Keeping Your Intelligence Local (and Cheap)

by Nicholas Coats

Brain in a Jar 2

Artifact

1, ↷: Put a charge counter on Brain in a Jar, then you may cast an instant or sorcery card with converted mana cost equal to the number of charge counters on Brain in a Jar from your hand without paying its mana cost.

3, ↷, Remove X charge counters from Brain in a Jar: Scry X.

252/297 R
SOI • EN DANIEL LJUNGGREN

% whoami

Name: Nicholas (Nick) Coats

Career: Sr. Consultant @ Nationwide

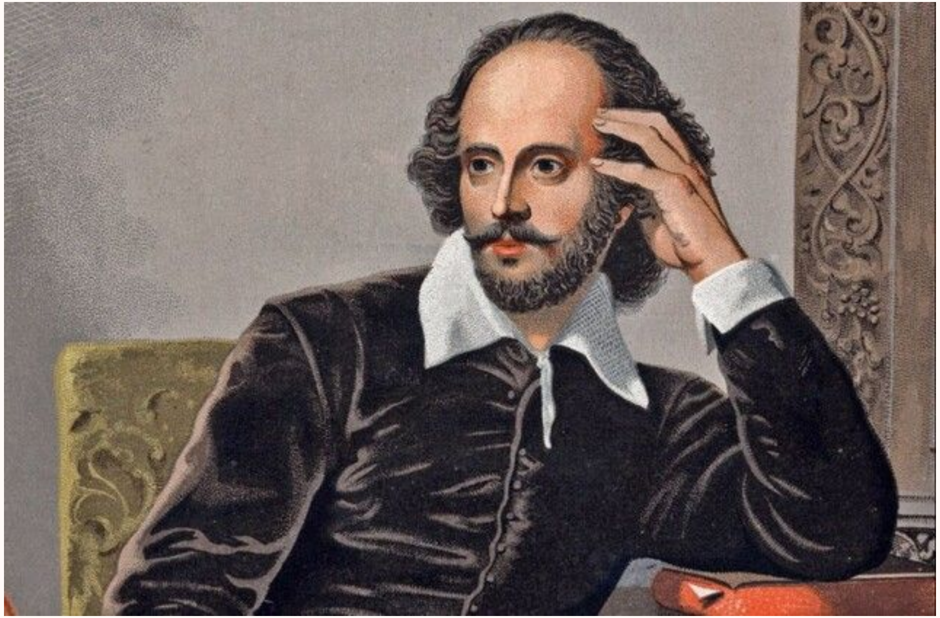
Specialties: Cloud Computing, Software Development and Architecture, DevSecOps

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

Relevant Experience: Not much, honestly

Generative AI

“To be or” ~=> “not to be”



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

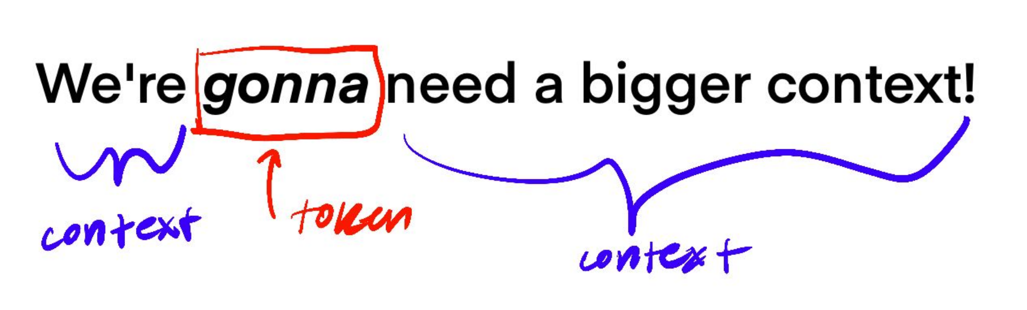
All Presentation Materials on Github

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

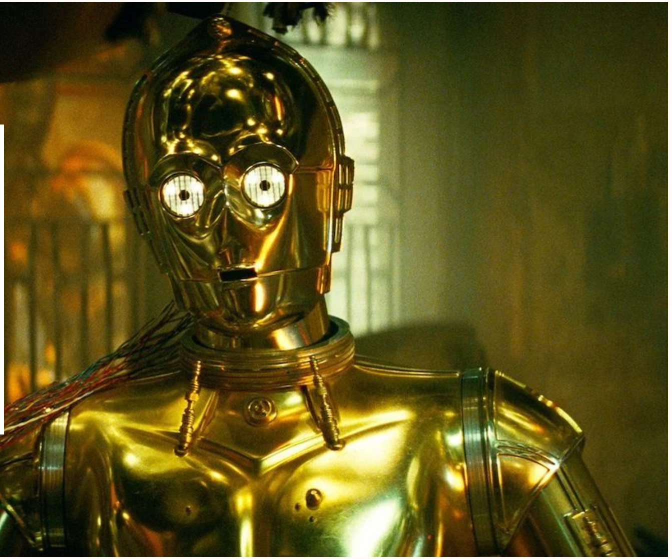
A Few Quick Definitions

- 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”

Token

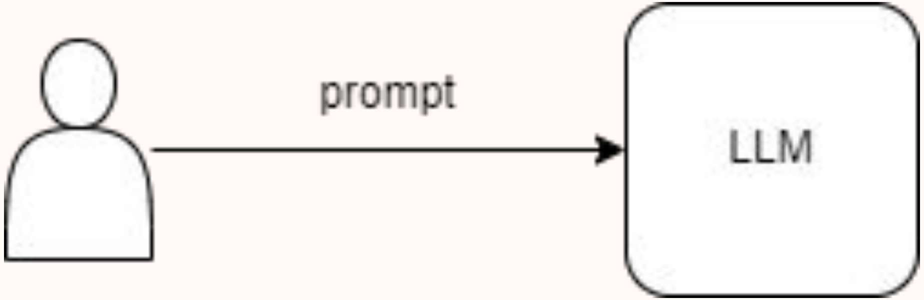


We're *gonna* need a bigger context!

