

LLM Adventure!

with your host, John Rauser
Director of Engineering, Cisco Systems

Welcome to the Adventure



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Confidential

Agenda

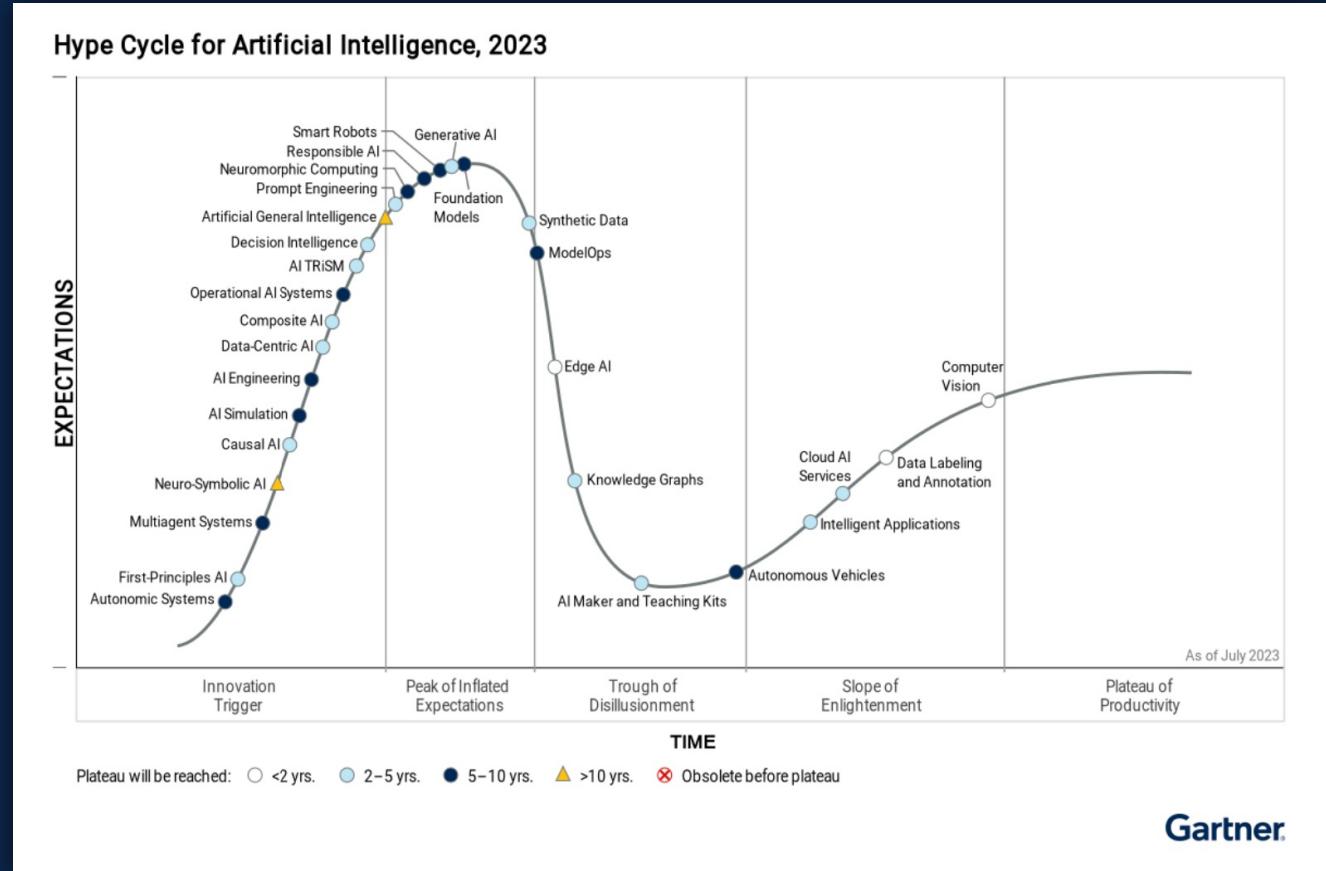
1. Prologue
2. The Council (LLM Product Patterns)
3. The Castle (LLM Delivery Model)
4. The Moat (Enterprise AI Platform)
5. Epilogue

September 2023



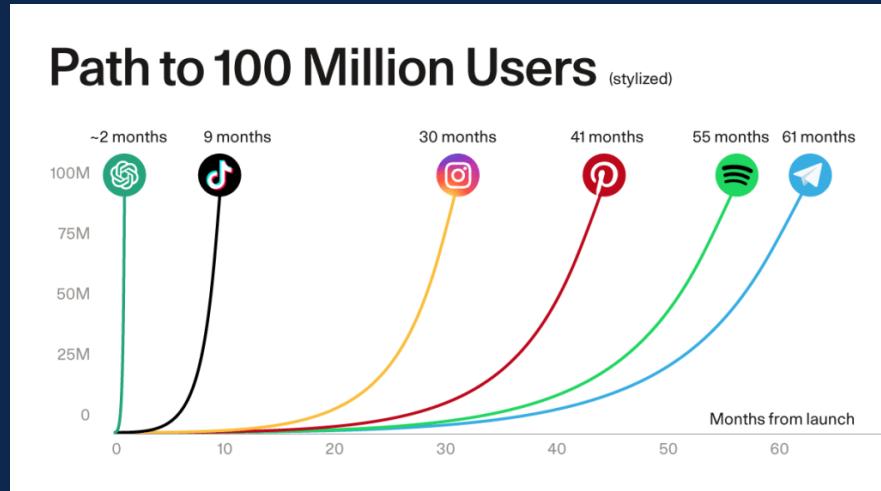
Based on the experiences of dozens of AI leaders across Cisco's broad portfolio of products and services

The Hype is on



<https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle>

The Impact is Real



The Fall of Stack Overflow

Over the past one and a half years, Stack Overflow has lost around 50% of its traffic. This decline is similarly reflected in site usage, with approximately a 50% decrease in the number of questions and answers, as well as the number of votes these posts receive.

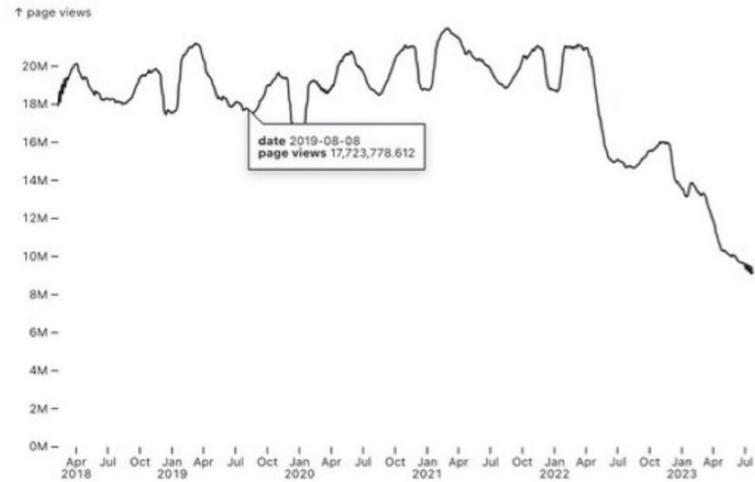
The charts below show the usage represented by a moving average of 49 days.

The data is from <https://stackoverflow.com/site-analytics>

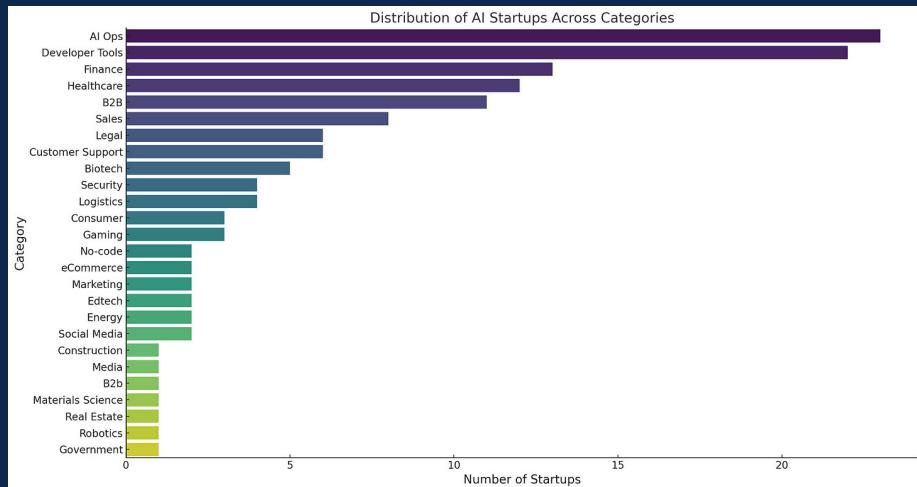
Window Size (days)

Traffic

Select one series for traffic



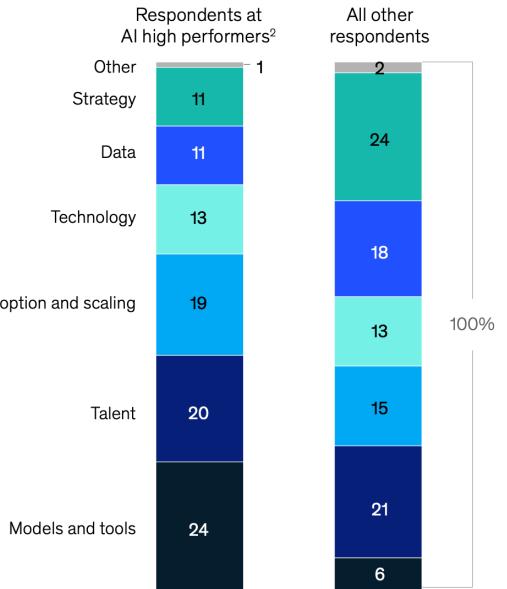
Ops is the Key



<https://www.ignorance.ai/p/5-lessons-from-139-yc-ai-startups>

Models and tools pose the biggest AI-related challenge for high performers, while strategy is a common stumbling block for others.

