Red Hat Summit

Connect

Why you want your AI to be Open Source

Business Track FSI – Financial Services and Insurances

Red Hat

Armin Warda

EMEA FSI Chief Technologist
Red Hat

Why is NOW a good time for companies to invest in AI?

Enterprises are taking the AI leap

AI technologies are becoming **more accessible and affordable** for businesses of all sizes

Companies can **realize the value** from AI-enabled applications and AI-support

Organizations are better prepared to manage, transform and **use their ever-increasing data**

AI has undergone significant evolution

The evolution of **Data-driven** AI: from Business Intelligence to Generative AI

Business Analysis & Intelligence

- Collecting data
- Storing & moving data
- Structured data
- Transforming data (ETL)

Data Warehouses

Advanced Analytics & Predictive AI

- Data science techniques
- Unstructured data
- Predictive analytics
- Real-time decision making

Big Data

Foundation Models & AI-enabled apps

- Deep learning techniques
- Model experimentation
- Model tuning
- Prompt engineering

Generative AI

Generative AI & Fact Databases

- Retrieval Augmented Generation (RAG)
- Generated content pointing to fact database
- Also filter hallucinations

RAG, Guardrails

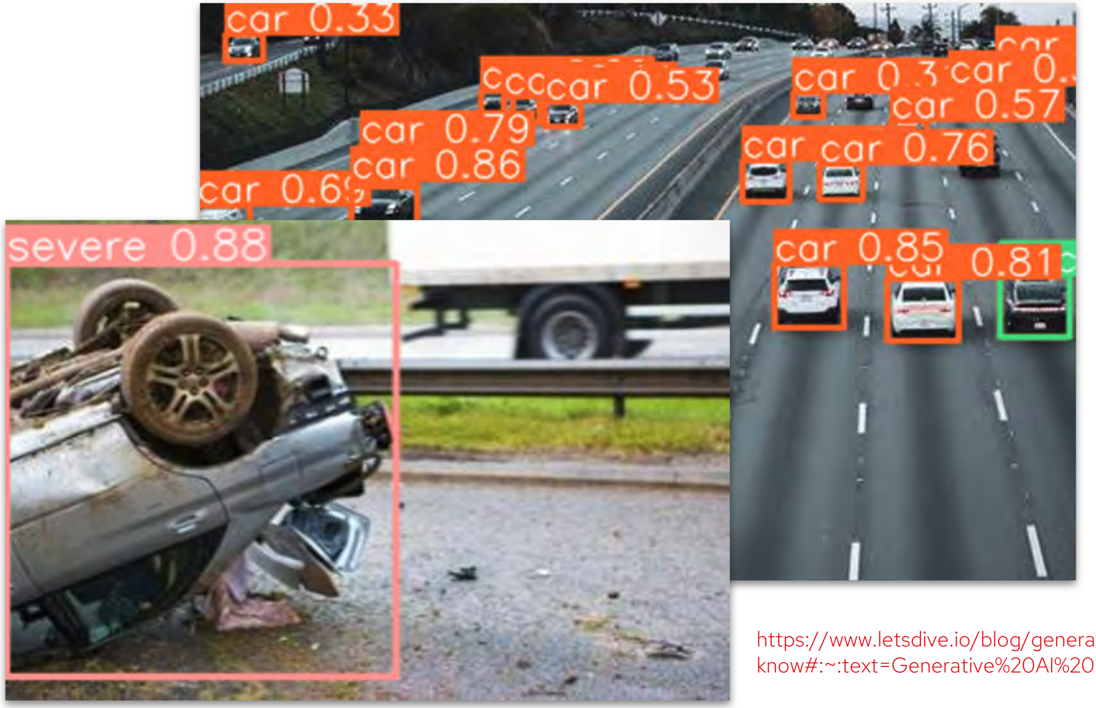
- Today, practical AI is **Data-driven**
- **Predictive AI** runs businesses today
- Foundation models provide a shortcut for realizing the value of **Generative AI**
- Models are embedded in and **guarded** by applications

