	40 GB	~20 GB	A100, RTX 4090	Large open-source LLM.
Bloom	350 GB	150 GB	A100/H100 clusters	Multilingual support.
Falcon	16–80 GB	8–40 GB	RTX 4090, multi-GPU	Efficient for deployment.
ChatGLM	12 GB	6–8 GB	RTX 3060, CPU for smaller tasks	Optimized for low-resource.
GPT-2	2–4 GB	N/A	Consumer GPUs (GTX 1080, RTX 3050)	Lightweight, smaller scale.

Let’s API Enable that Bad Boy



Risks of Cloud

- Vendor lock in
- Data gravity
- Data privacy (personal, IP, business process)
- Subject to vendor guardrails (censored)
- Variable Cost



Pricing

Model	Pricing	Pricing with Batch API*
gpt-4o	$2.50 / 1M input tokens	$1.25 / 1M input tokens
	$1.25 / 1M cached** input tokens	
	$10.00 / 1M output tokens	$5.00 / 1M output tokens

- https://openai.com/api/pricing/

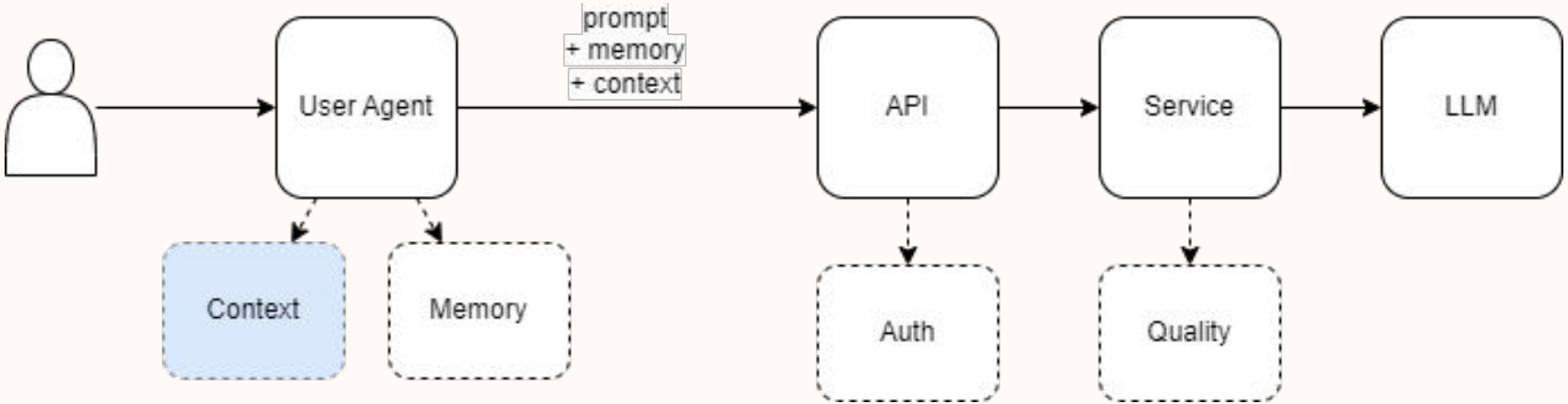
Coffee Numbers

For every 1 Million tokens, it costs about **$3.125**.