Element that poses the biggest challenge in capturing value from AI, % of respondents¹



Note: Figures do not sum to 100%, because of rounding.

¹Asked only of respondents whose organizations have adopted AI in at least 1 function.

²Respondents who said that at least 20 percent of their organizations' EBIT in 2022 was attributable to their use of AI. For respondents at AI high performers, n = 49; for all other respondents, n = 792.

Source: McKinsey Global Survey on AI, 1,684 participants at all levels of the organization, April 11–21, 2023

McKinsey & Company

<https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year>

Cisco and AI



Predictive AI

Maturity in 2017

Prediction, Recommendation

Predictive AI focuses on analyzing existing data that can be used **for predictions and automation**

Generative AI

Maturity in 2022

Creation, Reinvention

Generative AI focuses on learning a representation of artifacts from data and using it to **generate original artifacts** based on predicted sequences of information from a given prompt

Cisco's AI-driven portfolio today

Generative AI

Predictive AI

Networking

- Enabled by Silicon One Scheduled Fabric Ethernet Solution

In development:

- AIOps
- Assurance

Long-term:

- Virtual Network Assistant (incl network config. generator)

- Network Management
- Network Assurance
- Network Deployment
- Anomaly Detection
- Change Automation
- Predictive Internet

Security

- Email text threat analysis

In development:

- Simplifying Security (incl. Policy Assistant for Firewall)
- Sophisticated attack prevention (incl. SOC assistant)
- Secure use of LLMs (e.g., DLP)

- Statistical modeling
- Threat Analysis
- Endpoint, Policy, and Trust Analysis
- Advanced Endpoint Malware Protection/Prediction
- Cloud-based infra, application, and data protection

Collaboration

- Inclusive Meetings (transcription, live translation)

- Real-time speech enhancements

In development:

- Webex GPT
- Webex Assistant

Long-term:

- Intelligent Contact Center

- Noise Removal

- Peopler Focus (video)
- Personal Insights, Team Insights

In development:

- Cinematic Meetings, Meeting Zones, low-bandwidth video

Observability

- Model observability

In development:

- Prompt interface
- AI assistant for summarization

- Statistical Modeling

- Baselining
- Anomaly Detection
- Intelligent Automation

Long-term:

- Data Tagging

CX/Sales

In development:

- AI framework and generated test cases
- AI assisted automation

Long-term:

- Proactive protection of devices and services
- Proactive sustainability

- TAC Support Assistant

- Sentiment Prediction
- SW Anomaly Detection
- HW Failure Prediction
- Text Scraping
- Statistical Modeling
- Change Automation
- Problem Diagnosis

How do we build these products?



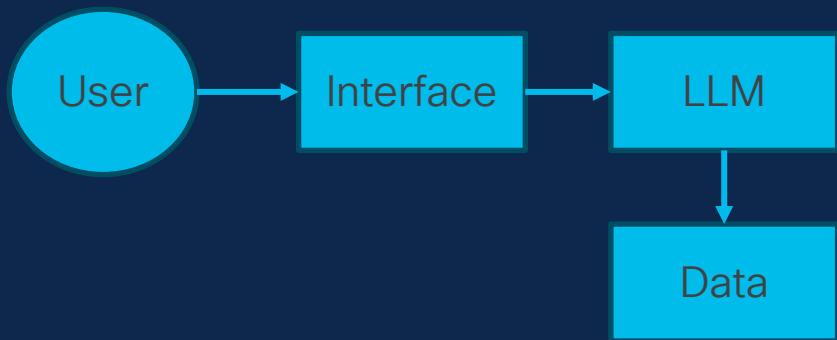


The Council



The Vizier

- The adviser: ask me a question and I will do my best
- Just keep asking until you are satisfied



Examples

- Conversational Agents
- Chatbots
- Co-Pilots

How many ports does the 6536 Fabric Interconnect (FI) have?

The Cisco UCS 6536 Fabric Interconnect has a total of 36 ports.

References

- <https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs6536-fabric-interconnect-ds.html>
- <https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/cisco-ucs-6536-fabric-interconnect-spec-sheet.pdf>
- <https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-b-series-blade-servers/6400-specsheet.pdf>
- <https://www.cisco.com/c/en/us/products/collateral/interfaces-modules/unified-computing-system-adapters/ucs-vic-15000-series-ds.html>
- <https://www.cisco.com/c/en/us/products/collateral/interfaces-modules/unified-computing-system-adapters/datasheet-c78-741130.html>

Cisco Nexus 9508 Switch
Specifications Overview

Series Cisco Nexus 9000 Series Switches
Overview Product Overview
Product Type Data Center Switches
Status Available Order
Release Date 05-NOV-2013
Product ID View All PIDs



Documentation Downloads Community Specifications

Search This Product's Documentation

Saved Content

You can now save documents for easier access and future use. Saved documents for this product will be listed here, or visit the [My Saved Content](#) view and manage all saved content from across Cisco.com.

Log in to see your Saved Content.

Latest Security Notices

23-Aug-2023 Security Advisory: Cisco Nexus 3000 and 9000 Series Switches IS-IS Protocol Denial of Service Vulnerability

23-Aug-2023 Security Advisory: Cisco NX-OS Software TACACS+ or RADIUS Remote Authentication Directed Request Denial of Service Vulnerability

23-Aug-2023 Security Advisory: Cisco Nexus 3000 and 9000 Series Switches SFTP Server File Access Vulnerability

10-Apr-2023 Field Notice: FN72006 - N9K-C9508-FAN/N9K-C9516-FAN: Fan Flow Direction Reversed - Hardware Upgrade Available

07-Mar-2023 Field Notice: FN - 72495 - Nexus 9000 Series Switch Stuck in Boot Loop Due to "policyitem hap" Reset After Conversion to ACI - Workaround

22-Feb-2023 Security Advisory: Cisco NX-OS Software CLI Command Injection Vulnerability

Thread ← Ian Adams @NetEng_Ian ...

OMG, OMG, OMG. This is not a joke. Our Cisco TAC engineer's name is Sherlock Holmes.We are in VERY good hands. He's on the case!

I don't think I've ever been this happy working on a Cisco issue.

SR [REDACTED] - [REDACTED] keeps rebooting without notice

 Sherlock Holmes <[REDACTED]@cisco.com>
To: [REDACTED] attach@cisco.com

Hello [REDACTED],
My name is **Sherlock** and I have ownership of your Service Request: [REDACTED]. I am sen of contact.
From your problem description, it sounds like the device might have crashed
Please do the following:

12:10 PM · Apr 27, 2022

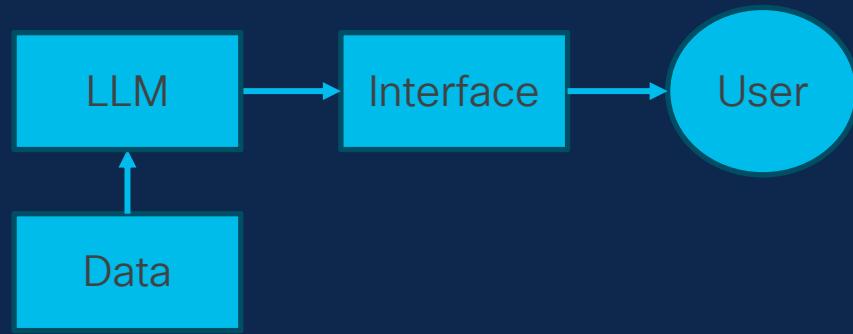
38 Retweets 6 Quotes 352 Likes 8 Bookmarks

Cisco Nexus 9000 Series Switches Overview (PDF - 2.19 MB)
Accelerating 40G Adoption (PDF - 2 MB)
Cisco Nexus 9000 Series ThousandEyes Integration At-a-Glance
Creating a global, zero-downtime network environment (PDF - 1 MB)

Data Sheets
Most Recent
[Cisco Nexus 9300-GX2 Series Fixed Switches Data Sheet](#) 28-Aug-2023
[Cisco Nexus 9300-FX2 Series Switches Data Sheet](#) 25-Aug-2023
[Cisco Nexus 9300-EX and EX 24 Port Series Switches Data Sheet](#) 11-Aug-2023

The Judge

- I'll take a complex situation and boil it down for you
- I think I'm right, but it's up to you to decide



Examples

- Support Cases, Sales Cases Summarizer
- Incident Report Drafting
- CVE Summarizer
- Publication Analysis
- Meeting notes
- Root Cause Analysis

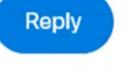
Best regards,
Cisco Virtual Engineer

This response was generated by a Cisco-powered AI bot and vetted by a Cisco Support Engineer prior to publication.

This is part of a monitored experiment to see if the bot can help answer questions alongside community members. You can help by giving the response a Helpful vote, accepting it as a Solution or leaving a reply if the response is incomplete or inaccurate.

