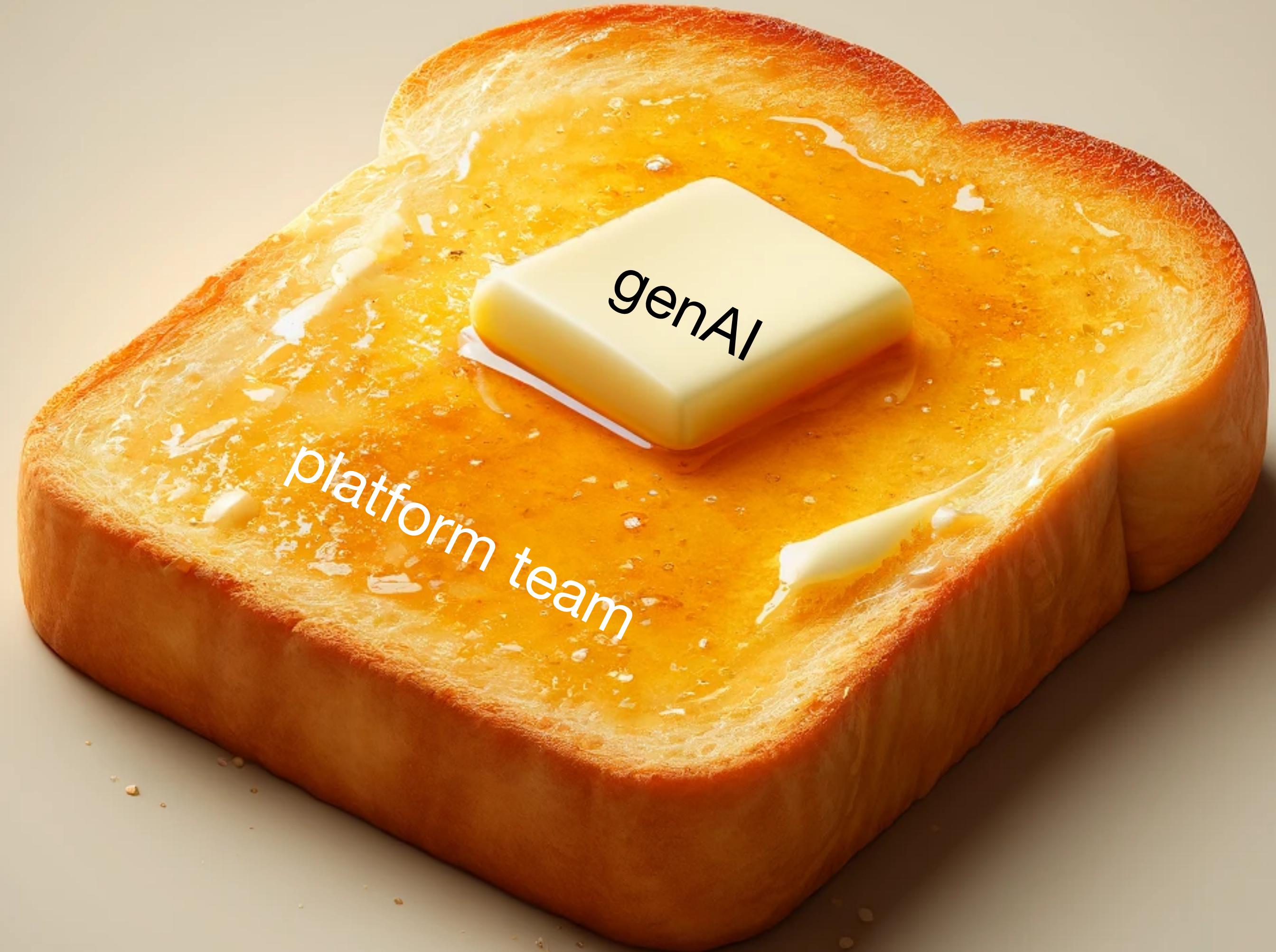


Every AI engineer
deserves an
AI platform (team)

organising for the next big thing



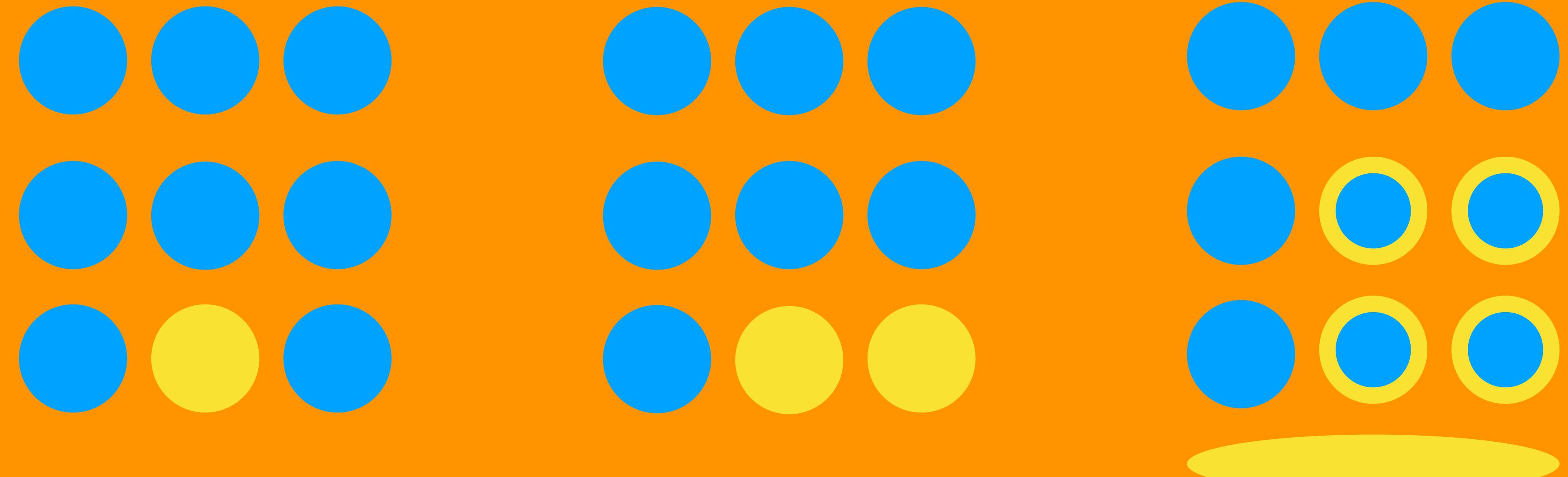
@patrickdebois

A close-up portrait of a man with grey hair and a beard, wearing glasses, looking slightly to the right.

Who am I

- Patrick Debois
- Started DevOps(Days)
- Co-Author of the DevOps Handbook
- Independent Industry Advisor
- Dev, Sec, Ops and now GenAI
- **Automation Intelligence**

