Every AI engineer deserves an AI platform (team)

organising for the next big thing

@patrickdebois

Who am I

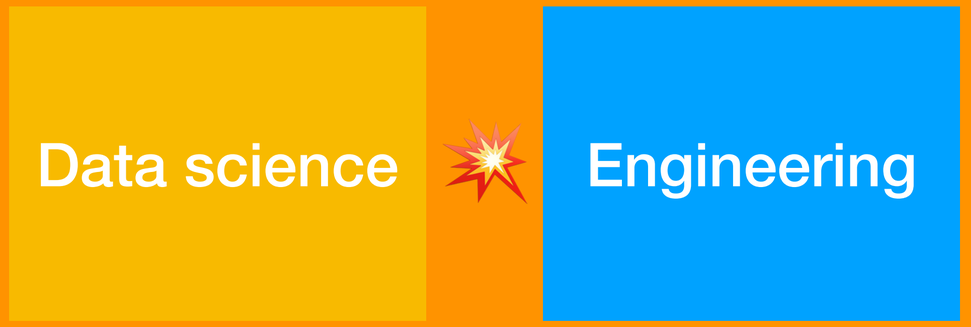
- Patrick Debois
- Started DevOps(Days)
- Co-Author of the DevOps Handbook
- Independent Industry Advisor
- Dev, Sec, Ops and now GenAI
- **Automation Intelligence**