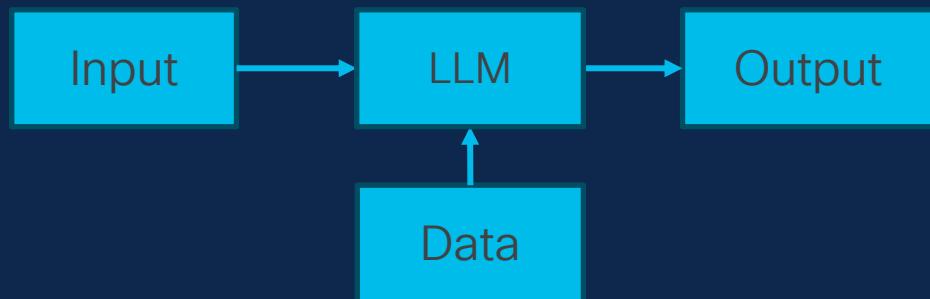
 1 Helpful

 Share

 Reply

The General

- The doer, give me a job and I will get it done
- The job needs to be done correctly,
there is a lot on the line!



Examples

- Query authoring, translating
- Threat Labeling
- Content Categorization
- “Chain of Thought”

* Langchain, Sematic Kernel

6.1 ***Does this model generate output that results in a consequential decision affecting a user?**

An example of a model that does not make consequential decisions affecting human users is a model to optimally route data in a data center.

Risk: If the model affects natural persons, then the model will inherently be higher risk because most legal regimes directed at ML/AI models are concerned with risks for users. In that case, the model should be carefully evaluated for bias, even if the specific industry or domain is not directly regulated. Unwanted bias can cause both legal liability when it runs afoul of antidiscrimination law and also reputational risk, as numerous public examples of biased algorithms have demonstrated.

Justify Your Answer

B I U S Ix " ≡ ≡ ≡ ≡ ≡ ≡ ≡ ≡ ≡ ≡

Enter Justification here.

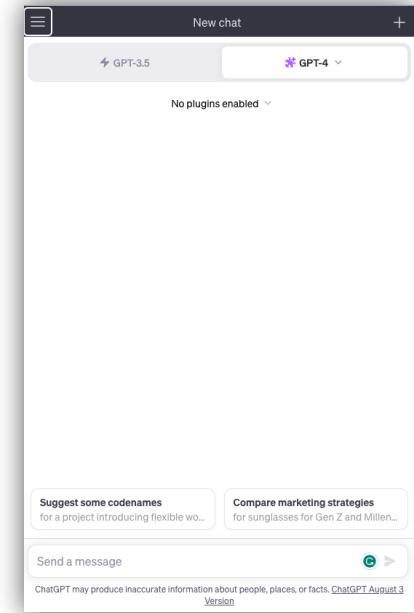
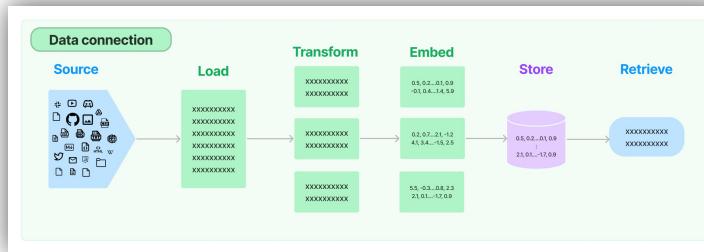
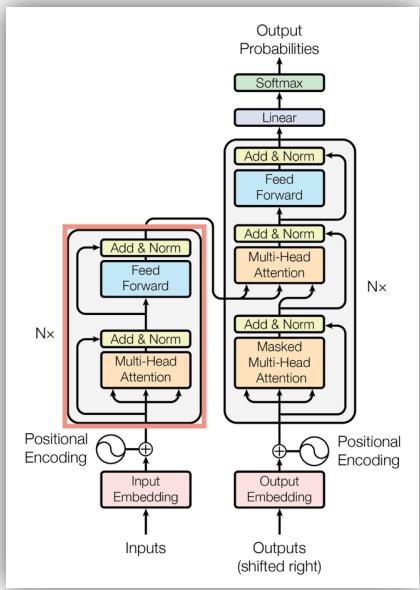
 0  0 



The Castle



The Main Elements

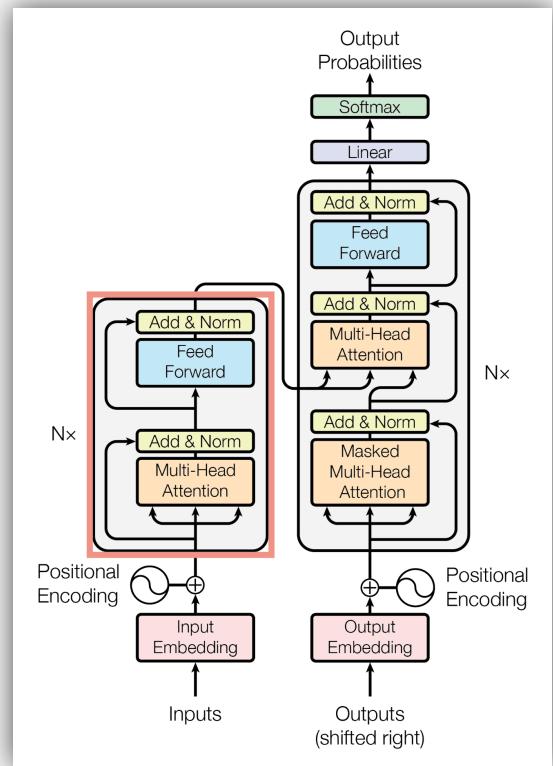


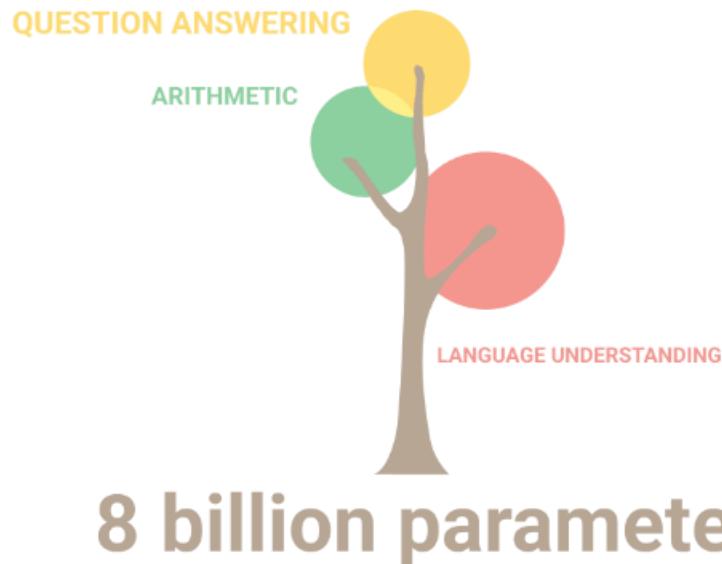
Model

Data

Interface

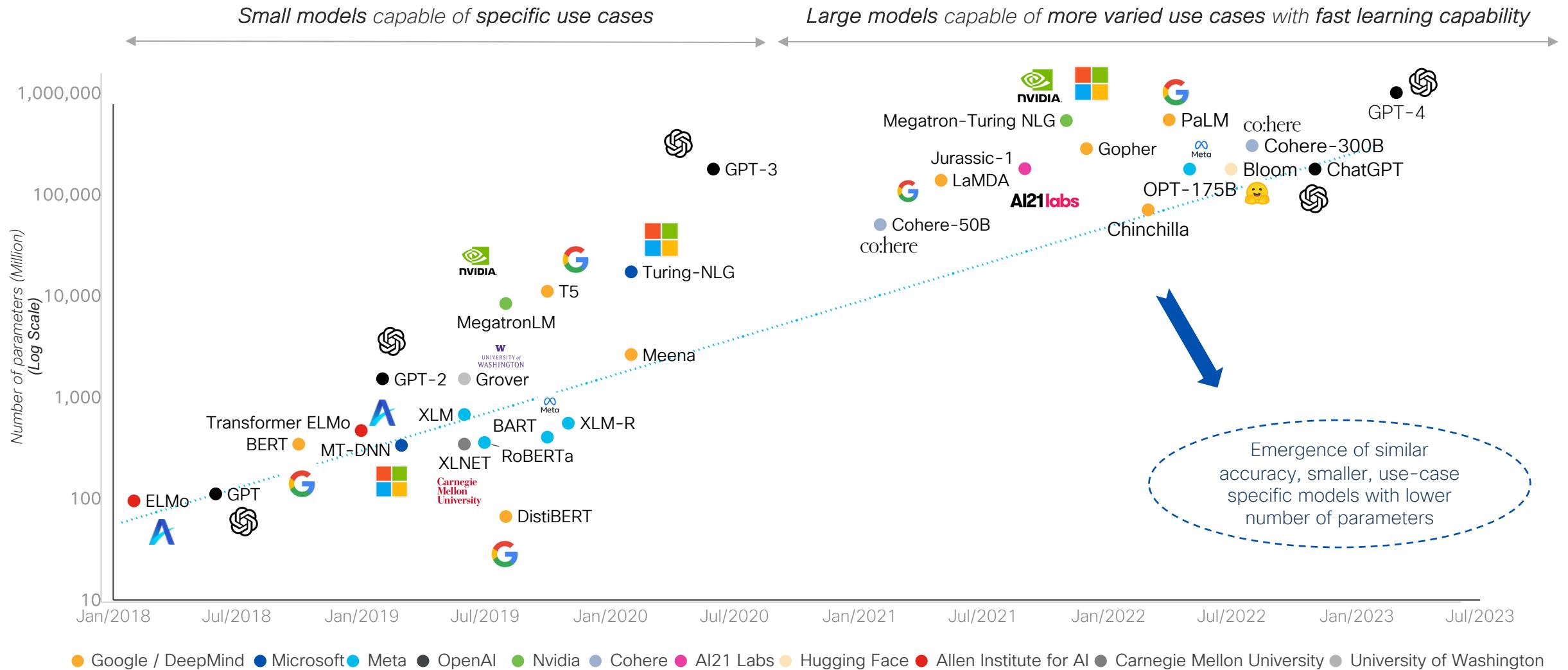
The Model





<https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html>

Generative AI models have crossed the threshold of tech viability



Llama 2

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture: Pretraining Tokens: 2 Trillion	Data collection for helpfulness and safety: Supervised fine-tuning: Over 100,000
13B	Context Length: 4096	Human Preferences: Over 1,000,000
70B		

<https://ai.meta.com/llama/>

✖ Chatbot Arena ✖ : Benchmarking LLMs in the Wild

| [Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Twitter](#) | [Discord](#) |

Rules

- Chat with two anonymous models side-by-side and vote for which one is better!
- You can do multiple turns of conversations before voting.
- The names of the models will be revealed after your vote. Conversations with identity keywords (e.g., ChatGPT, Bard, Vicuna) or any votes after the names are revealed will not count towards the leaderboard.
- Click "Clear history" to start a new round.