Predictive AI vs. Generative AI

Most common types of AI for business applications

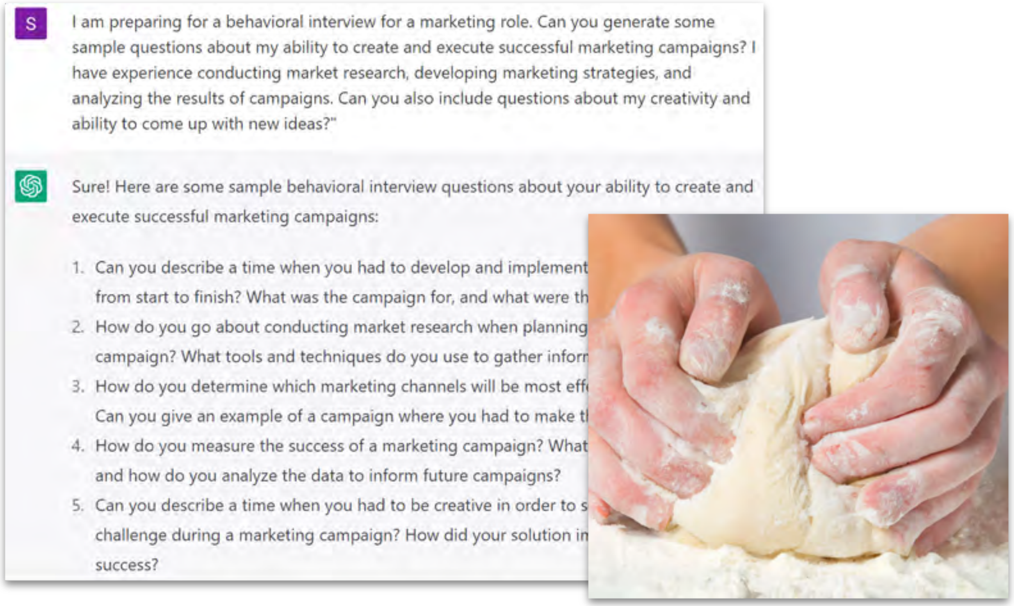
Predictive AI

Predicts or classifies outcomes with models trained on use-case specific data sources



Generative AI

Generates new content with models trained on vast amounts of data from many sources



I am preparing for a behavioral interview for a marketing role. Can you generate some sample questions about my ability to create and execute successful marketing campaigns? I have experience conducting market research, developing marketing strategies, and analyzing the results of campaigns. Can you also include questions about my creativity and ability to come up with new ideas?"

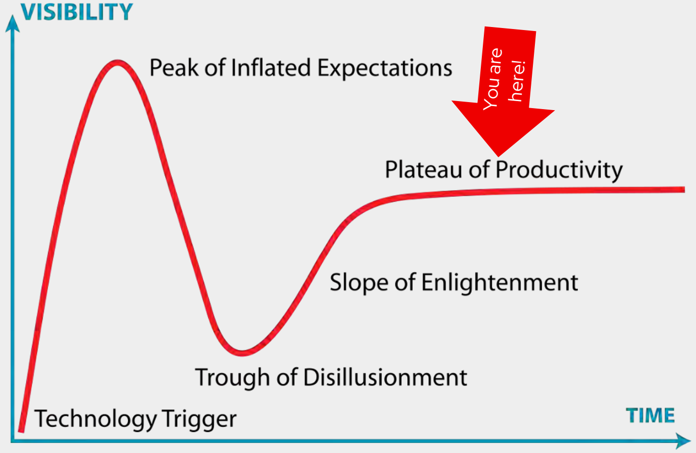
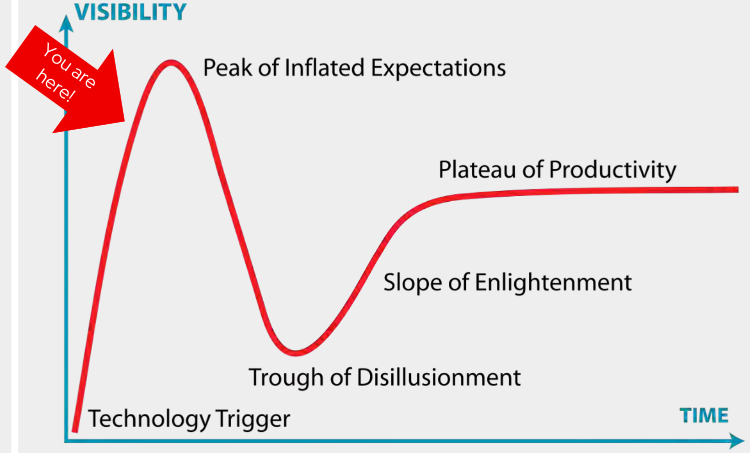
Sure! Here are some sample behavioral interview questions about your ability to create and execute successful marketing campaigns:

1. Can you describe a time when you had to develop and implement
   from start to finish? What was the campaign for, and what were th
2. How do you go about conducting market research when planning
   campaign? What tools and techniques do you use to gather infor
3. How do you determine which marketing channels will be most eff
   Can you give an example of a campaign where you had to make t
4. How do you measure the success of a marketing campaign? What
   and how do you analyze the data to inform future campaigns?
5. Can you describe a time when you had to be creative in order to s
   challenge during a marketing campaign? How did your solution in
   success?

https://www.letsdive.io/blog/generative-ai-vs-predictive-ai-all-you-need-to-know#:~:text=Generative%20AI%20is%20primarily%20focused,events%20based%20on%20historical%20data.