Introducing Org & Tech Change

Incubate **Repeat** **Scale**

Maturity

Change Agents want an identity



AI Engineer
DevOps
Agile
ITIL

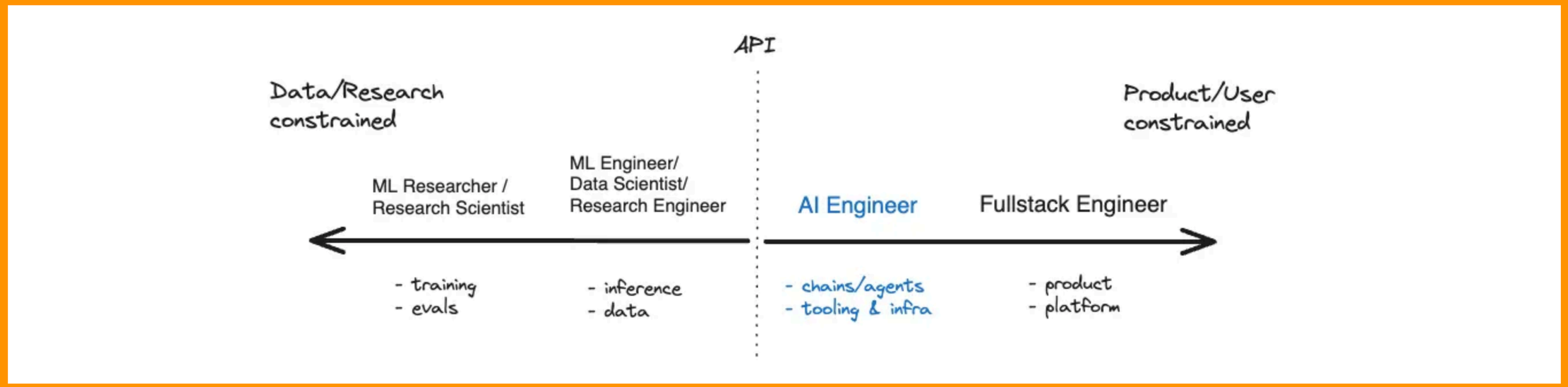
Names are ambiguous

Yet resonate with a pain

Have a matching technology shift

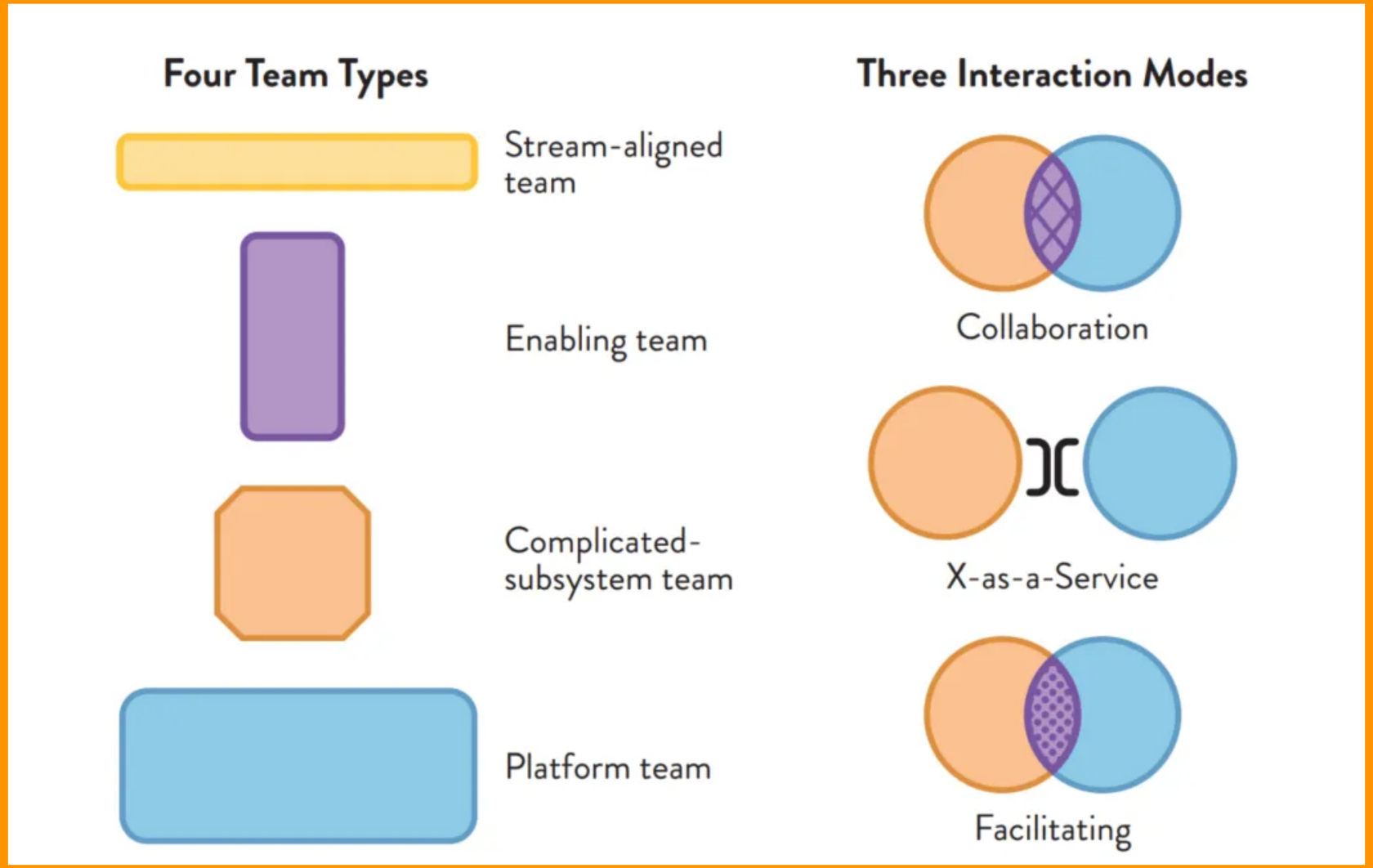
Act a label to get the stories out

Org shift - Shift Right - Breaking Silos



https://www.latent.space/p/ai-engineer

”Platform” team pattern



https://teamtopologies.com/

Steps to scale the change

 Platform

 Enablement

 Governance

Platform

What services go into your platform

Access to Models

Enterprise SaaS

OpenAI

Microsoft Azure
OpenAI

MISTRAL AI_

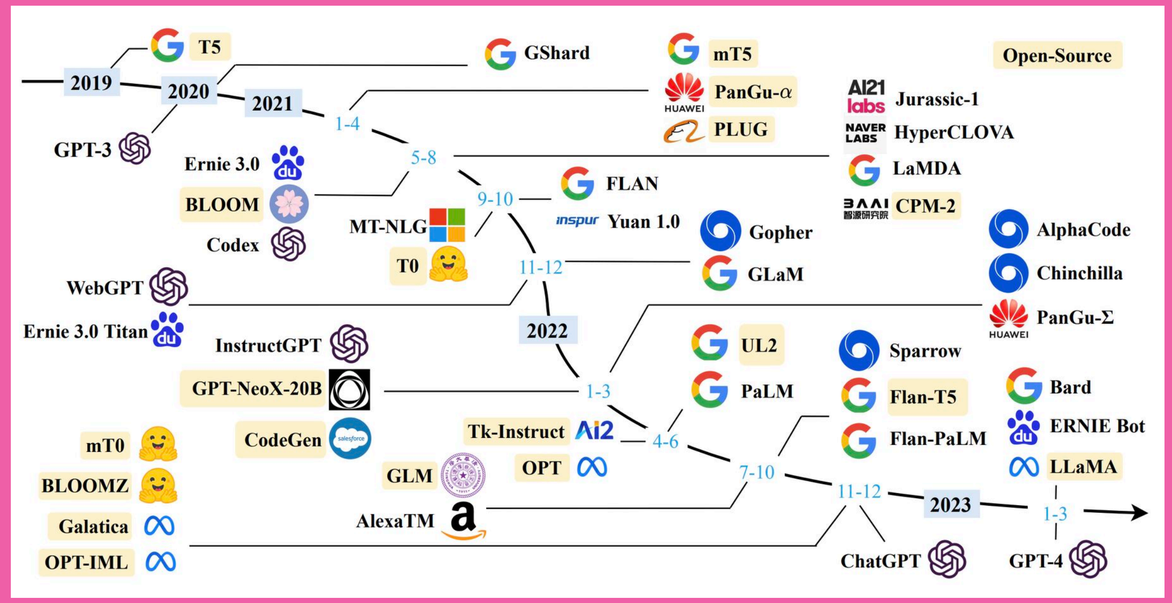
Vertex AI

Claude
BY ANTHROPIC

Amazon Bedrock

Model Broker

Open Source



Cloud Vendors

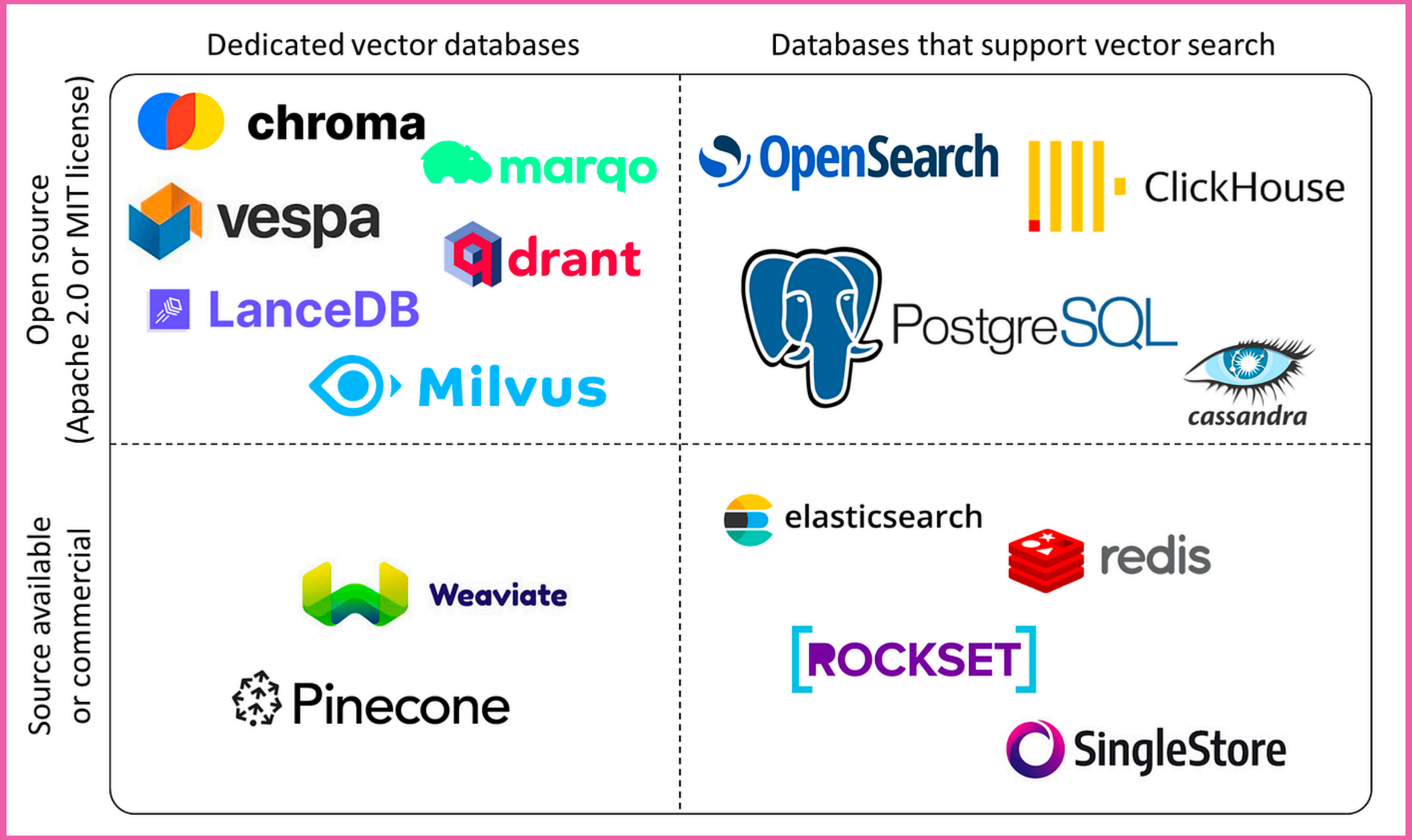
Independent Models

Meta
Llama 3

Microsoft
Phi-3

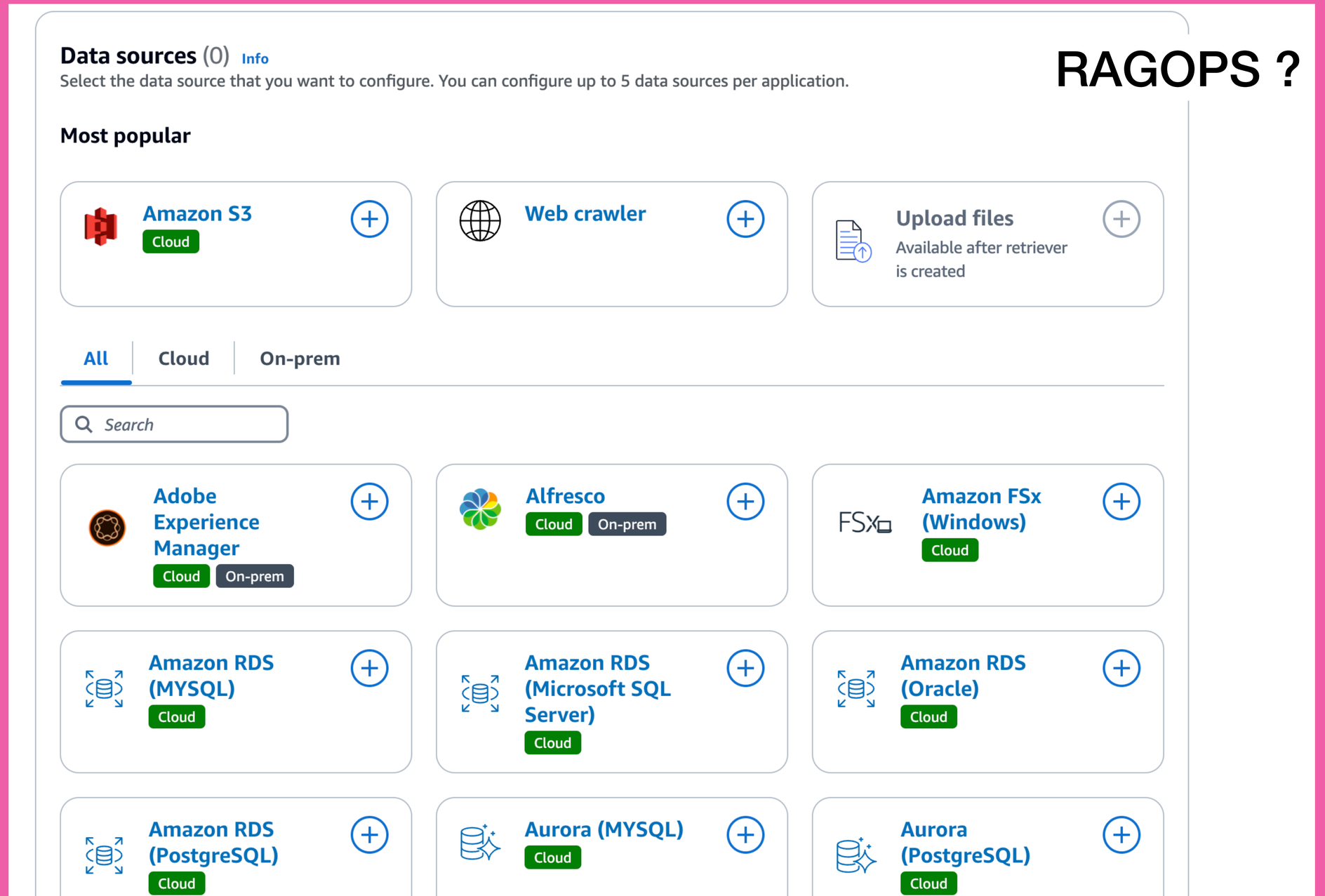
Bring your own Model

Unstructured data - Vector Databases



RAG as a Service

RAGOPS ?



