RICE

RICE UNIVERSITY
Ken Kennedy Institute
Responsible AI and Computing for Global Impact

Foundation Models, LLMs, Generative AI, …

Anshumali Shrivastava
Associate Professor, Computer Science and Ken Kennedy Institute
Rice University
Founder:- ThirdAI Corp. and xmad Corp.
anshumali@rice.edu

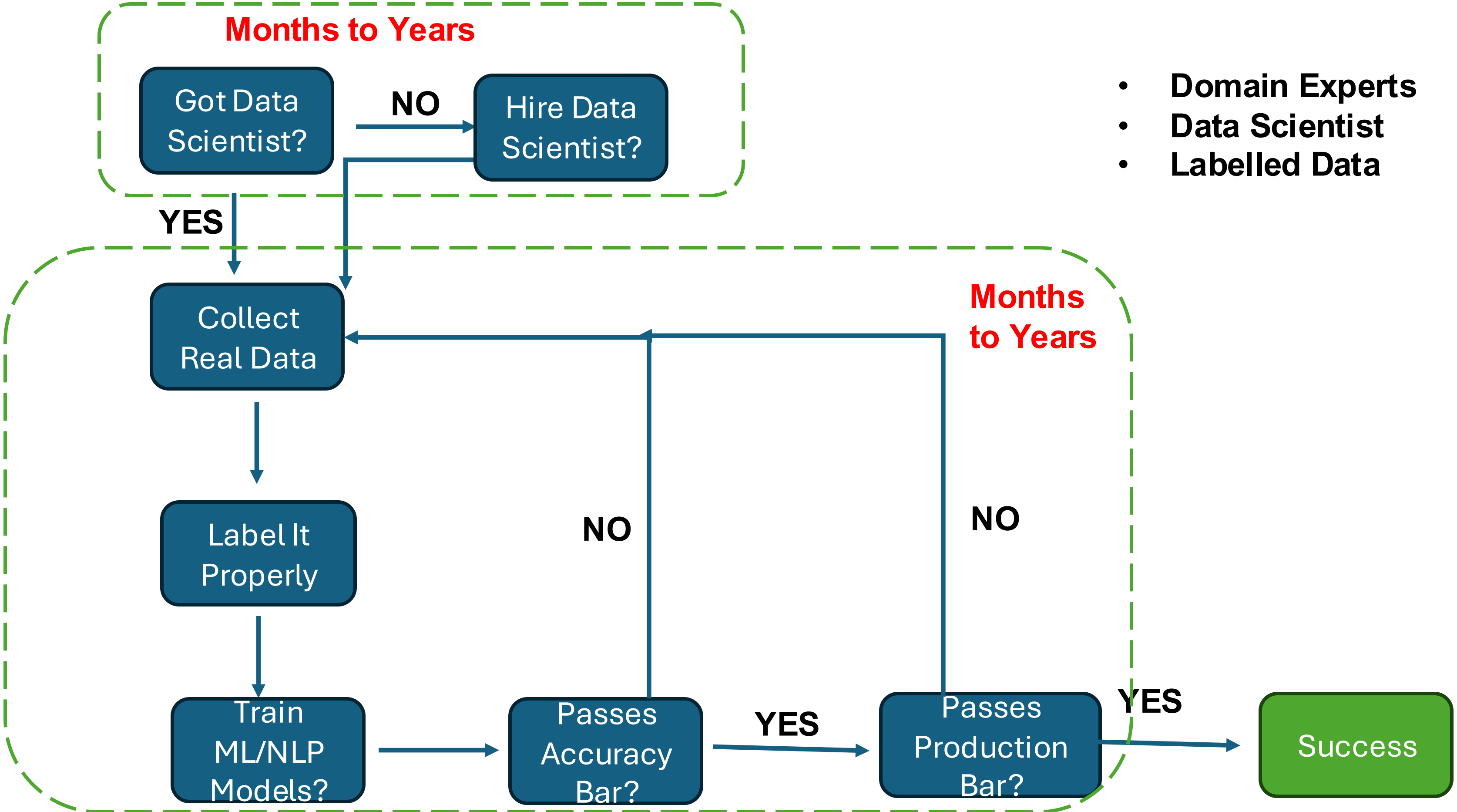
K2I Boot Camp
9th May 2025

Why and How is GenAI Different? – Experience it Yourself!

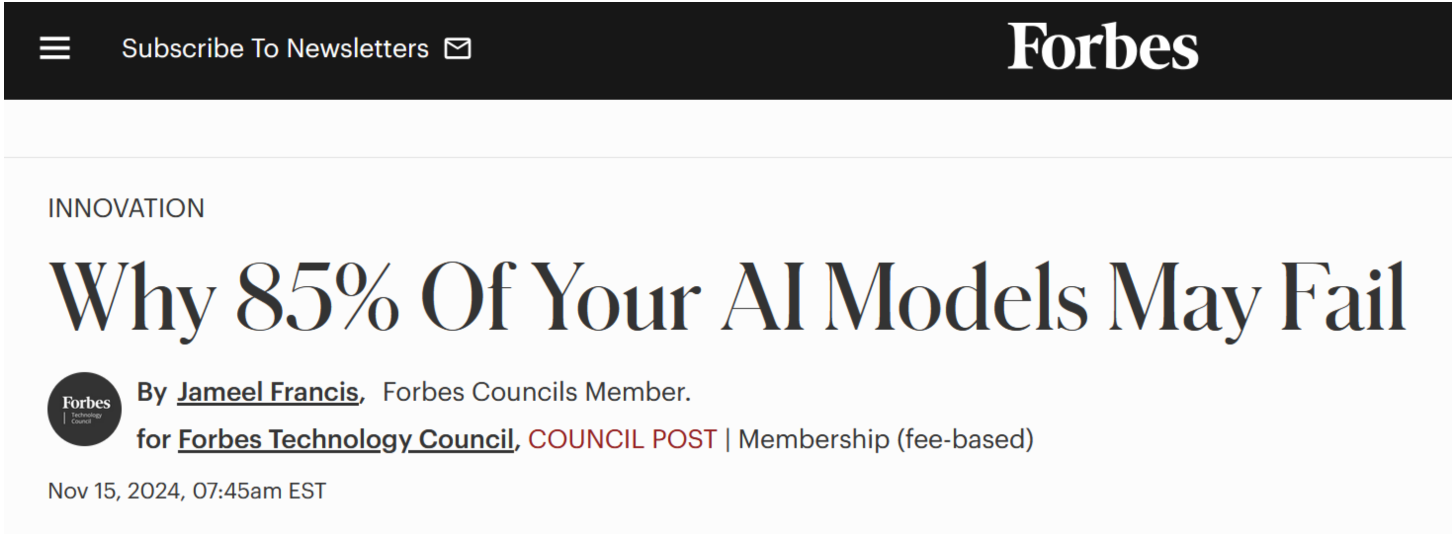
- **Goal**: Design a system that can detect address in any text!
- **Few Examples**:
  - Duncan Hall is located at 6100 Main Street, Houston, TX, 77005
  - My friend is staying at Aarambh Apartment, Shivaji Nagar, Pune, Maharashtra, India, PIN Code 411005
  - I am visiting the place in Vietnam Đường Quang Trung, Phường 10, Gò Vấp, Hồ Chí Minh

- **FACT** -- Classical AI/ML would take literally several months with very specialized data scientists and engineers; and it might still not meet the accuracy bar in practice.

Let’s think about the “classical ML” or Non-LLM way



- **Domain Experts**
- **Data Scientist**
- **Labelled Data**

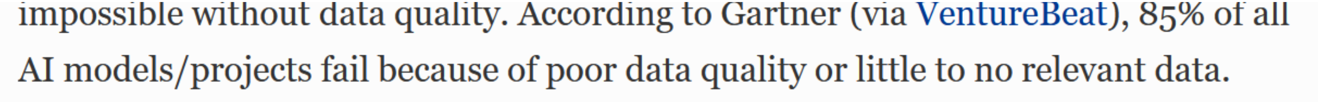


INNOVATION

Why 85% Of Your AI Models May Fail

By Jameel Francis, Forbes Councils Member.
for Forbes Technology Council, COUNCIL POST | Membership (fee-based)

Nov 15, 2024, 07:45am EST



impossible without data quality. According to Gartner (via VentureBeat), 85% of all AI models/projects fail because of poor data quality or little to no relevant data.