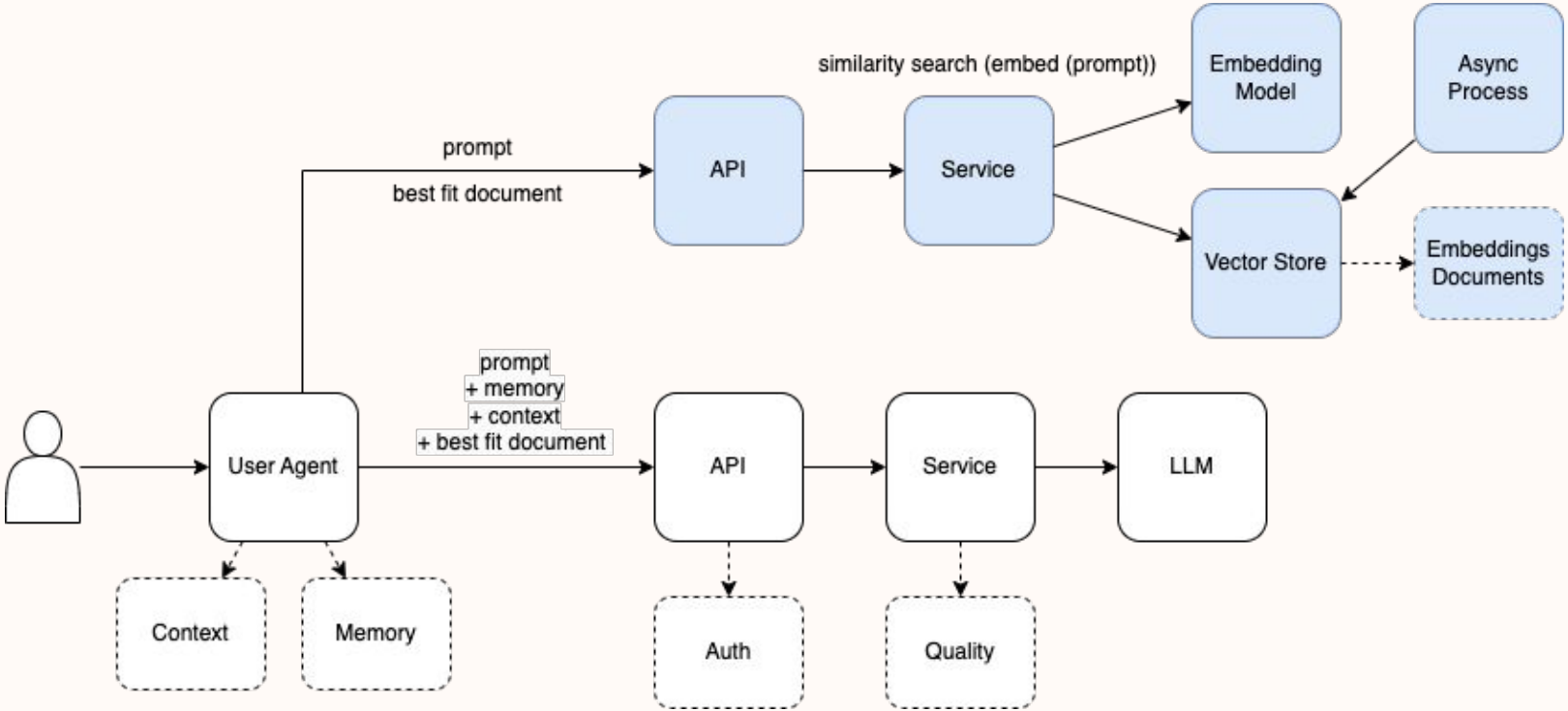
If we ran locally instead, we could eventually buy a coffee.

TTC (Time to Coffee) = How many times it would take to recoup a coffee worth of savings.