Products Solutions Pricing Developers Resources Partners En

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:

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

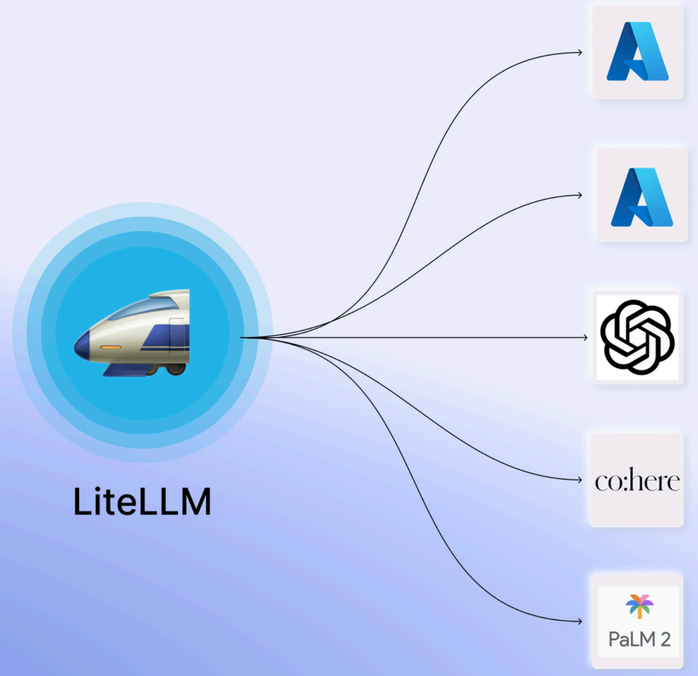
LiteLLM

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

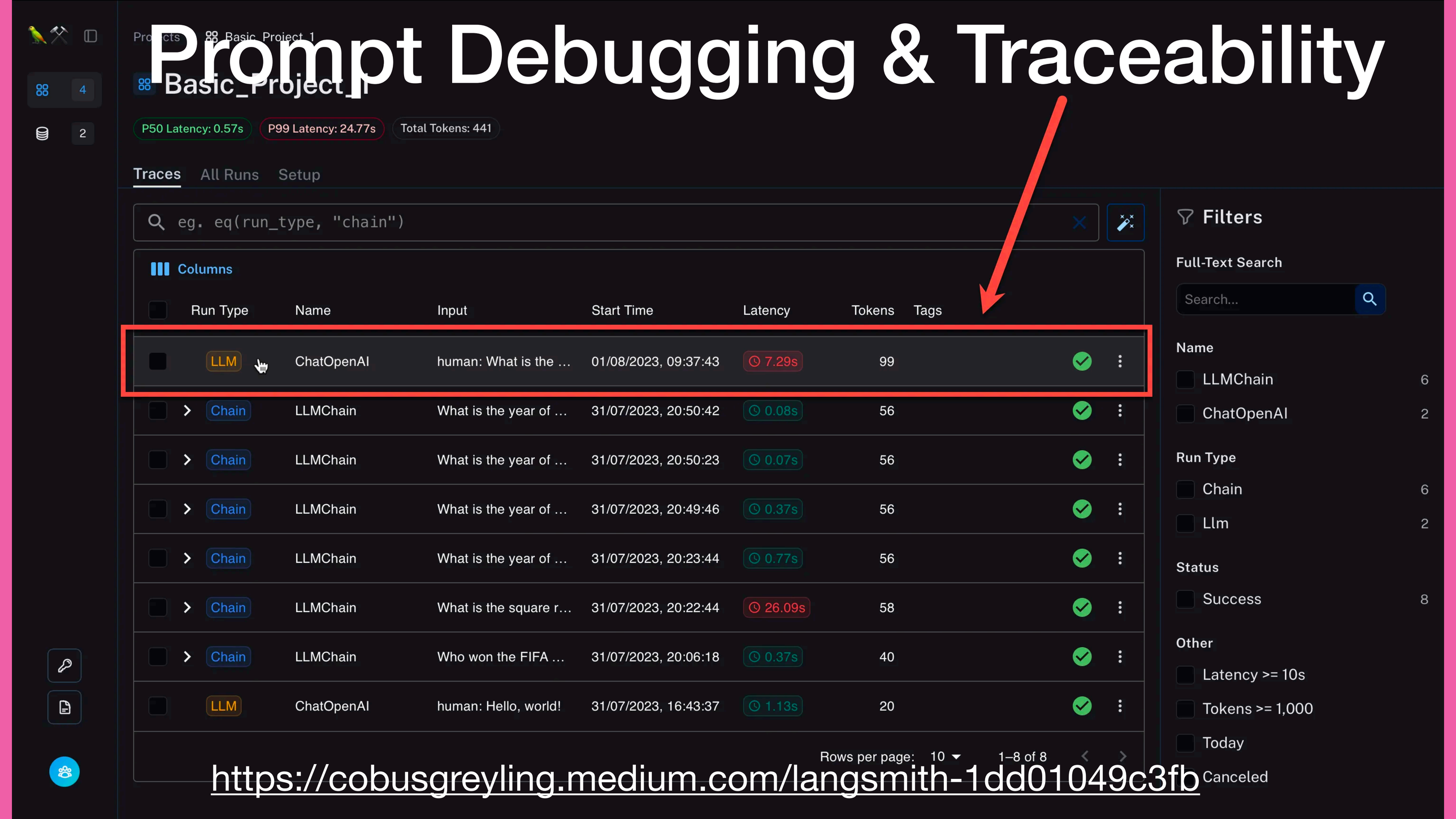
Try LiteLLM Cloud Free Deploy LiteLLM Open Source →