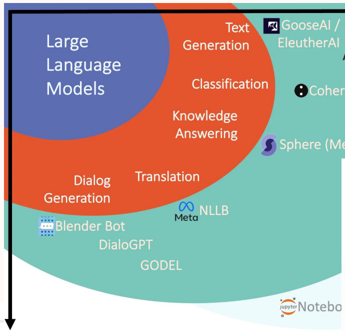


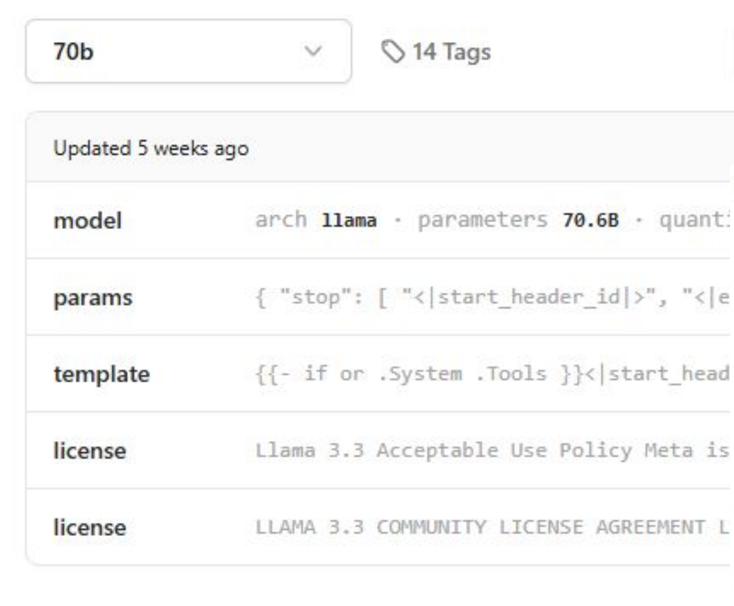
A Basic LLM

- LLM: Large Language Model



Choosing a Model



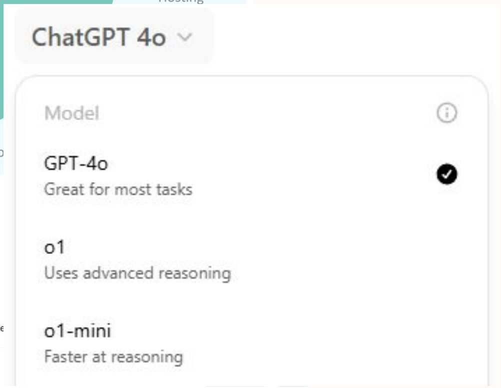


70b ⌄ | 14 Tags

```
ollama run llama3.3
```

Updated 5 weeks ago — a6eb4748fd29 · 43GB

```
model      arch llama · parameters 70.6B · quant
params     { "stop": [ "<|start_header_id|>", "<|e
template   {{- if or .System .Tools }}<|start_head
license    Llama 3.3 Acceptable Use Policy Meta is
license    LLAMA 3.3 COMMUNITY LICENSE AGREEMENT L
```



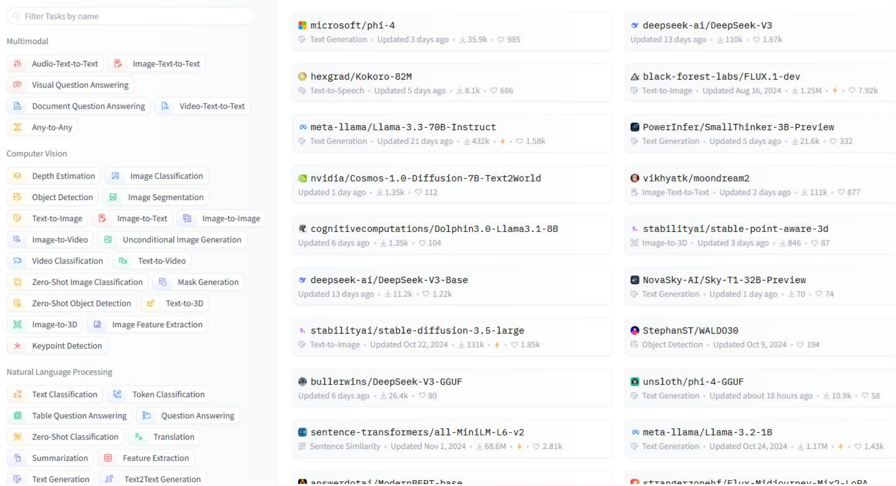
ChatGPT 4o ⌄

Model

GPT-4o ✓
Great for most tasks

o1
Uses advanced reasoning

o1-mini
Faster at reasoning



Hugging Face — Search models, datasets, users... — Models — Datasets — Space

Tasks | Libraries | Datasets | Languages | Licenses | Other

Filter Tasks by name

Multimodal

- Audio-Text-to-Text
- Image-Text-to-Text
- Visual Question Answering
- Document Question Answering
- Video-Text-to-Text
- Any-to-Any

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Text-to-Image
- Image-to-Text
- Image-to-Image
- Image-to-Video
- Unconditional Image Generation
- Video Classification
- Text-to-Video
- Zero-Shot Image Classification
- Mask Generation
- Zero-Shot Object Detection
- Text-to-3D
- Image-to-3D
- Image Feature Extraction
- Keypoint Detection

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Feature Extraction
- Text Generation
- Text2Text Generation

Models 1,275,414 — Filter by name

microsoft/phi-4
Text Generation • Updated 3 days ago • 35.9k • ♡ 985

deepseek-ai/DeepSeek-V3
Updated 13 days ago • 110k • ♡ 1.67k

hexgrad/Kokoro-82M
Text-to-Speech • Updated 5 days ago • 8.1k • ♡ 686

black-forest-labs/FLUX.1-dev
Text-to-Image • Updated Aug 16, 2024 • 1.25M • ♡ 7.92k

meta-llama/Llama-3.3-70B-Instruct
Text Generation • Updated 21 days ago • 432k • ♡ 1.58k