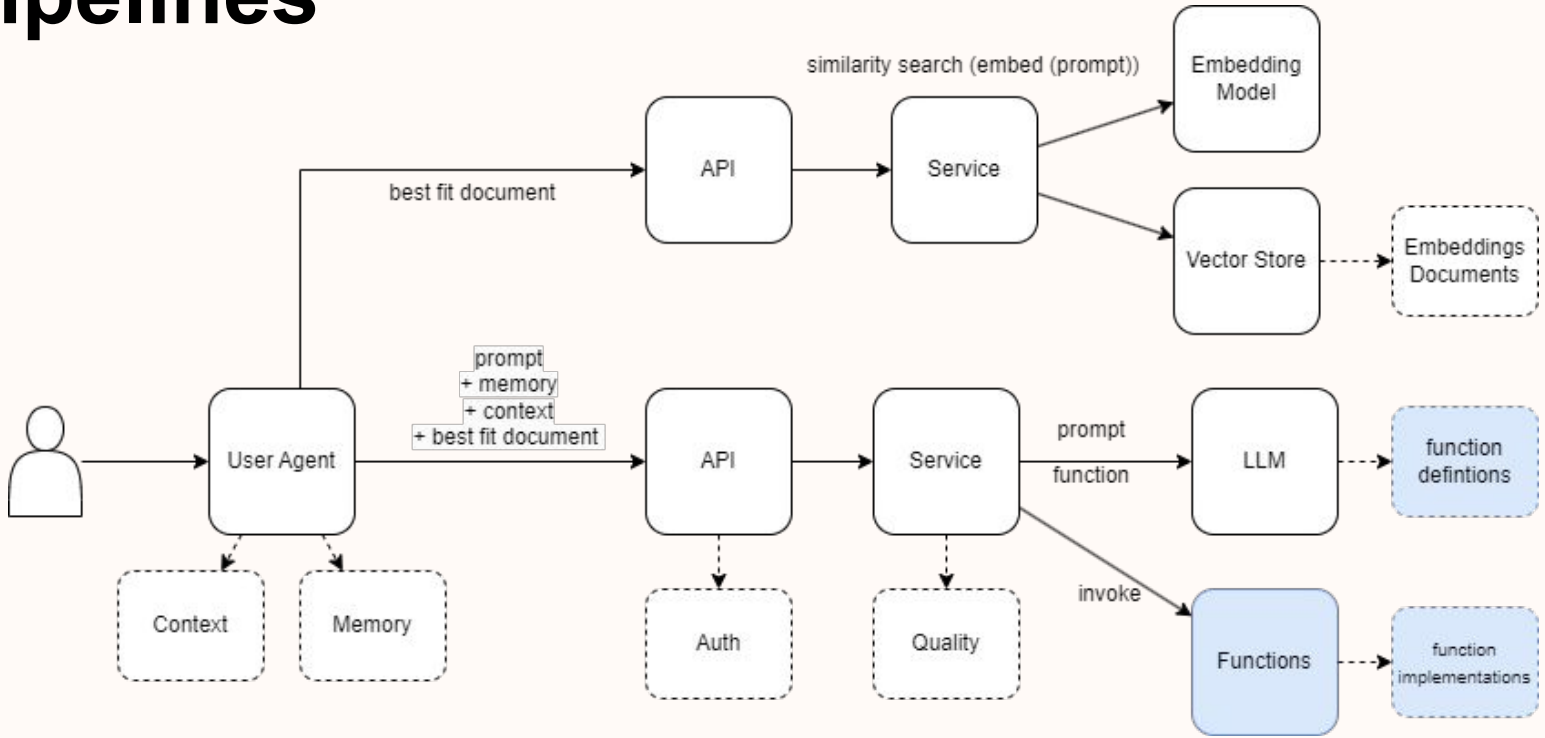
Context Augmented Generation



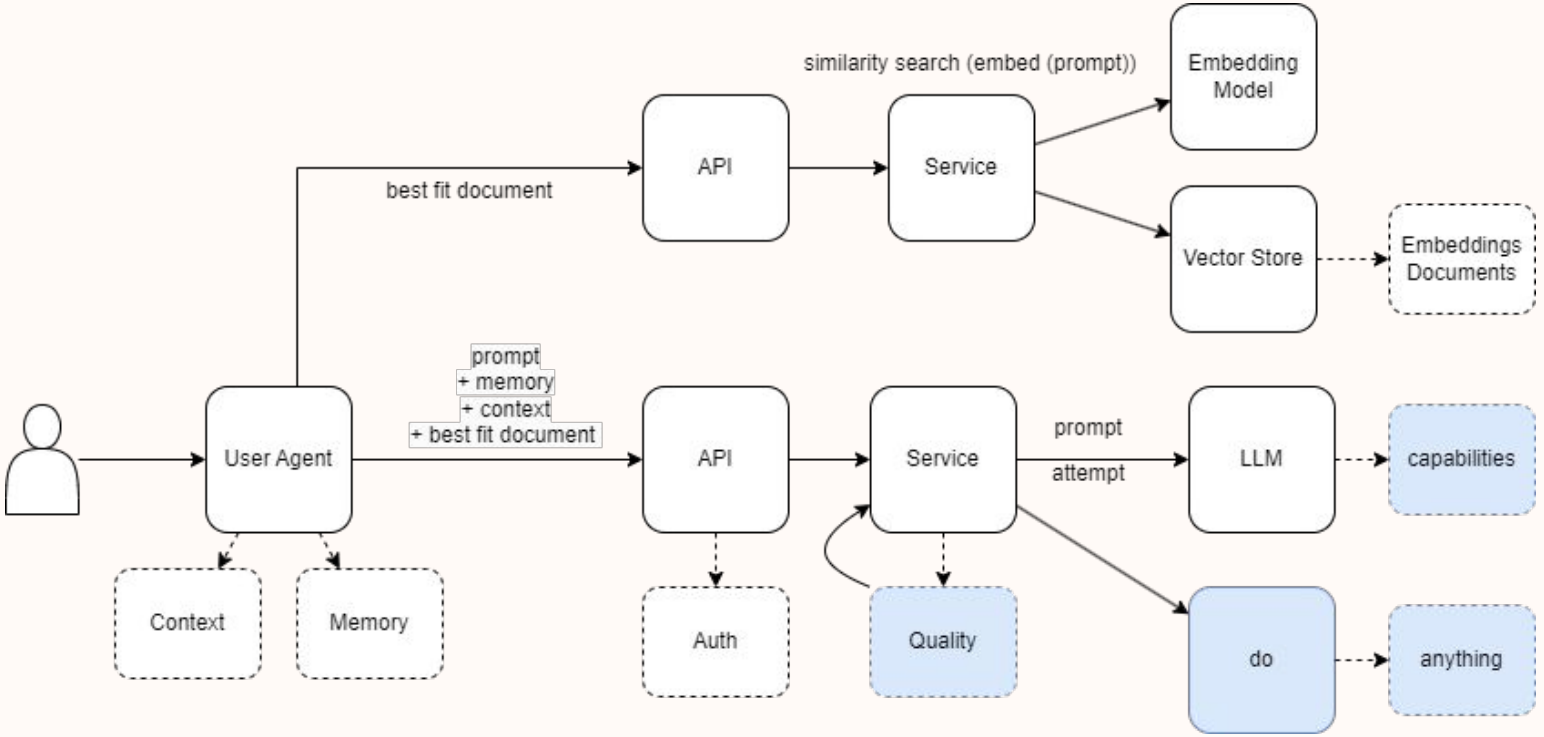
Retrieval Augmented Generation



Agentic Systems - Workflows AKA Pipelines



Agentic Systems - Agents



OSS Frameworks

LangChain

griptape

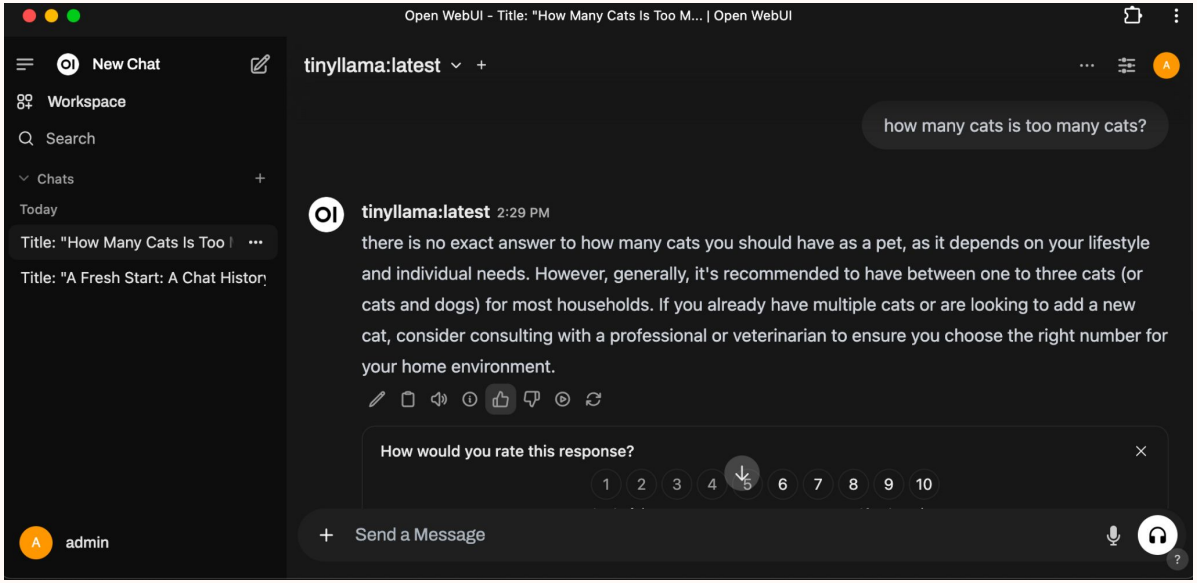
LangGraph

Haystack

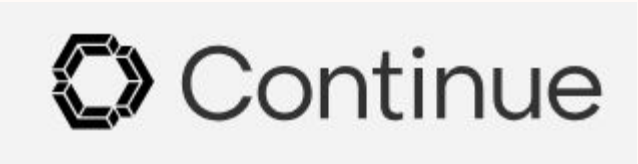
by deepset

Open Web UI

- https://docs.openwebui.com/
- http://localhost:3000/



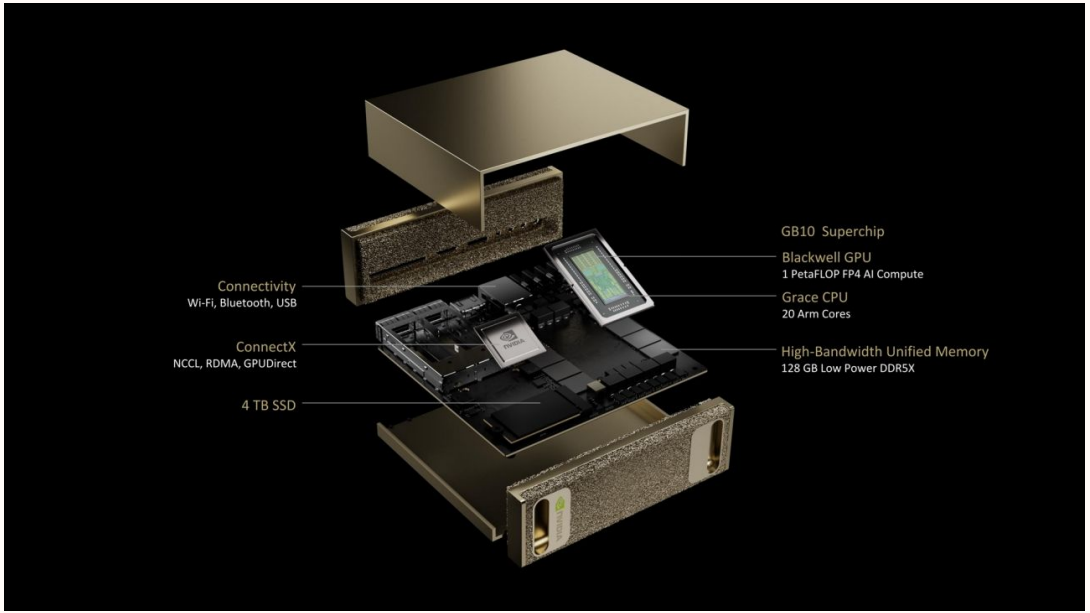
IDE Integration



https://www.continue.dev/

Some Recent Advancements

- NVidia's DIGITS ($3k)
- Jetson Nano ($250)



Summary

- Gen AI Fundamentals
    - Token Generation
    - Chat Completion
    - Image Analysis
    - Structured Output
    - Function Calling
    - RAG Fundamentals
- Products are iterating abstractions on top
    - AWS Bedrock, GCP VertX, Azure Open AI
- SaaS Products inherit the risk of the cloud
    - Vendor Constraints
    - Security
    - Variable Pricing
- Hyperscalers have a severe advantage in terms of competition
    - And there is no end in sight
- Local AI