https://github.com/BerriAI/litellm

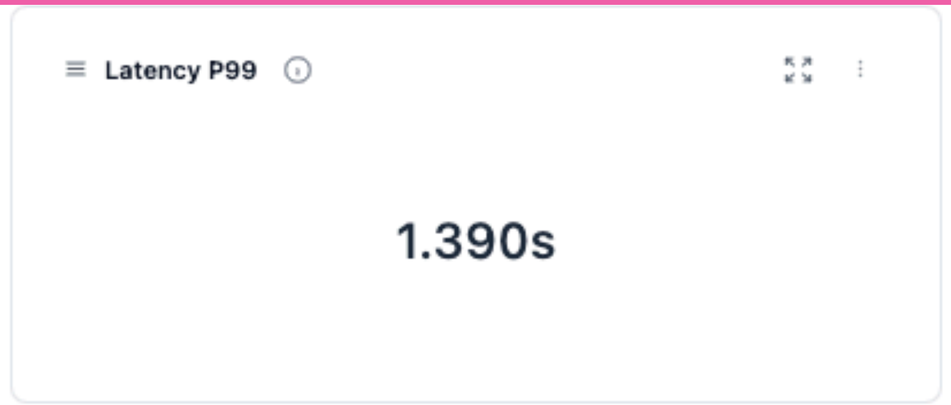
https://www.litellm.ai/

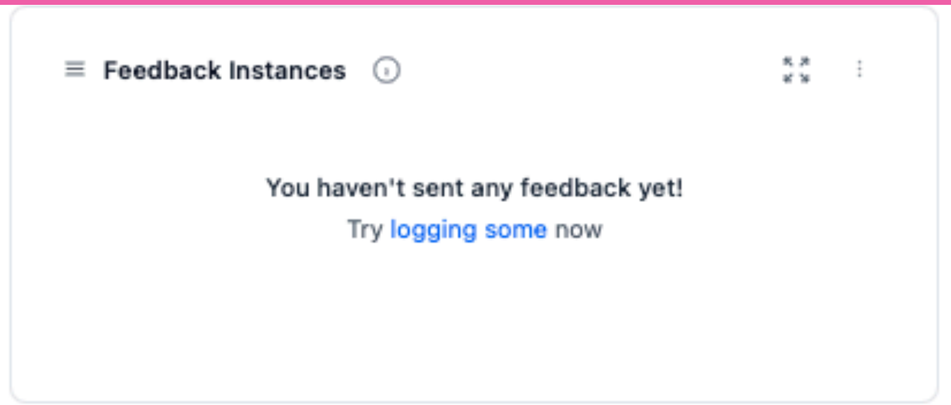
Prompt Debugging & Traceability

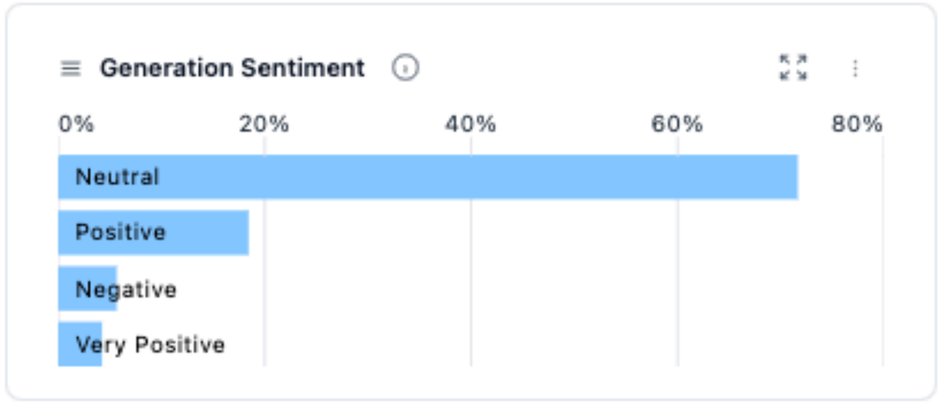


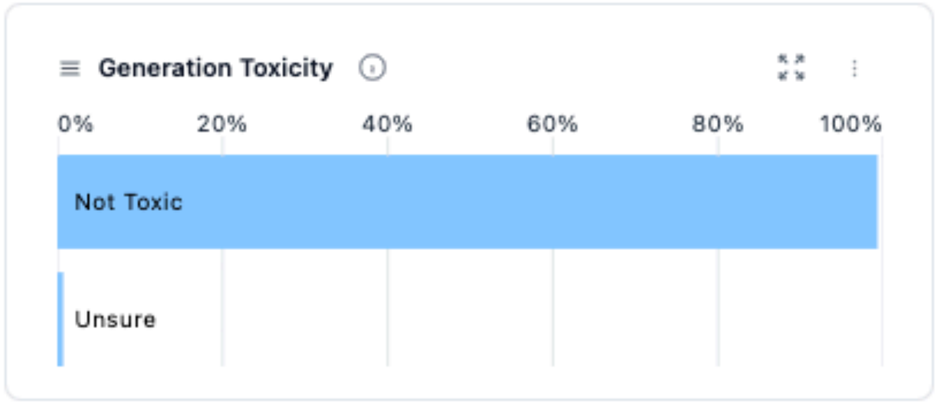
https://cobusgreyling.medium.com/langsmith-1dd01049c3fb