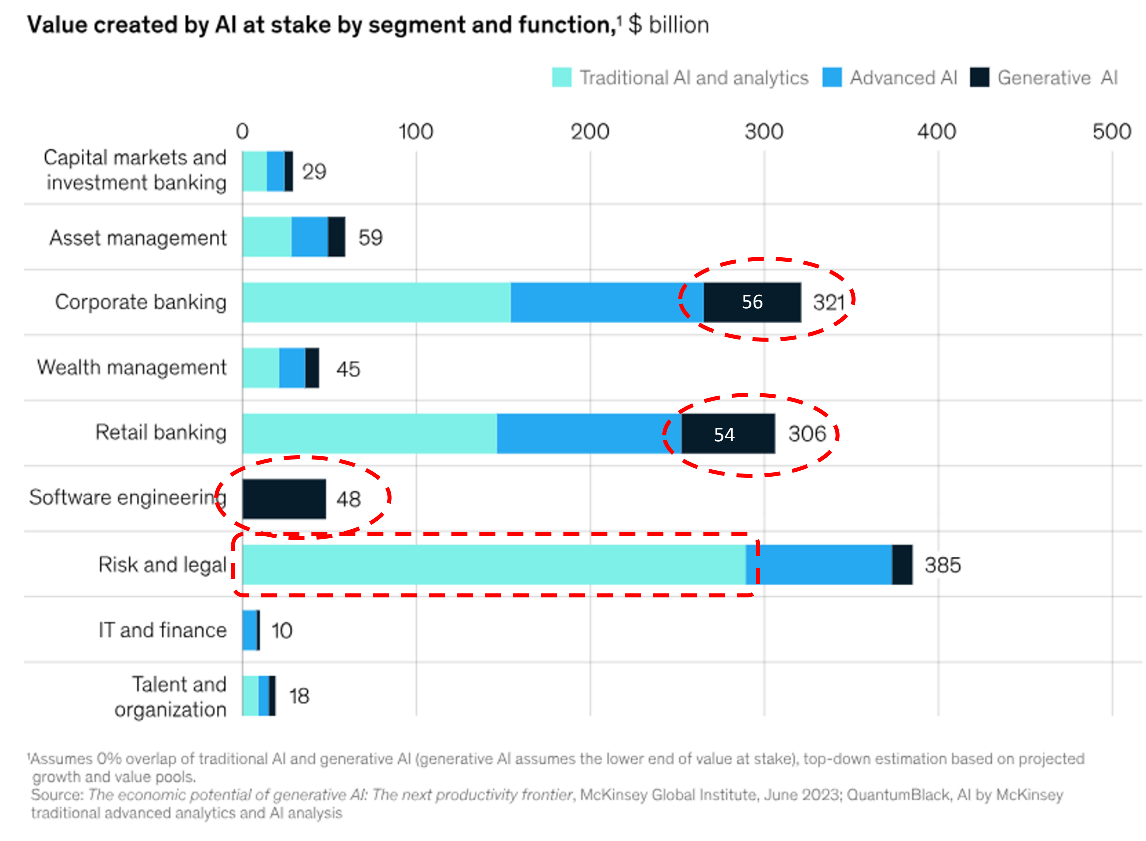
Predictive AI vs. Generative AI

Most common types of AI for business applications

	Predictive AI	**Generative AI**
What is it for?	Predicts or classifies outcomes with models trained on use-case specific data sources	Generates new content with models trained on vast amounts of data from many sources
Penetration	90%	10%
Maturity	 VISIBILITY Peak of Inflated Expectations You are here! Plateau of Productivity Slope of Enlightenment Trough of Disillusionment Technology Trigger TIME	 VISIBILITY You are here! Peak of Inflated Expectations Plateau of Productivity Slope of Enlightenment Trough of Disillusionment Technology Trigger TIME