    - We rely on the blessings of Meta and other open researchers
    - Start saving now for your Gen AI home appliance

A Tale of Two Cities (Closing Thoughts)

The Outwardly Optimistic Capitalists that Watch Gotham from the Skyscraper

- “We believe we’ll have agents acting as Mid-Level Engineers in 2025” - Zuckerberg (Meta)
- “We are no longer hiring humans.” - Sebastian Siemiatkowski (Klarna)
- “The TAM (Total Addressable Market) is in the trillions.” - Marc Benioff (Salesforce)
- “We believe that, in 2025, we may see the first AI agents “join the workforce” and materially change the output of companies.” - Sam Altman (Open AI)

The (Mostly) Silent Pragmatists Walking Through the Market

- “When the poor have nothing to eat, they will eat the rich.” - French Proverb
- “In a worse case, AI trillionaires have near-unlimited and unchecked power, and there's a permanent aristocracy that was locked in based on how much capital they had at the time of labour-replacing AI.” - L Rudolf (Captial Will Matter after AGI)
- “Our dreamtime will be a time of legend, a favorite setting for grand fiction, when low-delusion heroes and the strange rich clowns around them could most plausibly have changed the course of history. Perhaps most dramatic will be tragedies about dreamtime advocates who could foresee and were horrified by the coming slow stable adaptive eons, and tried passionately, but unsuccessfully, to prevent them.” - Robin Hanson (This is the Dream Time)


CodeMash

Appendix

- Presentation Materials Repository (github)
- CAG vs RAG
- Understanding RAG
- Capital Will Matter with AGI
- Building Effective Agents

Gaming Computer Results

Basic Prompt

```
Time to process request: 8 seconds
Tokens: 347
Tokens/s: 60.000000000
TTC (Time to Coffee): 2506.265664160
```

Structured Output

```
Time to process request: 5 seconds
Tokens: 282
Tokens/s: 69.000000000
TTC (Time to Coffee): 2949.852507374
```

Cats

```
Time to process request: 34 seconds
Tokens: 119
Tokens/s: 9.000000000
TTC (Time to Coffee): 8333.333333333
```

Embed

```
Time to process request: 0 seconds
Tokens: 256
Tokens/s: 878.000000000
TTC (Time to Coffee): 1953.125000000
```