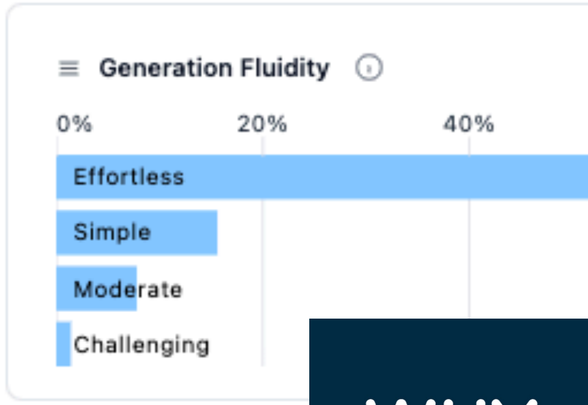
Continuous Data Quality Monitoring

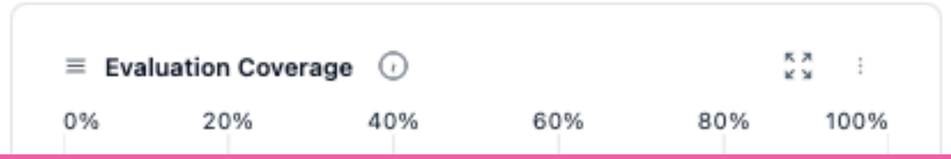










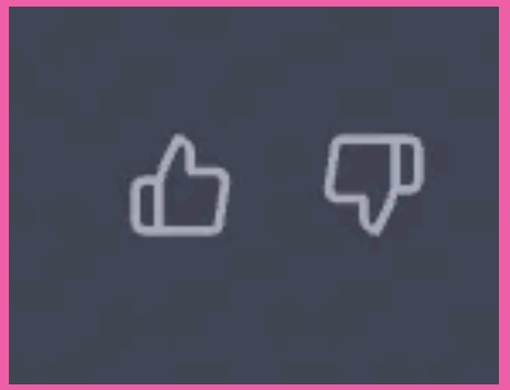


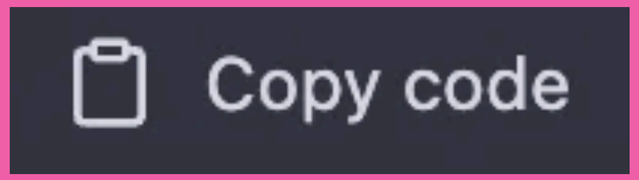
Use Small & Faster Models

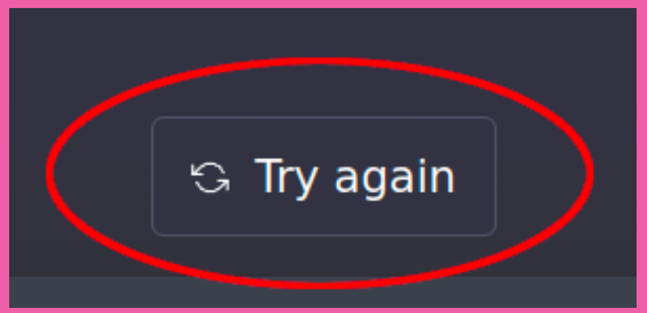
WHY LABS

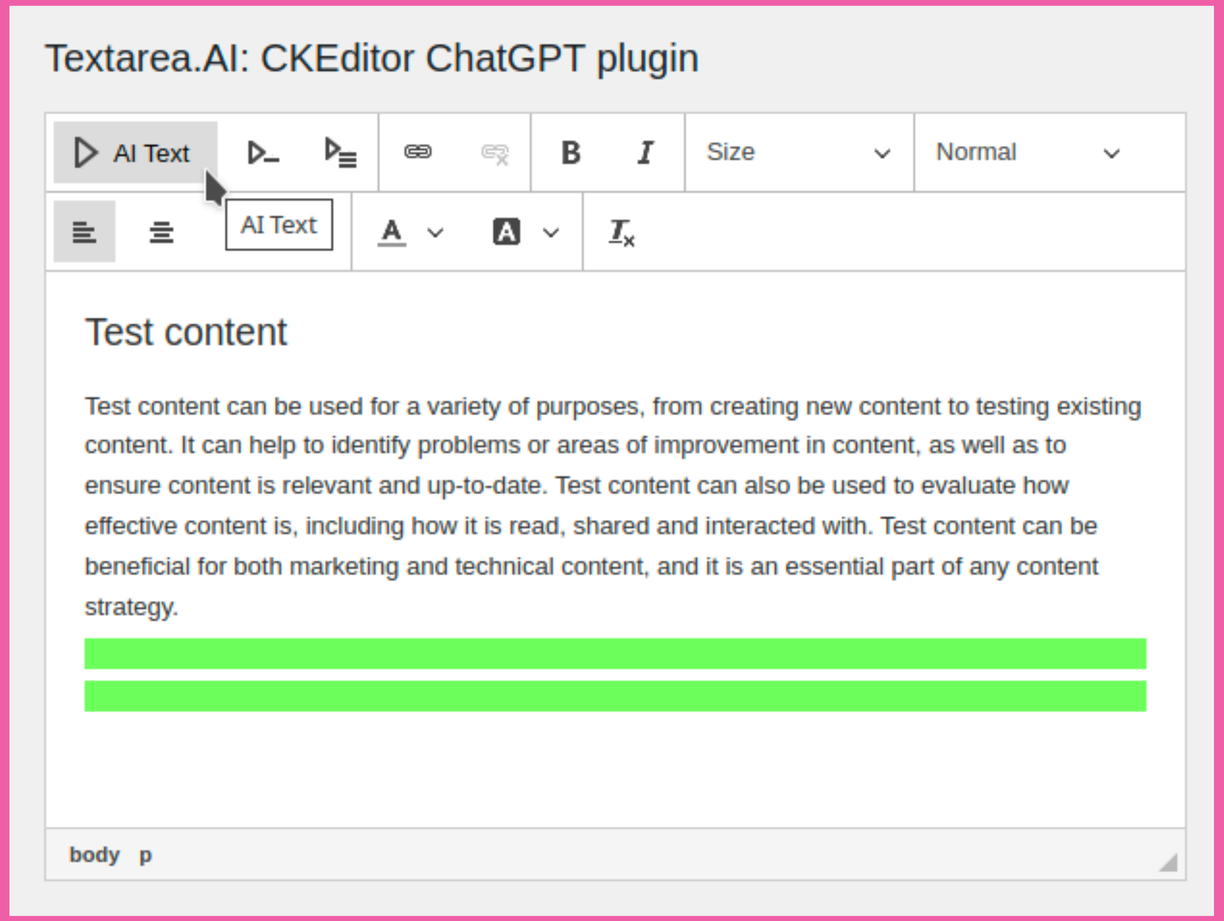
https://whylabs.ai/

Feedback as a service



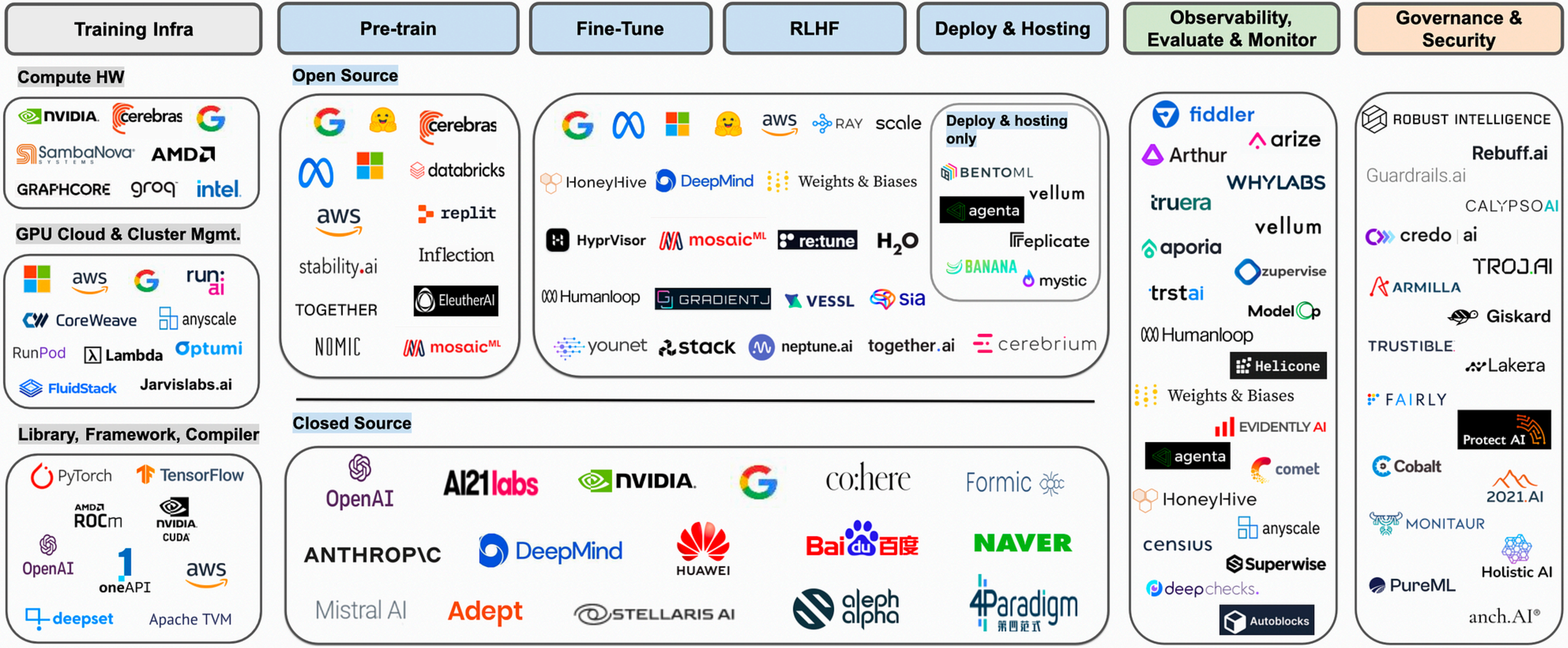






+ capture context
(tech & business) to improve Test Data Set to improve Confidence

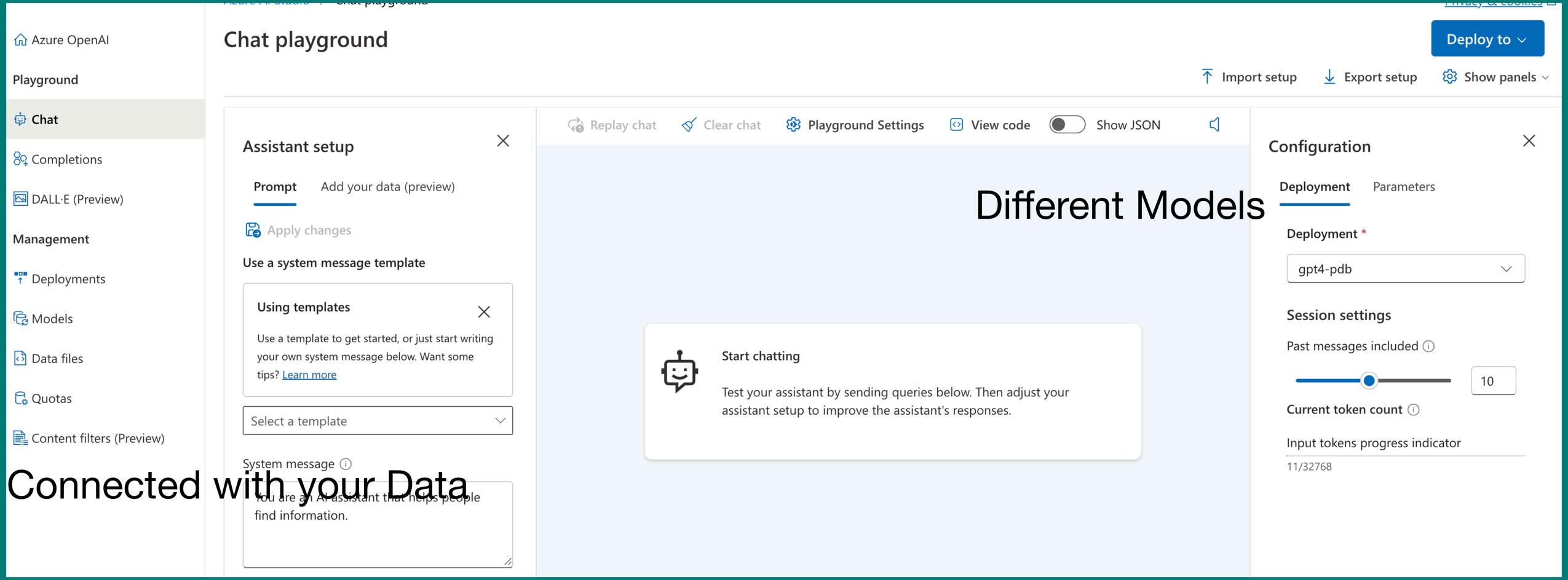
Ever expanding toolset



✨ Enablement

How can we help the customers of our platform ?

