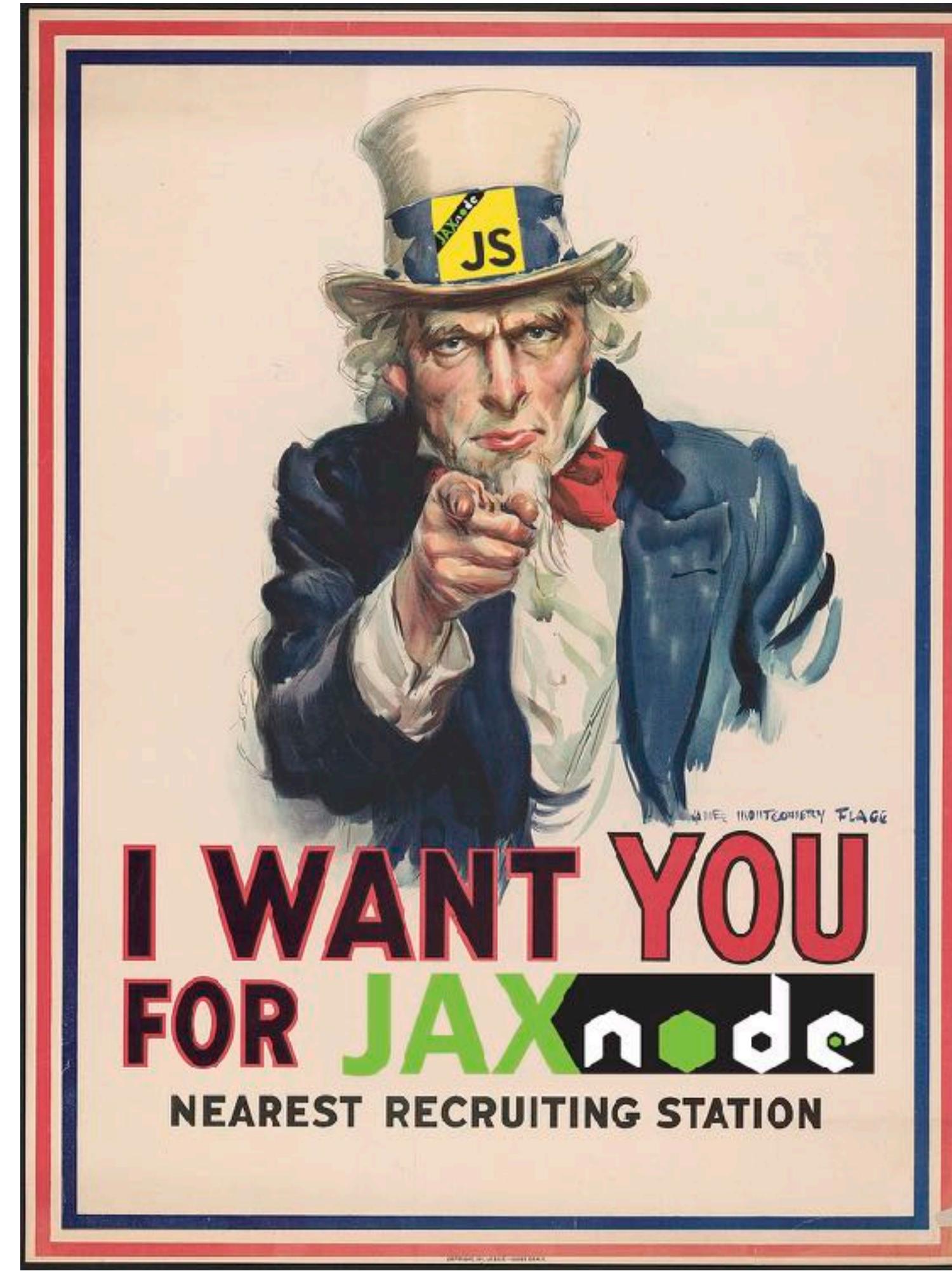


LangChain

Build AI apps with Python and LangChain

PyJax Feb 2025

Come to JaxNode next week!