Introducing Org & Tech Change



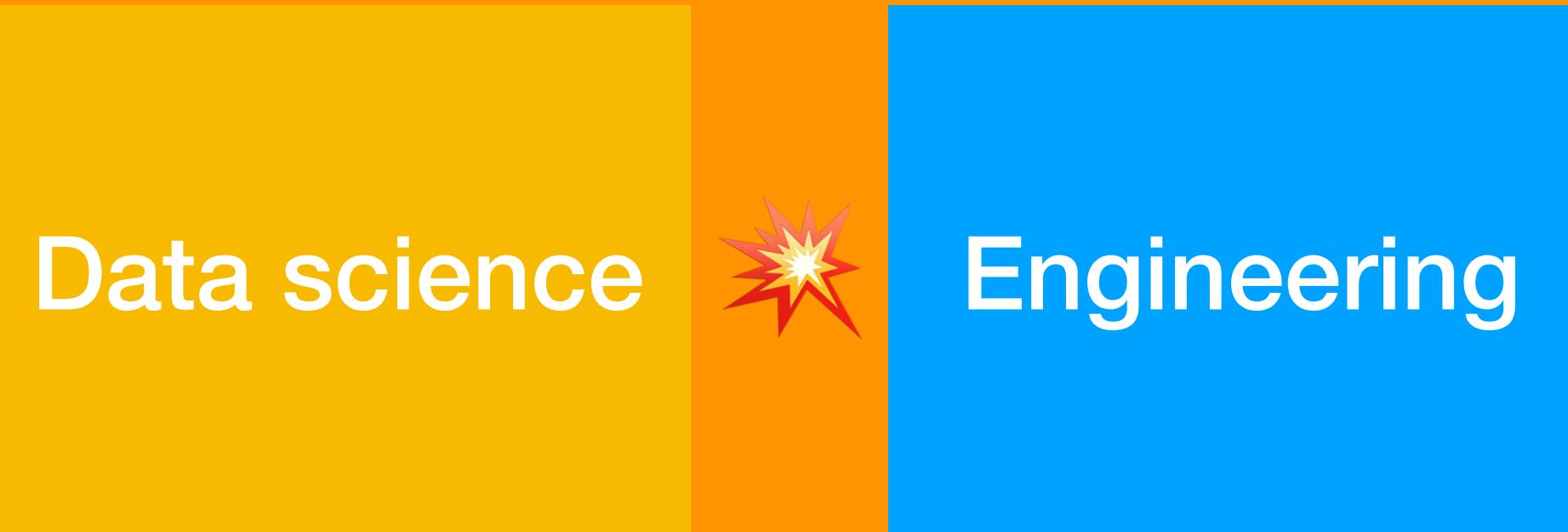
Incubate

Repeat

Scale

Maturity

Change Agents want an identity



Data science



Engineering

Names are ambiguous

Yet resonate with a pain

Have a matching technology shift

AI Engineer

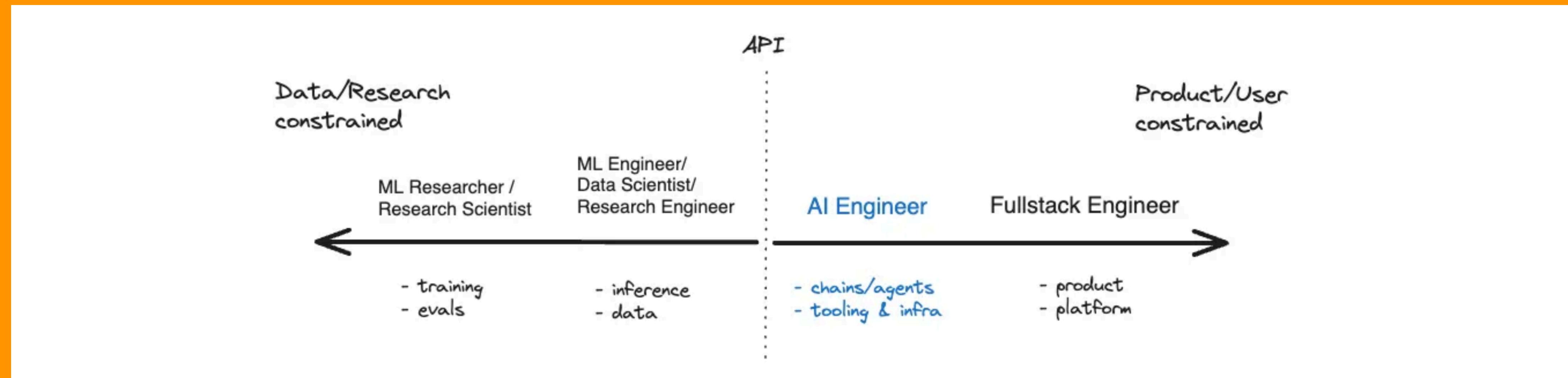
DevOps

Agile

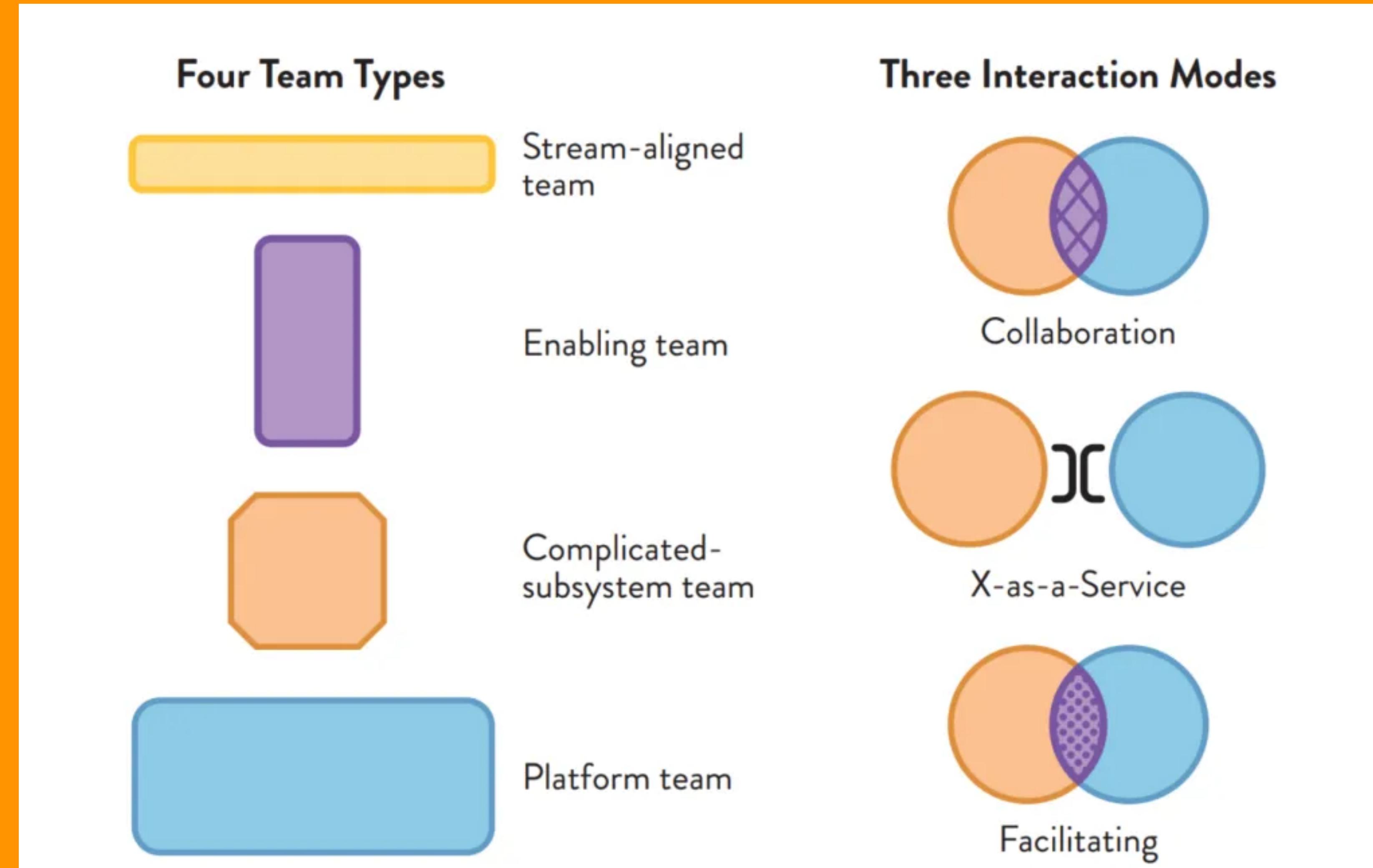
ITIL

Act a label to get the stories out

Org shift - Shift Right - Breaking Silos



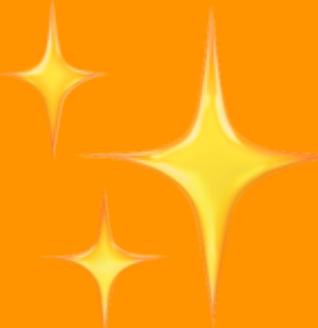
“Platform” team pattern



Steps to scale the change



Platform



Enablement



Governance

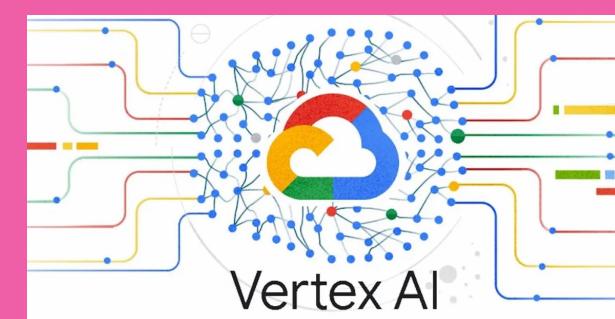
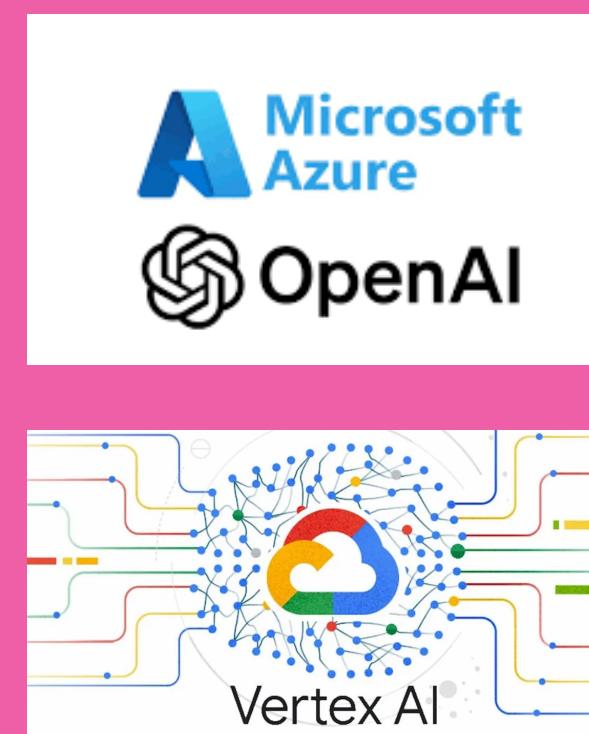
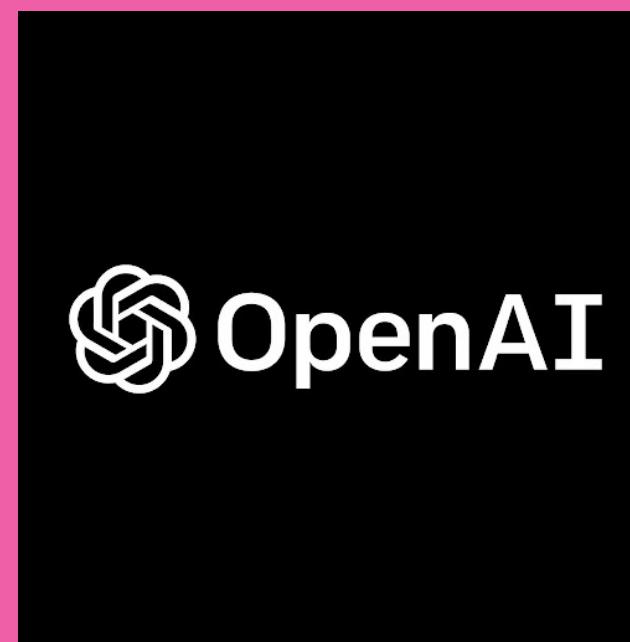


Platform

What services go into your platform

Access to Models

Enterprise SaaS



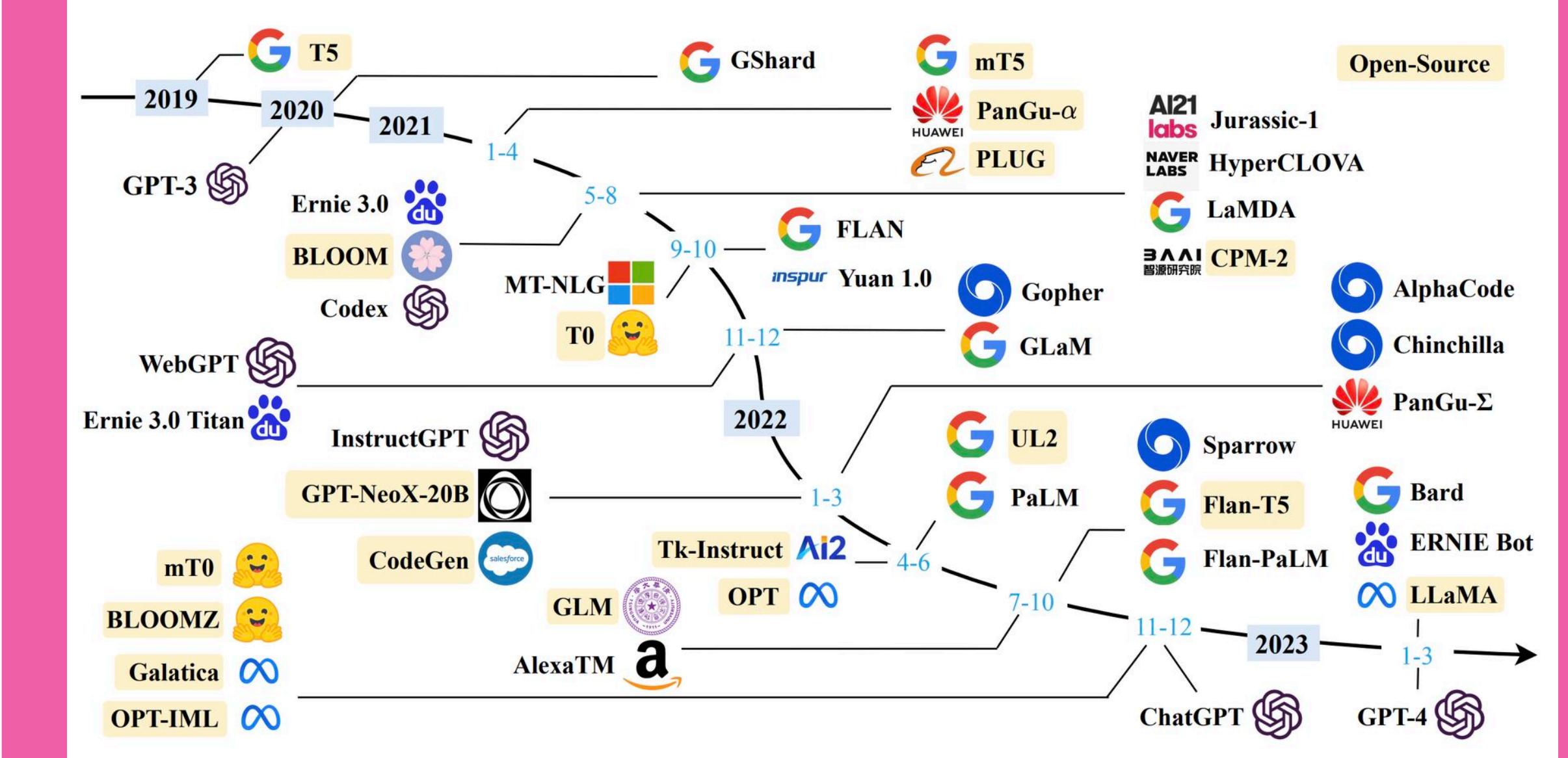
Model Broker

Cloud
Vendors

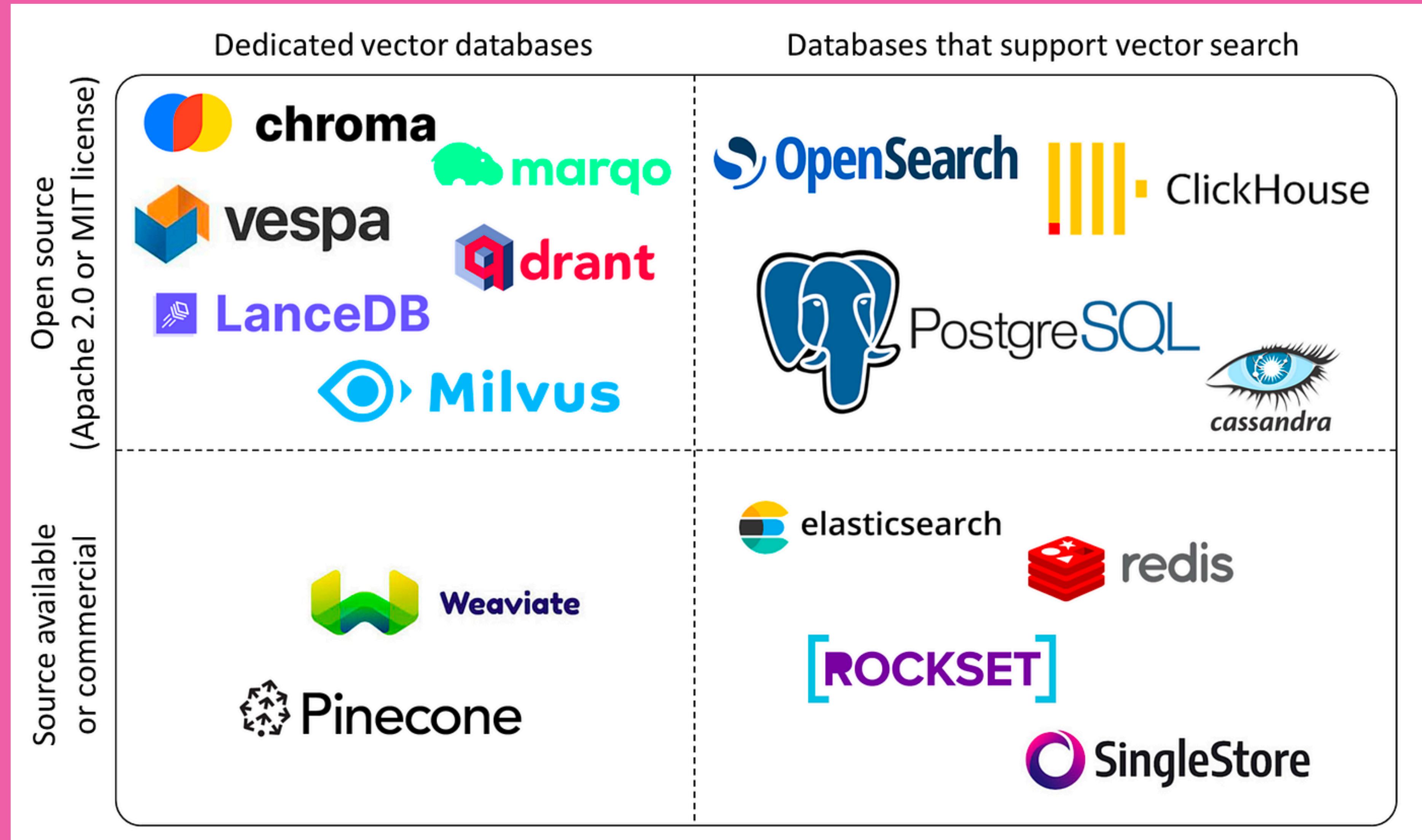
Independent
Models

Bring your own Model

Open Source



Unstructured data - Vector Databases



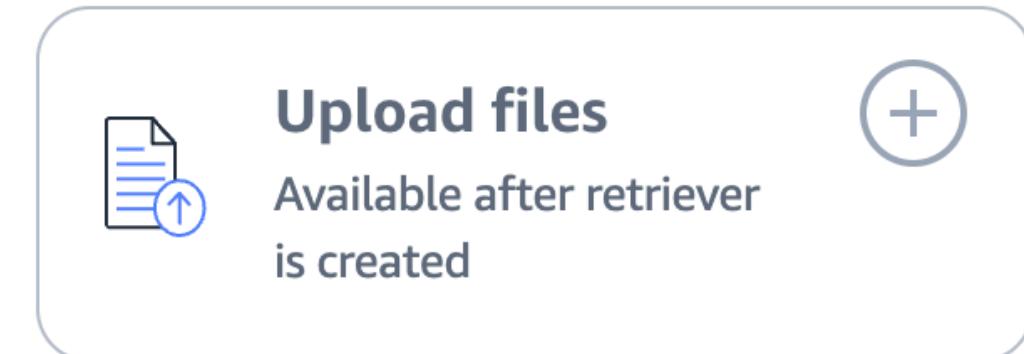
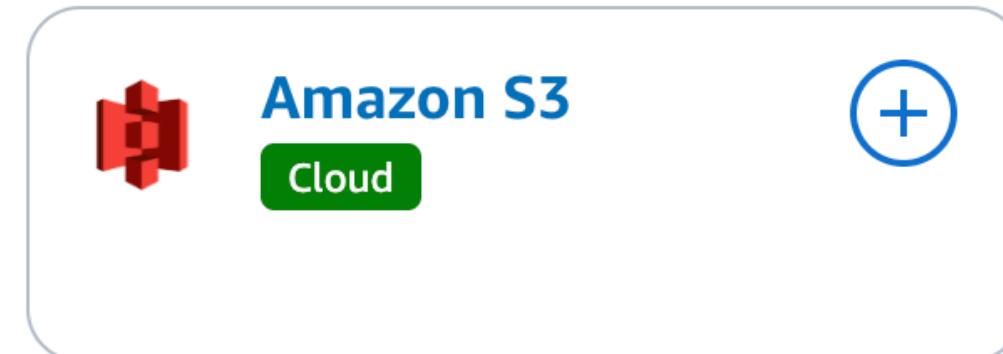
RAG as a Service

Data sources (0) Info

Select the data source that you want to configure. You can configure up to 5 data sources per application.