PowerInfer/SmallThinker-3B-Preview
Text Generation • Updated 5 days ago • 21.6k • ♡ 332

nvidia/Cosmos-1.0-Diffusion-7B-Text2World
Updated 1 day ago • 1.35k • ♡ 112

vikhyatk/moondream2
Image-Text-to-Text • Updated 2 days ago • 111k • ♡ 877

cognitivecomputations/Dolphin3.0-Llama3.1-8B
Updated 6 days ago • 1.35k • ♡ 104

stabilityai/stable-point-aware-3d
Image-to-3D • Updated 3 days ago • 846 • ♡ 87

deepseek-ai/DeepSeek-V3-Base
Updated 13 days ago • 11.2k • ♡ 1.22k

NovaSky-AI/Sky-T1-32B-Preview
Text Generation • Updated 1 day ago • 70 • ♡ 74

stabilityai/stable-diffusion-3.5-large
Text-to-Image • Updated Oct 22, 2024 • 131k • ♡ 1.85k

StephanST/WALDO30
Object Detection • Updated Oct 9, 2024 • ♡ 194

bullerwins/DeepSeek-V3-GGUF
Updated 6 days ago • 26.4k • ♡ 80

unsloth/phi-4-GGUF
Text Generation • Updated about 18 hours ago • 10.9k • ♡ 58

sentence-transformers/all-MiniLM-L6-v2
Sentence Similarity • Updated Nov 1, 2024 • 68.6M • ♡ 2.81k

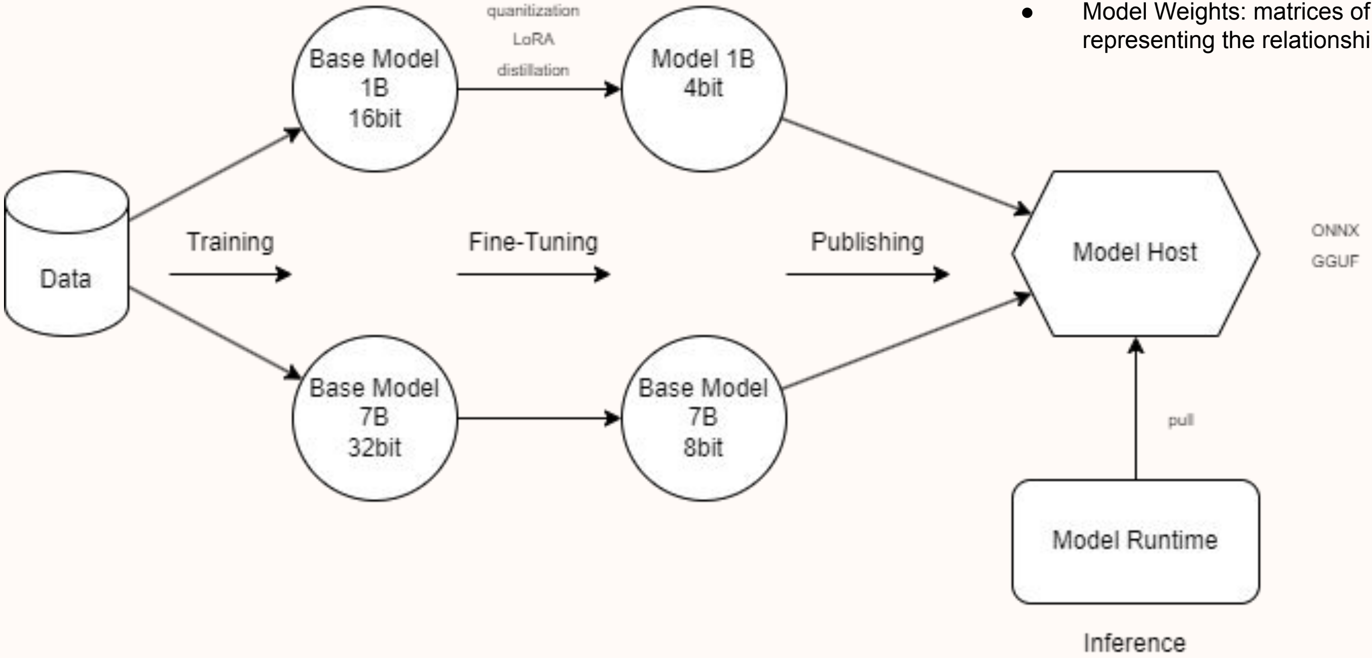
meta-llama/Llama-3.2-1B
Text Generation • Updated Oct 24, 2024 • 1.17M • ♡ 1.43k

answerdotai/ModernBERT-base

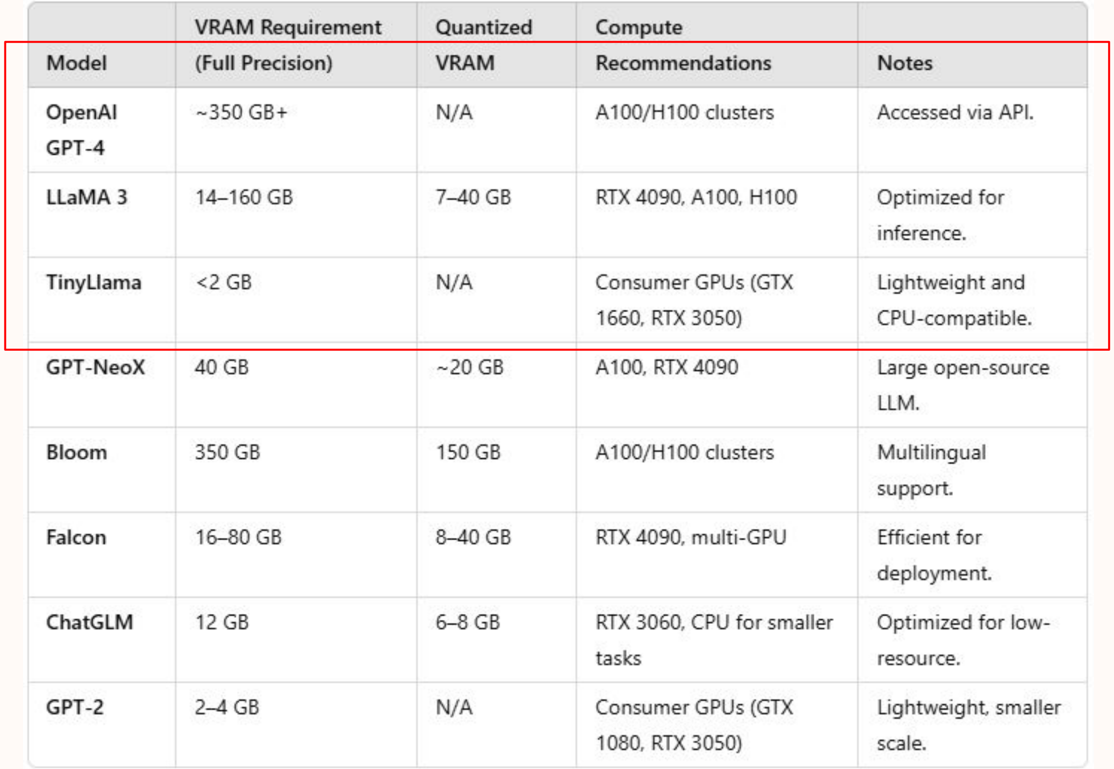
strangerzonehf/Flux-Midjourney-Mix2-LoRA