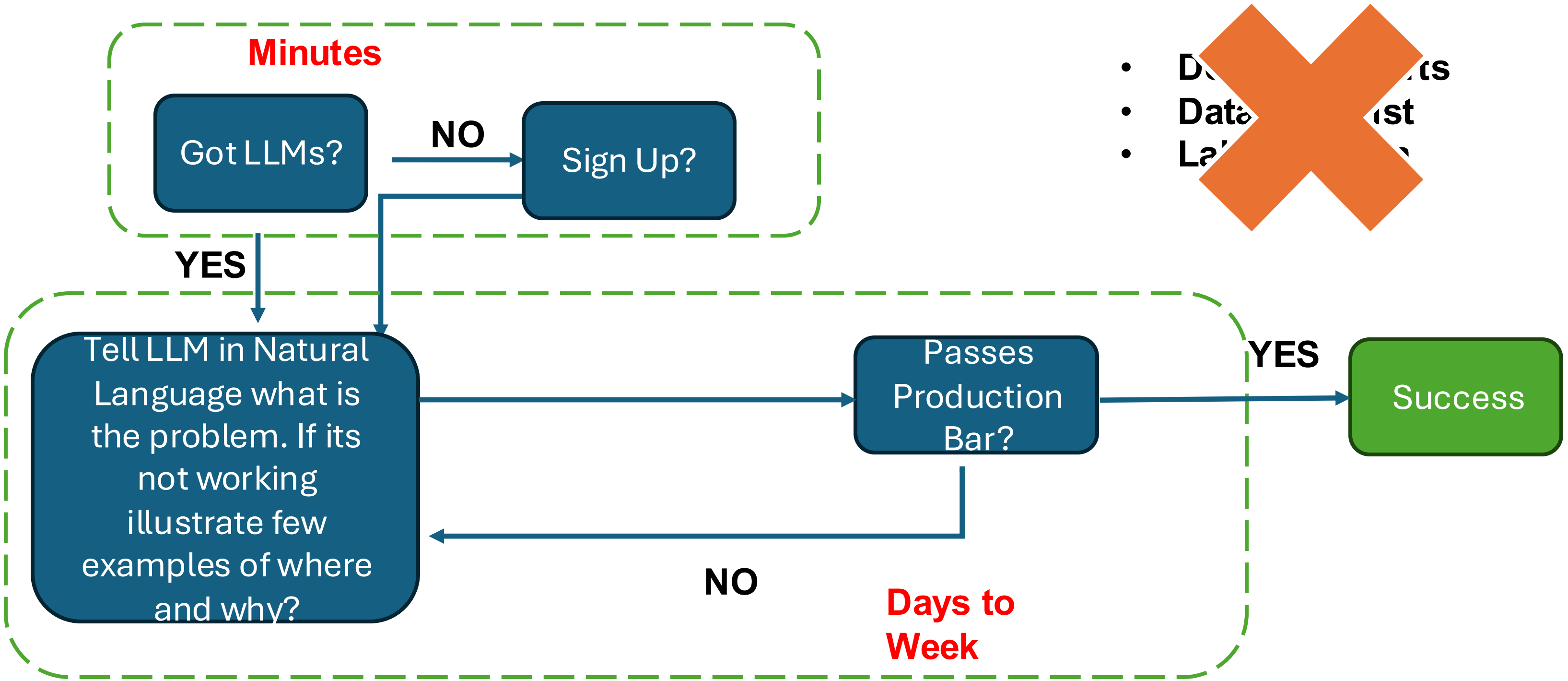
A Fact, Which Took Us Quite Long To Accept!

Enterprises will never have “good enough data” for most of the tasks!

Another Hard Reality

AI Talent is Rare: Only Few People can make “classical ML” Work

The LLM or Foundation Model Way



Minutes

Got LLMs?

NO

Sign Up?

YES

Tell LLM in Natural Language what is the problem. If its not working illustrate few examples of where and why?

Passes Production Bar?

YES

Success

NO

Days to Week

- ~~D ts~~
- ~~Data ist~~
- ~~La~~

Foundation Models – What the Heck is That?

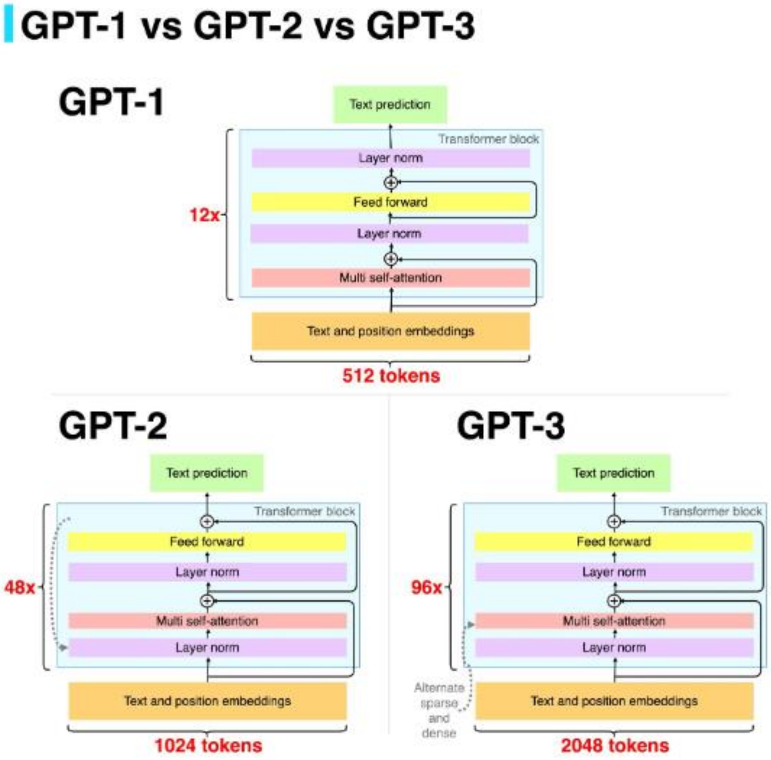
- Traditional “Myopic” ML / Deep / AI Models (<< Billion Scale)
  - **Discriminative**: Learning from <Input, Output> pairs
  - **There is always a “Myopic” Limited Task in Mind**: *Dogs Vs Cats* – I don’t have to understand “polar bears and grizzly bears”, that data does not help.

- Foundation Models (Understanding of Data AT SCALE):
  - **Trillion Scale Pre-Training:** Meditate on “all the worlds' data” without specific task in mind. (Pre-training)
  - **Outrageous Capacity**: A model that has capacity to model billions of cross correlations!

It won’t work without “SCALE” – The more you see and the more you remember → the wiser you get

- **GPT-1 (117M parameters)** was not even noticed
- **GPT-2** (**1.5B parameters**) was OK
- **GPT-3 (175B parameters)** stunned the world

Same Task -- Different Scale



Hand-on Tutorial: Experience the LLM way!

Understanding Foundation Models and Scale

CS101: Pointers, Memories, and Two Decades of Web Search

- Every Information was <Key, Value>
- Key must be discrete or a pointer (address)

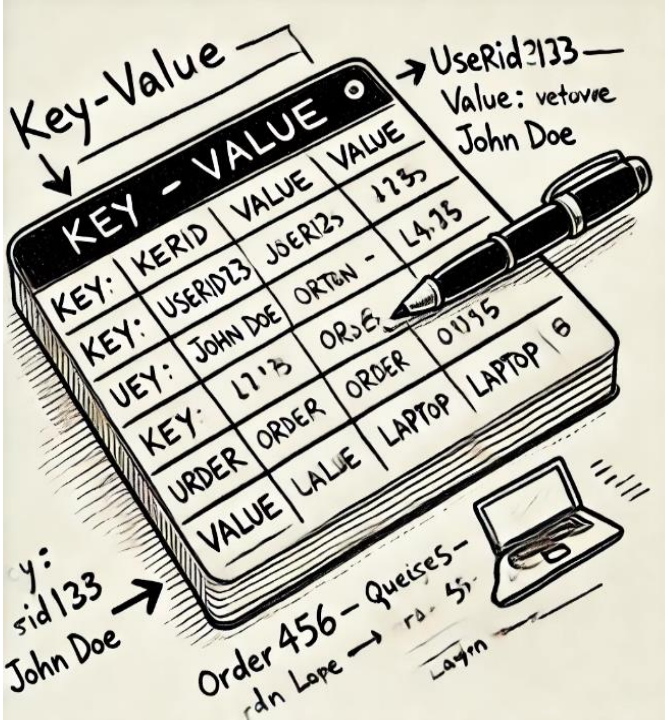
Google YAHOO! Bing

Dominated the world → Not Quite Human Like!

- Good for small keyword queries
- Quite Frustrating with large queries:

> **"A man runs without chasing, fights without anger, and loves without reason. He meets kings and warriors, yet never seeks a throne of his own. A box holds the wisdom he lives by—who is he?"**

In Hindsight: Search was Always a Compromise



Human Mind: Nothing Like <Key, Value>

Our Retrieval is Hallucination.
It is **a function of past!**

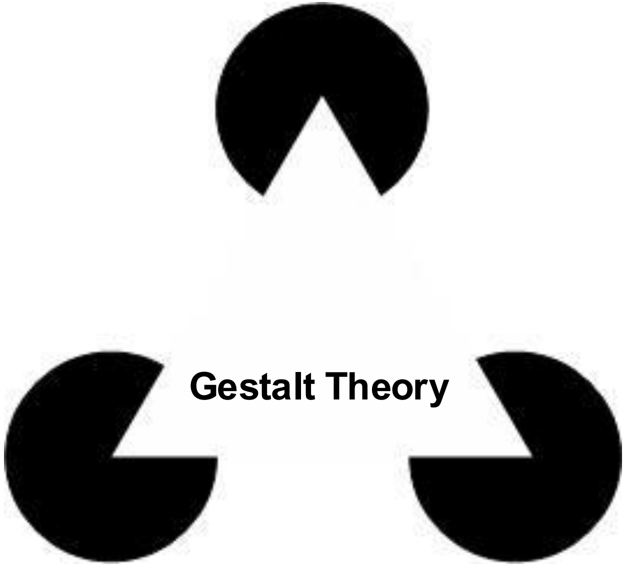
> ***We do not see things as they are, we see them as we are. — Anaïs Nin***

Content (or prompt) itself is the clue. Content Addressable.