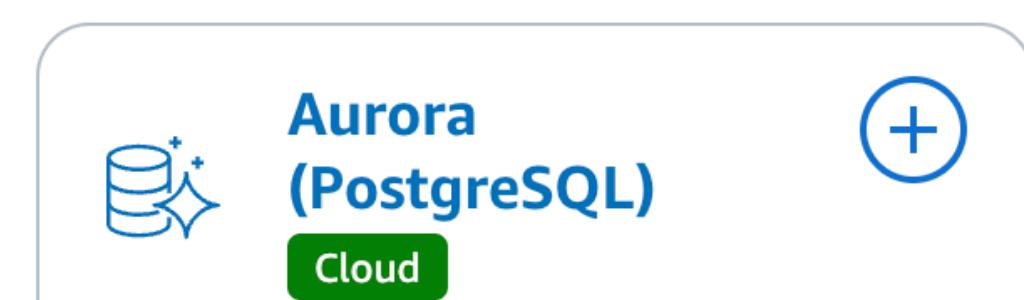
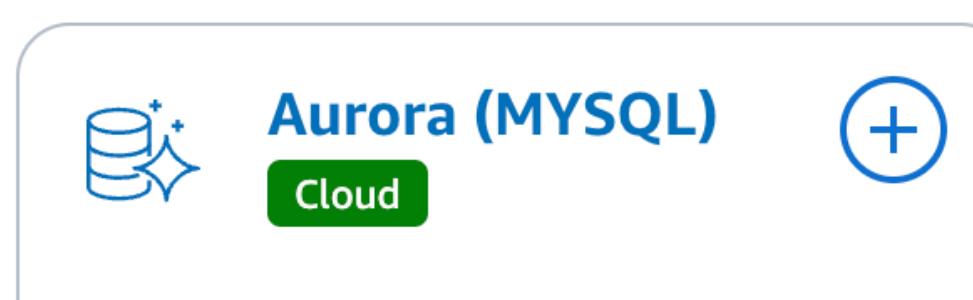
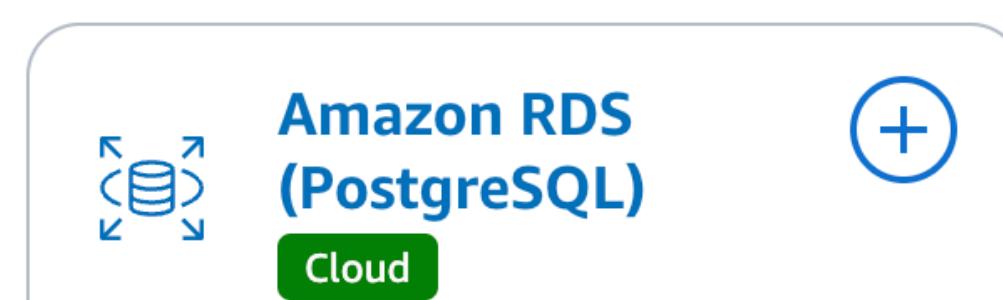
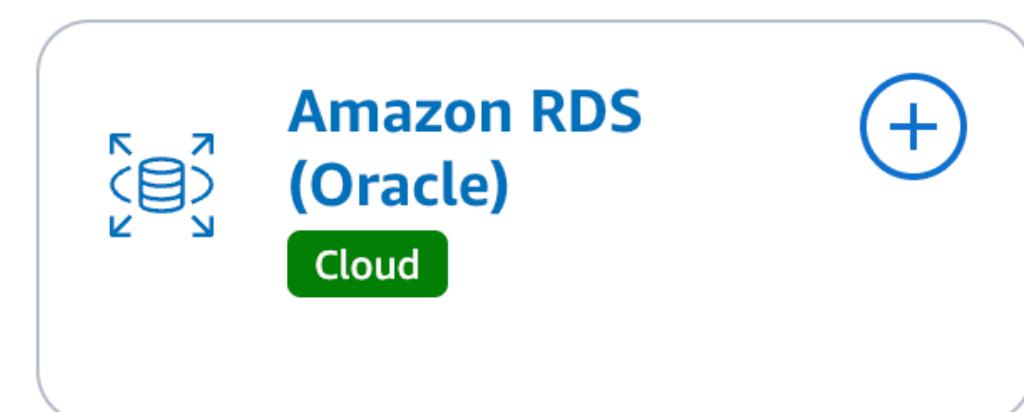
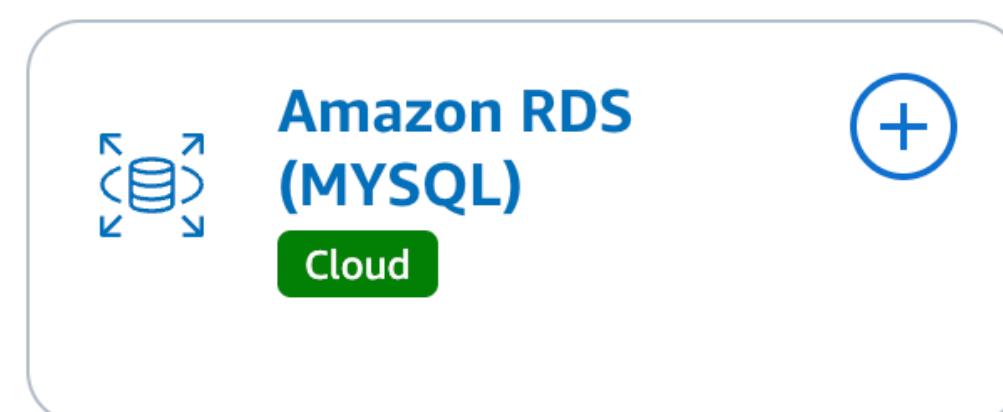
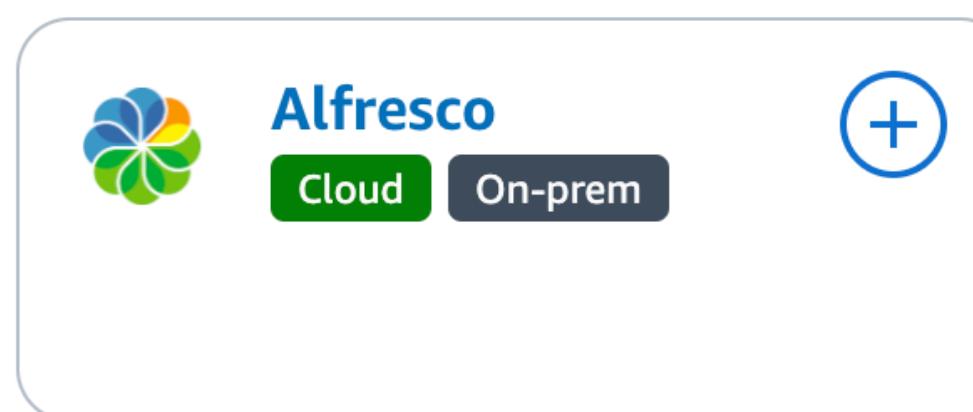
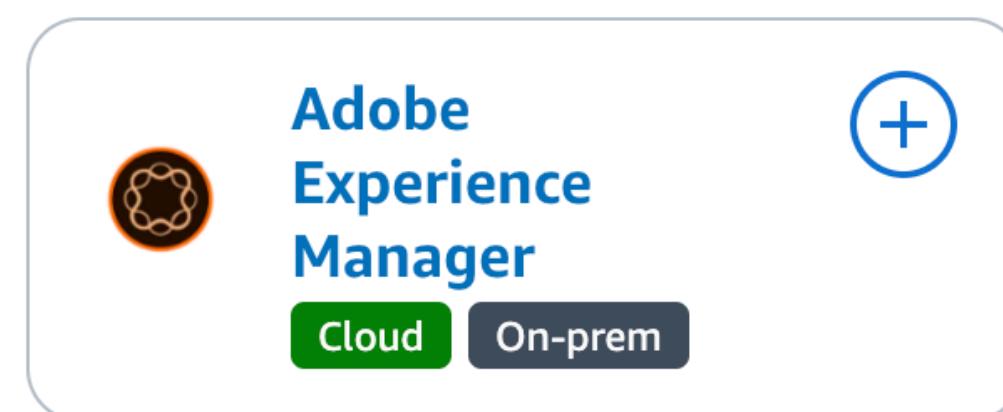
RAGOPS ?

Most popular



All Cloud On-prem

Search





ML MODEL MANAGEMENT

Create a single system of record for ML models that brings ML/AI development in line with your existing secure SSC.

SHARE: [f](#) [in](#) [X](#)

OVERVIEW

More development organizations are building and leveraging ML/AI models for use in software applications. However, a lack of standardized best practices on how to incorporate MLOps into the broader software supply chain has led ML model development to largely occur in isolation from the rest of software development. Further, the use of open source models poses similar challenges to using OSS packages – security, availability, versioning, etc. – particularly as the open source model ecosystem is still relatively new and the threat landscape uncertain.

ML Model Management with JFrog is an industry-first solution allowing organizations to bring development and security of AI/ML models alongside their other software components for a unified view of the software assets they're building and releasing. It delivers the same best practices organizations have benefited from for secure package management with JFrog to model management – control, availability, visibility, security, traceability/auditing.

BENEFITS

- **Manage ALL your Software Artifacts in One Place**

Store and manage models alongside the other components that make up modern software applications for better visibility and insight into the status of your software and its development.

- **Bring DevOps Best Practices to ML Development**

The DevOps practices developed over a decade of experience with OS package management, pipeline automation, and quality/feedback loops can now be applied to ML model management.

- **Ensure Integrity and Security of ML Models**

Manage your models in a system that introduces important controls including RBAC, versioning, license and security scanning so that ML, Security, and DevOps teams feel confident in the models used and be ready for the inevitable regulation to come.

- **A Single Platform for DevOps, SecOps, and MLOps**

Consolidate disparate tools and eliminate point solutions with a single system of truth that can manage ML Models and the technologies that package them into applications used by consumers. Seamlessly combine workflows of ML Engineers and DevOps teams without needing to change the way either party works.