Leaderboard

See [lmsys/chatbot-arena-leaderboard](#) or the 4th tab above on this page.

Battle

Please scroll down and start chatting. The models include both closed-source models (e.g., ChatGPT) and open-source models (e.g., Llama).

Model A

What is the meaning of life?

The meaning of life is a philosophical question that has been debated by thinkers and scholars for centuries. Different perspectives and beliefs exist regarding the purpose and significance of human existence. Some find meaning in religious or spiritual beliefs, others in personal fulfillment, relationships, or contributing to society. Ultimately, the meaning of life can be a subjective and personal interpretation that each individual may define for themselves.

Model A: gpt-3.5-turbo

Model B

What is the meaning of life?

The meaning of life is a question that has puzzled philosophers, theologians, scientists, and many other thinkers for centuries. There are many different perspectives on what constitutes the meaning of life, and there is no one definitive answer. However, here are some possible approaches to understanding the meaning of life:

1. Religious or spiritual perspective: Many people believe that the meaning of life is to fulfill a divine or spiritual purpose. According to this view, life has a higher purpose or destiny that is connected to a deity or a higher power. The purpose of life is to fulfill this divine plan or to follow the will of the higher power.
2. Personal fulfillment: From this perspective, the meaning of life is to find personal fulfillment and happiness. This can involve pursuing one's passions, developing one's talents, and achieving one's goals. The meaning of life is to live a life that is rich in experience, personal growth, and satisfaction.
3. Social or cultural perspective: Some people believe that the meaning of life is tied to one's social or cultural context. According to this view, the meaning of life is to contribute to the greater good of society, to uphold cultural traditions, or to fulfill one's responsibilities to others.

Model B: llama-2-13b-chat

Evals

Measure how well the product is doing and detect regressions



MMLU

Evaluates proficiency across wide variety of domains

Composed of 15,908 multiple-choice questions sourced from academic examinations, university course materials, and specialized texts



HumanEval

Assess functional correctness in algorithmic tasks, tests LLMs on comprehension, reasoning, and algorithmic synthesis

164 hand-crafted programming problems, which include function signatures, docstrings, and unit tests,



Folklore

Community collected tacit tests
Sally's Brothers , etc

Meet Sally ...

Prompt asked:

Sally (a girl) has 3 brothers. Each brother has 2 sisters. How many sisters does Sally have?

Note: Correct answer is 1.

[Back to home](#)

Model	Answer	Latency	Chars / s
Airoboros L2 70B	<p>3 sisters. Here's the reasoning: 1. We know that Sally has 3 brothers. 2. Then we are told that each of her brothers has 2 sisters. 3. Since Sally is a girl, she is one of the sisters for each of her brothers. 4. Therefore, Sally has 3 brothers * 2 sisters (each brother) = 3 sisters.</p>	12643ms	22.62
Alpaca (7B)	6</s>	509ms	9.82
Chronos Hermes (13B)	3 brothers x 2 sisters per brother = 6 sisters. So Sally has 6 sisters.	1940ms	36.6
Claude Instant v1	Sally has 0 sisters. The question provides no information about Sally having any sisters herself.	2805ms	34.94
Claude v1	* Sally has 3 brothers * Each of those 3 brothers has 2 sisters * So in total there are $3 * 2 = 6$ sisters * Since Sally is one of those sisters, and she has 3 brothers, Sally must have $6 - 1 = 5$ sisters.	3338ms	61.11
Claude v1.2	* Sally has 3 brothers * Each of those 3 brothers has 2 sisters * So each brother has 2 sisters, and there are 3 brothers, so there are $3 * 2 = 6$ sisters that Sally's brothers have * Since those are Sally's brothers' sisters, they are also Sally's sisters * Therefore, Sally has 6 sisters in total So in summary, if Sally has 3 brothers and each brother has 2 sisters, then Sally has a total of 6 sisters.	4442ms	91.85
Claude v2	* Sally has 3 brothers * Each of Sally's 3 brothers has 2 sisters * So in total Sally's 3 brothers have $3 * 2 = 6$ sisters * Sally is also a sister to her 3 brothers * So the total number of sisters Sally has is $6 + \text{Sally} = 7$ sisters	4315ms	54
Code Llama Instruct (13B)	6	411ms	4.87

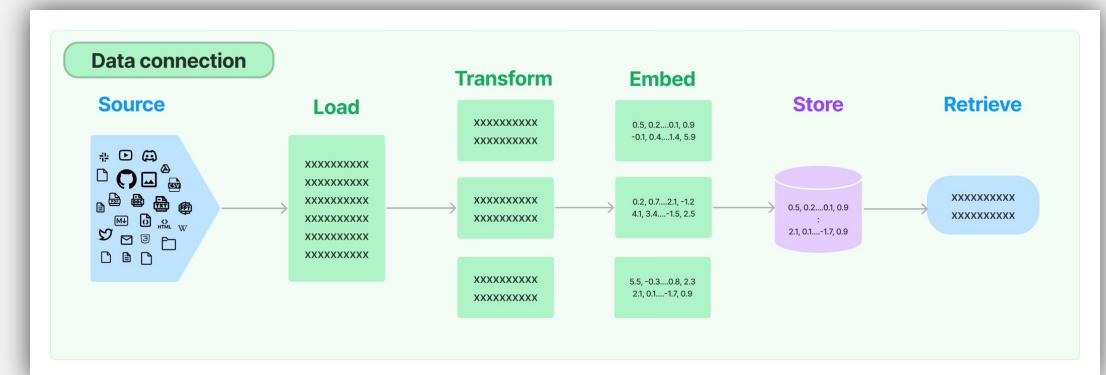
The Future of Foundational Models

- Bigger Models
- Smaller Models
- Reduced Memory
- Multi-language
- Multi-modal
- Inferencing Performance
- Hardware Optimization
- ...



The Data

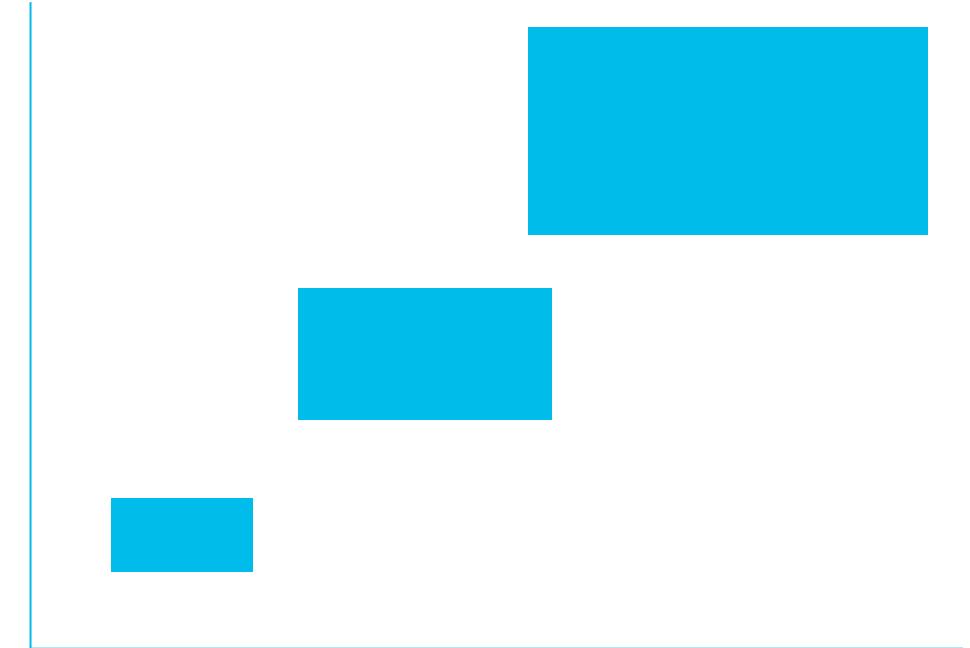
How do we get the data we need into the Model so that we can use it for our specific purpose