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

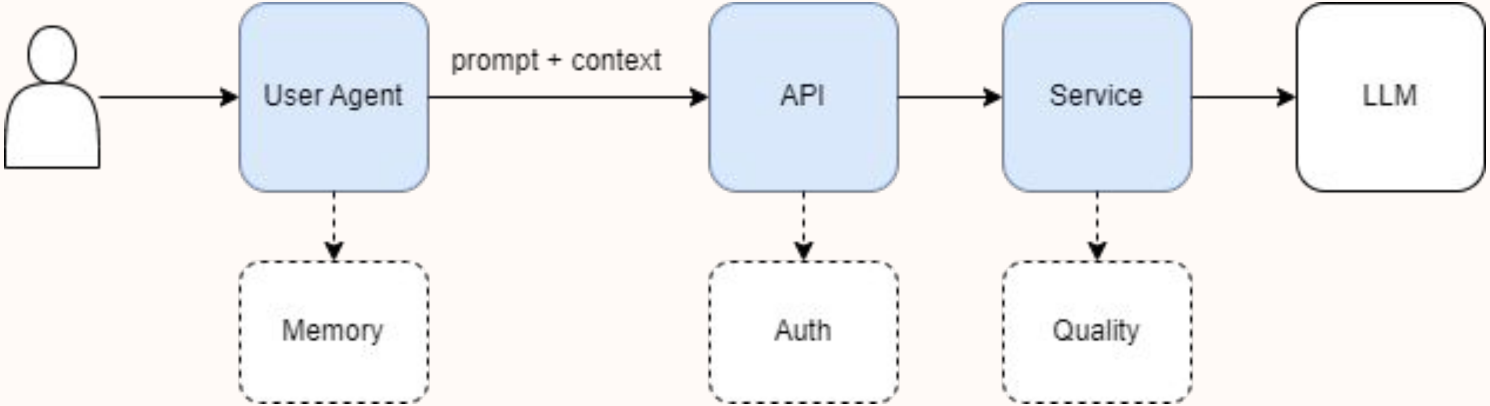


Requirements



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.