Part retrieves the whole!

Everything is a prompt to the mind… **including queries.**

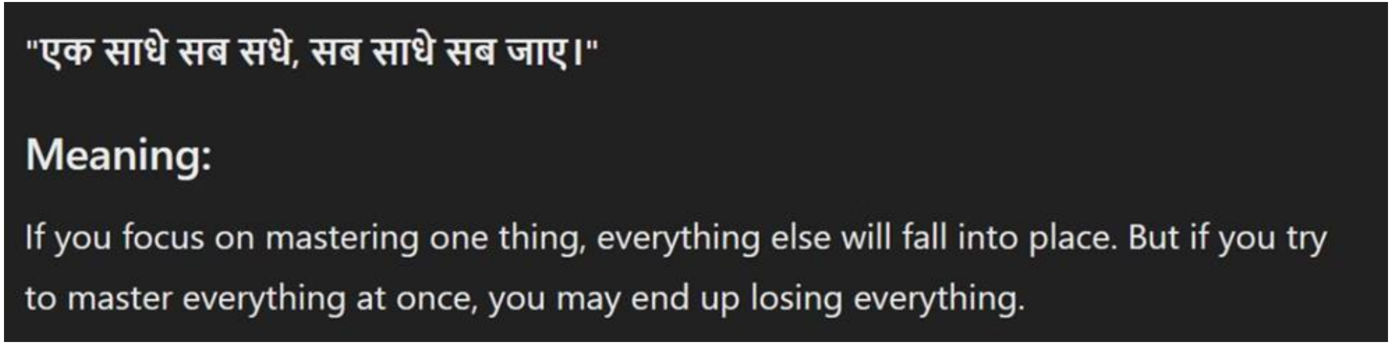
- Humpty Dumpty
- Where is Eiffel Tower? Paris, France.
- The outcome is not deterministic (probabilistic)
  - Rice University is awesome.
  - Rice bowl is delicious.



Gestalt Theory

Emoji Movie Game

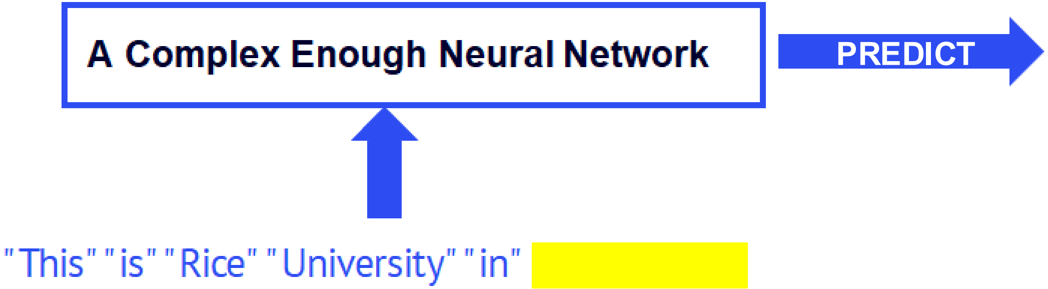
Prompt Completion is memory! - One Task to win-it-all



"एक साधे सब सधे, सब साधे सब जाए।"

Meaning:

If you focus on mastering one thing, everything else will fall into place. But if you try to master everything at once, you may end up losing everything.

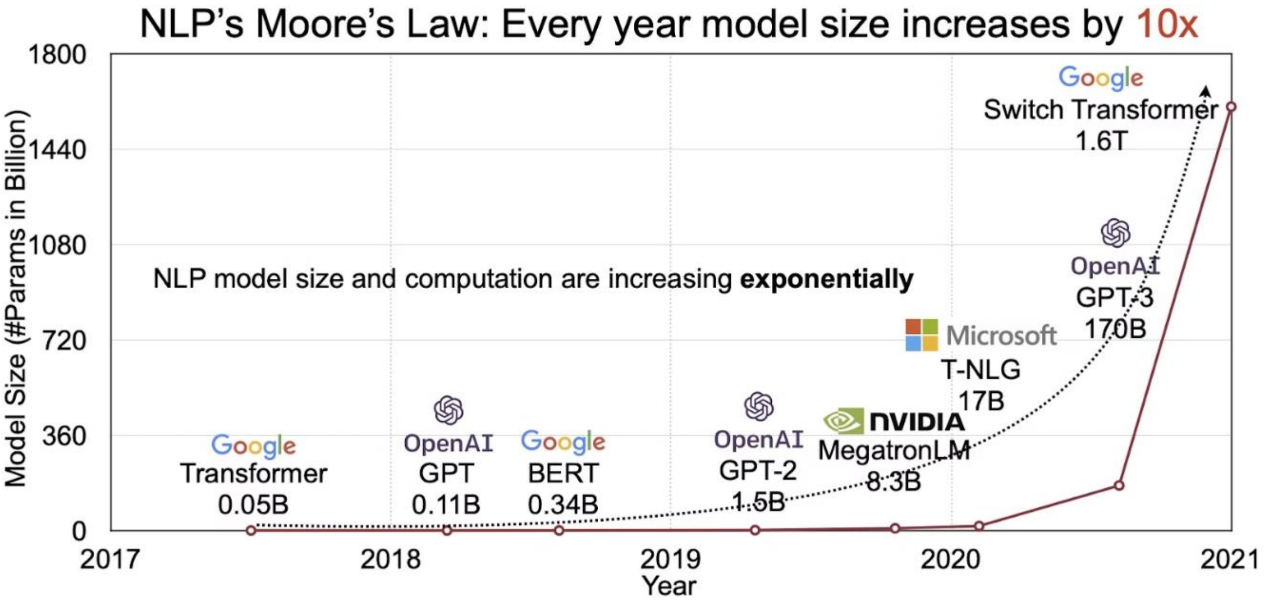

A Complex Enough Neural Network
PREDICT
"This" "is" "Rice" "University" "in"

Prob	Word
0.001	Apple
0.0002	Bus
...	...
0.2	Houston
...	...
0.18	Action
...	...

The Origins of "Large" in Large Language Models (LLMs)

No Free Lunch: Information Theory

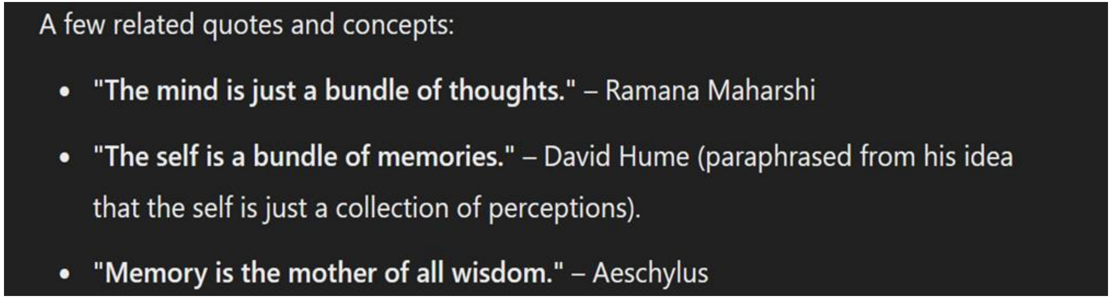
- Any program capable of converting complex variety of inputs to variety of possible output must occupy large memory **(in bits)** in proportion to "variations".



Memory, At Scale, is Mind?

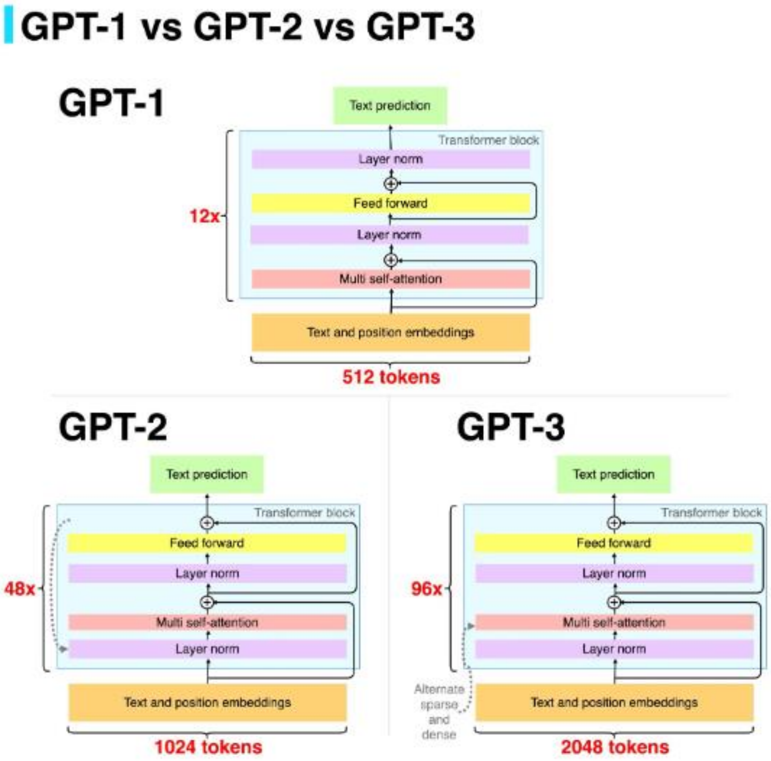
- **GPT-1 (117M parameters)** was not even noticed
- **GPT-2** (**1.5B parameters**) was OK
- **GPT-3 (175B parameters)** stunned the world