About me

David Fekke

- JaxNode user group
- Web and Mobile Developer
- JS, TS, React, C#, Swift, Obj-C, Kotlin, Java and SQL
- fek.io/blog/1
- youtube.com/c/polyglotengineer
- github.com/davidfekke
- @jaxnode @polyglotengine1





Whats new in AI?

▼ DJIA Futures 44700.00 0.00% ↑ S&P 500 Futures 6065.50 -0.55% ↓ Nasdaq Futures 21571.25 -0.25% ↓ SSE Index 3250.60 -0.06% ↓ Stoxx 600 539.53 0.13% ↑ Crude < >

THE WALL STREET JOURNAL.

DAVID FEKKE ▾

DON'T MISS YOUR STORY

English Edition ▾ | Print Edition | Video | Audio | Latest Headlines | More ▾

Latest World Business U.S. Politics Economy Tech Markets & Finance Opinion Arts Lifestyle Real Estate Personal Finance Health Style Sports



TECHNOLOGY | ARTIFICIAL INTELLIGENCE [Follow](#)

Tech Leaders Pledge Up to \$500 Billion in AI Investment in U.S.

OpenAI, Oracle and SoftBank unveil AI infrastructure plans at White House

By [Deepa Seetharaman](#) [Follow](#) and [Tom Dotan](#) [Follow](#)

Updated Jan. 21, 2025 8:08 pm ET



Gift unlocked article



Listen (4 min)





What to Know About China's DeepSeek AI

The Chinese upstart says it has trained high-performing AI models cheaply, without using the most advanced chips

By [Sam Schechner](#) [Follow](#) and [Stu Woo](#) [Follow](#)