Encourage Experimentation / Prototyping



Connected with your Data

Different Models

Including Product Owners

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

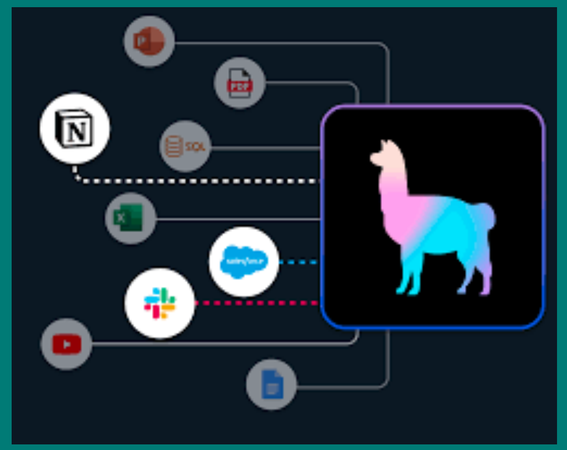
Learn from Orchestration Frameworks

LangChain

https://www.langchain.com/

{{guidance}}

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



https://www.llamaindex.ai/

griptape

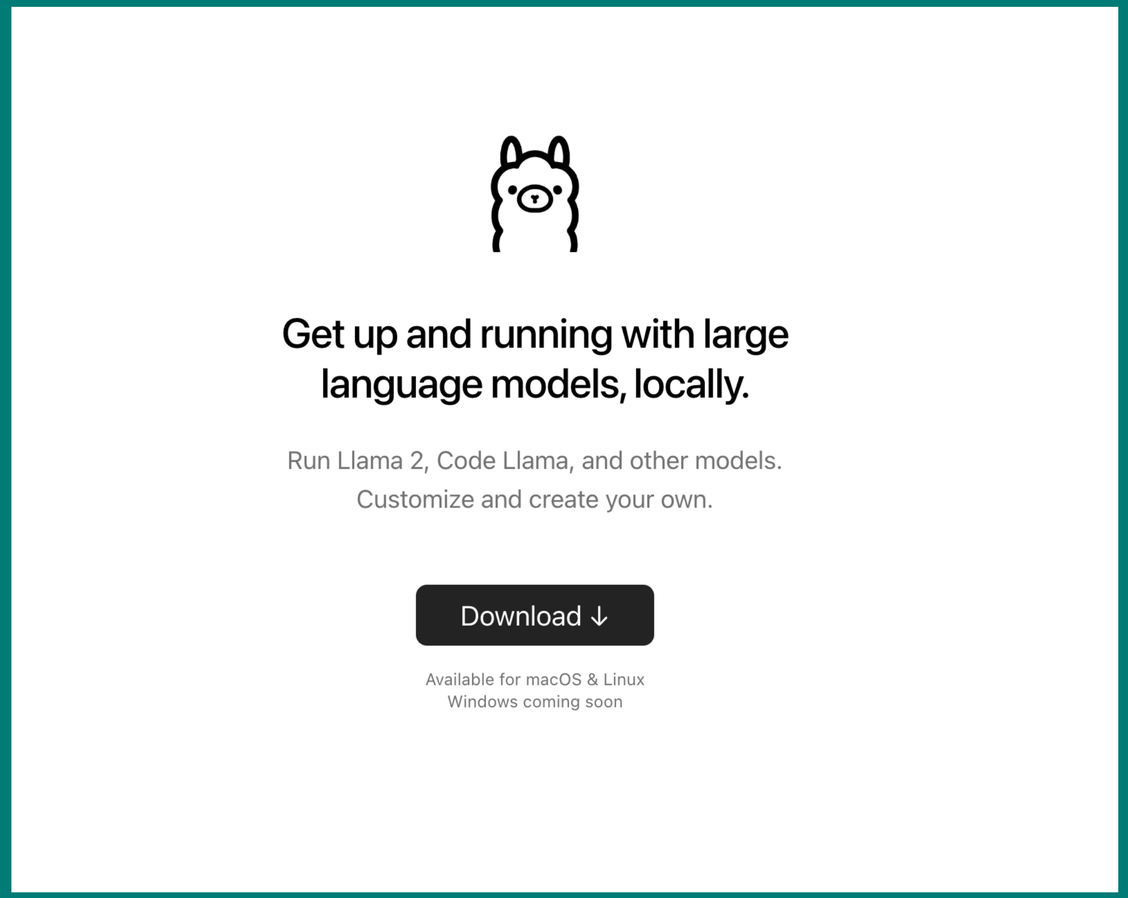
https://www.griptape.ai/

DSPy

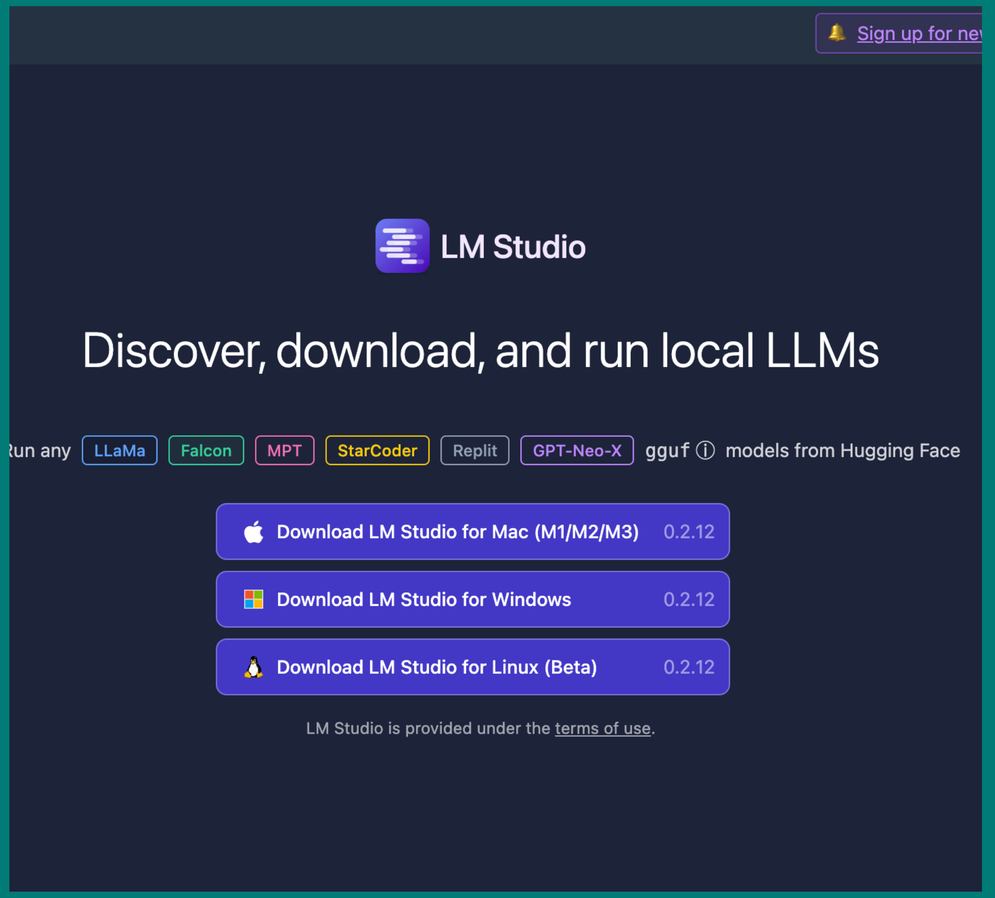
https://github.com/stanfordnlp/dspy

(Local) dev environment

https://ollama.ai/



https://lmstudio.ai/



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

@patrickdebois

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**

@patrickdebois

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?

@patrickdebois

https://www.reddit.com/r/ProgrammerHumor/comments/10bl4xn/get_up_and_start_integrating_ai_in_your_code/

Coding time goes up but so do review times

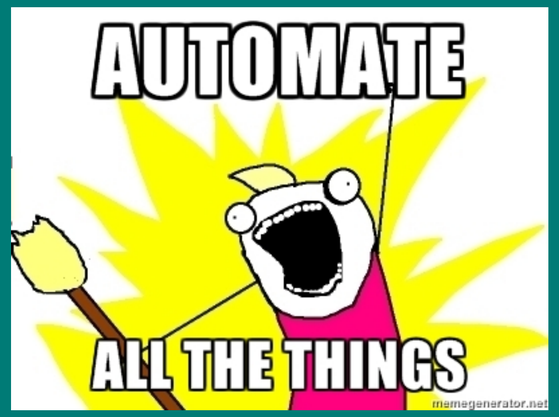
https://linearb.io/resources/measuring-impact-the-genai-code-report

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

https://arxiv.org/abs/2402.11364v1

@patrickdebois



20% automation
80% prepare for failure

Infrastructure as code | CI/CD | Monitoring | Resilient Engineering | Observability | Chaos engineering

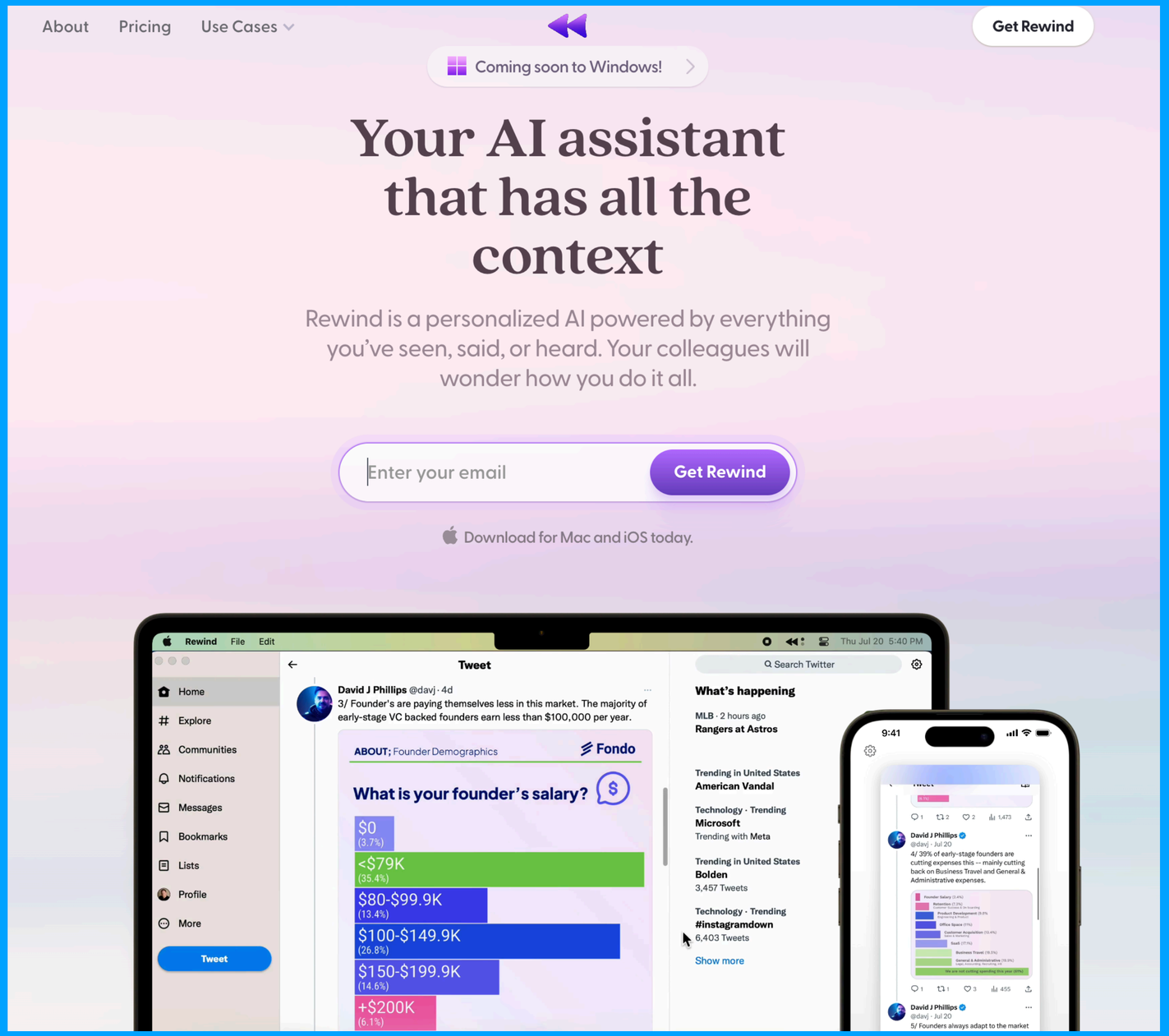
Automate | *Prevent Failure* | *Detect Failure* | *Design for failure* | *Deal & Predict failure* | *Train failure situations*

Reduce Cognitive Load

@patrickdebois

Governance

Creating a safe environment for our users to do their work



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 documention 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.