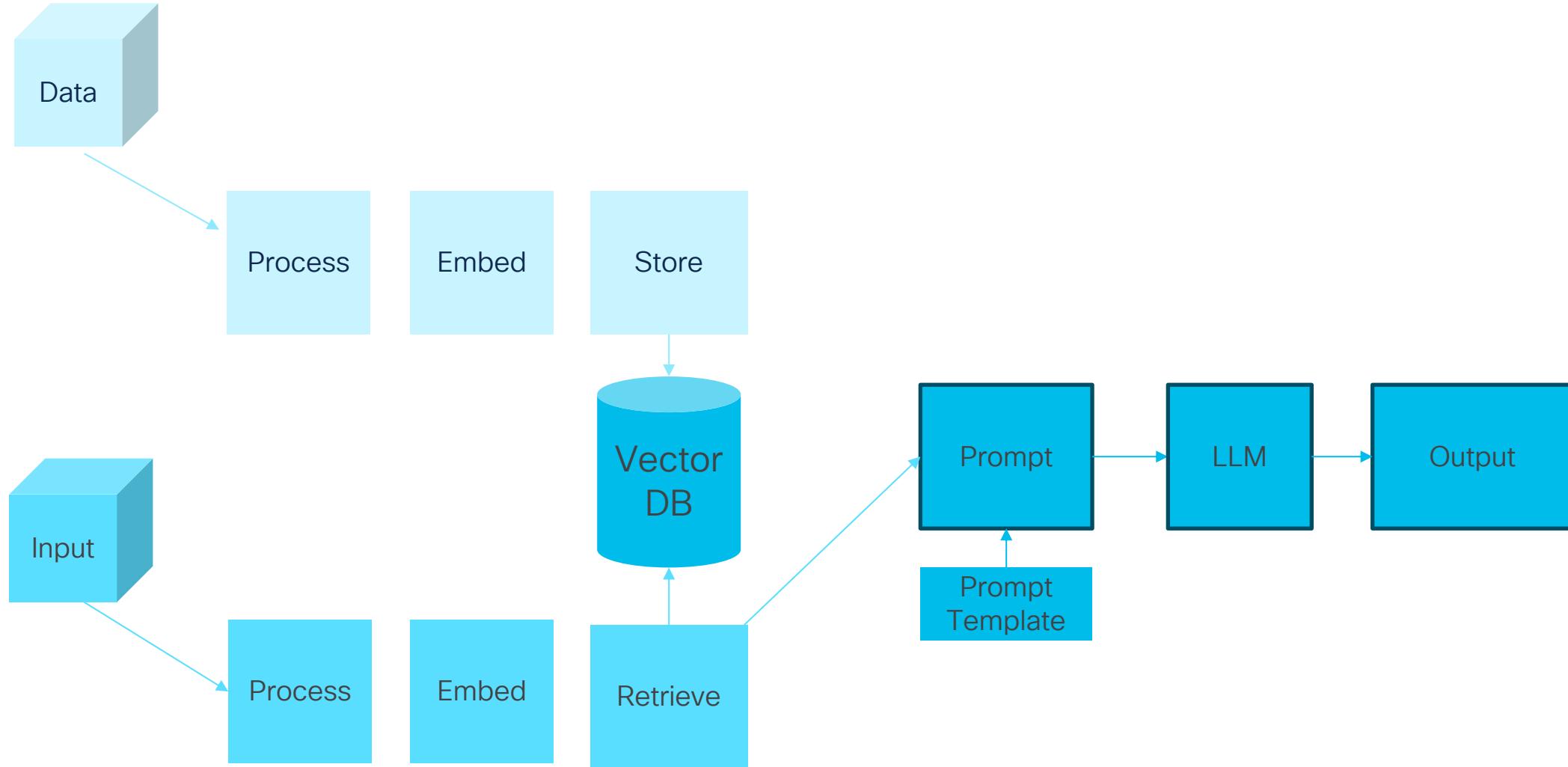
“Grounding” the model

- ~~Training~~(too expensive)

- Fine-Tuning
 - Using your dataset to tune the model for specific use cases or custom data sets
- Retrieval Augmentation (RAG)
 - Enriching the prompt using embeddings and semantic search
- Prompt Engineering
 - Stuffing the prompt itself with additional context

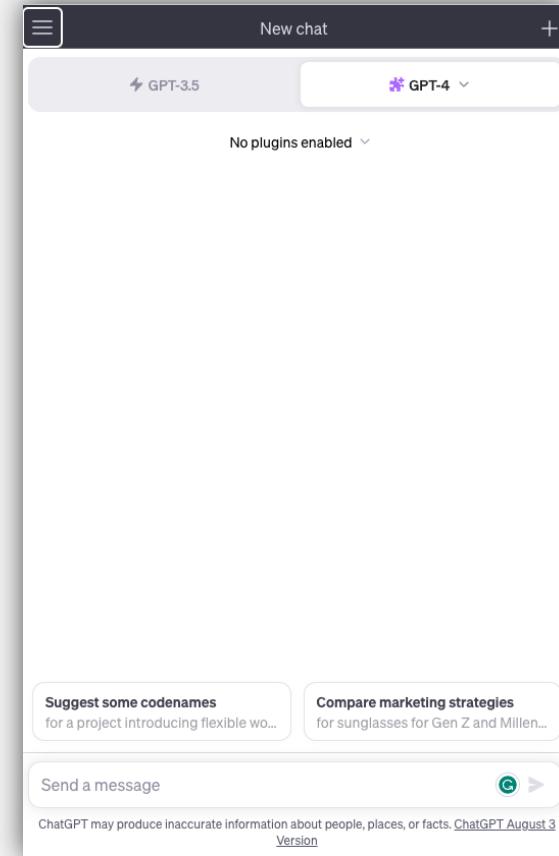


RAG (Retrieval Augmented Generation)



The Interface

How do we interact with the LLM using a structured input management system



Prompt Design

Instruction

→ Your goal is to only answer questions with prefix 'question:' only when the provided context with prefix 'context:' has relevant information to be answered.

Negative response instruction

→ Otherwise respond with message
'Content doesn't have enough information to address this inquiry'.

In-context learning examples

→ Here are some examples for you to learn:
question: <sample question 1>
context: <sample context 1>
answer: <sample answer 1>

...<more examples>

Question

→ question: {_question}

Candidate answers

→ context: {_context}

Guardrails

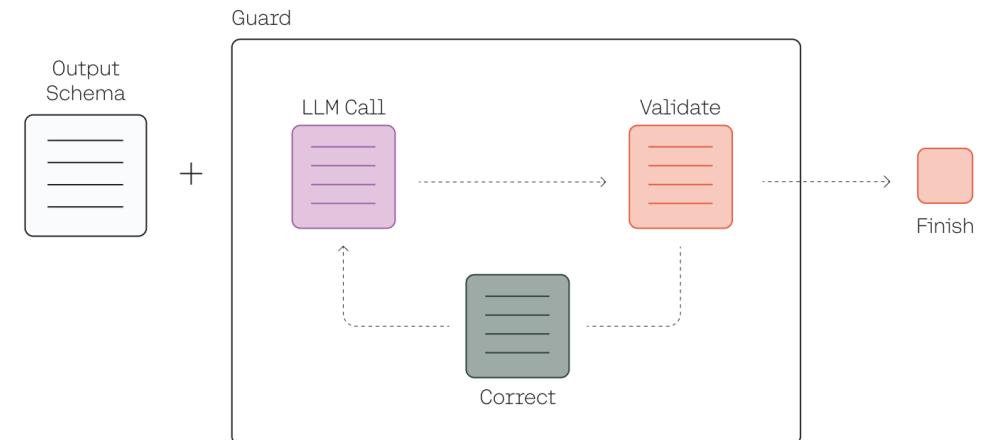
Syntactic Validation – check generated content is valid for the structure of the expected format

Semantic Distance – confirm output is semantically similar to input

Safety Checks – Filter for content and toxicity

Second Opinion – validate using a secondary model

Input checking – ensure the user is not mis-prompting



Data

Docs

Data

Preparation

Embedding

Training

VectorDB



Guardrails

Prompt

Init

Template

Context

Interface

Interface



LLM

LLM

LLM

LLM

Plugins

Evals

Model Serving Infra (vLLM)

Model Hardware (AWS, GCP, DGX, UCS)

Observability

Model

Operations

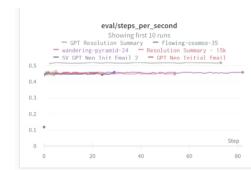
Running the system in production

$$\sum_x [-P(x) \log(P(x))]$$

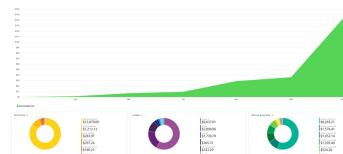
Accuracy



Performance



Observability



Cost



Security



Accuracy

- The Problem of Accuracy
 - Accuracy is still an open area of research
 - There is no way to completely prevent hallucinations
 - We don't fully understand why LLMs work!
- Maintaining accuracy is a difficult problem
 - Explainability
 - It is hard to explain what the model is doing, how it is changing as you add data, change algorithms, etc
 - Model Drift, Data Drift, Reality Drift
- Custom Evals are needed for your Dataset!

**The Reversal Curse:
LLMs trained on “A is B” fail to learn “B is A”**

Lukas Berglund* Meg Tong^{†1} Max Kaufmann^{‡1} Mikita Balesni^{§1}

How Is ChatGPT’s Behavior Changing over Time?