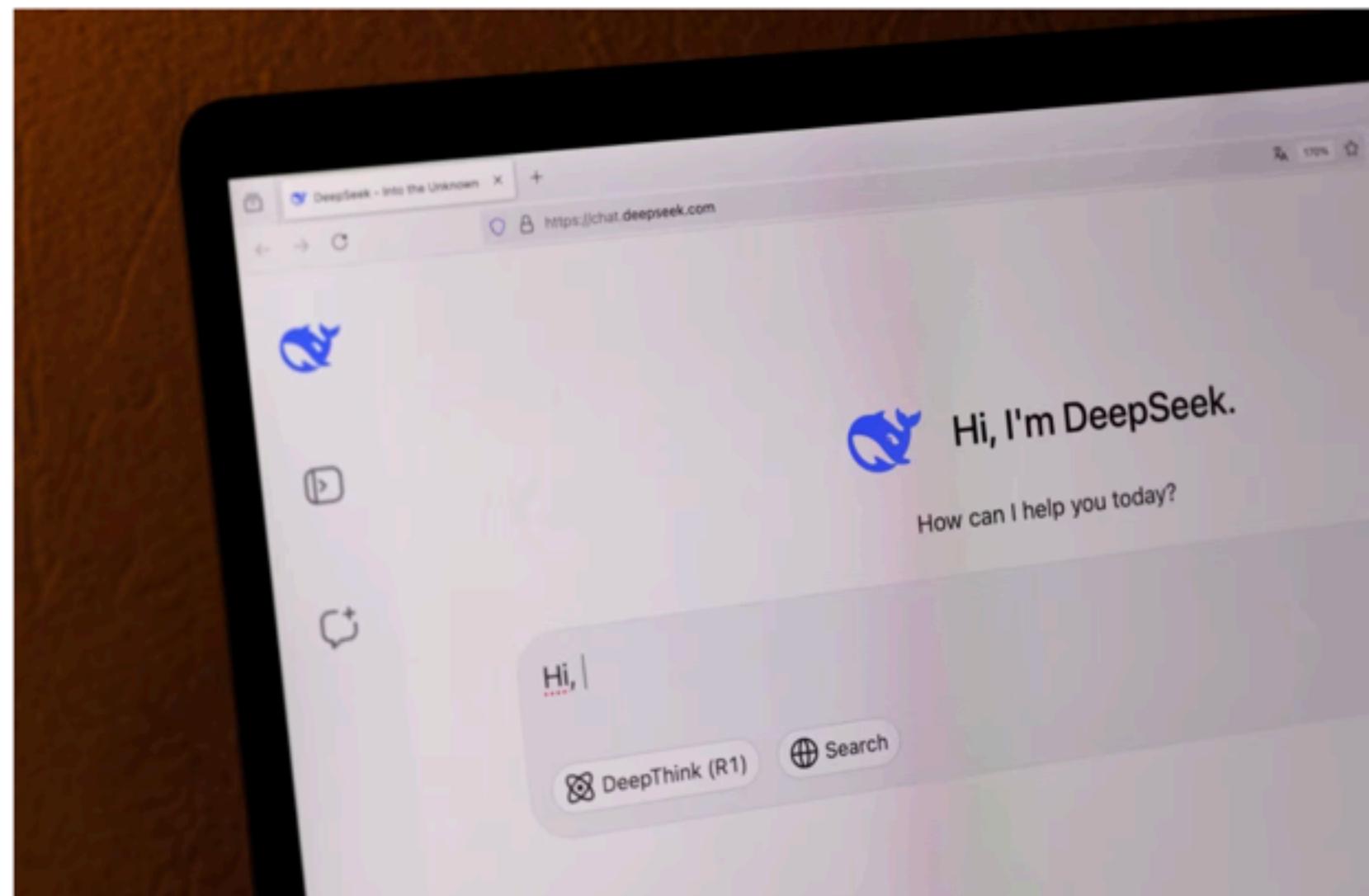
Updated Jan. 28, 2025 12:44 pm ET



Gift unlocked article



Listen (8 min)



What do we know about Deepseek?

Did they really train it for only \$6 million

- Deepseek used to be a Hedge Fund
- Created an Open model that is on the same level as OpenAI and Anthropic
- They spent \$1 billion on infrastructure
- Used a technique called “distillation” to train their model
- Innovation from combining smaller models into a larger model

How does Deepseek Work

What are the innovations

- Like OpenAI's o3 model, It is a reasoning model
- On the training it uses distillation and reinforcement learning
- Extreme Context Extension
- Mixture-of-Experts with Dynamic Load Balancing using outed experts
- Optimized inter-GPU communication without using CUDA



Jiayi Pan
@jiayi_pirate · [Follow](#)

X

We reproduced DeepSeek R1-Zero in the CountDown game, and it just works

Through RL, the 3B base LM develops self-verification and search abilities all on its own

You can experience the Ahah moment yourself for < \$30
Code: github.com/Jiayi-Pan/Tiny...

Here's what we learned  [Show more](#)

User: Using the numbers [19, 36, 55, 7], create an equation that equals 65.

Assistant: Let me solve this step by step.

<think> We need to find an equation using the numbers 19, 36, 55, and 7 exactly once, with basic arithmetic operations, that equals 65. One possible combination is $55 + 36 - 19 + 7$. Let's check: $55 + 36 = 91$, $91 - 19 = 72$, and $72 + 7 = 79$. However, this doesn't equal 65. Another combination is $55 + 36 + 7 - 19$. Let's check: $55 + 36 = 91$, $91 + 7 = 98$, and $98 - 19 = 79$. This also doesn't equal 65. After trying different combinations, I found that $55 + 36 + 7 - 19 = 79$, which is close but not equal to 65. Finally, I found that $55 + 36 - 19 + 7 = 79$, which is still not equal to 65. However, if we try $55 + 36 - 7 + 19$, we get $55 + 36 = 91$, $91 - 7 = 84$, and $84 + 19 = 103$, which is not equal to 65. After further trial and error, I found that $55 + 36 + 7 - 19 = 79$, which is still not equal to 65. However, if we try $55 + 36 - 7 - 19$, we get $55 + 36 = 91$, $91 - 7 = 84$, and $84 - 19 = 65$. This equals 65. </think>