Models registries & version control

KEY CAPABILITIES

- Secure ML model registry
- Store and manage proprietary and modified OSS models
- Simplified, intuitive ML versioning
- Proxy Hugging Face for always available open source models
- Detect malicious models and enforce license compliance
- Standardize MLOps processes across teams
- Integrated with ML tools such as Jupyter Notebooks and Amazon SageMaker

Unified Model API proxies - ACL



Schedule Demo LiteLLM Docs

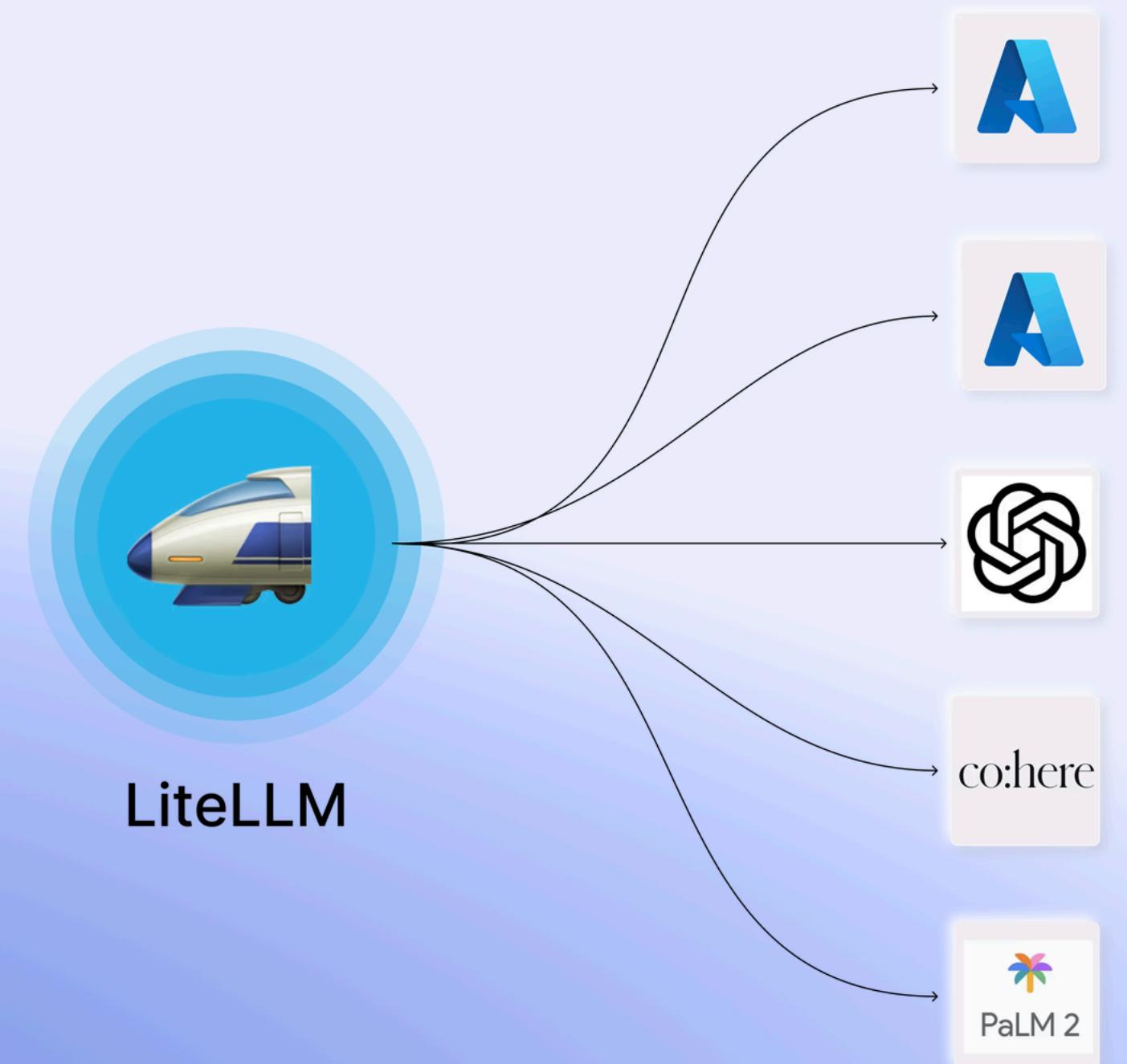
Github 8.4K ⚡

Load Balance across OpenAI

LiteLLM handles loadbalancing, fallbacks and spend tracking across 100+ LLMs. All in the OpenAI format

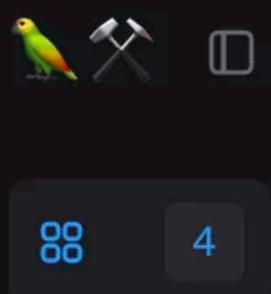
Try LiteLLM Cloud Free

Deploy LiteLLM Open Source →



<https://github.com/BerriAI/litellm>

<https://www.litellm.ai/>



Projects

Basic Project 1

Basic_Project_1

P50 Latency: 0.57s

P99 Latency: 24.77s

Total Tokens: 441

Traces All Runs Setup

eg. eq(run_type, "chain")

Columns

Run Type

Name

Input

Start Time

Latency

Tokens

Tags



Filters

Full-Text Search

Search...



Name

LLMChain

6

ChatOpenAI

2

Run Type

Chain

6

Llm

2

Status

Success

8

Other

Latency >= 10s

Tokens >= 1,000

Today

Canceled

Run Type	Name	Input	Start Time	Latency	Tokens	Tags
LLM	ChatOpenAI	human: What is the ...	01/08/2023, 09:37:43	⌚ 7.29s	99	:
Chain	LLMChain	What is the year of ...	31/07/2023, 20:50:42	⌚ 0.08s	56	:
Chain	LLMChain	What is the year of ...	31/07/2023, 20:50:23	⌚ 0.07s	56	:
Chain	LLMChain	What is the year of ...	31/07/2023, 20:49:46	⌚ 0.37s	56	:
Chain	LLMChain	What is the year of ...	31/07/2023, 20:23:44	⌚ 0.77s	56	:
Chain	LLMChain	What is the square r...	31/07/2023, 20:22:44	⌚ 26.09s	58	:
Chain	LLMChain	Who won the FIFA ...	31/07/2023, 20:06:18	⌚ 0.37s	40	:
LLM	ChatOpenAI	human: Hello, world!	31/07/2023, 16:43:37	⌚ 1.13s	20	:

Rows per page: 10 ▾ 1-8 of 8 < >

<https://cobusgreyling.medium.com/langsmith-1dd01049c3fb>

Continuous Data Quality Monitoring

≡ Latency P99 ⓘ

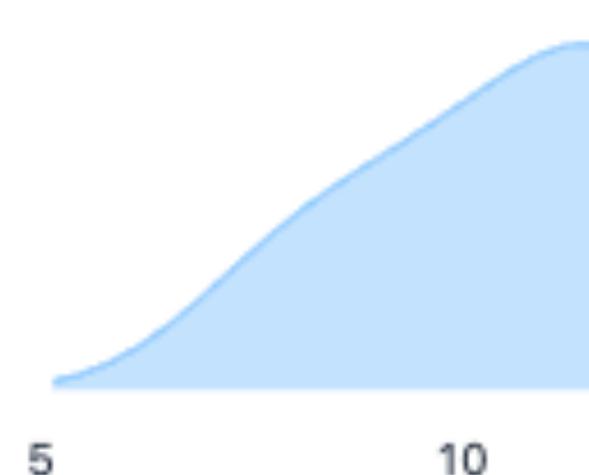
1.390s

≡ Feedback Instances ⓘ

You haven't sent any feedback yet!