Lingjiao Chen[†], Matei Zaharia[‡], James Zou[†]

[†]Stanford University [‡]UC Berkeley

Abstract

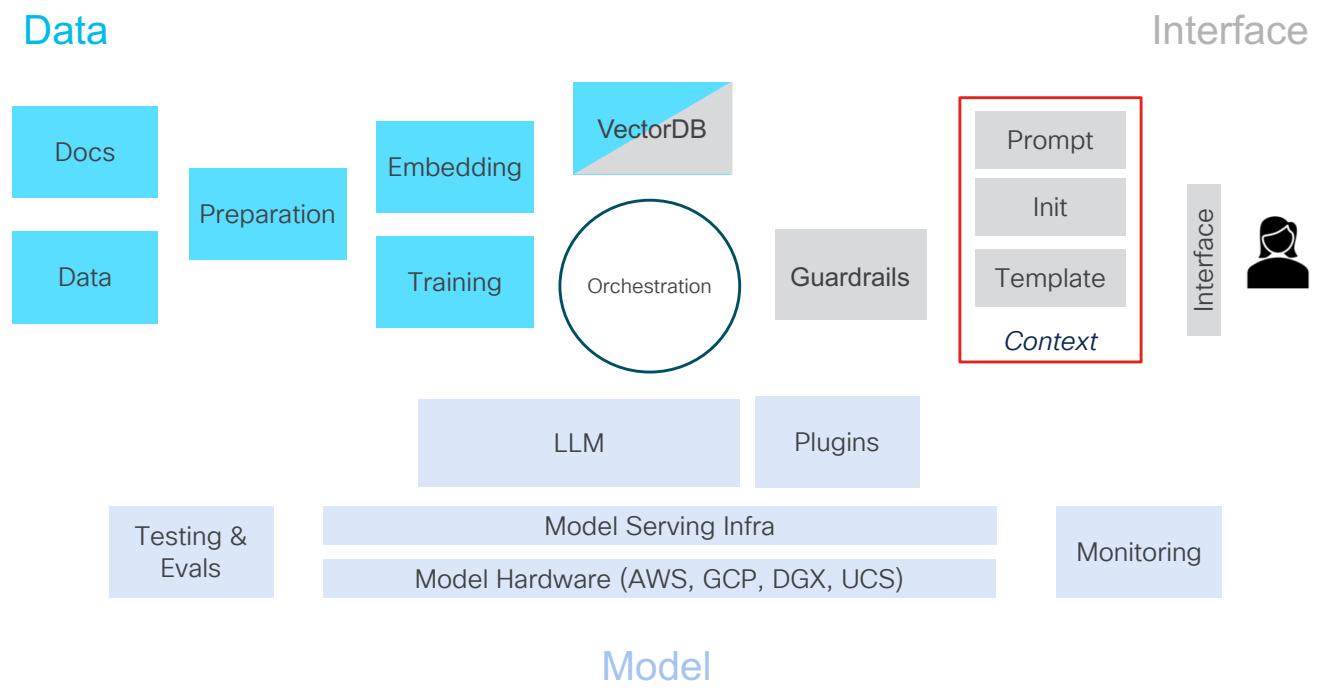
GPT-3.5 and GPT-4 are the two most widely used large language model (LLM) services. However, when and how these models are updated over time is opaque. Here, we evaluate the March 2023 and June 2023 versions of GPT-3.5 and GPT-4 on several diverse tasks: 1) math problems, 2) sensitive/dangerous questions, 3) opinion surveys, 4) multi-hop knowledge-intensive questions, 5) generating code, 6) US Medical License tests, and 7) visual reasoning. We find that the performance and behavior of both GPT-3.5 and GPT-4 can vary greatly over time. For example, GPT-4 (March 2023) was reasonable at identifying prime vs. composite numbers (84% accuracy) but GPT-4 (June 2023) was poor on these same questions (51% accuracy). This is partly explained by a drop in GPT-4’s amenity to follow chain-of-thought prompting. Interestingly, GPT-3.5 was much better in June than in March in this task. GPT-4 became less willing to answer sensitive questions and opinion survey questions in June than in March. GPT-4 performed better at multi-hop questions in June than in March, while GPT-3.5’s performance dropped on this task. Both GPT-4 and GPT-3.5 had more formatting mistakes in code generation in June than in March. Overall, our findings show that the behavior of the “same” LLM service can change substantially in a relatively short amount of time, highlighting the need for continuous monitoring of LLMs.

1 Introduction

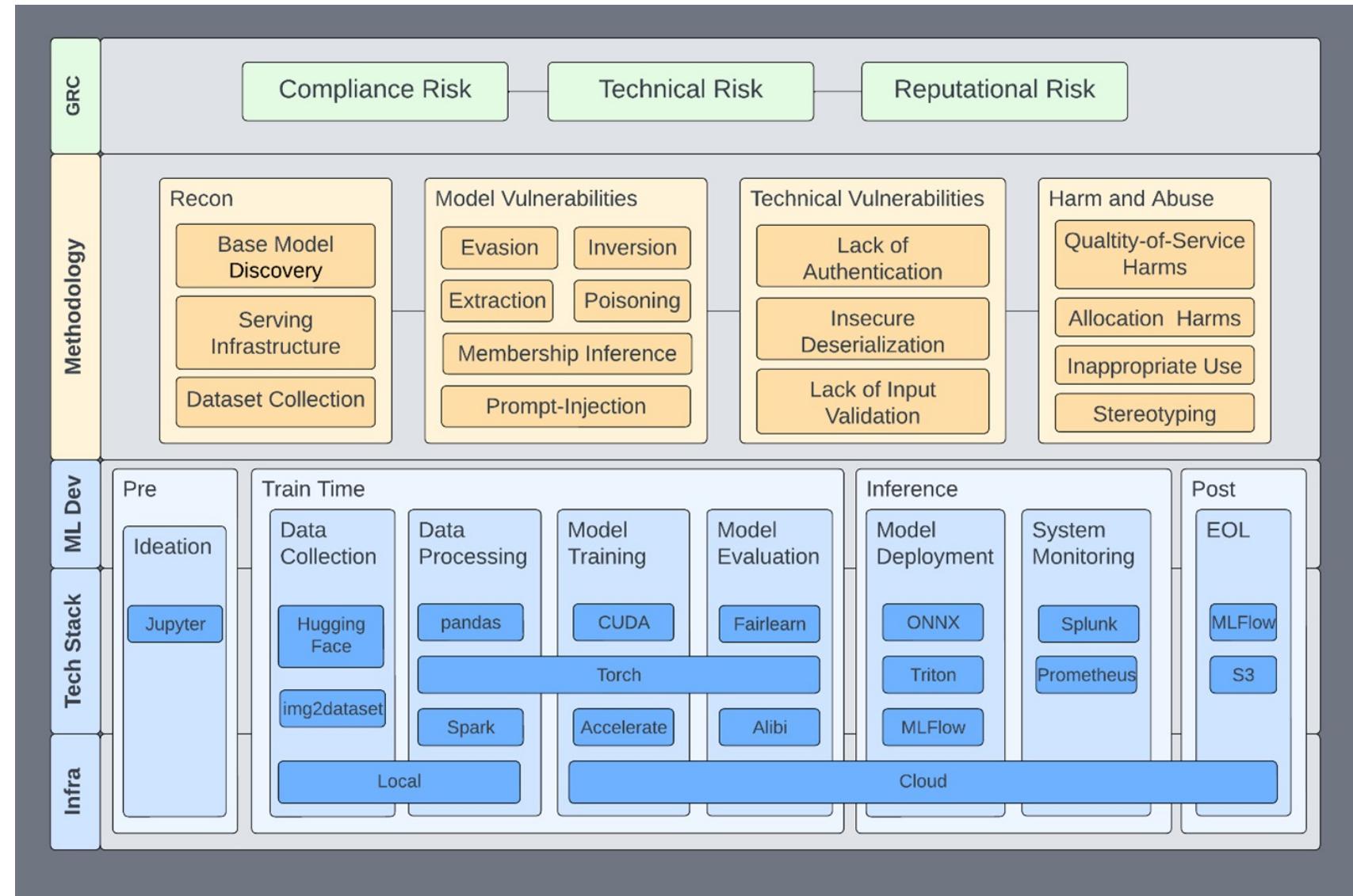
Large language models (LLMs) like GPT-3.5 and GPT-4 are being widely used. A LLM like GPT-4 can be updated over time based on data and feedback from users as well as design changes. However, it is currently opaque when and how GPT-3.5 and GPT-4 are updated, and it is unclear how each update affects the behavior of these LLMs. These unknowns makes it challenging to stably integrate LLMs into larger workflows: if LLM’s response to a prompt (e.g. its accuracy or formatting) suddenly changes, this might break the downstream pipeline. It also makes it challenging, if not impossible, to

Elements of Accuracy

- Data Hygiene
- Embedding Algorithm
- Vector Indexing Algorithm
- Fine-tuning method
- Over/Underfitting
- The Model!
- Model hyperparameters
- Model training
- Inferencing Platform
- Prompt template
- Guardrails