<answer> $55 + 36 - 7 - 19$ </answer>

12:14 PM · Jan 24, 2025

i



5.5K



Reply



Copy link

[Read 153 replies](#)

New AI Czar

David Sacks

- David Sacks is the new AI and Crypto Czar
- Co-host of the All In Podcast
- Tech investor who has invested in AI startups
- Was part of the PayPal Mafia



OpenAI

New o1 and o3 models

- The o1 and o3 models are more than just a LLM
- These are reasoning models
- They are optimized logical reasoning
- Better at solving problems
- Tend to not be as creative, but lower chance of hallucination

Gemini 2.0 Flash

Google's AI model

- Support for over 1M tokens
- 8k of tokens for output
- It has a reasoning mode
- Accessible through the Gemini API

Elon Musk-Led Group Makes \$97.4 Billion Bid for Control of OpenAI

Unsolicited offer complicates Sam Altman's plans to convert OpenAI to a for-profit company

By *Jessica Toonkel* [Follow](#) and *Berber Jin* [Follow](#)

Updated Feb. 10, 2025 4:37 pm ET



598



Gift unlocked article



Listen (6 min)



LangChain

Lets build AI apps with LangChain

- What is LangChain
- Different Large Language Models
- RAG apps
- Vector Databases
- Text processing
- Chaining Prompts, Data and LLMs

LangChain

- LangChain is a framework for chaining LLMs together with data and functions
- Both Python and JavaScript are the official frameworks
- There is a Java framework
- LangChain company also makes LangGraph and LangSmith

LLMs

Large Language Models

- OpenAI's GPT (Generative Pre-trained Transformer)
- Google Palm 2, Gemini and Gemini 2
- Claude by Anthropic (Offshoot of former OpenAI engineers)
- Cohere
- Llama 2, 3 by Meta (Open Source)
- Mistral.ai
- Many others

OpenAI APIs

- OpenAI has become a default
- Completion, Chat and Assistant APIs
- Dall-E imaging API
- Text to Speech API
- Charge per token
- GPT 4.0-mini most affordable

Google Deepmind

- Palm 2
- Gemini 2 recently released
- Multi-model can generate text, images, video and audio
- APIs available
- Safety parameters part of the API

Llama 3.3

Open model

- Released by Meta last year
- Allows developers to pre-train
- Already many different versions of Llama
- Run locally with Ollama
- Find other on Hugging Face
- GGUF or GPT-Generated Unified Format



Customizing LLMs and Reasoning models

- Create an LLM from scratch (Prohibitively expensive)
- Pre-train or retrain existing models (Still expensive)
- RAG (Retrieval Augmented Generation) Apps

RAG Apps

Low cost customization of LLM

- Make data available to LLM based on prompt
- Typically text is pulled from documents and stored in a vector database
- Documents have to be split into chunks or small blobs
- Embeddings are created from text and stored into database
- Text can be retrieved and used in prompts or passed as embedding



**I DECIDED TO QUIT MY JOB
AND START A COMPANY THAT MAKES
A CHATBOT THAT READS PDF FILES**

LangChain

- LangChain has all of the tools needed to build RAG apps
- Python and TypeScript/JavaScript libraries
- LangChain allows developers to build context aware AI applications
- Chain prompts, data and functions to LLMs
- Built in tools for text and document splitting
- Support for Vector databases
- Supports almost all LLMs