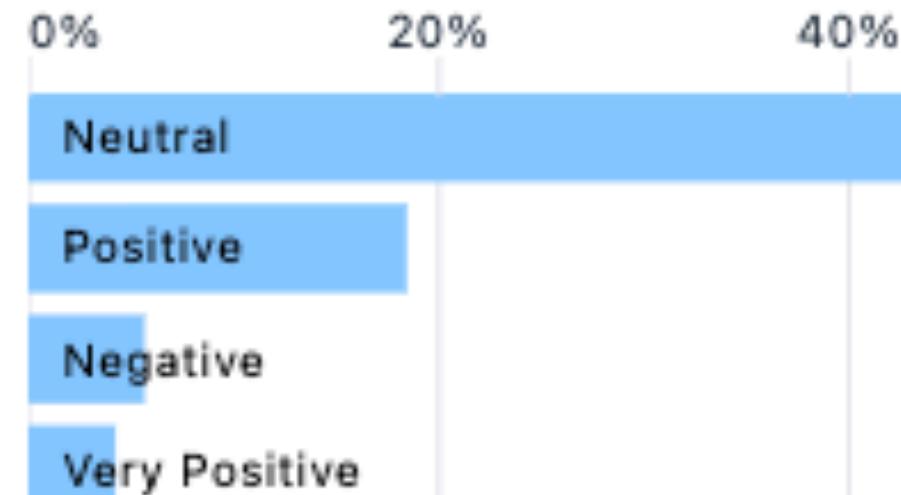
Try [logging some](#) now

≡ Completion Length ⓘ

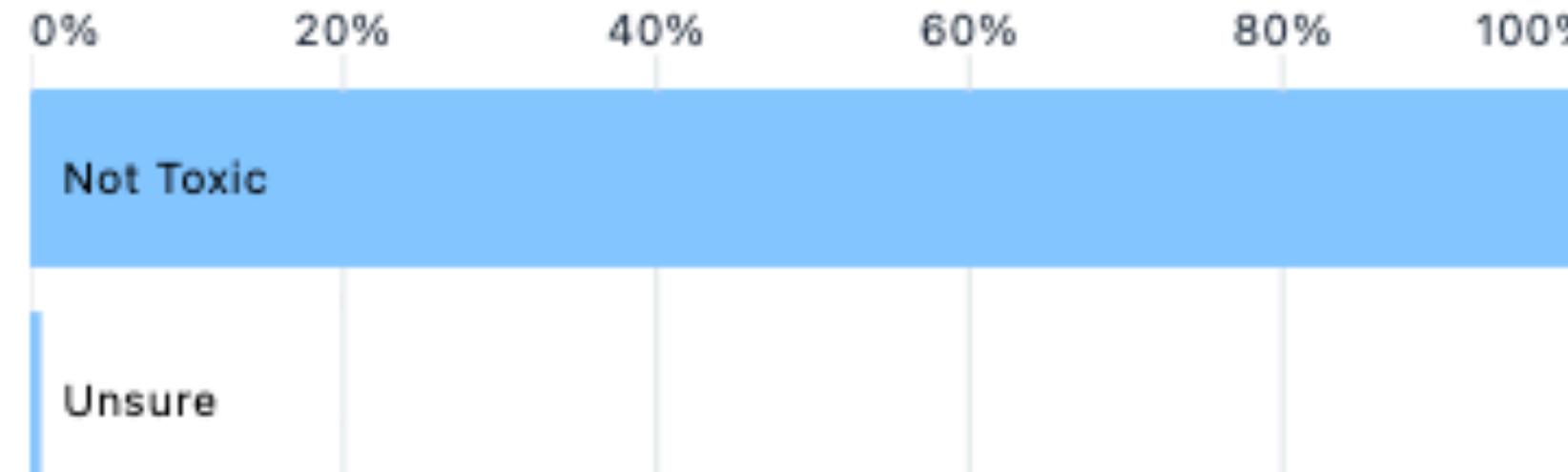


0 5 10

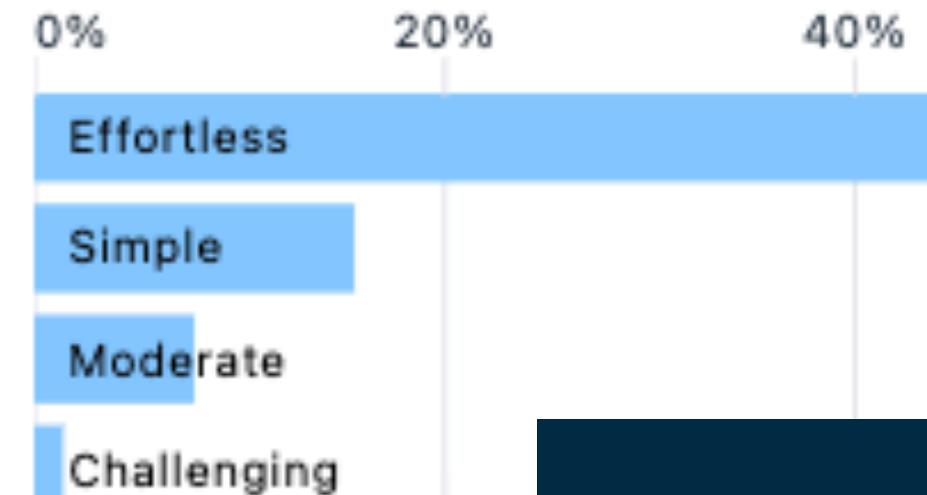
≡ Generation Sentiment ⓘ



≡ Generation Toxicity ⓘ



≡ Generation Fluidity ⓘ



≡ Evaluation Coverage ⓘ

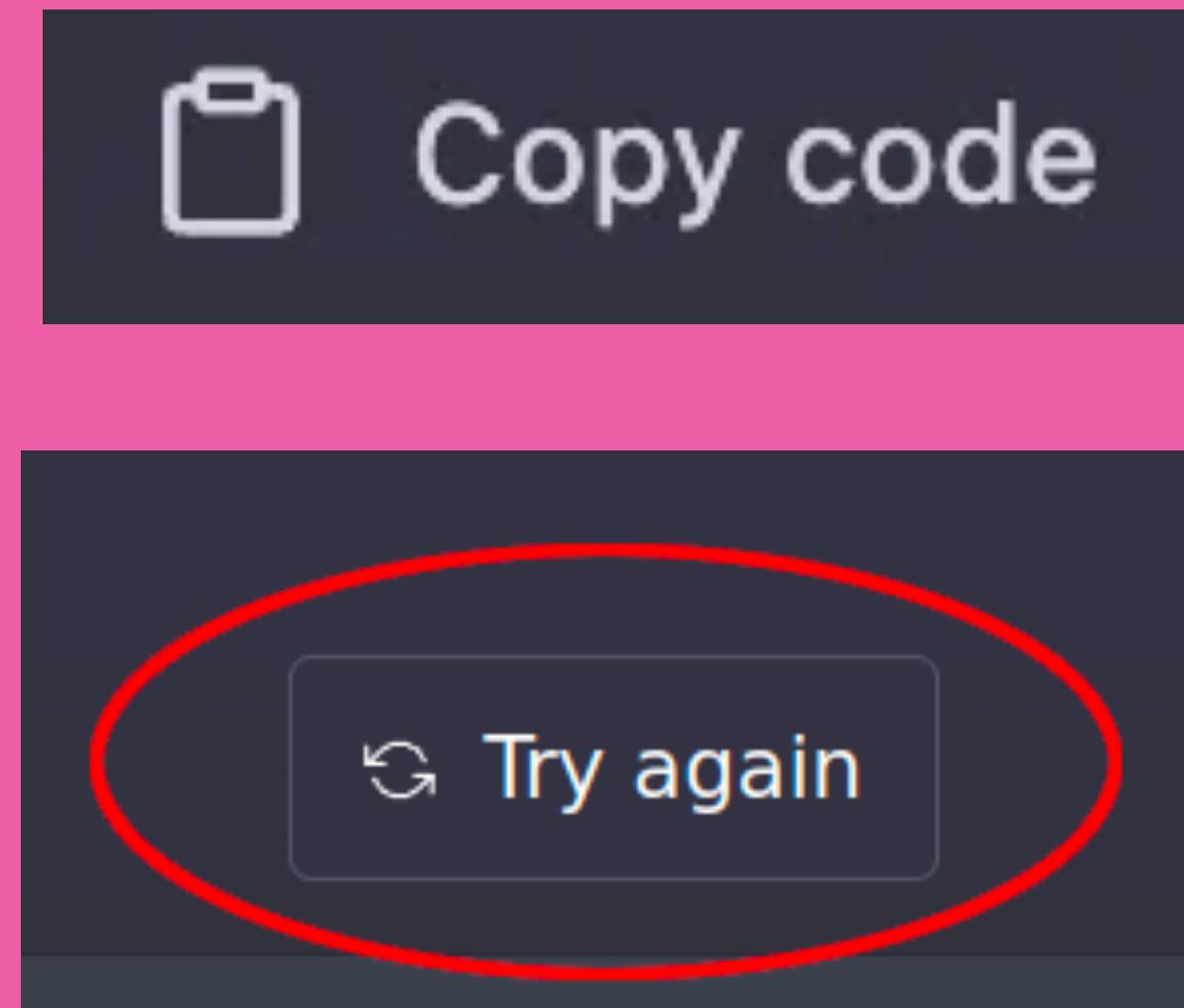
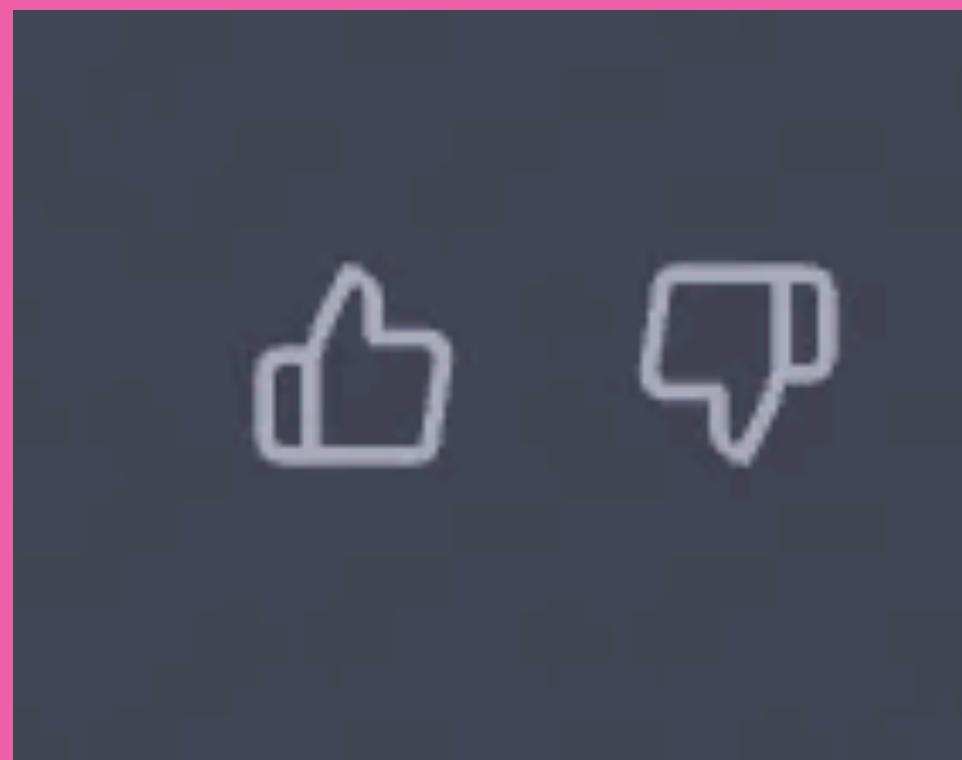
0% 20% 40% 60% 80% 100%

Use Small & Faster Models

WHY
LABS

<https://whylabs.ai/>

Feedback as a service



Textarea.AI: CKEditor ChatGPT plugin

The CKEditor toolbar is shown with various buttons for bold, italic, underline, and alignment. The 'AI Text' button is highlighted with a blue border and has a tooltip 'AI Text'. A cursor arrow points to this button.

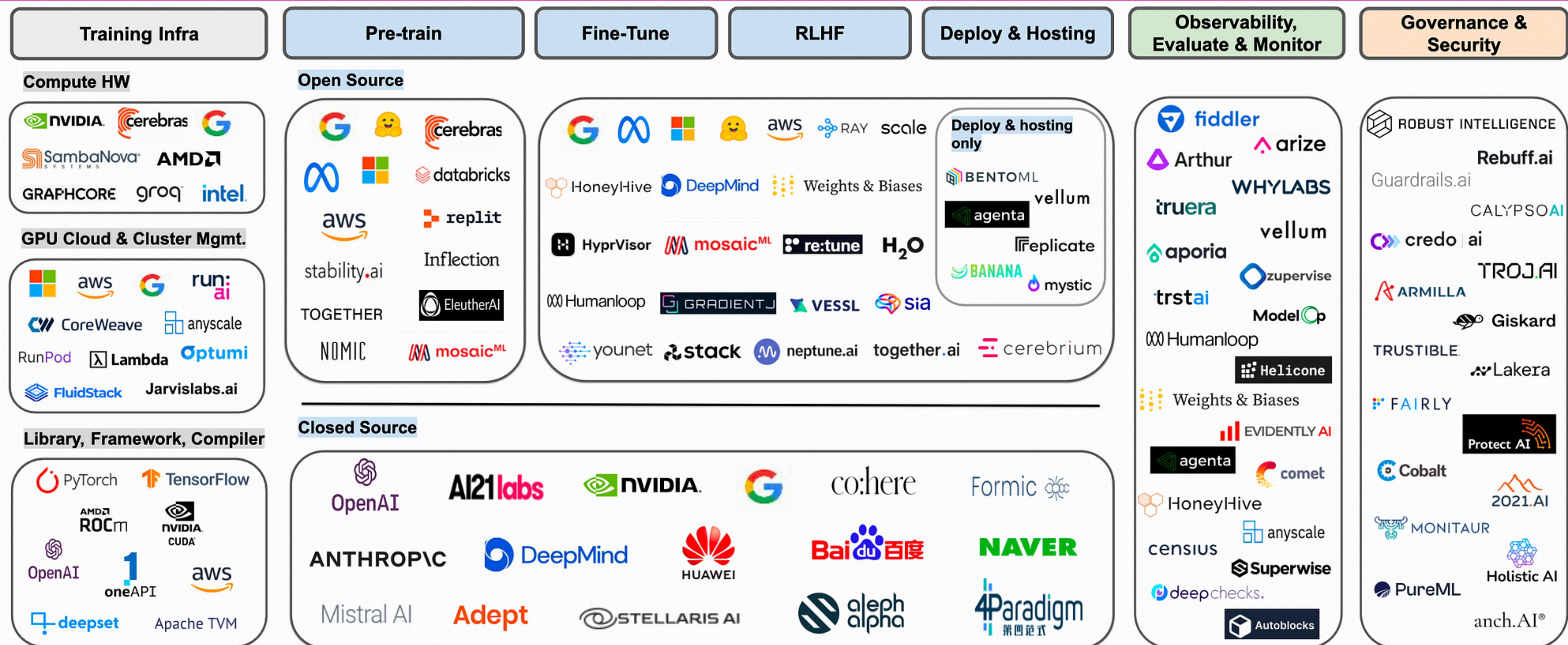
Test content

Test content can be used for a variety of purposes, from creating new content to testing existing content. It can help to identify problems or areas of improvement in content, as well as to ensure content is relevant and up-to-date. Test content can also be used to evaluate how effective content is, including how it is read, shared and interacted with. Test content can be beneficial for both marketing and technical content, and it is an essential part of any content strategy.

body p

+ capture context
(tech & business) to improve Test Data Set to improve Confidence

Ever expanding toolset





Enablement

How can we help the customers of our platform ?

Encourage Experimentation / Prototyping

The screenshot shows the Azure OpenAI Chat playground interface. On the left, a sidebar lists various AI services: Azure OpenAI, Playground, Chat, Completions, DALL-E (Preview), Management, Deployments, Models, Data files, Quotas, and Content filters (Preview). The Chat option is selected. The main area is titled "Chat playground" and contains several configuration panels:

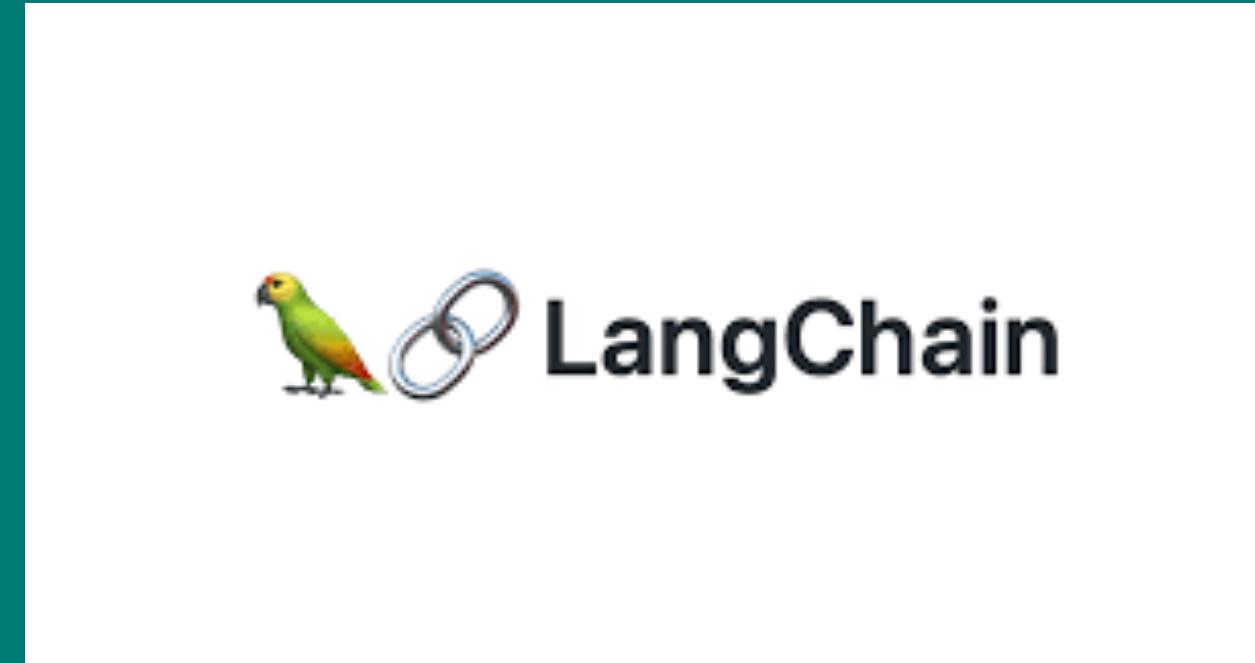
- Assistant setup**: A panel with tabs for "Prompt" (selected) and "Add your data (preview)". It includes an "Apply changes" button and a "Use a system message template" section.
- Different Models**: A central panel showing a "Start chatting" section with a robot icon and instructions to test the assistant by sending queries.
- Configuration**: A panel on the right showing deployment settings for "gpt4-pdb". It includes sections for "Deployment", "Parameters", "Deployment *", "Session settings" (with a slider for "Past messages included" set to 10), and "Current token count" (11/32768).

Large text overlays are present: "Connected with your Data" on the left and "Different Models" in the center.