Predictive AI and Generative AI in banking

McKinsey Insights: Capturing the full value of generative AI in banking



¹Assumes 0% overlap of traditional AI and generative AI (generative AI assumes the lower end of value at stake), top-down estimation based on projected growth and value pools.
Source: *The economic potential of generative AI: The next productivity frontier*, McKinsey Global Institute, June 2023; QuantumBlack, AI by McKinsey traditional advanced analytics and AI analysis

Among industry sectors, banking is expected to have one of the largest opportunities, largely from increased productivity

- The economic impact will likely benefit all banking segments and functions, with the greatest absolute gains through Generative AI in the corporate and retail sectors with $56 billion and $54 billion, respectively
- No surprise: software engineering ❤️ LLMs
- Risk and legal get largest value from AI, but that’s mostly traditional AI

https://www.mckinsey.com/industries/financial-services/our-insights/capturing-the-full-value-of-generative-ai-in-banking

Proven AI Use-Cases in Financial Services

Fraud Management

Anomaly detection, Countering financial crime such as money laundering, terror financing, tax evasion

SWIFT

Hyper-Personalization

Improve customer and employee experience, Customer Next Best Offer, Chatbots, Onboarding

İŞBANK

Galicia

Operational Efficiency

Branch Location & Staff Planning, ATM Cash on Hand, Call Routing, Workflow Automation