High risk

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

1) AI systems that are used in products falling under the EU's product safety legislat
aviation, cars, medical devices and lifts.

2) AI systems falling into eight specific areas that will have to be registered in an EU

- 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 be

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

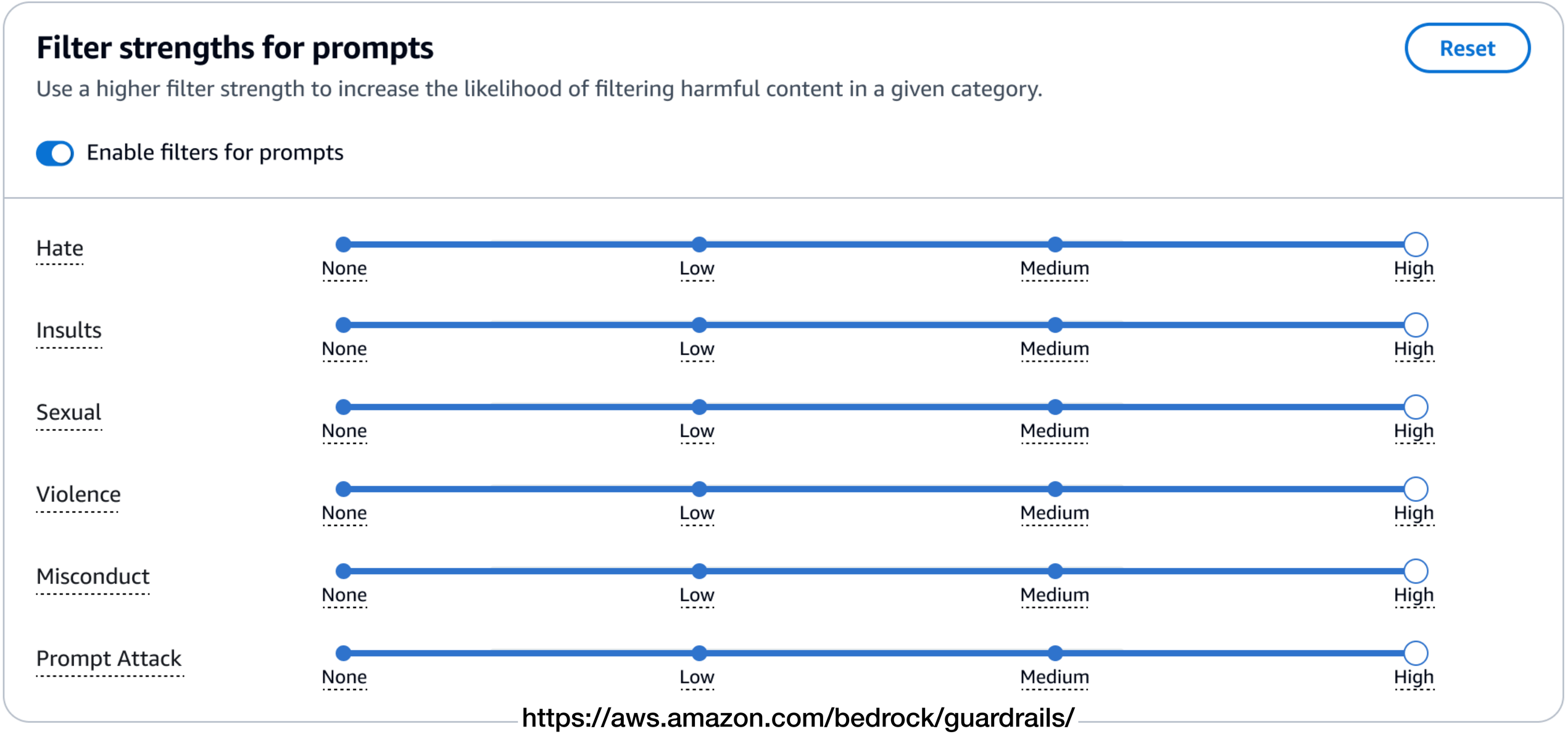
Some exceptions may be allowed: For instance, "post" remote biometric identification systems where identification occurs after a significant delay will be allowed to prosecute serious crimes but only after court approval.

l border control management

erpretation and application of the law.