Including Product Owners

<https://azure.microsoft.com/en-us/products/ai-studio>

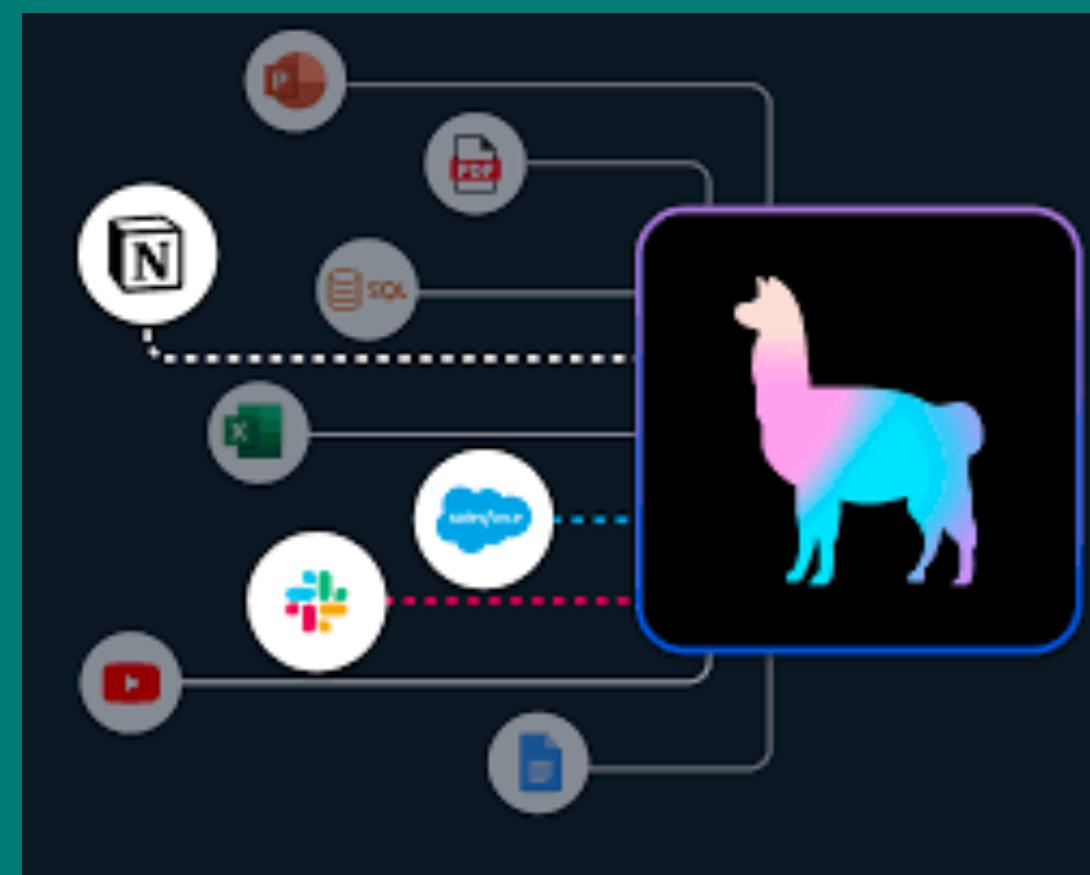
Learn from Orchestration Frameworks



<https://www.langchain.com/>

`{{guidance}}`

<https://github.com/guidance-ai/guidance>



<https://www.llamaindex.ai/>



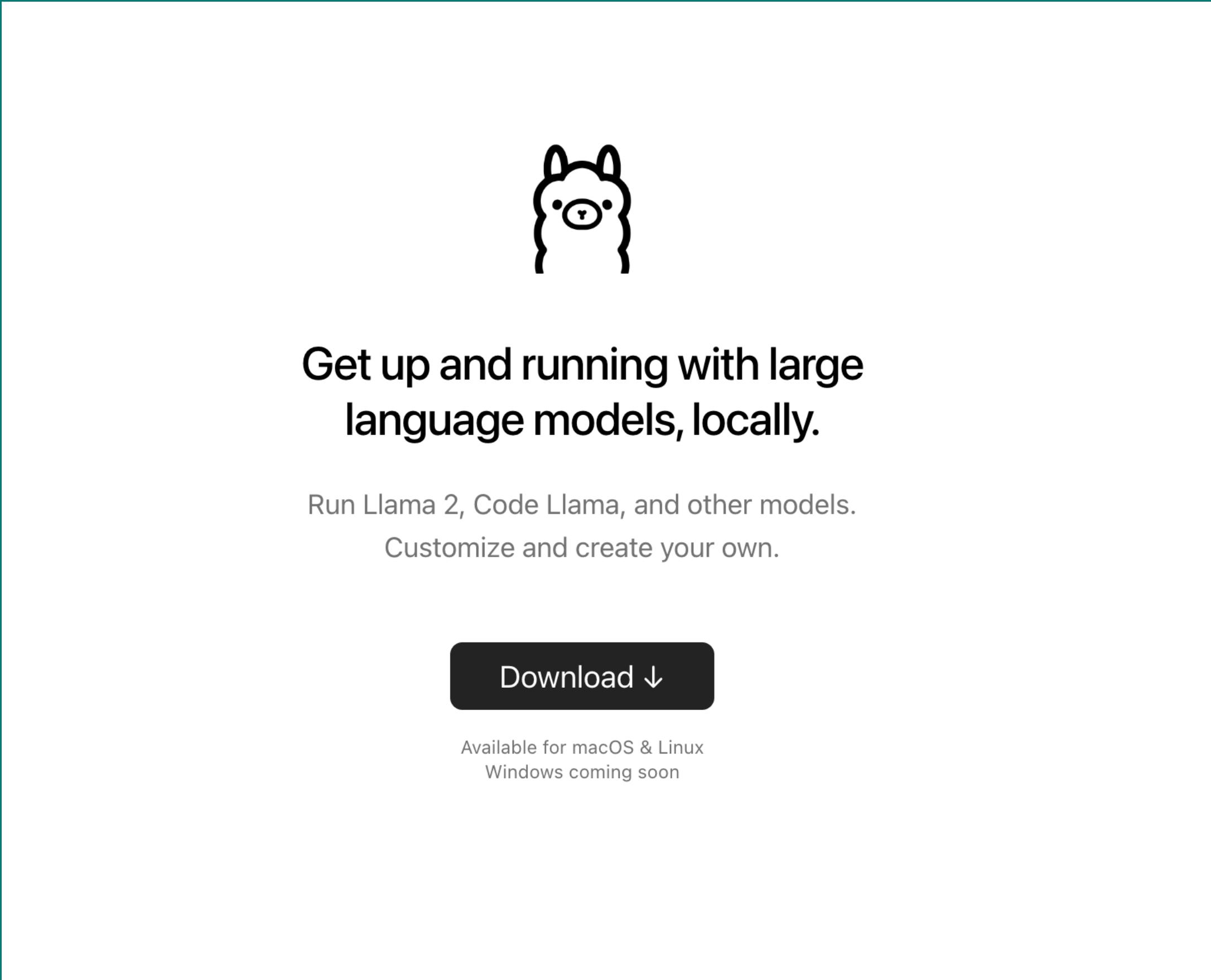
<https://www.griptape.ai/>



<https://github.com/stanfordnlp/dspy>

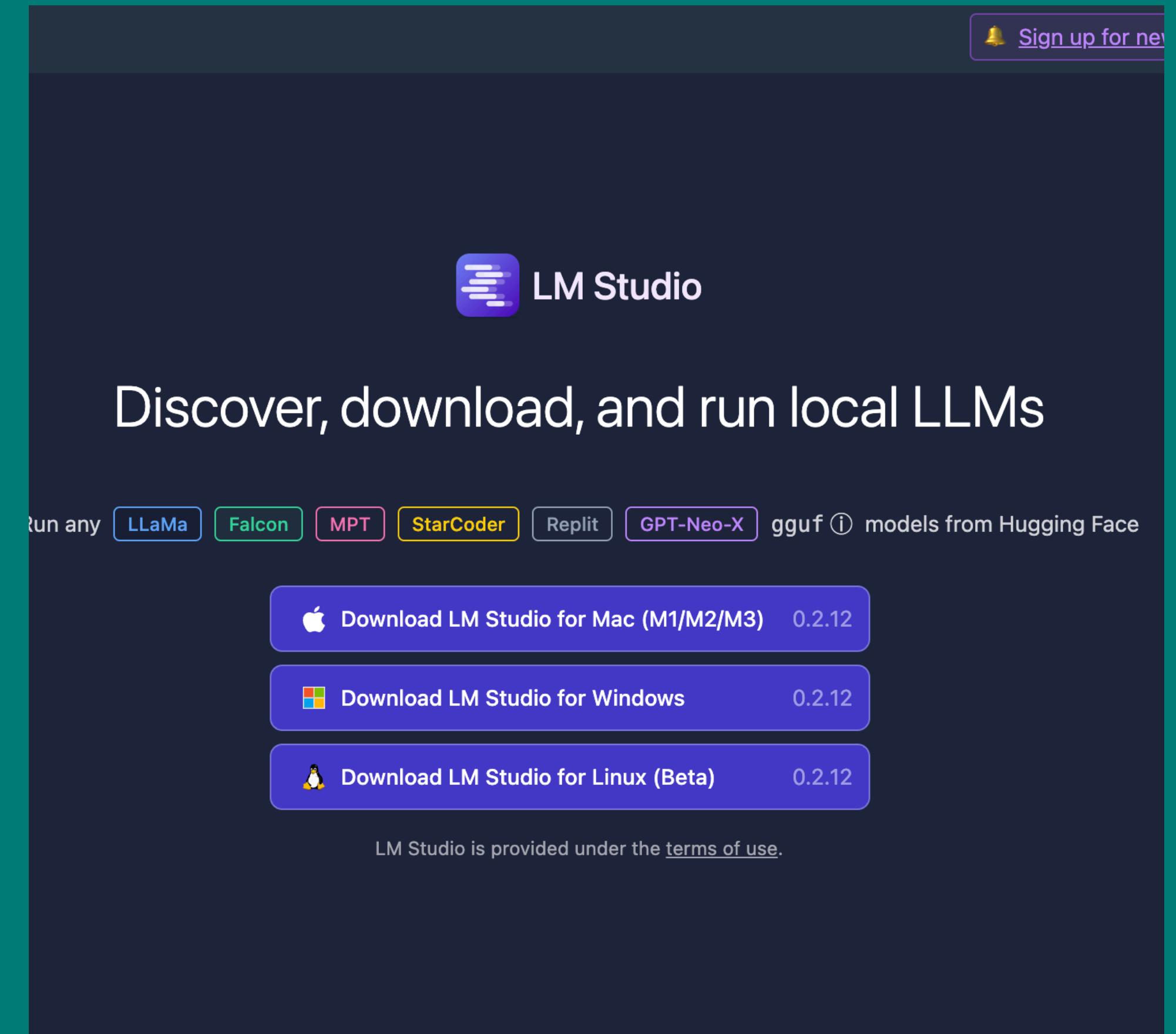
(Local) dev environment

<https://ollama.ai/>



The screenshot shows the Ollama AI website. At the top center is a large white llama icon. Below it, the text reads: "Get up and running with large language models, locally." In the middle, there's a section for "Run Llama 2, Code Llama, and other models." Below that, a "Customize and create your own." section is shown. A prominent "Download ↓" button is at the bottom left. At the very bottom, it says "Available for macOS & Linux" and "Windows coming soon".

<https://lmstudio.ai/>



The screenshot shows the LM Studio website. At the top right is a "Sign up for new features" button. The main heading is "Discover, download, and run local LLMs". Below it, a "Run any" section lists several model names: LLaMa (blue), Falcon (green), MPT (pink), StarCoder (yellow), Replit (purple), and GPT-Neo-X (orange). A note mentions "gguf ⓘ models from Hugging Face". At the bottom, there are three download buttons: "Download LM Studio for Mac (M1/M2/M3) 0.2.12" (with an Apple icon), "Download LM Studio for Windows 0.2.12" (with a Windows icon), and "Download LM Studio for Linux (Beta) 0.2.12" (with a Linux icon). A small note at the bottom right says "LM Studio is provided under the [terms of use](#)".

GenAI Project Pitfalls



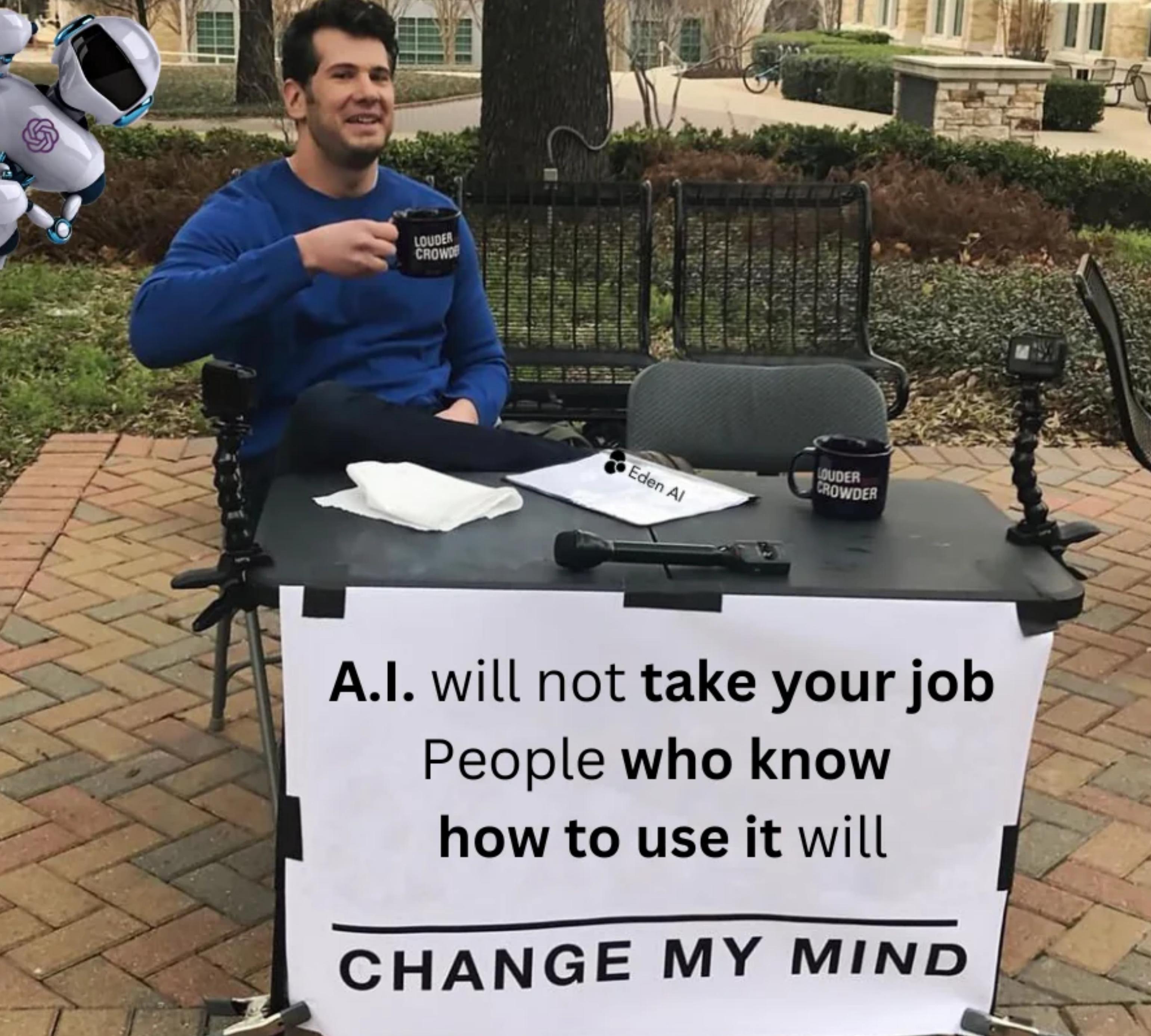
- 💡 No real business use case - Avoid AI for the sake of AI
- 🧪 Too focused on model training & fine tuning - Needs focus on customer
- 💰 Premature cost optimization - it will only get cheaper
- 👂 Enduser feedback not actionable - Needs ownership

Developer Experience Pains 😡

- 🤷 Model / embedding selection - Most just use OpenAI , will decide later
- 💾 Access to data in Test environments - Trick is to expose Data using API's
- 🚧 LLM Framework Bloat , Python centric - drives DIY , but not sustainable
- 📄 Outdated documentation - Rapid Change hard to keep up with
- 😅 Updating llms & prompts is painful & costly - Refactoring needs testing
- 🔎 Writing tests & evaluations data - Unsure how to deal with undeterminism

Test Mindset - Evaluation types

Exact Testing	Pattern Matching	Ask a Model	Semantic Distance	Ask an LLM	Ask a Human
Is this text > 20 characters	regex(/ keyword/)	Sentiment , Quality, Toxic, ...	Related concepts ?	Do you think this is a toxic answer ?	Got any Feedback?