Vector Databases

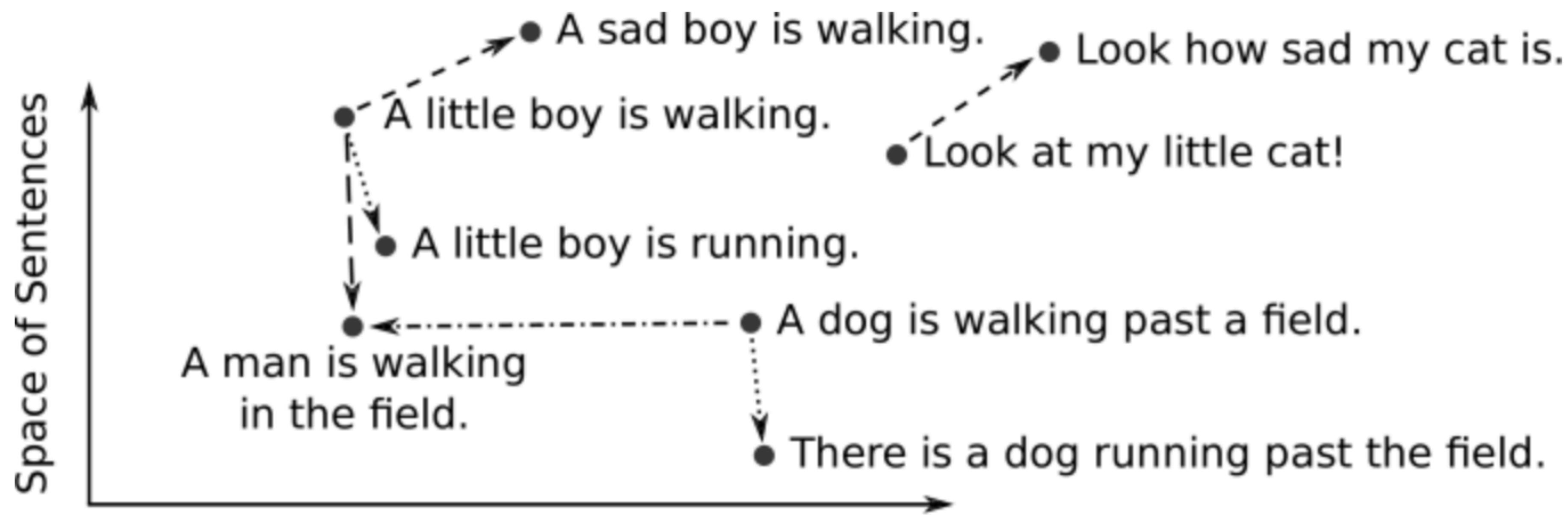
Vector storage and retrieval

- Vector DBs can be used to store embeddings
- Many vendors available
- Pinecone, Supabase, Chroma, Mills and Weaviate
- Vector extension for Postgres
- Azure Cosmos DB has vector search