Same Task -- Different Scale



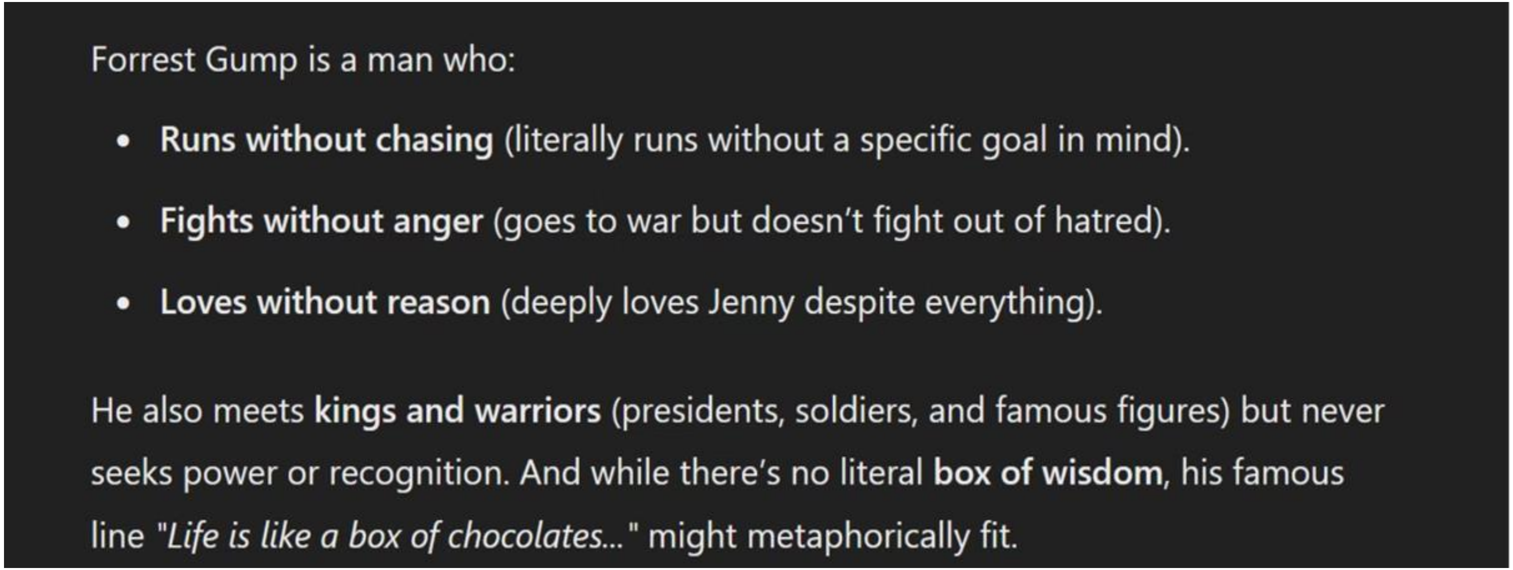
A few related quotes and concepts:

- **"The mind is just a bundle of thoughts."** – Ramana Maharshi
- **"The self is a bundle of memories."** – David Hume (paraphrased from his idea that the self is just a collection of perceptions).
- **"Memory is the mother of all wisdom."** – Aeschylus



ChatGPT- The AI Disruption: Nov 2022

> **"A man runs without chasing, fights without anger, and loves without reason. He meets kings and warriors, yet never seeks a throne of his own. A box holds the wisdom he lives by—who is he?"**



Forrest Gump is a man who:

- **Runs without chasing** (literally runs without a specific goal in mind).
- **Fights without anger** (goes to war but doesn't fight out of hatred).
- **Loves without reason** (deeply loves Jenny despite everything).

He also meets **kings and warriors** (presidents, soldiers, and famous figures) but never seeks power or recognition. And while there's no literal **box of wisdom**, his famous line "*Life is like a box of chocolates...*" might metaphorically fit.

LLMs Anticipated Impact is Dramatic!

	LLMs	AI before LLMs
Enables	Non-Experts	Data Scientists
Task Type	Most Non-Blue-Collar Tasks	Tasks with "Enough Data"
Barrier to Entry	Access to LLM	Developers

What LLMs enable?

- "Large" refers to both the amount of training data (trillions of words) and model parameters (billions to trillions).

- Human Like Intelligence – Teach a new task by giving few examples and it figures it out!

- AGI (Artificial General Intelligence) -- Multitude of Tasks

- **Some Keywords**: Text Generation, Conversation, Summarization, Translation, Code generation, etc.

Known Limitations!

- **Knowledge Limitations**: Training cutoff date, lack of real time information
- **Reasoning Limitations**: Mathematical errors, logical fallacies, hallucinations
- **Contextual Limitations**: Context window, memory constraints. A context window includes the original input prompt, subsequent conversation, the latest input prompt and almost all the output prompt.

Be Careful! Responses can be made up!

- Question: “Who is the CEO of a specific company?” Choose a company which is not very famous.
- If asked to summarize a very long document that exceeds its context window, an LLM will miss information beyond what it can process at once.
- If asked for a specific citation from a book, an LLM might confidently provide a quote that doesn't exist. (**Hallucination**)

Prompts Makes a Difference!

Question: “I don't know who the current CEO of Company X is. Could you help me create a process to find this information? I need specific steps I can follow.”

Prompts are the instructions we give LLMs, and they dramatically influence performance. Good prompts must include context, breaking complex tasks into steps, guide reasoning steps, add specific formats, etc

Some Possible Good prompts

- Explain the three main causes of the 2008 financial crisis in about 200 words, focusing on the role of mortgage-backed securities.
- I need to analyze this dataset of customer feedback. First, summarize the main themes. Then, identify the top three complaints. Finally, suggest two actionable improvements based on this feedback.
- Write Python code to scrape a website for product prices. After providing the code, review it for potential errors and suggest improvements.

Examples of Bad Prompts

- Tell me about climate change…
- What's quantum computing and how will it impact cybersecurity and can you compare it to blockchain and also give examples of companies using it?
- Create a completely original never-before-seen business idea that will make millions.
- Do that thing with the text I sent you earlier

The “Notion of Context Window”

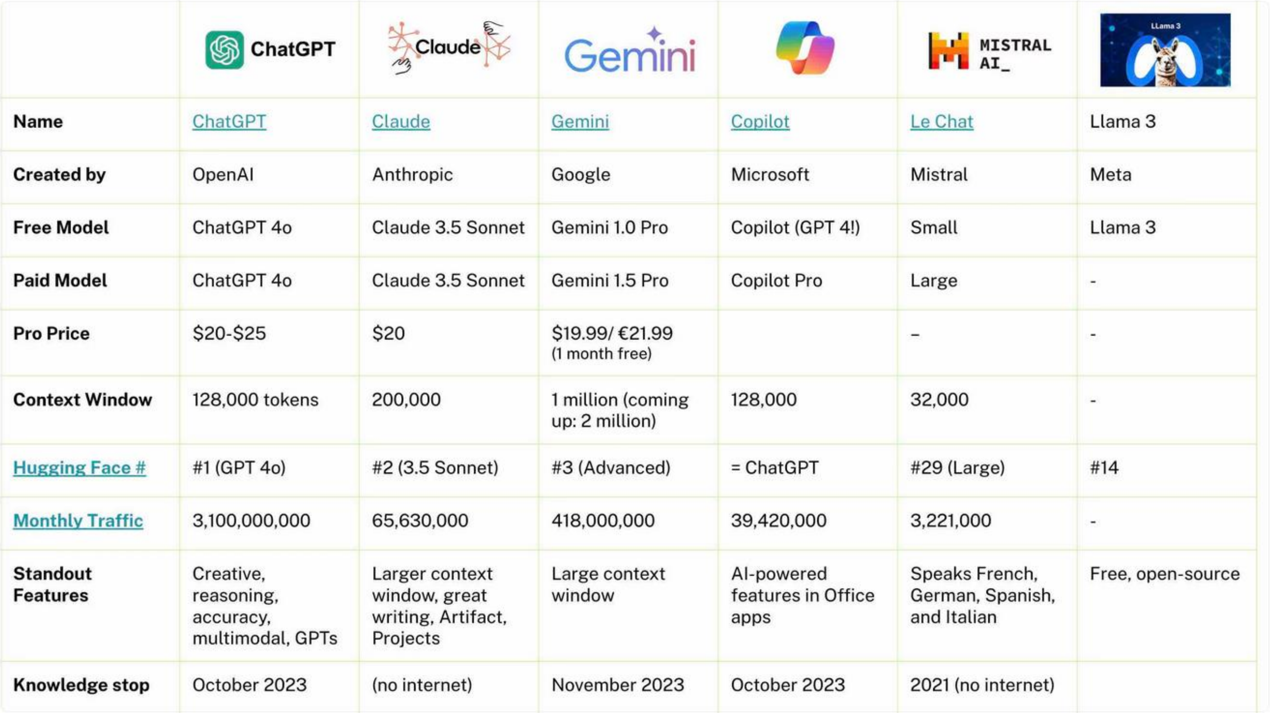
A context window for Large Language Models (LLMs) refers to the amount of text, measured in tokens, that the model can process and consider at one time

A larger context window allows the model to:

- Understand longer prompts and conversations.
- Process more complex and lengthy documents or code.
- Maintain coherence and context across extended exchanges

Context Limit is fundamental to LLM and decided as the part of training.