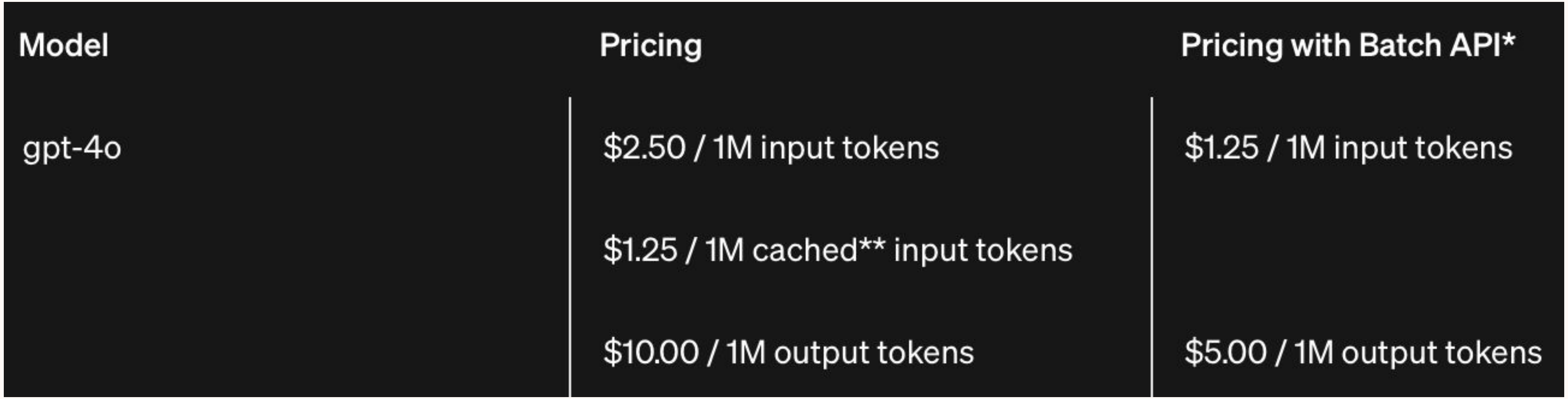
Let’s API Enable that Bad Boy



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

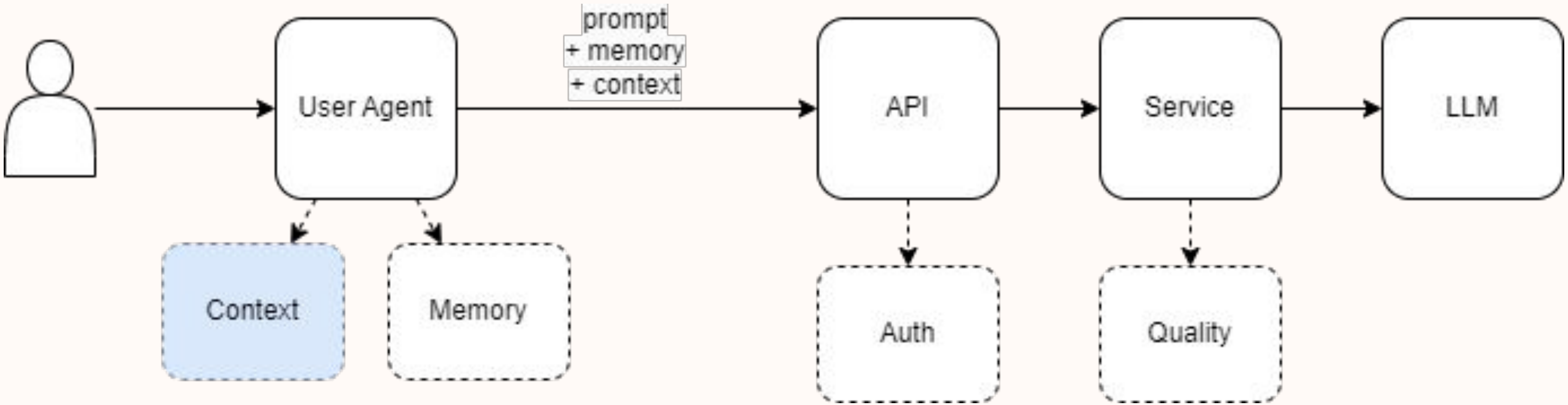
- https://openai.com/api/pricing/

Coffee

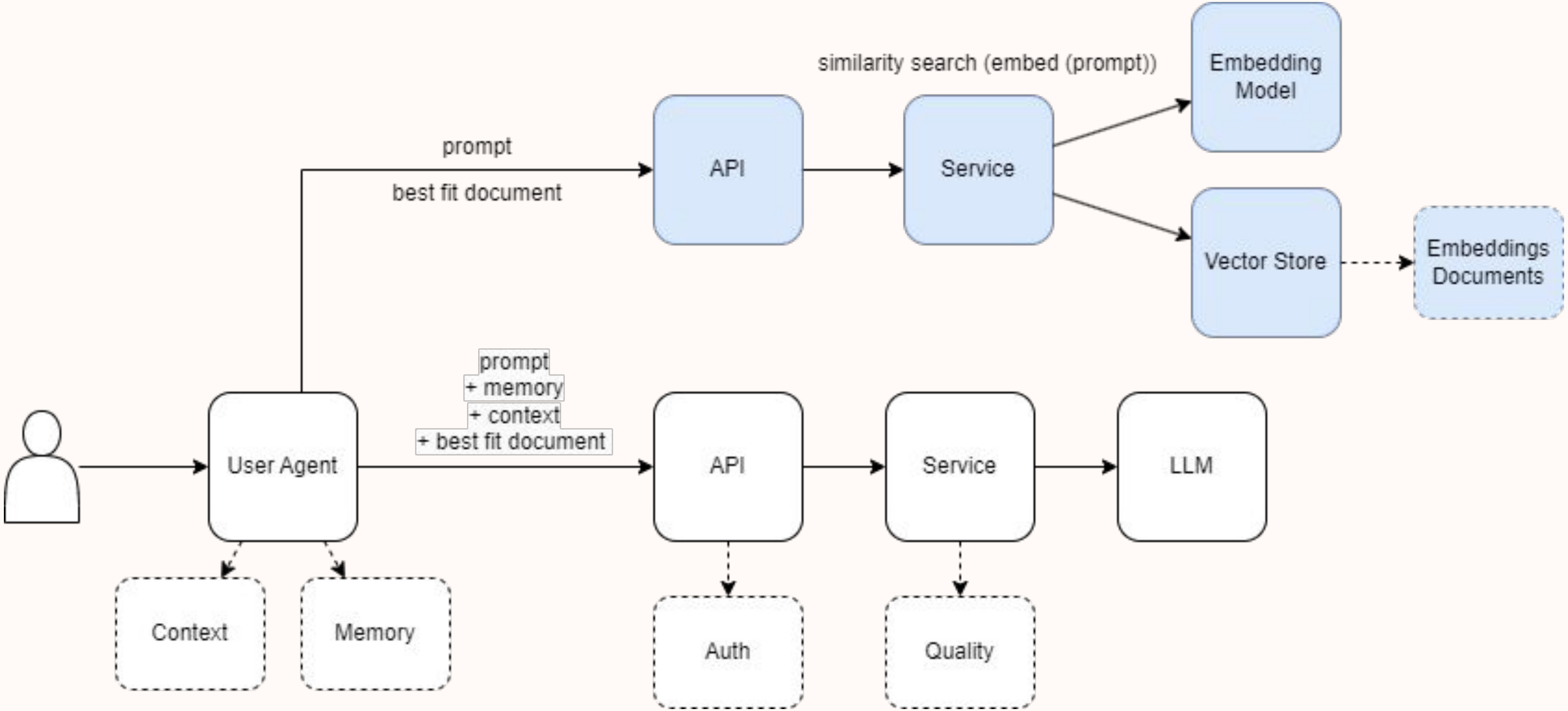
For every 1 Million token, if we ran

TTC = Time to Coffee

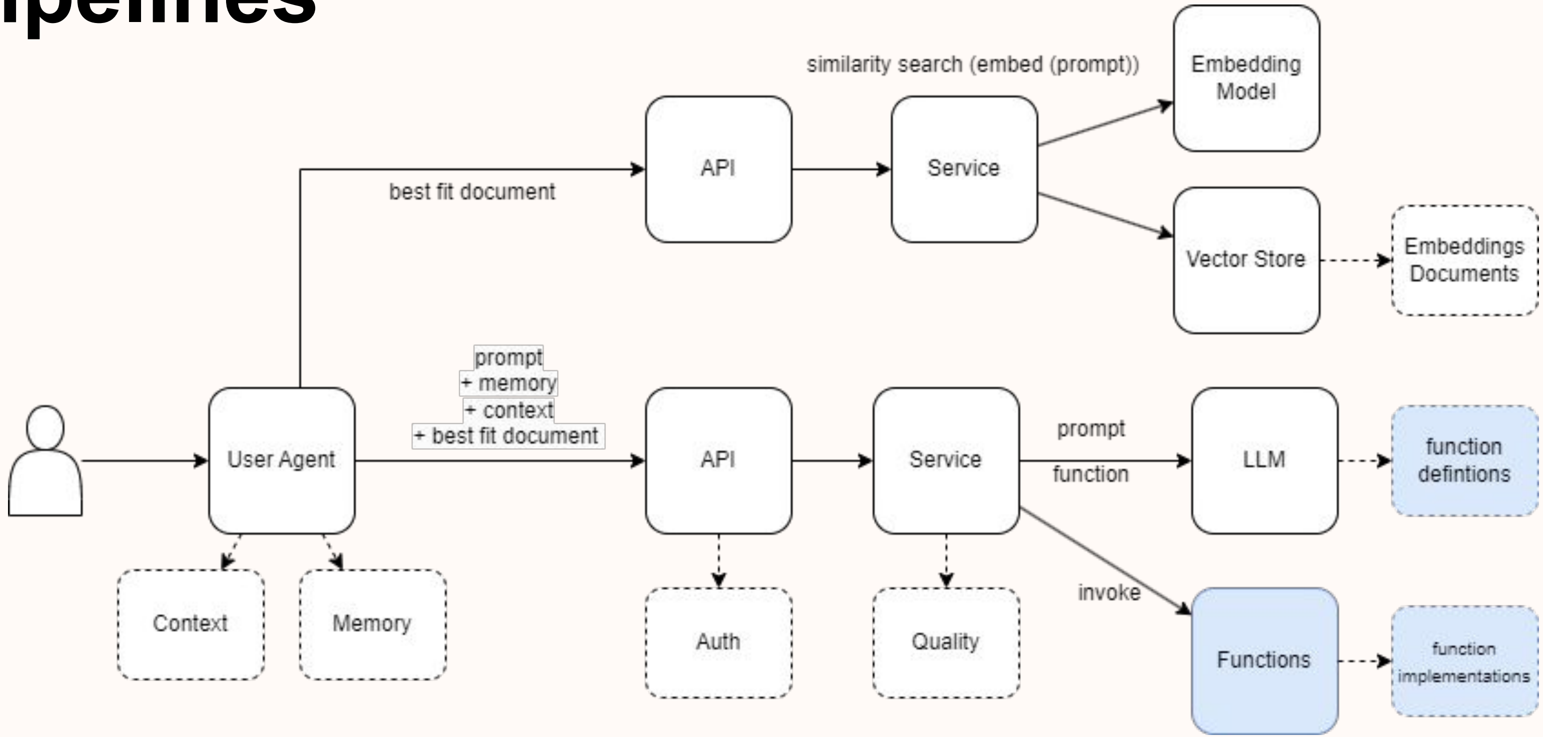
Context Augmented Generation



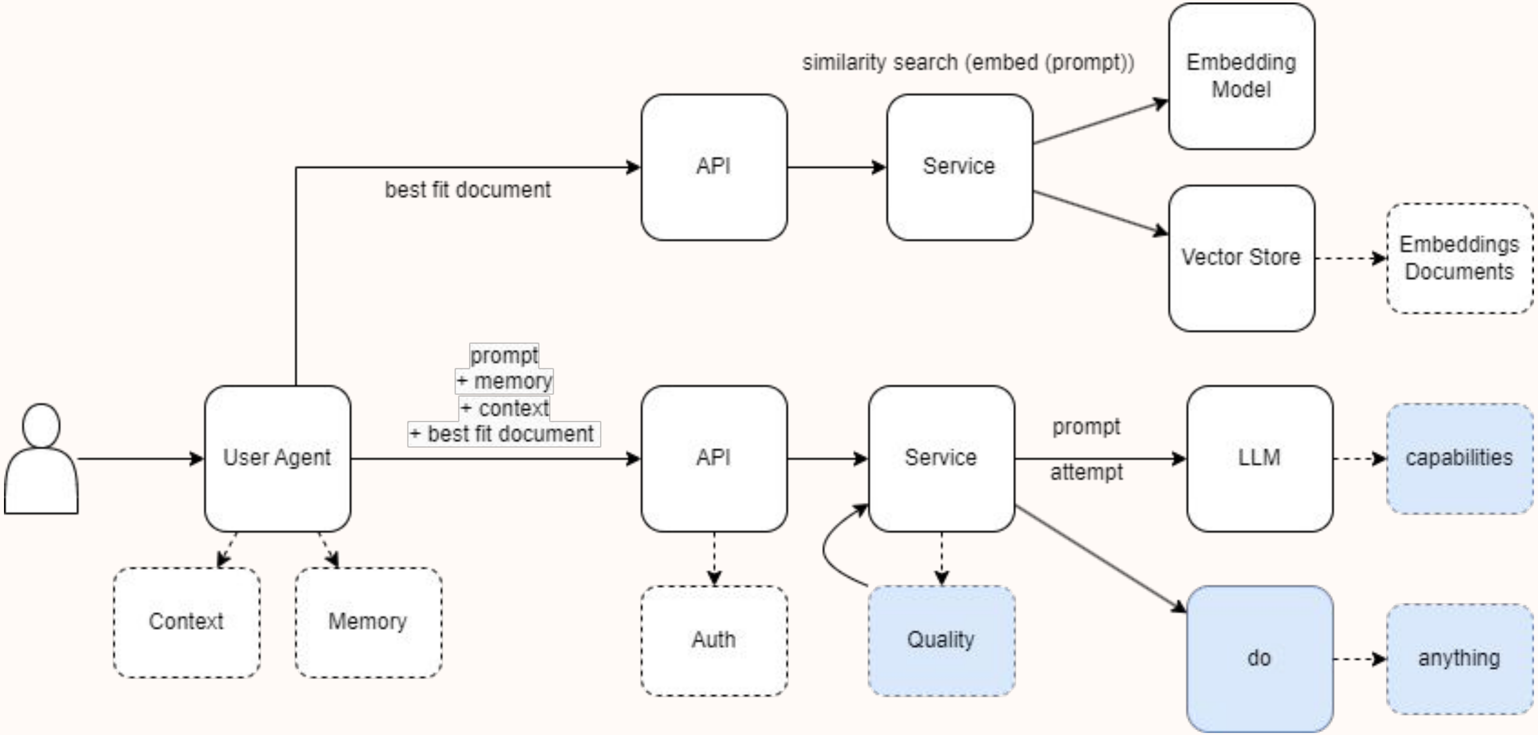
Retrieval Augmented Generation