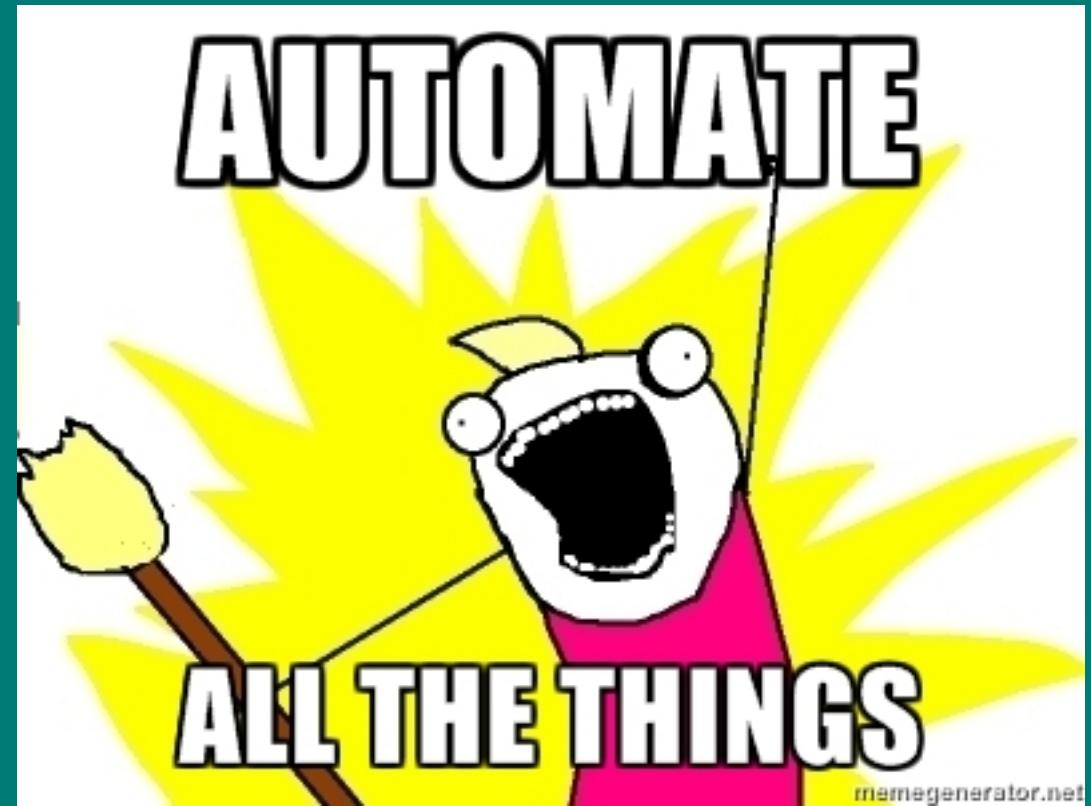
Coding time goes up but so do review times



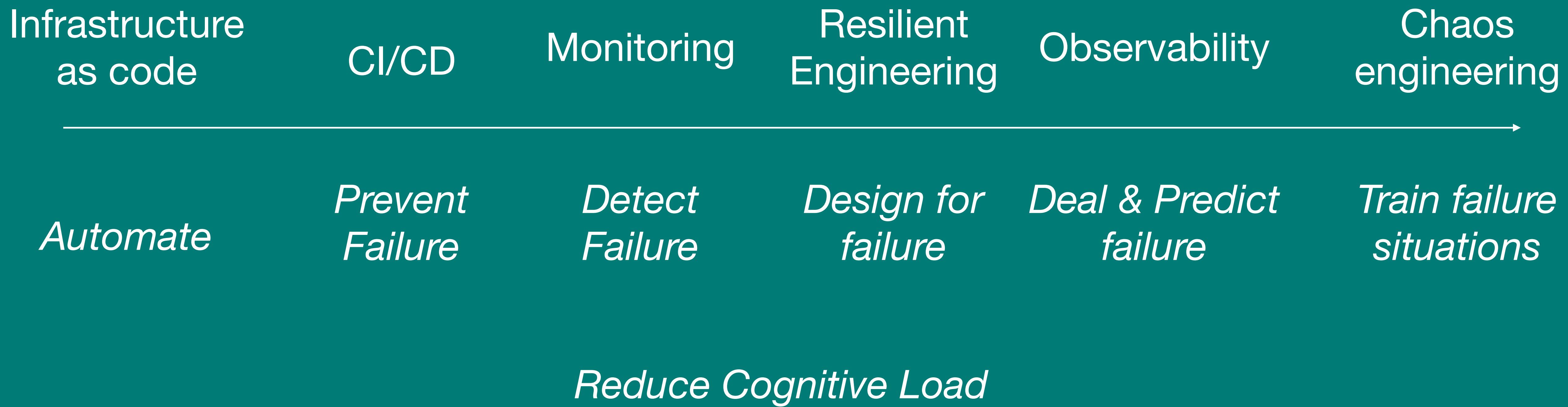
<https://linearb.io/resources/measuring-impact-the-genai-code-report>

Ironies of GenAI Automation

-  **role of user changes from producing to managing**
-  The easy things get easy, the hard parts (exceptions) get harder
-  Too many decisions causes decision fatigue
-  Tool needs to adapt to mental model of User
-  Gradual loss of situational awareness, needs training



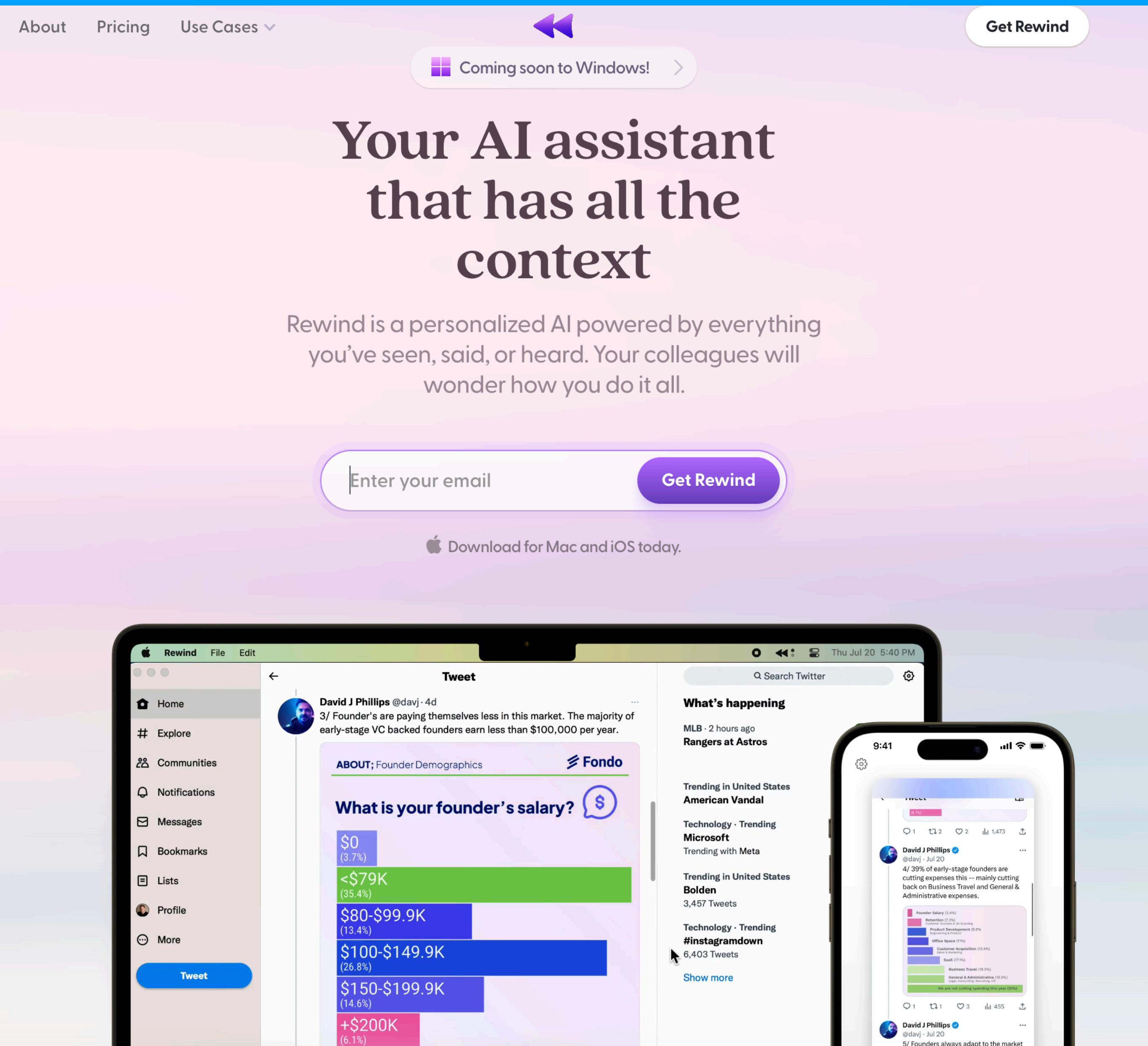
20% automation
80% prepare for failure





Governance

Creating a safe environment for our users to do their work



The image shows the Rewind AI website and its mobile application. The top half of the image displays the website's landing page with a pink-to-white gradient background. It features a navigation bar with 'About', 'Pricing', and 'Use Cases' links, a purple double arrow icon, a 'Coming soon to Windows!' button, and a 'Get Rewind' button. The main headline reads 'Your AI assistant that has all the context'. Below it is a subtext: 'Rewind is a personalized AI powered by everything you've seen, said, or heard. Your colleagues will wonder how you do it all.' A large input field for 'Enter your email' and a 'Get Rewind' button are at the bottom. A small note says 'Download for Mac and iOS today.' The bottom half of the image shows two phones side-by-side. The left phone is a Mac screen displaying the Rewind desktop application. The right phone is an iPhone displaying the Rewind mobile app. Both screens show a tweet from David J Phillips (@davj) about founder salaries, with a bar chart overlay showing salary distribution percentages.

Coming soon to Windows! >

Your AI assistant that has all the context

Rewind is a personalized AI powered by everything you've seen, said, or heard. Your colleagues will wonder how you do it all.

Enter your email

Get Rewind

Download for Mac and iOS today.

What's happening

MLB · 2 hours ago
Rangers at Astros

Trending in United States
American Vandal

Technology · Trending
Microsoft

Trending with Meta

Trending in United States
Bolden

Technology · Trending
#instagramdown

Show more

What is your founder's salary? \$

Salary Range	Percentage
\$0	(3.7%)
<\$79K	(35.4%)
\$80-\$99.9K	(13.4%)
\$100-\$149.9K	(26.8%)
\$150-\$199.9K	(14.6%)
+\$200K	(6.1%)

4/39% of early-stage founders are cutting expenses this — mainly cutting back on Business Travel and General & Administrative expenses.

Category	Percentage
Retention (7.7%)	
Product Development (9.8%)	
Office Space (11%)	
Customer Acquisition (13.4%)	
Sales & Marketing (15.1%)	
General & Administrative (18.3%)	
Logistics, Accounting, Recording, Inf.	

We are not cutting spending this year (21%)

David J Phillips @davj · Jul 20
4/ Founders always adapt to the market

Personal AI awareness programs

<https://www.rewind.ai/>

<https://github.com/jasonjmcghee/rem>

Opt-Out of AI Training

Getting started with AI services opt-out policies

Follow these steps to get started using Artificial Intelligence (AI) services opt-out policies.

1. [Enable AI services opt-out policies for your organization.](#)
2. [Create an AI services opt-out policy.](#)
3. [Attach the AI services opt-out policy to your organization's root, OU, or account.](#)
4. [View the combined effective AI services opt-out policy that applies to an account.](#)

For all of these steps, you sign in as an AWS Identity and Access Management (IAM) user, assume an IAM role, or sign in as the root user ([not recommended](#)) in the organization's management account.

How to opt out of OpenAI using your content

OpenAI has [documentation](#) on how to opt out of permitting their web scraper, **GPTbot**, from cataloging your site. You do this with a small text file at the root of your domain, `robots.txt`.

Learn all about this file type [here](#) on the robotstxt.org site. The robots.txt file is a public statement regarding who and what scrapes your site's content.

<https://supergeekery.com/blog/ai-and-your-content-how-to-opt-to-opt-out>

https://docs.aws.amazon.com/organizations/latest/userguide/orgs_manage_policies_ai-opt-out.html

What are Model Cards?

Model cards are files that accompany the models and provide handy information. Under the hood, model cards are simple Markdown files with additional metadata. Model cards are essential for discoverability, reproducibility, and sharing! You can find a model card as the README.md file in any model repo.