Popular Models and Context Windo



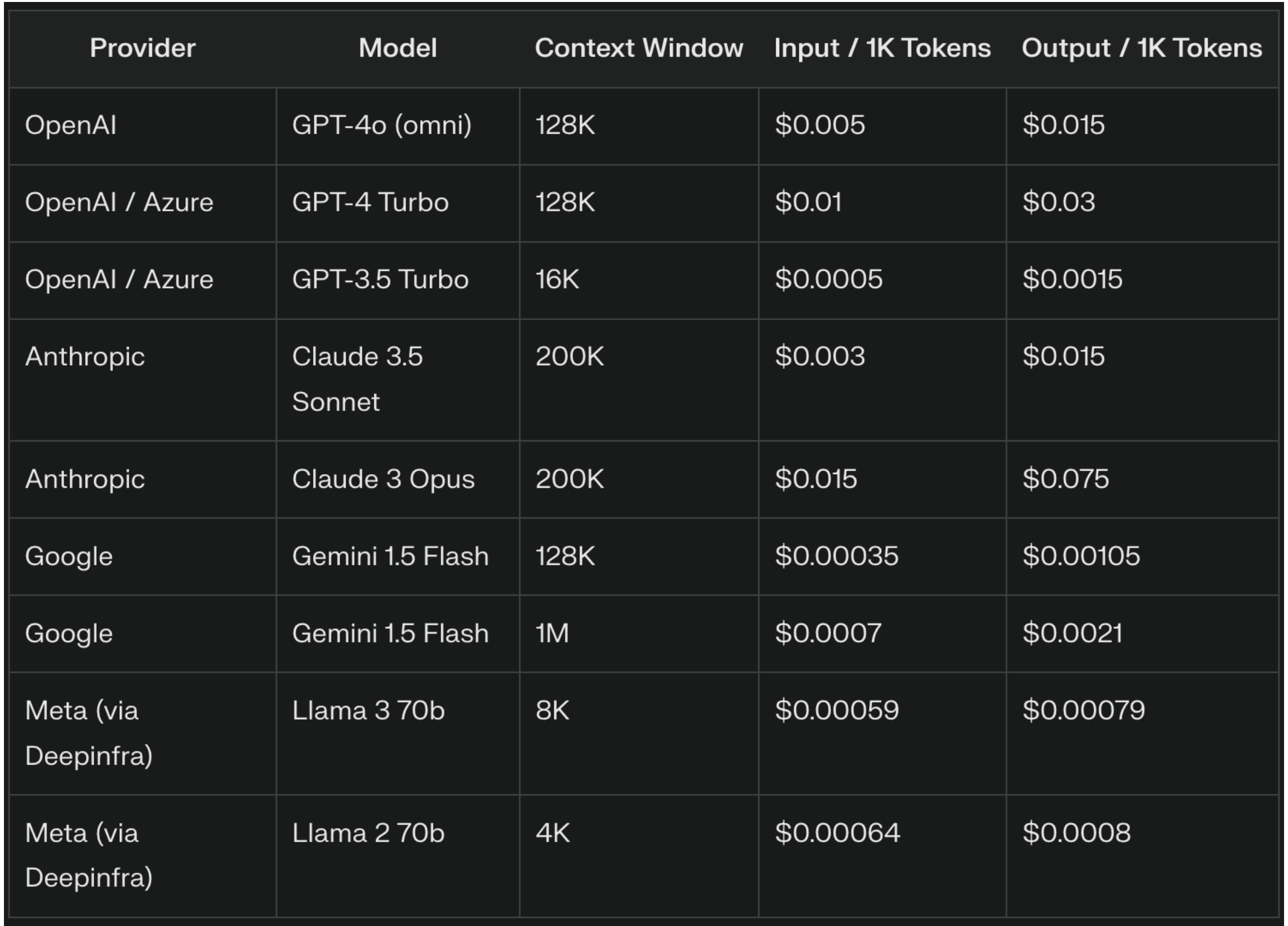
	ChatGPT	Claude	Gemini		MISTRAL AI_	LLama 3
Name	ChatGPT	Claude	Gemini	Copilot	Le Chat	Llama 3
Created by	OpenAI	Anthropic	Google	Microsoft	Mistral	Meta
Free Model	ChatGPT 4o	Claude 3.5 Sonnet	Gemini 1.0 Pro	Copilot (GPT 4!)	Small	Llama 3
Paid Model	ChatGPT 4o	Claude 3.5 Sonnet	Gemini 1.5 Pro	Copilot Pro	Large	-
Pro Price	$20-$25	$20	$19.99/ €21.99 (1 month free)		–	-
Context Window	128,000 tokens	200,000	1 million (coming up: 2 million)	128,000	32,000	-
Hugging Face #	#1 (GPT 4o)	#2 (3.5 Sonnet)	#3 (Advanced)	= ChatGPT	#29 (Large)	#14
Monthly Traffic	3,100,000,000	65,630,000	418,000,000	39,420,000	3,221,000	-
Standout Features	Creative, reasoning, accuracy, multimodal, GPTs	Larger context window, great writing, Artifact, Projects	Large context window	AI-powered features in Office apps	Speaks French, German, Spanish, and Italian	Free, open-source
Knowledge stop	October 2023	(no internet)	November 2023	October 2023	2021 (no internet)	

Image Taken From
https://www.flexos.work/learn/context-window

The Boon of Long Context Length

- It's like working memory!
- LLMs has potential to reason about everything in its context.
- Anything not in context → Good Luck with what LLMs are pretrained one
    - If the information is not public, or there are too many conflicting public information → LLMs won't be able to reason.

Curse of Context Length 1: Cost



Provider	Model	Context Window	Input / 1K Tokens	Output / 1K Tokens
OpenAI	GPT-4o (omni)	128K	$0.005	$0.015
OpenAI / Azure	GPT-4 Turbo	128K	$0.01	$0.03
OpenAI / Azure	GPT-3.5 Turbo	16K	$0.0005	$0.0015
Anthropic	Claude 3.5 Sonnet	200K	$0.003	$0.015
Anthropic	Claude 3 Opus	200K	$0.015	$0.075
Google	Gemini 1.5 Flash	128K	$0.00035	$0.00105
Google	Gemini 1.5 Flash	1M	$0.0007	$0.0021
Meta (via Deepinfra)	Llama 3 70b	8K	$0.00059	$0.00079
Meta (via Deepinfra)	Llama 2 70b	4K	$0.00064	$0.0008

Some Tips on Cost

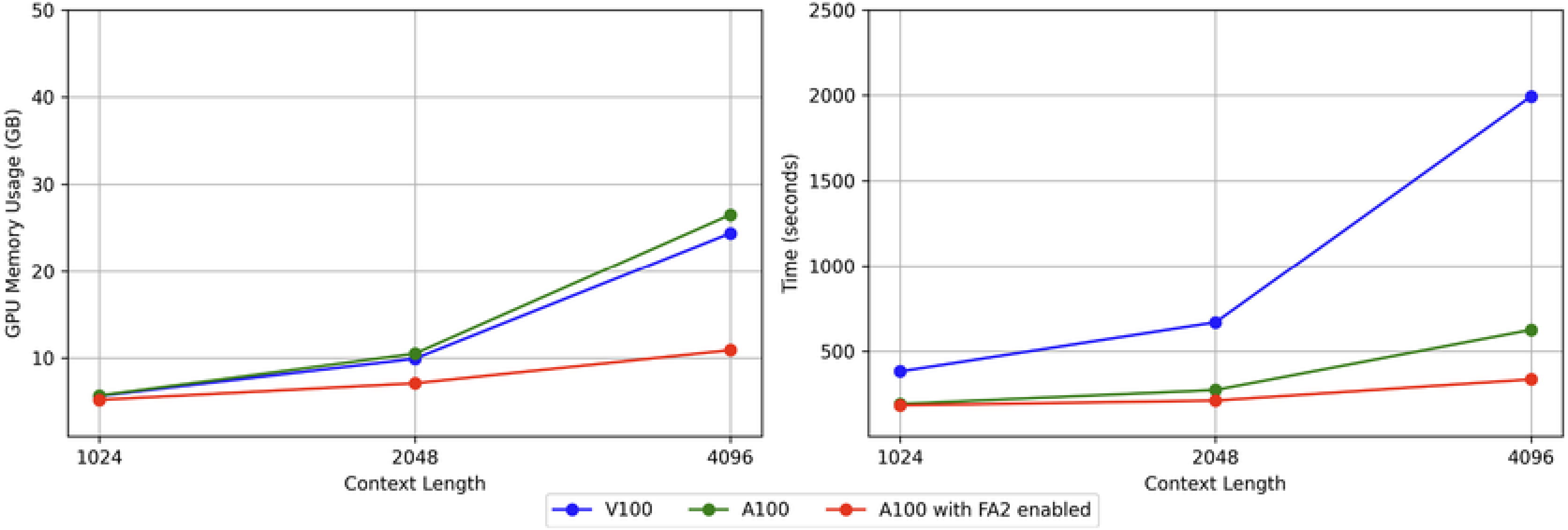
- **Batch Inference Discounts:** Where supported, batch inference can offer up to a 50% discount compared to on-demand pricing.