Agentic Systems - Workflows AKA Pipelines

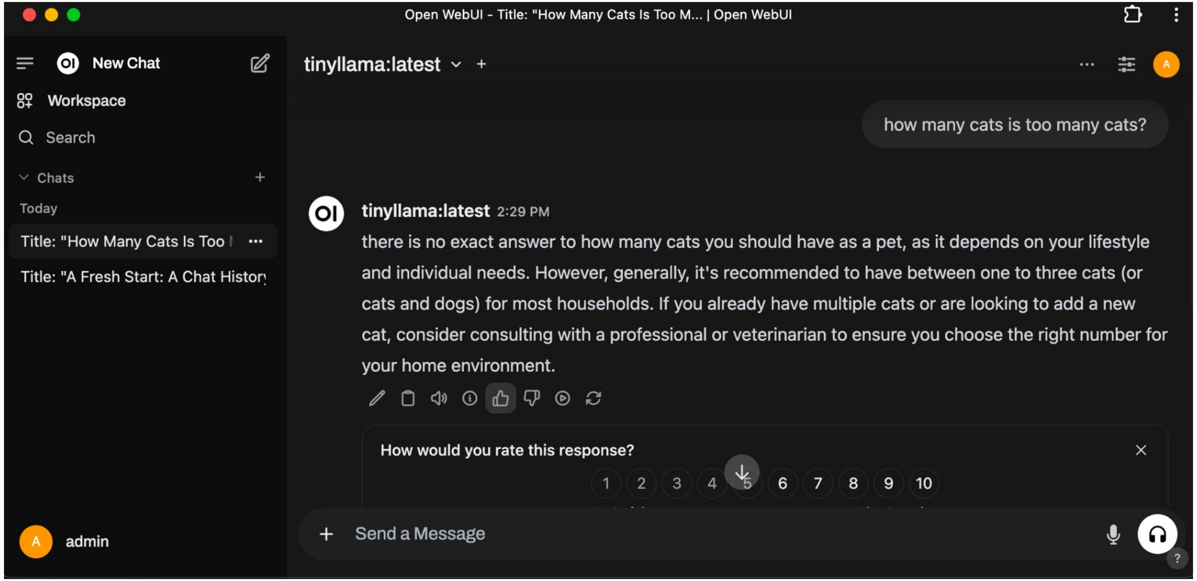


Agentic Systems - Agents



Open Web UI

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



Open WebUI - Title: "How Many Cats Is Too M... | Open WebUI
New Chat
Workspace
Search
Chats
Today
Title: "How Many Cats Is Too I
Title: "A Fresh Start: A Chat History
admin
tinyllama:latest

how many cats is too many cats?

tinyllama:latest 2:29 PM