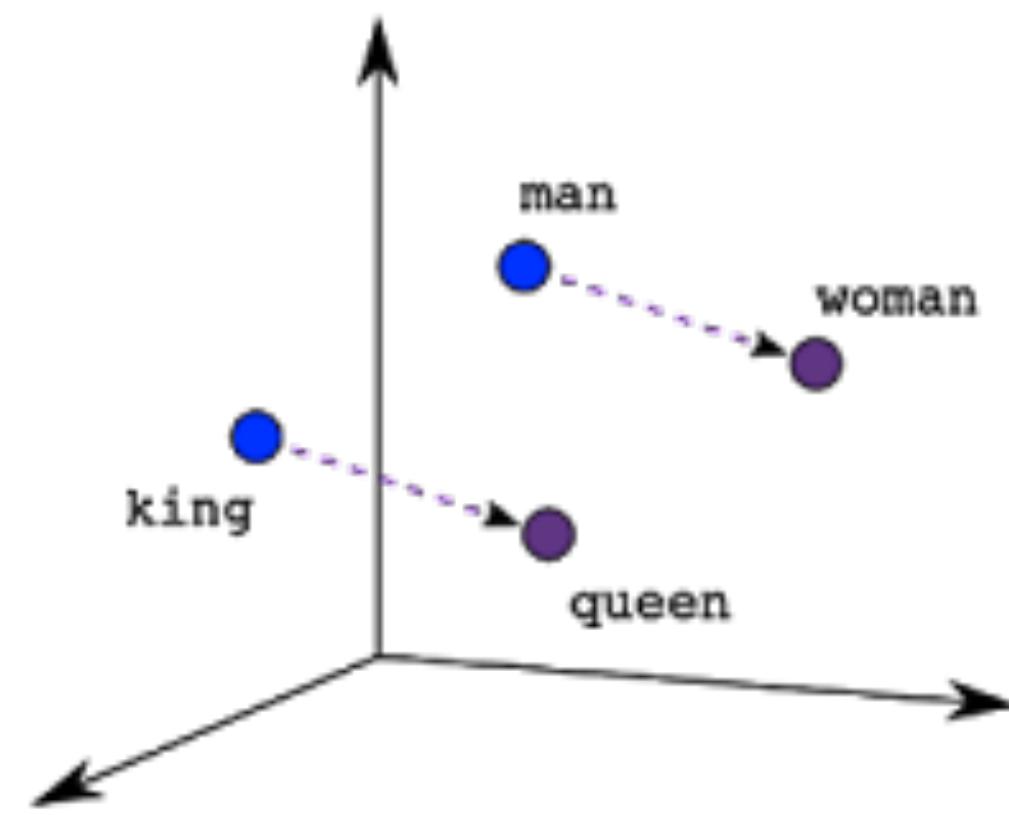
Embeddings

What is an embedding

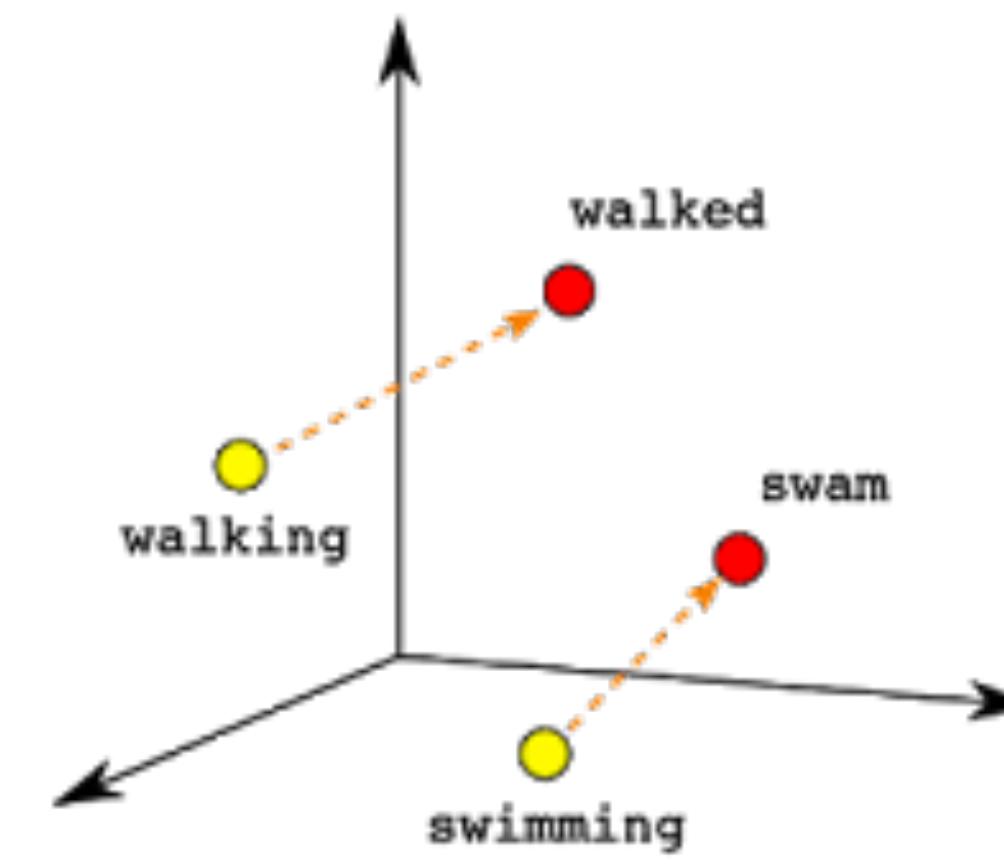
- Vectors or embeddings are just arrays of numbers
- LLMs convert text into tokens that can be stored as a numerical representation
- OpenAI's embedding is an array of floats with a length of 1536
- Tokens are typically four characters in length
- About three words for every four tokens
- OpenAI charges by the token