The model card should describe:

- the model
- its intended uses & potential limitations, including biases and ethical considerations as detailed in [Mitchell, 2018](#)
- the training params and experimental info (you can embed or link to an experiment tracking platform for reference)
- which datasets were used to train your model
- the model's evaluation results

The model card template is available [here](#).

How to fill out each section of the model card is described in [the Annotated Model Card](#).

Model Cards on the Hub have two key parts, with overlapping information:

- [Metadata](#)
- [Text descriptions](#)

Model Origins & Licenses Cards ~ SBOMs

<https://huggingface.co/docs/hub/en/model-cards>

EU Legislation - Risk Levels

<https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

Generative AI

Generative AI, like ChatGPT, would have to comply with transparency requirements:

- Disclosing that the content was generated by AI
- Designing the model to prevent it from generating illegal content
- Publishing summaries of copyrighted data used for training

Limited risk

Limited risk AI systems should comply with minimal transparency requirements that would allow users to make informed decisions. After interacting with the applications, the user can then decide whether they want to continue using it. Users should be made aware when they are interacting with AI. This includes AI systems that generate or manipulate image, audio or video content, for example deepfakes.

Unacceptable risk

Unacceptable risk AI systems are systems considered a threat to people and will be banned. They include:

- Cognitive behavioural manipulation of people or specific vulnerable groups: for example voice-activated toys that encourage dangerous behaviour in children
- Social scoring: classifying people based on behaviour, socio-economic status or personal characteristics
- Real-time and remote biometric identification systems, such as facial recognition

Some exceptions may be allowed: For instance, “post” remote biometric identification systems where identification occurs after a significant delay will be allowed to prosecute serious crimes but only after court approval.

High risk

AI systems that negatively affect safety or fundamental rights will be considered high risk and fall into two categories:

- 1) AI systems that are used in products falling under [the EU's product safety legislation](#), such as those used in transport, aviation, cars, medical devices and lifts.
- 2) AI systems falling into eight specific areas that will have to be registered in an EU database:
 - Biometric identification and categorisation of natural persons
 - Management and operation of critical infrastructure
 - Education and vocational training
 - Employment, worker management and access to self-employment
 - Access to and enjoyment of essential private services and public services and border control management

border control management
erpretation and application of the law.

will be assessed before being put on the market and also th

Guard Rails as Service ~ WAF

Filter strengths for prompts

[Reset](#)

Use a higher filter strength to increase the likelihood of filtering harmful content in a given category.

Enable filters for prompts



Select resource

demo-llm-chatbot

PII & Security Metrics

Security

Performance

Cohorts

Select batch:

02/05/2023 00:00:00 UTC

Compare with

ChatGPT-3.5Turbo

Select comparison batch

02/05/2023 00:00:00 UTC

Go to anomaly feed

Show events

Show insights

demo-llm-chatbot

reponse.data_leakage

2 tags | When PII is produced by the model the count of items i... [Read more](#)

text count

56Based on selected batch:
02/05/2023 00:00:00 UTC

Frequent item count ?

Monitors:

has_patterns_monitor



ChatGPT-3.5-Turbo

reponse.data_leakage

3 tags | When PII is produced by the model the count of items i... [Read more](#)

text count

41Based on selected batch:
02/05/2023 00:00:00 UTC

Frequent item count ?

Monitors:

has_patterns_monitor



demo-llm-chatbot

p99 ?

⚠️ Not monitored

[Learn more](#)

prompt.jailbreak_risk

4 tags | When prompts that attempt to jailbreak the model are in... [Read more](#)

p99

0.7261Based on selected batch:
02/05/2023 00:00:00 UTC

p99



ChatGPT-3.5-Turbo

p99 ?

⚠️ Not monitored

[Learn more](#)

prompt.jailbreak_risk

2 tags | When prompts that attempt to jailbreak the model are in... [Read more](#)

p99

0.0507Based on selected batch:
02/05/2023 00:00:00 UTC

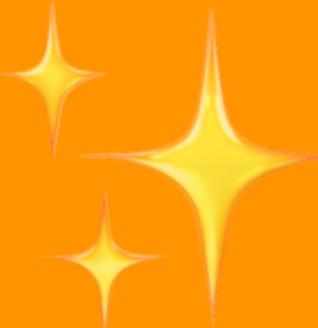
p99



Steps to scale the change



Platform

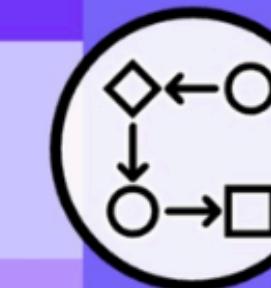


Enablement

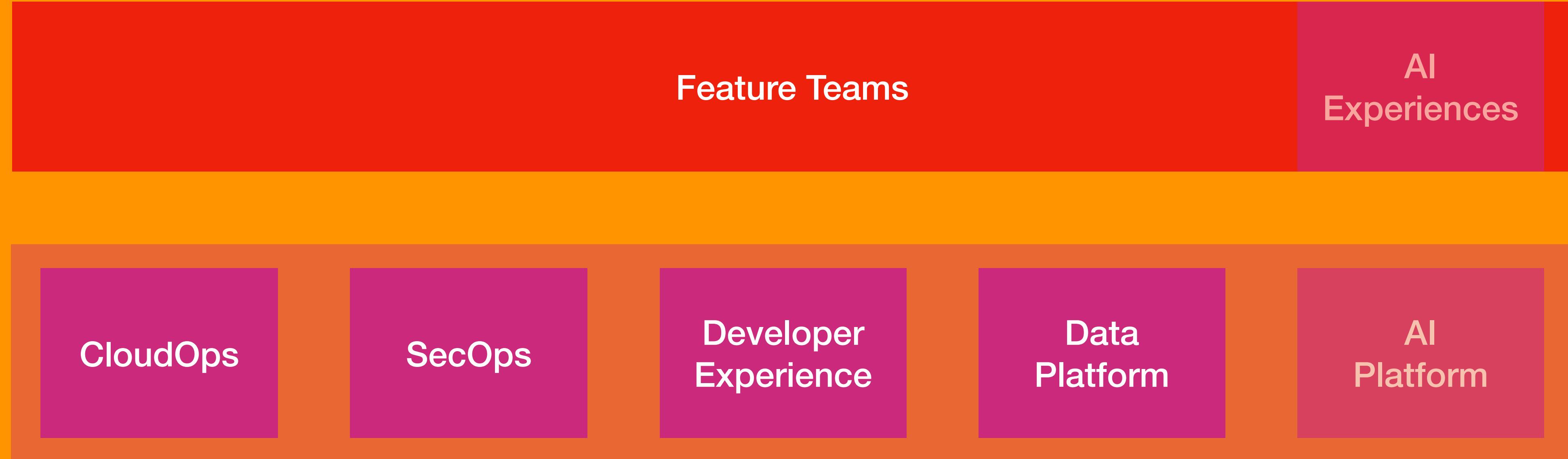


Governance

Beyond Platform teams

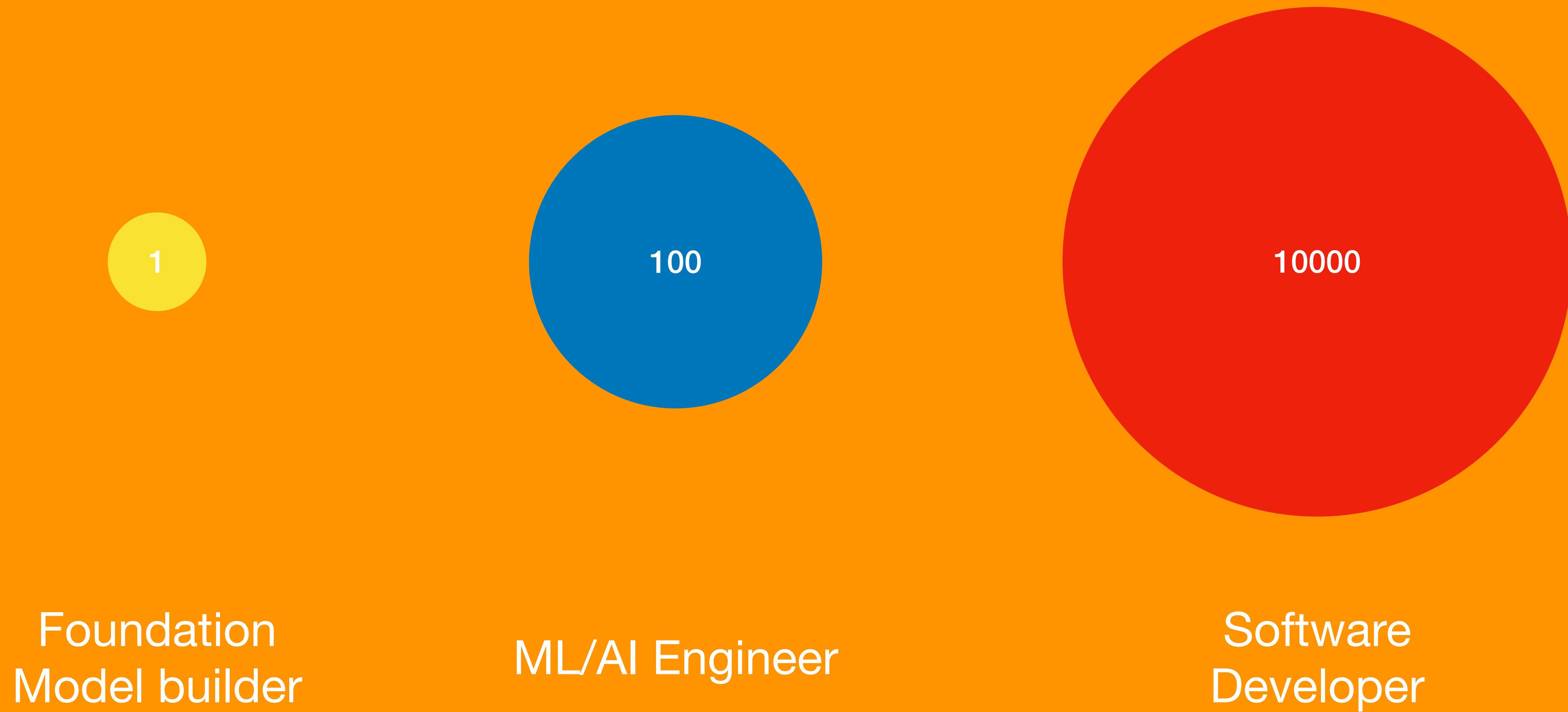
<p>Crew Type Value Stream Crew</p>  <p>unfix.com/value-stream-crew</p> <p>The Value Stream Crew has end-to-end responsibility for a value stream.</p> <p>001</p>	<p>Crew Type Governance Crew</p>  <p>unfix.com/governance-crew</p> <p>The Governance Crew is the team that sets constraints on self-organization.</p> <p>002</p>	<p>Crew Type Platform Crew</p>  <p>unfix.com/platform-crew</p> <p>The Platform Crew offers shared services to everyone else in the Base.</p> <p>003</p>	<p>Crew Type Facilitation Crew</p>  <p>unfix.com/facilitation-crew</p> <p>The Facilitation Crew enables other Crews to get their work done.</p> <p>004</p>
<p>Crew Type Partnership Crew</p>  <p>unfix.com/partnership-crew</p> <p>The Partnership Crew cares about vendors, freelancers, and gig workers.</p> <p>005</p>	<p>Crew Type Experience Crew</p>  <p>unfix.com/experience-crew</p> <p>The Experience Crew ensures that the customer experience is a great one.</p> <p>006</p>	<p>Crew Type Capability Crew</p>  <p>unfix.com/capability-crew</p> <p>The Capability Crew offers unique expertise to everyone in the Base.</p> <p>007</p>	<p>Unfix Model</p>

Cross Platform Team collaboration



Run the Platform
Enable users on the Platform
Govern the Platform

Talent ratio

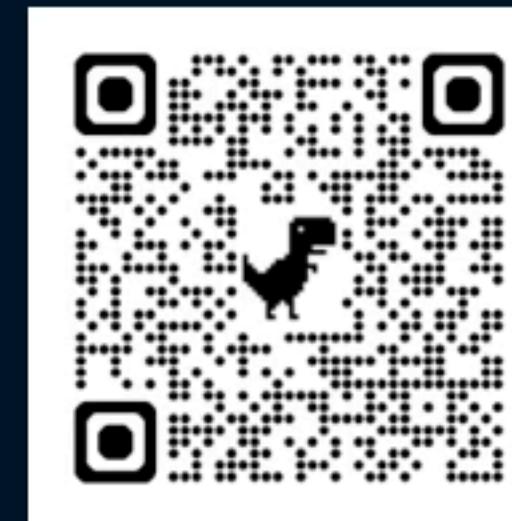


Hiring thesis

- A full-stack engineer can integrate our system with an LLM provider, and build the minimum infrastructure to connect user requests to the generated results from the LLM.
- A data scientist can understand the basics of evaluation, quality, and user data to continuously improve the AI product.
- A product person can focus our efforts, talk to users, and help us prioritize what's essential to learn the jobs to be done.
- A designer can help us identify the ineffable experience users have integrating with generative AI, and ensure it's delightful.
- An MLE can push our capabilities forward, ensuring we're not bound to commodity intelligence.

2 DAY WORKSHOP

Patrick Debois & John Willis



GenAI for DevOps engineers

www.jedi.be



Austin Texas, USA
12-13 September 2024

<https://jedi.be/workshops/genai-for-devops-engineers/>

I ❤️ to hear about your journey !

Come say hi or connect for the slides



<https://www.youtube.com/@jedi4ever>



<https://www.linkedin.com/in/patrickdebois/>