there is no exact answer to how many cats you should have as a pet, as it depends on your lifestyle and individual needs. However, generally, it's recommended to have between one to three cats (or cats and dogs) for most households. If you already have multiple cats or are looking to add a new cat, consider consulting with a professional or veterinarian to ensure you choose the right number for your home environment.

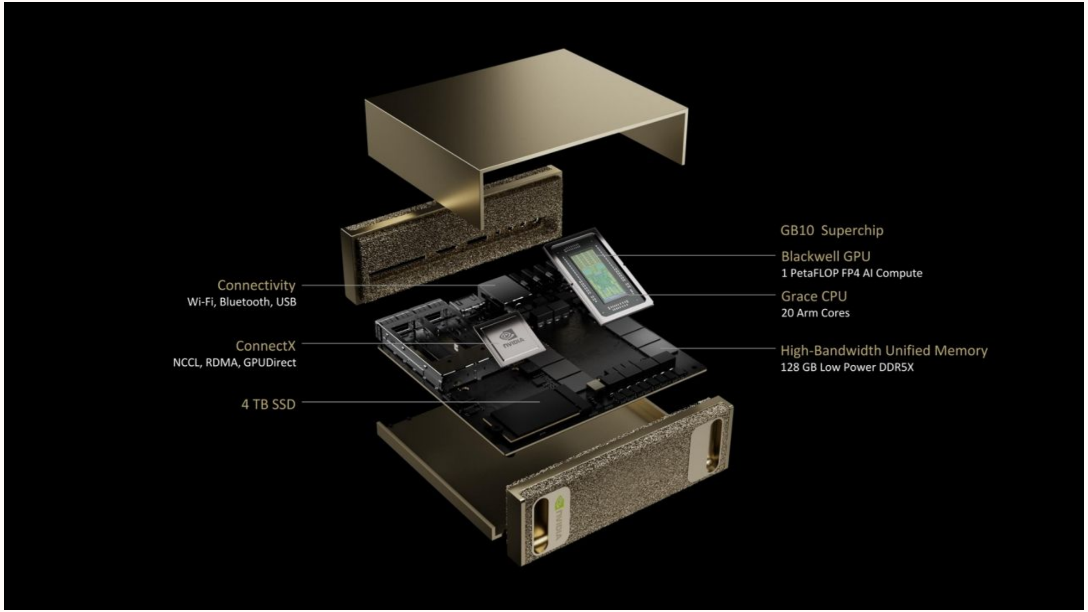
How would you rate this response?