Security



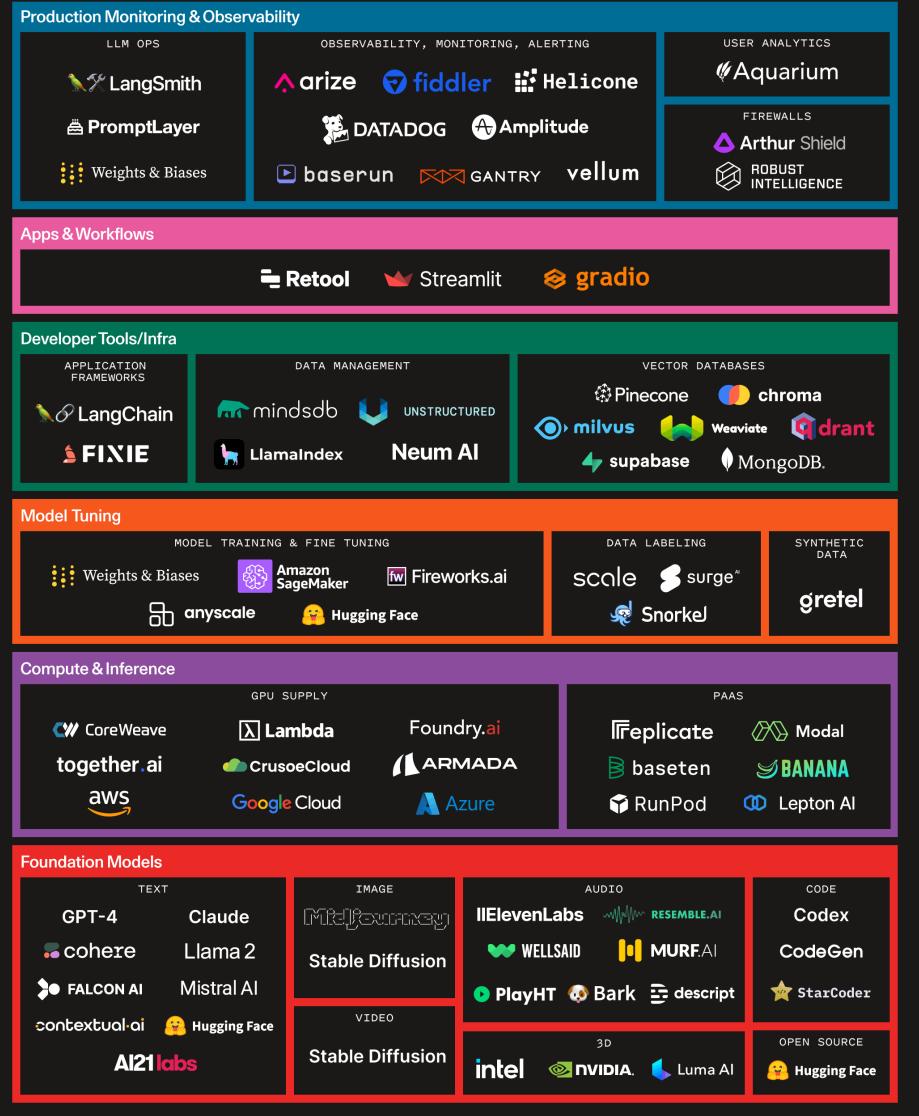
Nvidia Red Team Model

The “New” Stack

Running the system in production

The Generative AI Infrastructure Stack v1

A work in progress

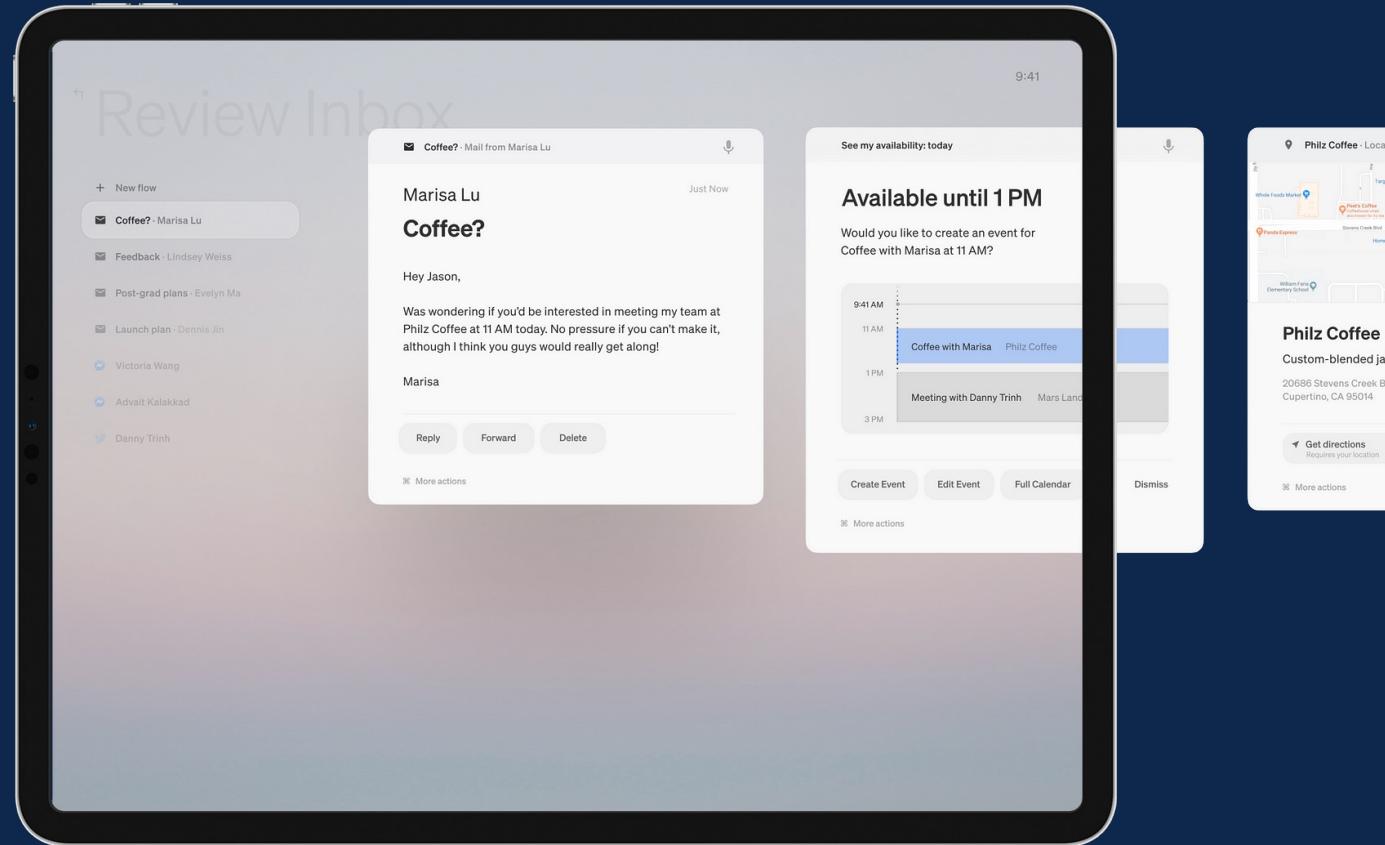




The Moat



Enterprise AI Platform Stack



Enabling many teams to work coherently using platform elements that enable the entire business to build and deploy AI applications

Platform Initiatives

Key Enterprise AI platforms launching at Cisco:

1. **Common Design System**: To provide a consistent experience
2. **Bot Framework**: To easily create chatbots
3. **Inferencing Infra**: To host models centrally
4. **Model Zoo**: To reduce latency & cost
5. **Corporate Corpus**: To train models on
6. **Prompt Library**: To share learning on prompt engineering
7. **Enterprise API**: To provide access to third SaaS providers

Model Zoo

- Sharing customized models

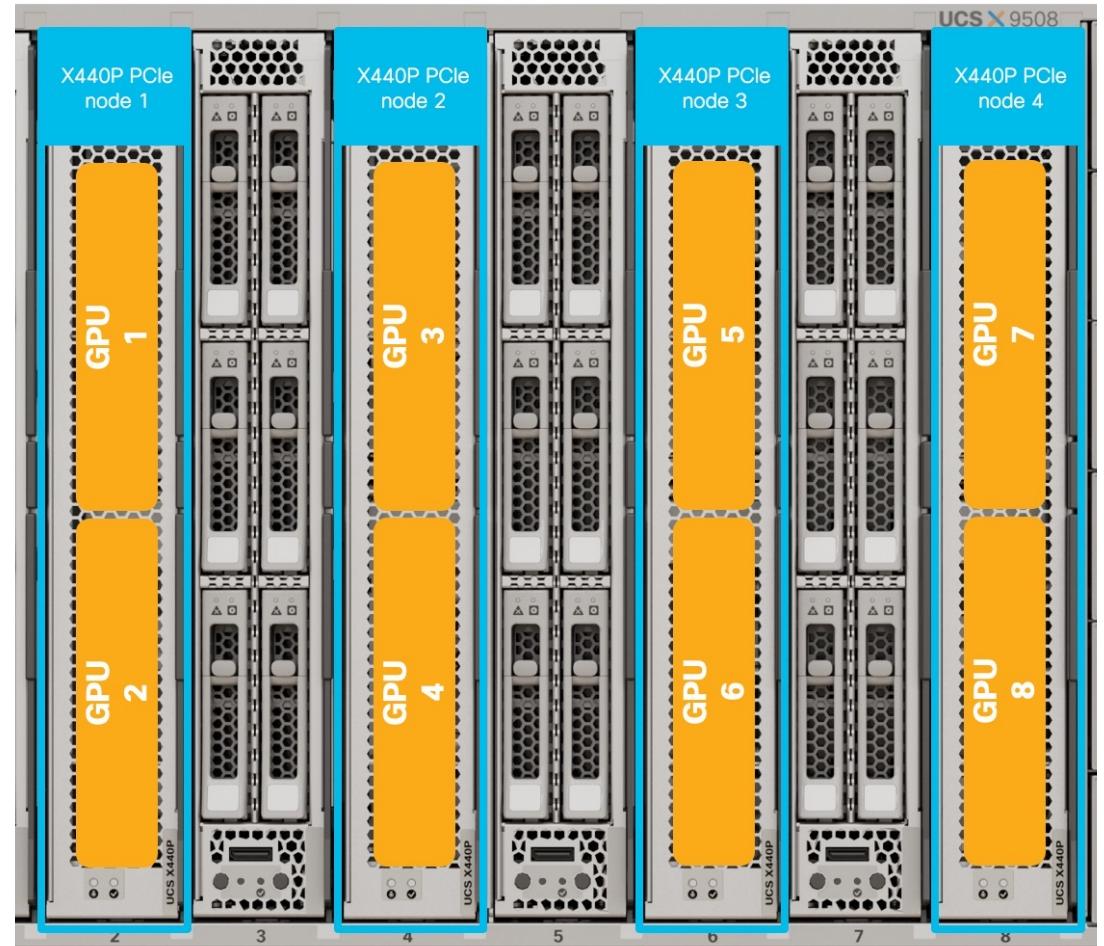
The screenshot shows the Hugging Face Model Zoo interface. At the top, there's a navigation bar with tabs for Tasks, Libraries, Datasets, Languages, Licenses, and Other. Below the navigation bar is a search bar labeled "Filter Tasks by name". The main content area is divided into sections based on tasks:

- Multimodal**: Feature Extraction, Text-to-Image, Image-to-Text, Text-to-Video, Visual Question Answering, Document Question Answering, Graph Machine Learning.
- Computer Vision**: Depth Estimation, Image Classification, Object Detection, Image Segmentation, Image-to-Image, Unconditional Image Generation, Video Classification, Zero-Shot Image Classification.
- Natural Language Processing**: Text Classification, Token Classification, Table Question Answering, Question Answering, Zero-Shot Classification, Translation, Summarization, Conversational, Text Generation, Text2Text Generation.