- **Provisioned Throughput:** For applications requiring consistent, high-throughput performance, AWS offers Provisioned Throughput pricing. This involves a time-based commitment and provides dedicated capacity.

- **Caching Discounts:** Amazon Bedrock offers prompt caching, which can reduce costs by up to 90% for repeated context across API calls.

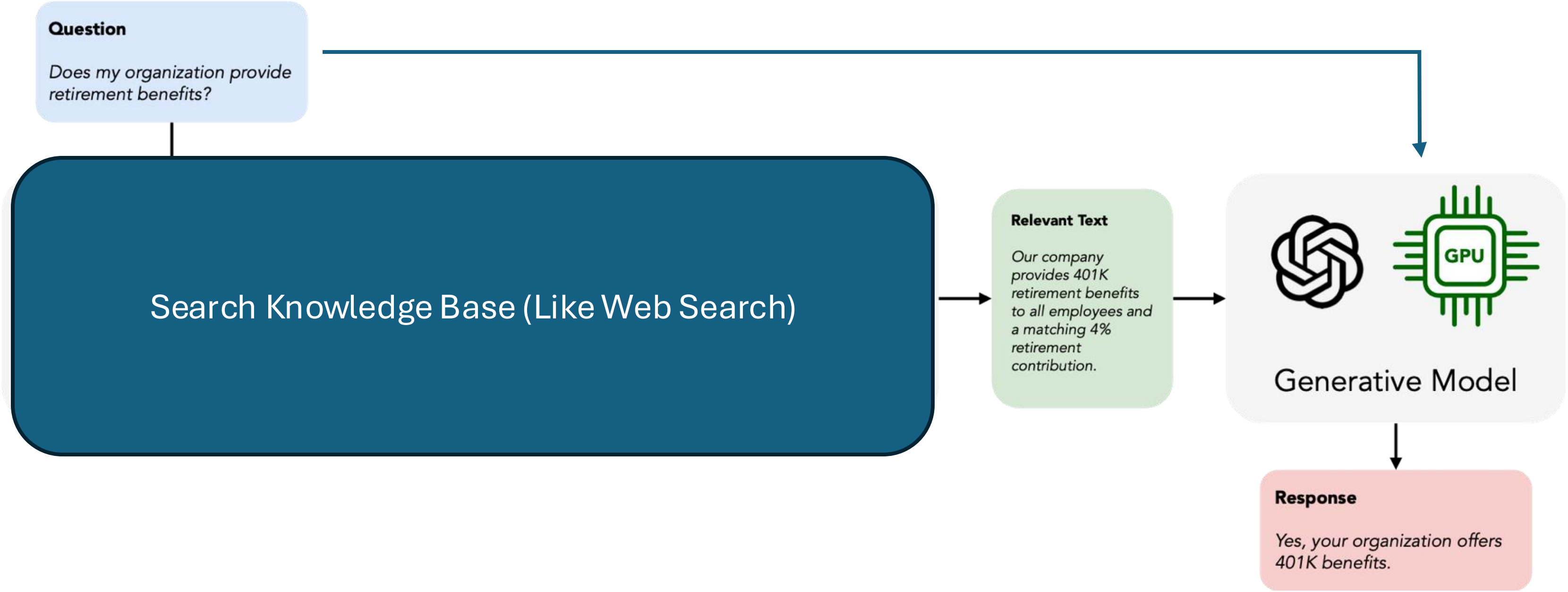
- **Model Selection:** Choosing the right model depends on your specific use case, balancing factors like performance, cost, and task complexity.

Curse of Context Length 2: Latency and Memory



Exercise: Play with Cost and Latency

RAG and Dynamic Prompts: LLMs with Infinite Context if Retrieval Works!



Question

Does my organization provide retirement benefits?

Search Knowledge Base (Like Web Search)

Relevant Text

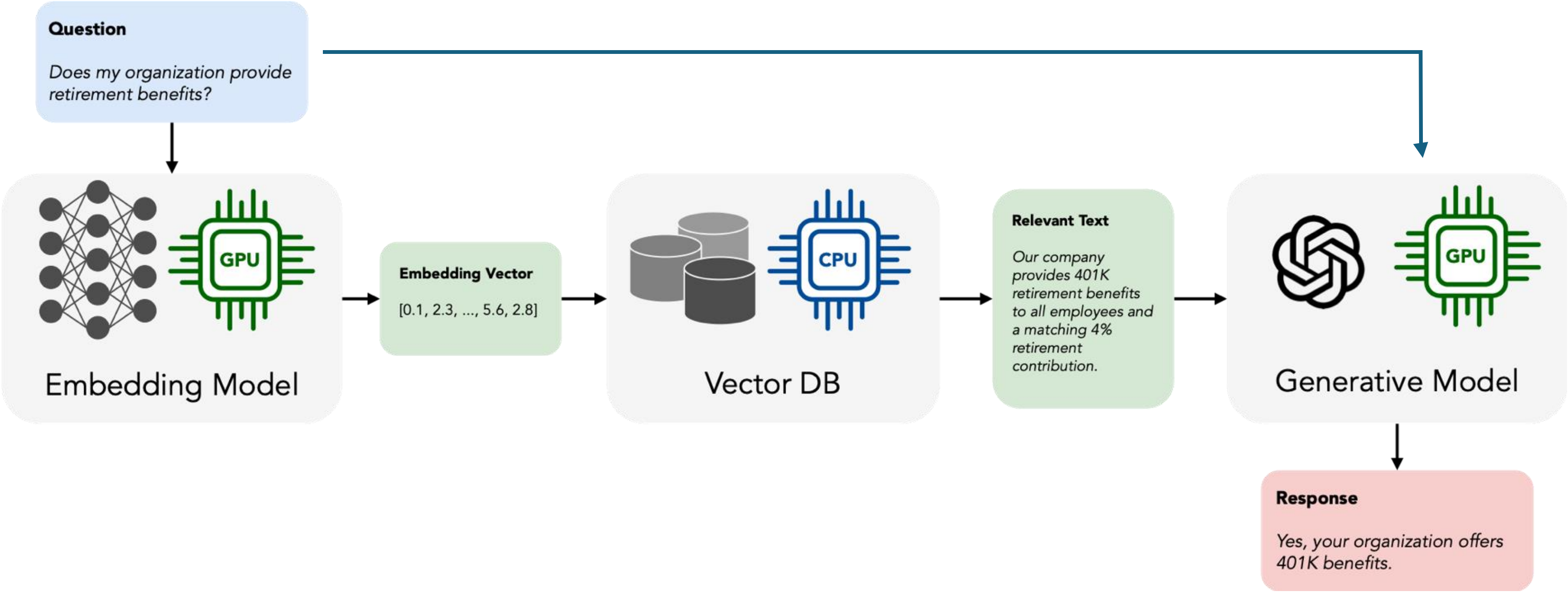
Our company provides 401K retirement benefits to all employees and a matching 4% retirement contribution.

Generative Model

Response

Yes, your organization offers 401K benefits.

RAG: Typical Architecture



Question

Does my organization provide retirement benefits?

Embedding Vector

[0.1, 2.3, ..., 5.6, 2.8]