1 2 3 4 5 6 7 8 9 10

Send a Message

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

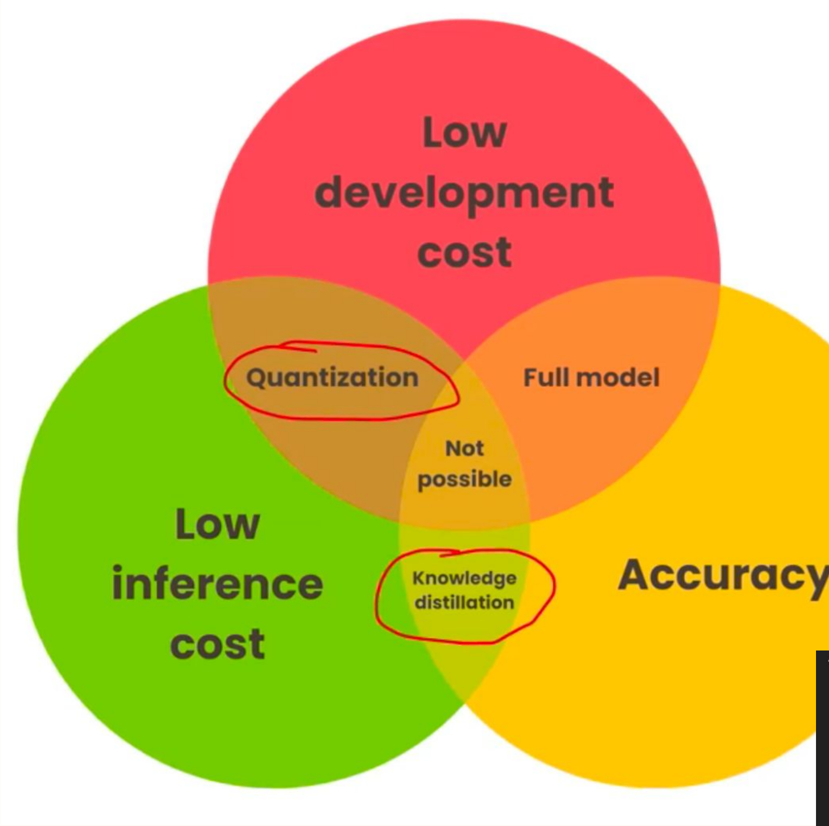
- “When the poor have nothing to eat, they will eat the rich.” - French Proverb
- “In a worse case, AI trillionaires have near-unlimited and unchecked power, and there's a permanent aristocracy that was locked in based on how much capital they had at the time of labour-replacing AI.” - L Rudolf (Captial 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)

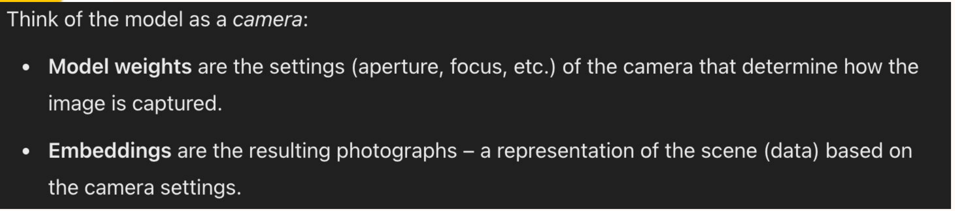
CodeMash

Appendix

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

OLD

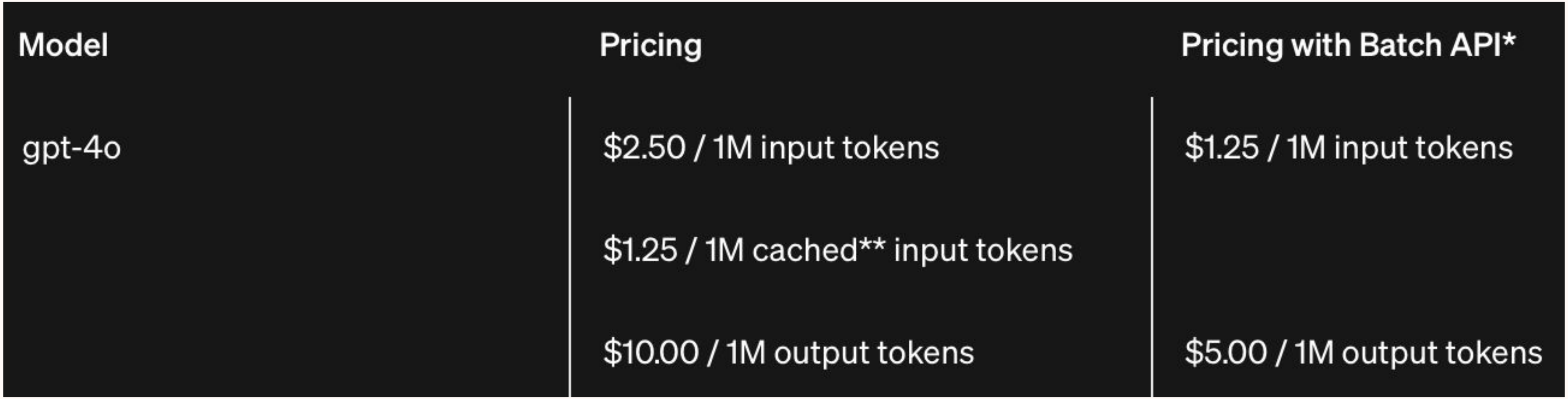




Think of the model as a *camera*:

- **Model weights** are the settings (aperture, focus, etc.) of the camera that determine how the image is captured.
- **Embeddings** are the resulting photographs – a representation of the scene (data) based on the camera settings.

pictures… put in right place



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

pricing

% what - does the future look like

- Robin Hanson calls the present "the dreamtime", following a concept in Aboriginal myths: the time when the future world order and its values are still liquid, not yet set in stone.

  - https://www.lesswrong.com/posts/KFFaKu27FNugCHFmh/by-default-capital-will-matter-more-than-ever-after-agi

Appendix

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