On the right side of the interface, there's a list of models with their details:

- microsoft/phi-1_5**: Text Generation, Updated about 7 hours ago, 25.5k, 491
- tiiuae/falcon-180B-chat**: Text Generation, Updated 10 days ago, 6.99k, 337
- stabilityai/stable-diffusion-xl-base-1.0**: Text-to-Image, Updated 11 days ago, 1.39M, 2.62k
- meta-llama/Llama-2-7b**: Text Generation, Updated Jul 19, 2.47k
- meta-llama/Llama-2-7b-chat-hf**: Text Generation, Updated Aug 9, 581k, 1.16k
- coqui/XTTS-v1**: Text-to-Speech, Updated about 19 hours ago, 4, 66
- THUDM/chatglm2-6b**: Updated 8 days ago, 468k, 1.7k
- monster-labs/control_v1p_sd15_qrcode_monster**: Updated Jul 21, 2.7k, 446
- tiiuae/falcon-180B**: Text Generation, Updated 10 days ago, 45.3k, 631
- llyasviel/sd_control_collection**: Updated 7 days ago, 335
- Deci/DeciLM-6b**: Text Generation, Updated 1 day ago, 401, 104
- Deci/DeciLM-6b-instruct**: Text Generation, Updated 1 day ago, 298, 81
- microsoft/phi-1**: Text Generation, Updated 4 days ago, 1.95k, 73
- baichuan-inc/Baichuan2-13B-Chat**: Text Generation, Updated 4 days ago, 280k, 145
- runwayml/stable-diffusion-v1-5**: Text-to-Image, Updated 23 days ago, 8.08M, 9.21k
- meta-llama/Llama-2-70b-chat-hf**: Text Generation, Updated Aug 9, 199k, 1.31k

Model Farm, Model Infra



Courtesy of Patrick Lemaistre

Internal Corpus

- At Cisco we have piles of Cisco information that we can use to train models
- We can assemble this together into a single corpus of information that anyone can use with their LLMs
- Need good data hygiene!

Latest Security Notices

23-Aug-2023	Security Advisory: Cisco Nexus 3000 and 9000 Series Switches IS-IS Protocol Denial of Service Vulnerability
23-Aug-2023	Security Advisory: Cisco NX-OS Software TACACS+ or RADIUS Remote Authentication Directed Request Denial of Service Vulnerability
23-Aug-2023	Security Advisory: Cisco Nexus 3000 and 9000 Series Switches SFTP Server File Access Vulnerability
10-Apr-2023	Field Notice: FN72006 - N9K-C9508-FAN/N9K-C9516-FAN: Fan Flow Direction Reversed - Hardware Upgrade Available
07-Mar-2023	Field Notice: FN - 72495 - Nexus 9000 Series Switch Stuck in Boot Loop Due to "policyelem hap" Reset After Conversion to ACI - Workaround Provided
22-Feb-2023	Security Advisory: Cisco NX-OS Software CLI Command Injection Vulnerability
06-Feb-2023	Field Notice: FN - 72150 - Nexus 9000/3000 Will Fail With SSD Read-Only Filesystem - Power Cycle Required - BIOS/Firmware Upgrade Recommended
16-Dec-2022	Field Notice: FN - 70277 - DC Power Supply Hardware Component Failure Results in Loss of Power Output - Hardware Upgrade Required
27-Sep-2022	Security Advisory: Vulnerabilities in Layer 2 Network Security Controls Affecting Cisco Products: September 2022
19-Sep-2022	Field Notice: FN - 72145 - Nexus ACI 9000 Will Fail With SSD Read-Only Filesystem - Power Cycle Required - BIOS/Firmware Upgrade Recommended

Document Categories

Configuration	Multimedia
Data Sheets and Product Information	Reference
Design	Release and Compatibility
Install and Upgrade	Security Notices
Literature	Troubleshooting
Maintain and Operate	

Data Sheets and Product Information

At-a-Glance
Build Hierarchical Fabrics with VXLAN EVPN Multi-Site At-a-Glance (PDF – 439 KB)
Cisco Services for Application Centric Infrastructure and Nexus 9000 Series Switches (PDF – 821 KB)
Simplifying the DevOps Model – Nexus switches and Puppet automation (PDF – 552 KB)
Cisco Nexus 9000 Series Switches Services (PDF – 219 KB)
Accelerating 40G Adoption (PDF – 2 MB)
Cisco Nexus 9000 Series ThousandEyes Integration At-a-Glance
Creating a global, zero-downtime network environment (PDF – 1 MB)

Data Sheets

Most Recent

Cisco Nexus 9300-GX2 Series Fixed Switches Data Sheet	28-Aug-2023
Cisco Nexus 9300-FX2 Series Switches Data Sheet	25-Aug-2023
Cisco Nexus 9300-FX and FX-24 Port Series Switches Data Sheet	11-Aug-2022

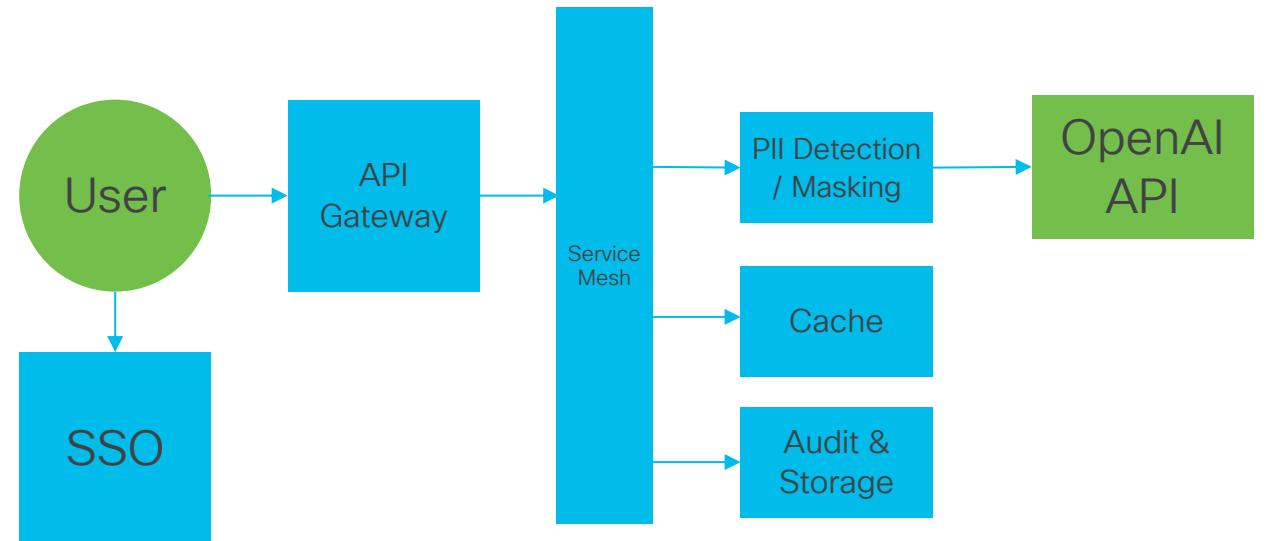


Enterprise API

Seamless Wrapper around popular AI frameworks

Approved Interface with,

- Conversation Storage
- Prompt Redaction
- Real-time Audit
- PII detection & masking
- Auth via SSO



Want to Learn More?
Ask a model!
(OpenAI does it)



Language models can explain neurons in language models

AUTHORS

Steven Bills*, Nick Cammarata*, Dan Mossing*, Henk Tillman*, Leo Gao*, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu*, William Saunders*

* Core Research Contributor; Author contributions statement below. Correspondence to interpretability@openai.com.

AFFILIATION

OpenAI

PUBLISHED

May 9, 2024

Introduction

Language models have become more capable and more widely deployed, but we do not understand how they work. Recent work has made progress on understanding a small number of circuits and narrow behaviors [2] but to fully understand a language model, we'll need to analyze millions of neurons. This paper applies automation to the problem of scaling an interpretability technique to all the neurons in a large language model. Our hope is that building on this approach of automating interpretability [3] [4] [5] will enable us to comprehensively audit the safety of models before deployment.

Our technique seeks to explain what patterns in text [6] [7] cause a neuron to activate. It consists of three steps:

Step 1 Explain the neuron's activations using GPT-4

Show neuron activations to GPT-4



The bridge to possible