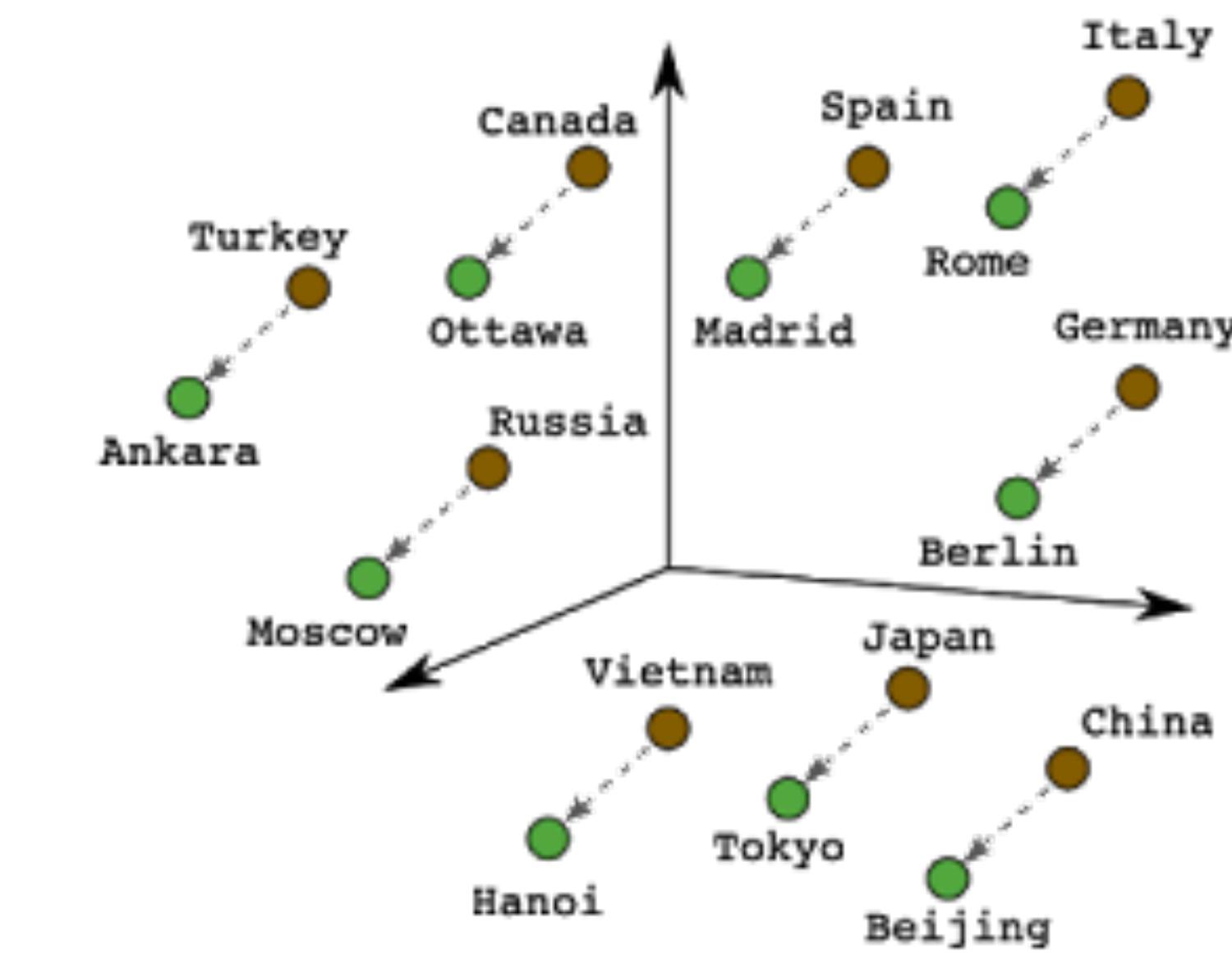
Source: DeepAI



Male-Female



Verb Tense



Country-Capital

Source Google

Token and Embedding cost

- GPT 4o-mini is \$0.15 per million input token and \$0.30/m for out put token
- o1 \$15.00 / 1M input tokens, \$60 / 1M output tokens
- o3-mini \$1.10 / 1M input tokens, \$4.40 / 1M output tokens
- Ada 2 embedding tokens are \$0.10 / 1M tokens, text-embedding-3-small \$0.020 / 1M tokens
- Open Source libraries you can use to generate embeddings for free
- Ollama will also generate embeddings

Store embedding

Add embeddings to Vector store

- Vector store usually stores text, metadata and the vector in a row
- LangChain has an API for creating documents and storing them as vectors
- LangChain has `Stores` for binding documents to vector database
- Most vendors are supported

PGVector

Storing vectors in Postgres

- Postgres is a very flexible RDBMS
- Can be extended with extensions
- PGVector extension allows Postgres to store vectors as special type
- Allows searches to be performed based on similarity
- Operators for distance <-> , inner product <#> and cosine distance <=>

Text Processing

The practice of breaking docs up into manageable pieces

- LangChain has multiple document processors
- Document Loaders for TXT, PDF, CSV, JSON, JSONL
- Text Splitting of documents allows for chunk size and overlap
- These chunks can be converted into embeddings and stored in a VectorDB
- Overlap is necessary to give best context to LLM

Chaining

Putting the Chain in LangChain

- The real power of LangChain comes from chaining prompts, data and LLMs
- RAG applications take advantage of these chaining features to retrieve data from vector stores and functions and chain this data to a LLM