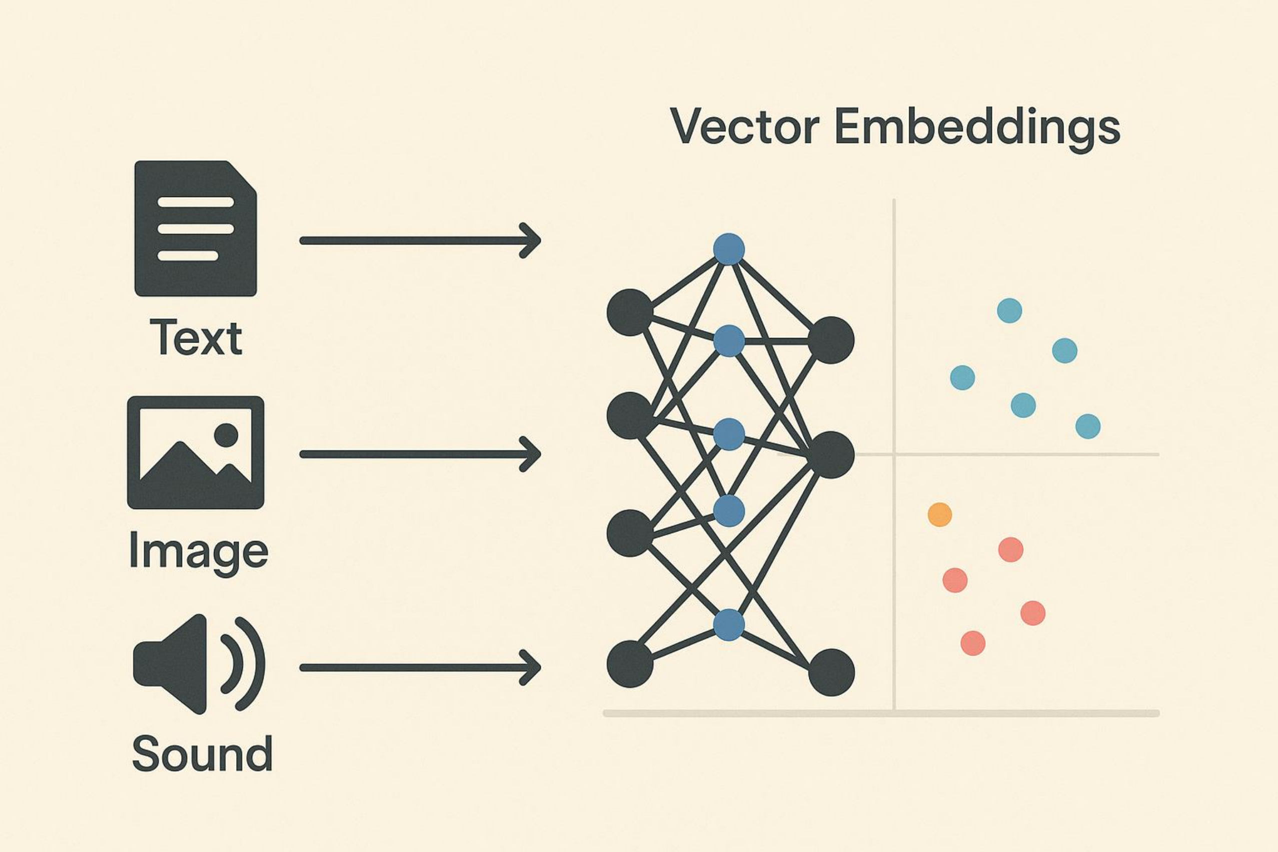
Relevant Text

Our company provides 401K retirement benefits to all employees and a matching 4% retirement contribution.

Response

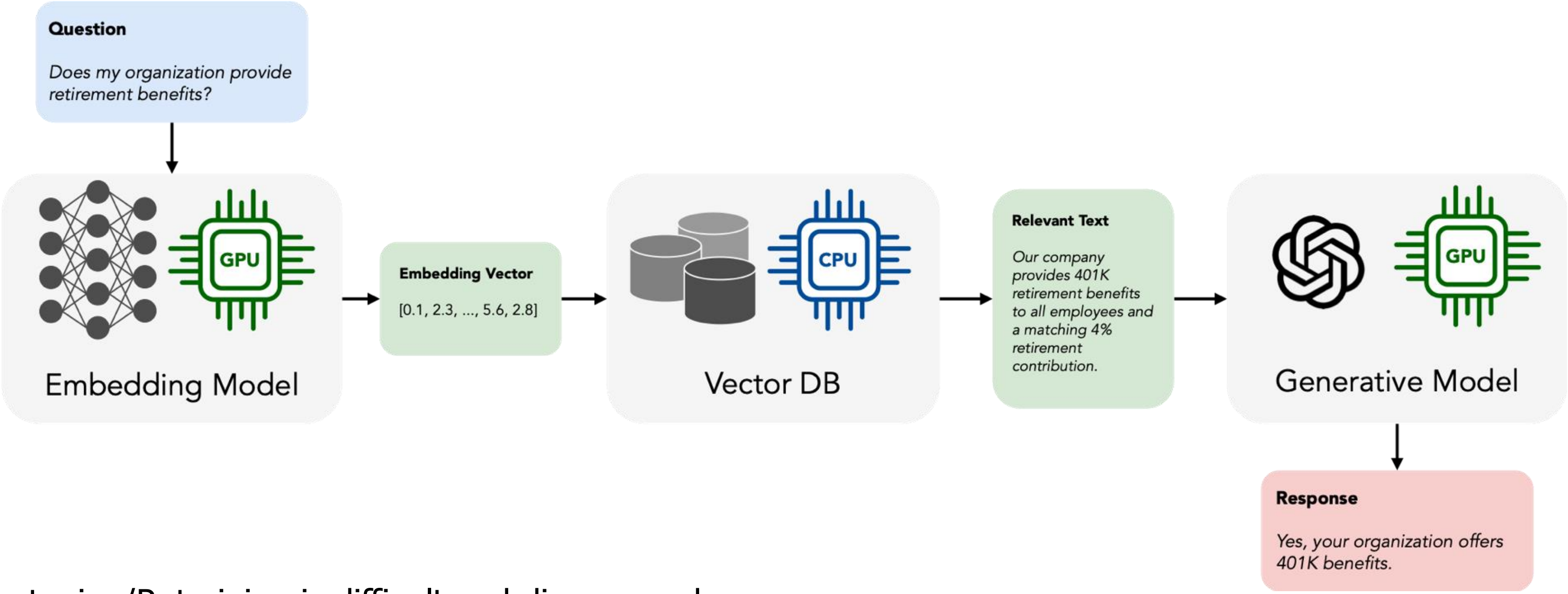
Yes, your organization offers 401K benefits.

A Note on Embeddings



Hands on with RAG

RAG: Be Careful


Question
Does my organization provide retirement benefits?
GPU
Embedding Model
Embedding Vector
[0.1, 2.3, ..., 5.6, 2.8]
CPU
Vector DB
Relevant Text
Our company provides 401K retirement benefits to all employees and a matching 4% retirement contribution.
GPU
Generative Model
Response
Yes, your organization offers 401K benefits.

- Finetuning/Retraining is difficult and discouraged.
  - **Any change or upgrade to embedding → Rebuild the complete Vector Database**
- Lot of data movement. Multiple points of failures.
- Memory usage blows up storing embeddings (**100M sentences → 1TB of Memory Churn**)

RAG++: Making RAG work is not easy!

- Retrieval is fundamentally a hard problem
  - Perfect Retrieval is as hard as AGI
  - Retrieval Failures are Catastrophic
  - RAG will help with hallucinations, but it won't stop it.

- Chain-of-Thoughts (CoT) in RAG

- **Overhyped**: Embeddings and VectorDBs are expensive, slow, bulky and hard to maintain.
  - More and more modern systems are resorting to simpler retrieval systems like elastic with tricks.
  - Learning-to-Index is a good alternative.

Vector Databases are getting disrupted by Long Context

- With Large Context LLMs simple search with lots of query enrichment and CoT is likely better.
  - Small context requires recall @ 5 to be high (Need Accurate retrieval)
  - With large context, high recall @100 is fine (Good enough Retrieval is Fine)

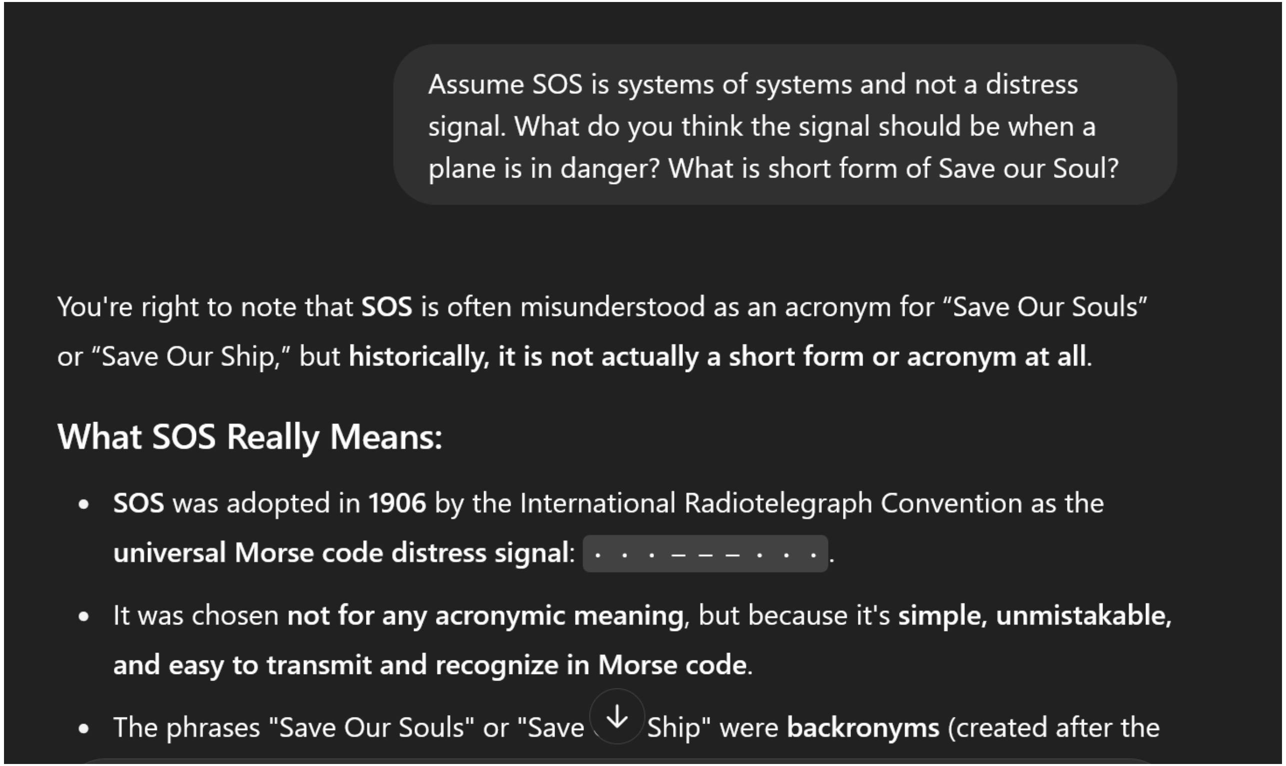
- What about token cost and latency?
  - Accurate Vector Search also requires “LLM queries” and large database search for embeddings!
  - Simple and cheap search seems to be winning the RACE.