FriendliAI

RBC

Risk Analytics

Automated Underwriting Decisions, Mortgage Prepayment Analytics, Credit Scoring

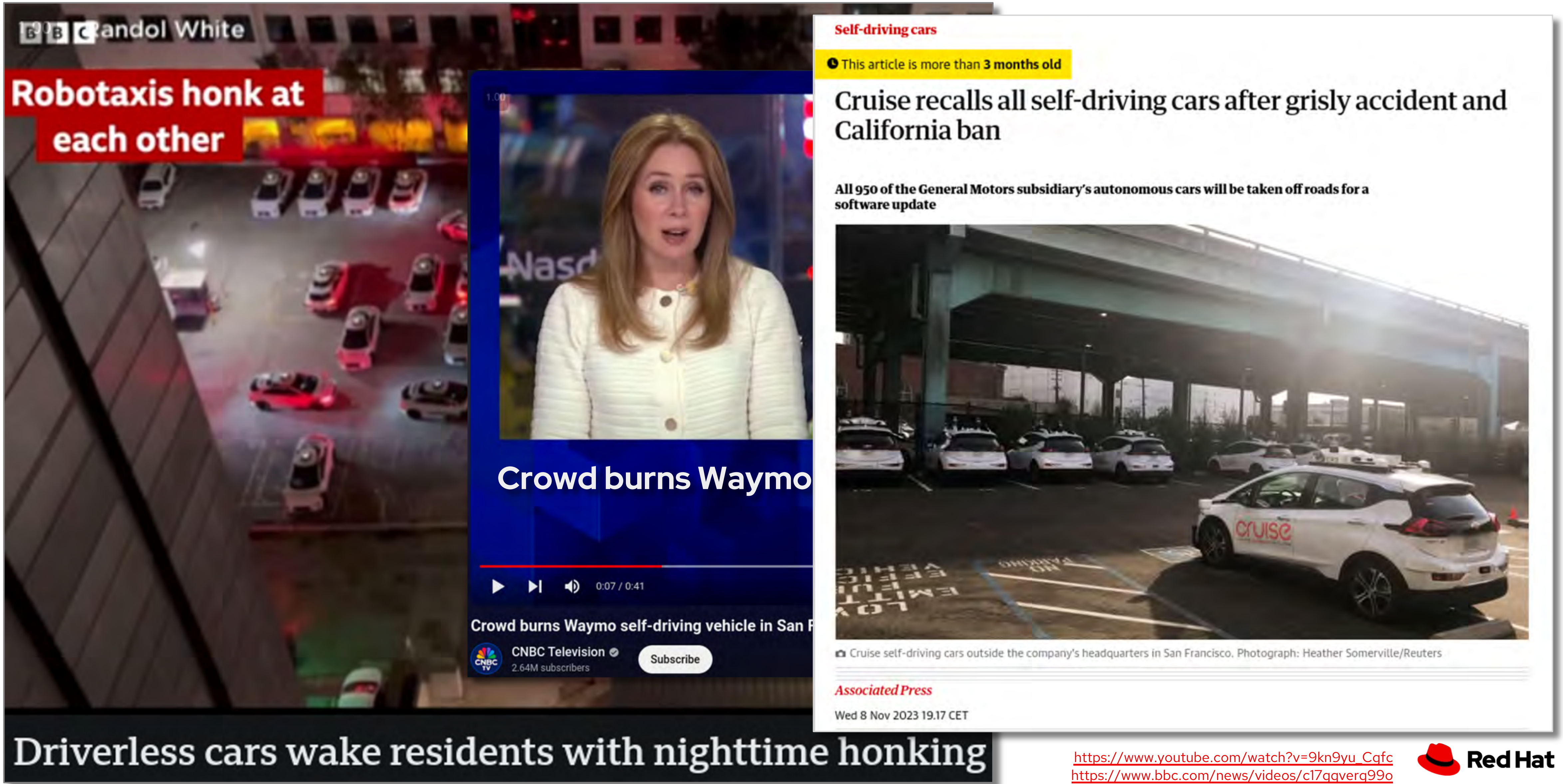
AXA FRANCE

JPMorganChase

These are great AI use-cases.

But there are also challenges.

Rage against the machine?



Driverless cars wake residents with nighttime honking

Crowd burns Waymo self-driving vehicle in San F

Self-driving cars

This article is more than **3 months old**

Cruise recalls all self-driving cars after grisly accident and California ban

All 950 of the General Motors subsidiary's autonomous cars will be taken off roads for a software update

Cruise self-driving cars outside the company's headquarters in San Francisco. Photograph: Heather Somerville/Reuters

Associated Press

Wed 8 Nov 2023 19.17 CET

https://www.youtube.com/watch?v=9kn9yu_Cqfc
https://www.bbc.com/news/videos/c17gqverq99o
https://www.theguardian.com/technology/2023/nov/08/cruise-recall-self-driving-cars-gm

Regulating AI: The EU-AI Act (March 13)

Unregulated, irresponsible or abusive use of AI could lead to negative consequences for individuals or the society, create public opposition and **hinder AI innovation in the EU**.

The EU is committed to strive for a balanced approach to AI

- Lawful
- Ethical
- Robust

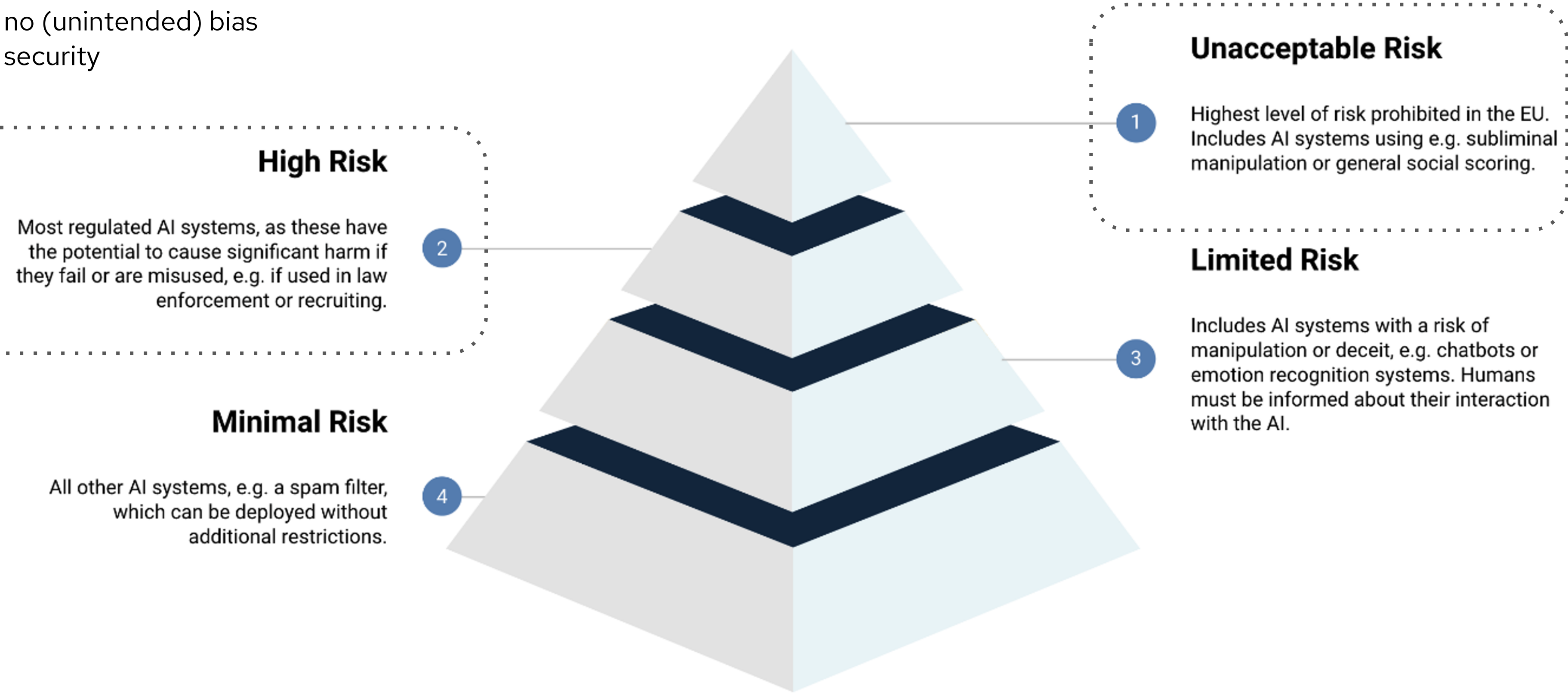
➔ accurateness
➔ transparency
➔ fairness
➔ no (unintended) bias
➔ security