- The ‘Chain’ can pass contents from a ‘vectorStore’ to a LLM along with the prompt to create an answer that is relevant
- Debugging tools: { verbose: true }
- Chains can be invoked or executed with agents

Chaining workflow

RAG Workflow

- Create a connection to vector store
- Create embeddings
- Create an LLM reference
- Use a prompt that can take context and question parameters
- Pass prompt, vector store as a retriever and LLM into a runnable
- Execute the runnable

Agents

- What is an Agent
- Agents can be used to reason between actions with LLMs
- Routers and State Machines can be used in chains
- Agents not as reliable as chains
- Agents are harder to debug
- OpenAI Deep Research is example of agent
- LangSmith is a tool for aiding in the development of Agents

Open LLMs

Llama 3

- Very easy to install and use locally
- Can install from ollama.ai, runs on Windows, Mac and Linux
- `$ ollama run <model-name>`
- Creates an endpoint you can access at <http://localhost:11434>
- Download different models from command line
- Models come in 7B, 13B, 70B and 405B parameters
- Inference on these models are very good

Hardware requirements

Running LLMs locally

- Helps to have modern processor technology like GPU and NPUs
- You should have at least 8 GB of RAM available to run the 7B models
- 16 GB to run the 13B models
- 32 GB to run the 33B models

OpenAI APIs

Ollama is compatible with the OpenAI API

- You can access the OpenAI compatible endpoint by using the following URL:
- <http://localhost:11434/v1>
- Supports the chat completions API
- Plans for support vision, embedding and function calling

Demo

Examples from tonight's Demo

- Source code can be found at the following URL:
- <https://github.com/davidfekke/langchain-pyjax>
- <https://langchain.com>

Questions?