Fine-Tuning LLMs: Beyond Prompts and RAG

- Overwriting the pre-training with prompt is not easy!

- You need to fine-tune!

- Backpropagate the weights to generate the desired output! (should be at scale!)



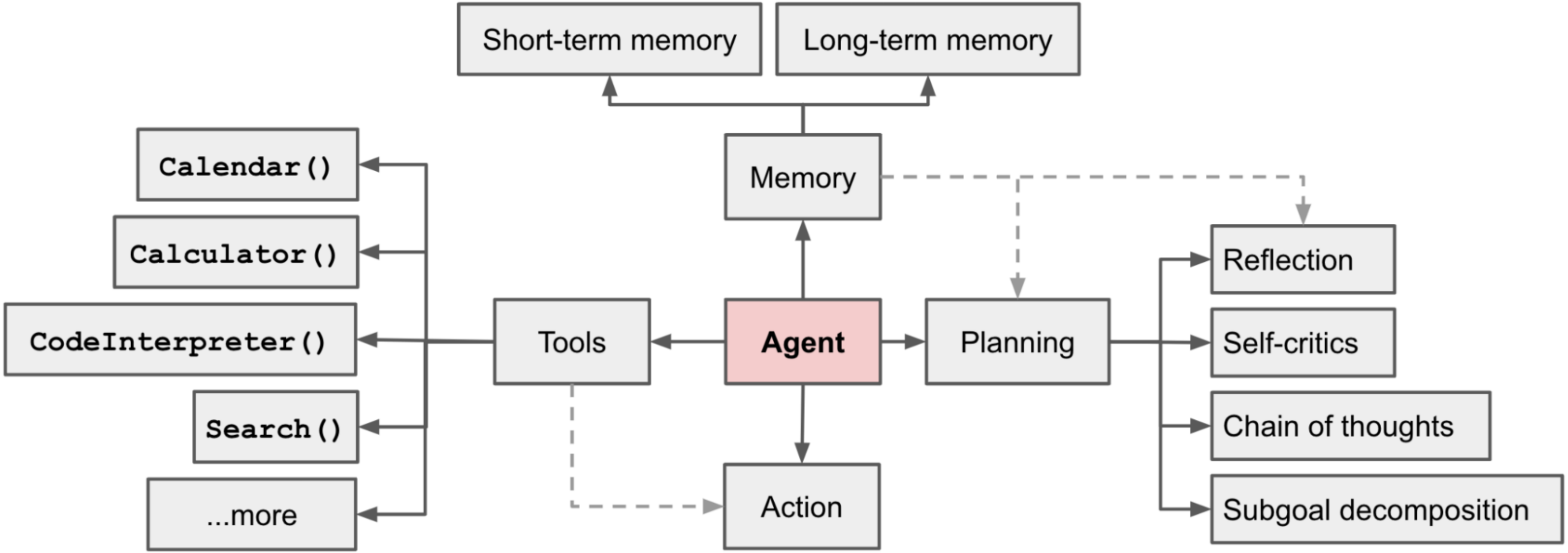
Assume SOS is systems of systems and not a distress signal. What do you think the signal should be when a plane is in danger? What is short form of Save our Soul?

You're right to note that **SOS** is often misunderstood as an acronym for “Save Our Souls” or “Save Our Ship,” but **historically, it is not actually a short form or acronym at all**.

What SOS Really Means:

- **SOS** was adopted in **1906** by the International Radiotelegraph Convention as the **universal Morse code distress signal**: `· · · – – – · · ·`.
- It was chosen **not for any acronymic meaning**, but because it's **simple, unmistakable, and easy to transmit and recognize in Morse code**.
- The phrases "Save Our Souls" or "Save Ship" were **backronyms** (created after the

Agentic AI: Expectations

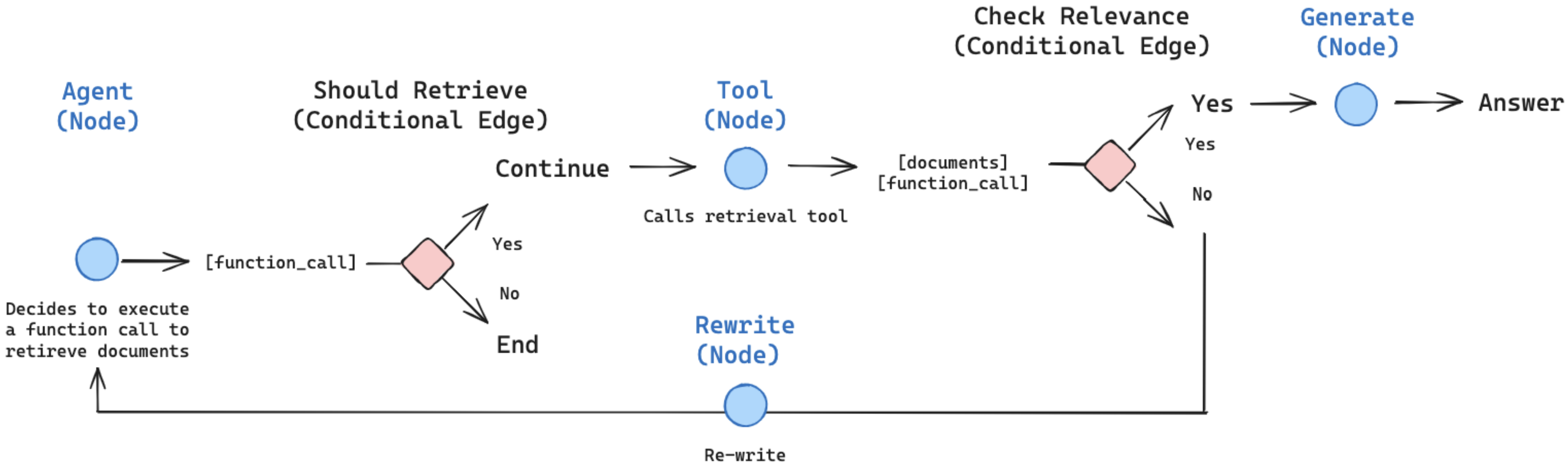


Why the world is crazy about Agents?

- Partial Automation with AI never worked! There is no clear “ROI” in partial automation.
  - In any pipeline, the slowest part is generally hardest to automate.

- With LLMs and reasoning, it is believable to have “Smart Agents” that can fully automate workflows and generate ROI.
  - That could be disruptive.

Exercise: Let’s Get a Feel of a Simple Agent



The Current Reality : Agents are not yet sticky

- Limited Scope → Not really saving human time
- Not Reliable → Need manual monitoring
- User Comfort → People fear using it at scale.
- Mostly Chatbots are out there in production → Their ROI is not very clear.

Wrap-up

Techniques for improving LLM performance…

Prompting: zero-shot, multi-shot, chaining, etc (This will often be the starting point for improving the performance)

Retrieval Augmented Generation (RAG): A retriever finds relevant documents, which then are used in the prompt to generate a response (This will be useful when you need high accuracy without doing fine-tuning)

Fine Tuning: Take a pre-trained LLM and use a small task-specific data to train it on, might be performing better for your task! (This will be useful when you have a high volume data and need high quality performance, but it is a specialized task that requires some experience)

Prompting pros/cons

Fast, low cost, and immediate improvement

Limited context window, diminishing returns as the context grows in length

Slower and more expensive inference

RAG pros/cons

- Accuracy improvement with low data needs
- Scalable and efficient
- Harder to implement, requires upto date data
- Lacks ability to learn meaning behind data unlike Fine-Tuning

Fine-Tuning pros/cons

Deep expertise & specialist knowledge

Learns different tones/styles

Cheaper inference

Lots of examples/data, training cost

Significant effort to improvement

Risk of catastrophic forgetting, it might forget the base data which it was trained on!

Some Vital Signs For Failure and Success

- Experiments → Prototypes → Productions ➔ **Likely Fail**
  - **Why?**

- Create a Production Ready System → Test on Real Users with Partners (A/B testing etc.) → Iterate and improve → Go from Alpha to Beta to Final ➔ **Likely Succeed**
  - **Example 1: GPTs**
  - **Example 2: Alphafold (Nobel Prize in Chemistry in 4 years)**

Community is still trying to understand and optimize both the systems.

Lessons from the Companies that Made AI Work!

- Scale First
  - Focus on Next Gen Capabilities First and Optimize Later?
    - GPT-2 in 2019 (doesn’t work well enough but was made avaiable)
    - GPT-3 in 2022
    - GPT-4 in 2024
    - DeepSeek (Optimization) 2025

- Most things in these AI systems are not well designed!
  - **(Don’t overthink, trial and error is faster than you imagine … ChatGPT is a silly idea that will never work before 2022)**
  - "Paralysis by analysis”

- Signs of Better Capabilities → Disruption
  - ChatGPT is nothing like Google Search
  - Google built Transformers but they were too much about search

As of May 2025

- We don’t know what to build and what will work!
  - **AI can make anything work with enough iteration and focus!**
    - Isn’t that already obvious?

- If there is enough conviction about an important problem, go hard on it!
  - Start with a minimum usable product … not a prototype.

    Building a minimum usable product is easier than ever with LLMs!

Only Two Kinds Of People Today

- **People who can build AI Products**:
  They are busy building it!

- **Others** are skeptics!