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

Guard Rails as Service ~ WAF



https://aws.amazon.com/bedrock/guardrails/

PII & Security Metrics

WHYLABS

Select resource

demo-llm-chatbot

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

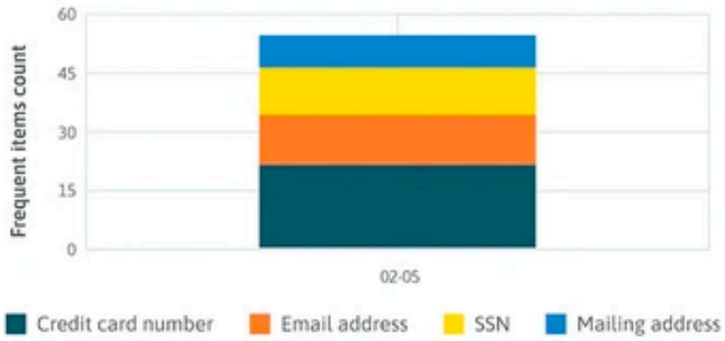
text count

56

Based 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

41

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

Frequent item count

Monitors: has_patterns_monitor

demo-llm-chatbot

prompt.jailbreak_risk

4 tags | When prompts that attempt to jailbreak the model are inp... Read more

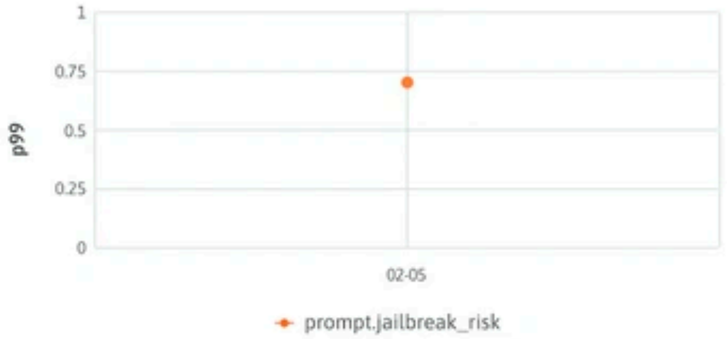
p99

0.7261

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

p99

Not monitored | Learn more



ChatGPT-3.5-Turbo

prompt.jailbreak_risk

2 tags | When prompts that attempt to jailbreak the model are inp... Read more

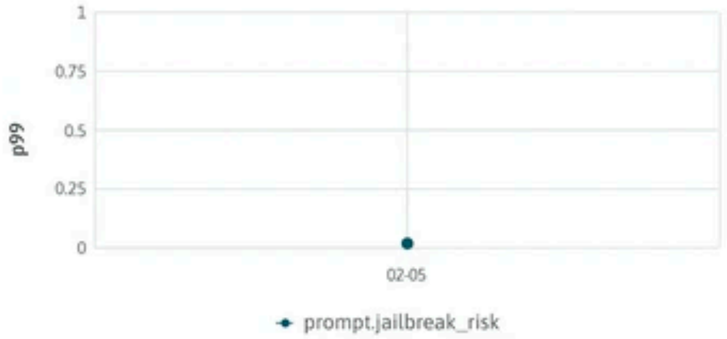
p99

0.0507

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

p99

Not monitored | Learn more



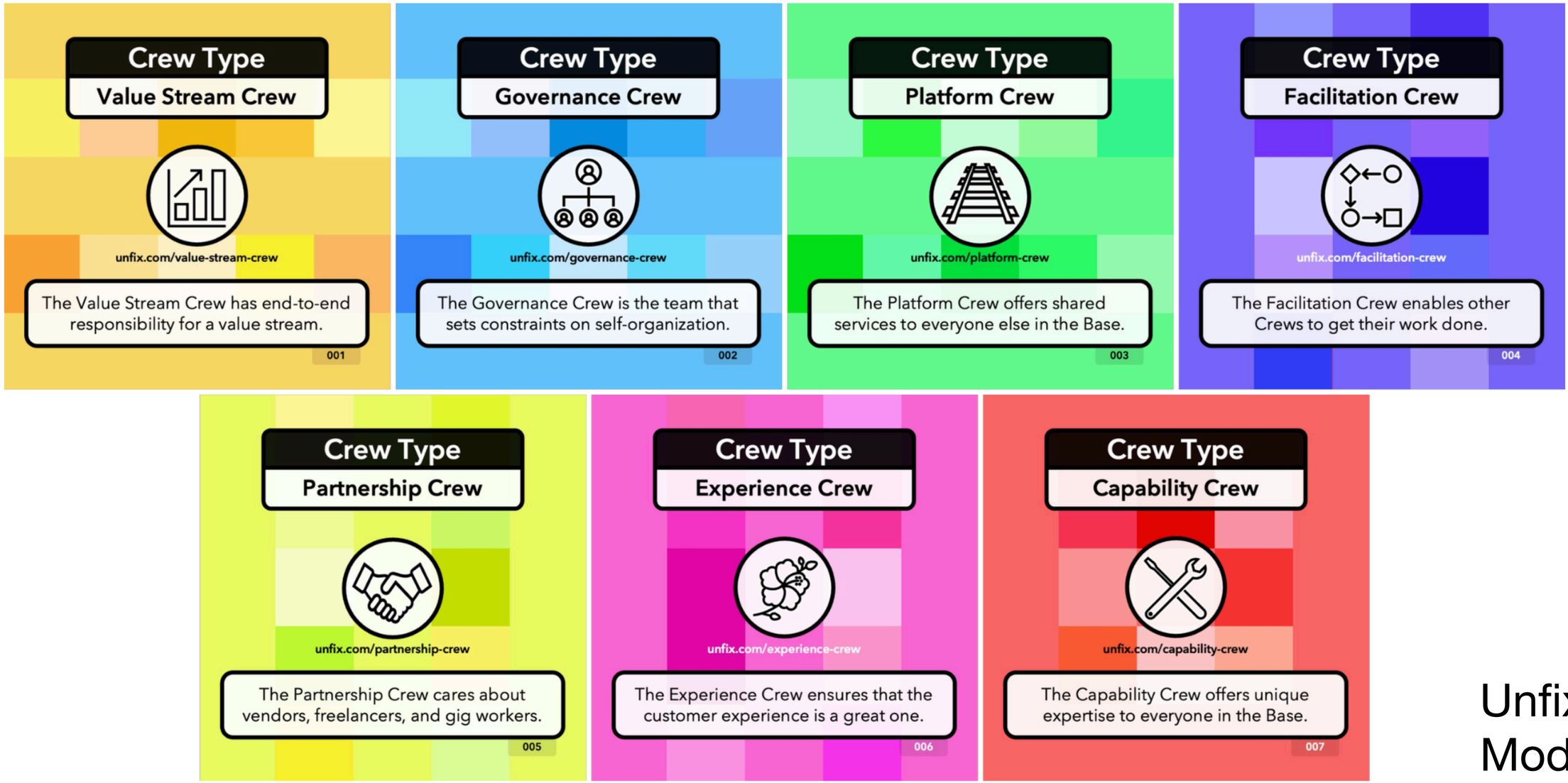
Steps to scale the change

 Platform

 Enablement

 Governance

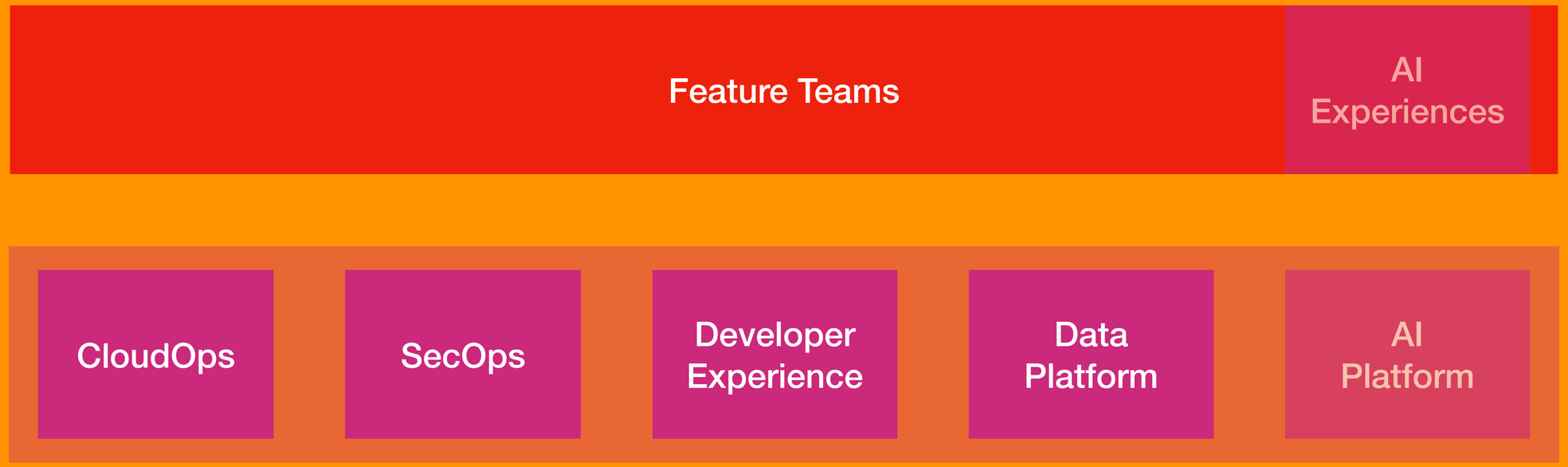
Beyond Platform teams



Unfix Model

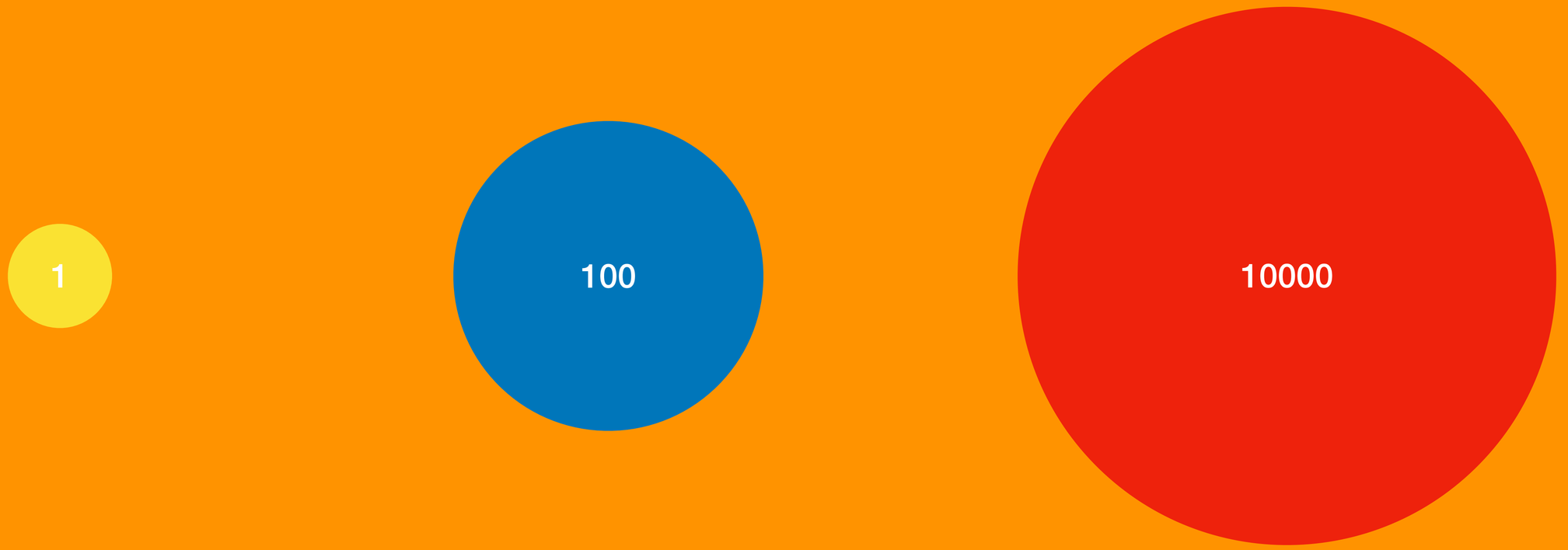
https://unfix.com/artificial-intelligence

Cross Platform Team collaboration



Run the Platform
Enable users on the Platform
Govern the Platform

Talent ratio



Foundation
Model builder

ML/AI Engineer

Software
Developer

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 impro the AI product.
- A product person can focus our efforts, talk to users, and help us prioritize what's essential to learn t jobs to be done.
- A designer can help us identify the ineffable experience users have integrating with generative AI, an 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/