EU AI Act Requirements:

Explainability, Documentation,
Process & Data Governance,
Human Oversight,
Risk Management, Auditability.

There are some exceptions for AI systems released under **Open Source** licenses.

banned:



Unacceptable Risk

1 Highest level of risk prohibited in the EU. Includes AI systems using e.g. subliminal manipulation or general social scoring.

High Risk

2 Most regulated AI systems, as these have the potential to cause significant harm if they fail or are misused, e.g. if used in law enforcement or recruiting.

Limited Risk

3 Includes AI systems with a risk of manipulation or deceit, e.g. chatbots or emotion recognition systems. Humans must be informed about their interaction with the AI.

Minimal Risk

4 All other AI systems, e.g. a spam filter, which can be deployed without additional restrictions.

Image: https://www.trail-ml.com/blog/eu-ai-act-how-risk-is-classified

What are **Open Source** AI models?

Open Source **Software** ✓

Open Source **Hardware** ✓

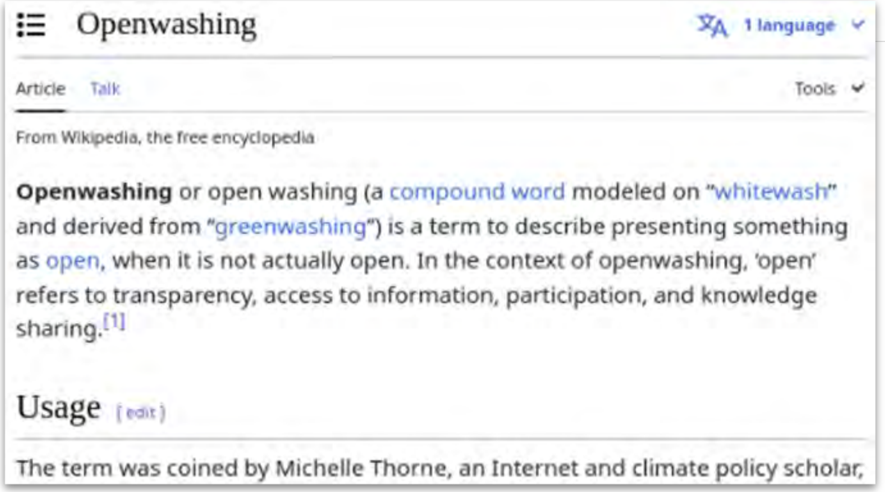
Open Source **AI/ML Models** ?

https://en.wikipedia.org/wiki/List_of_open-source_hardware_projects

How open are today's "Open Source" Models?

Open Source Software	Today's "Open Source" LLMs
Frequent releases (sometimes nightly)	**Irregular** releases (e.g. 1y between LLaMA versions)
Incremental contributions	**Monolithic** development
Feature roadmaps	"**Emergent** behaviour", no one knows what's coming
Community contributions (pull requests)	Largely **single-party** development (expensive collection of training data)
Contributions from many contributors can be merged and reconciled	Contributions to model, in the form of fine-tuning, are mutually incompatible between contributors, leading to **fragmentation** in model families (forks)
Almost any developer can, in principle, contribute	High **barrier** to contribution (clusters, GPUs for fine-tuning)

Openwashing?



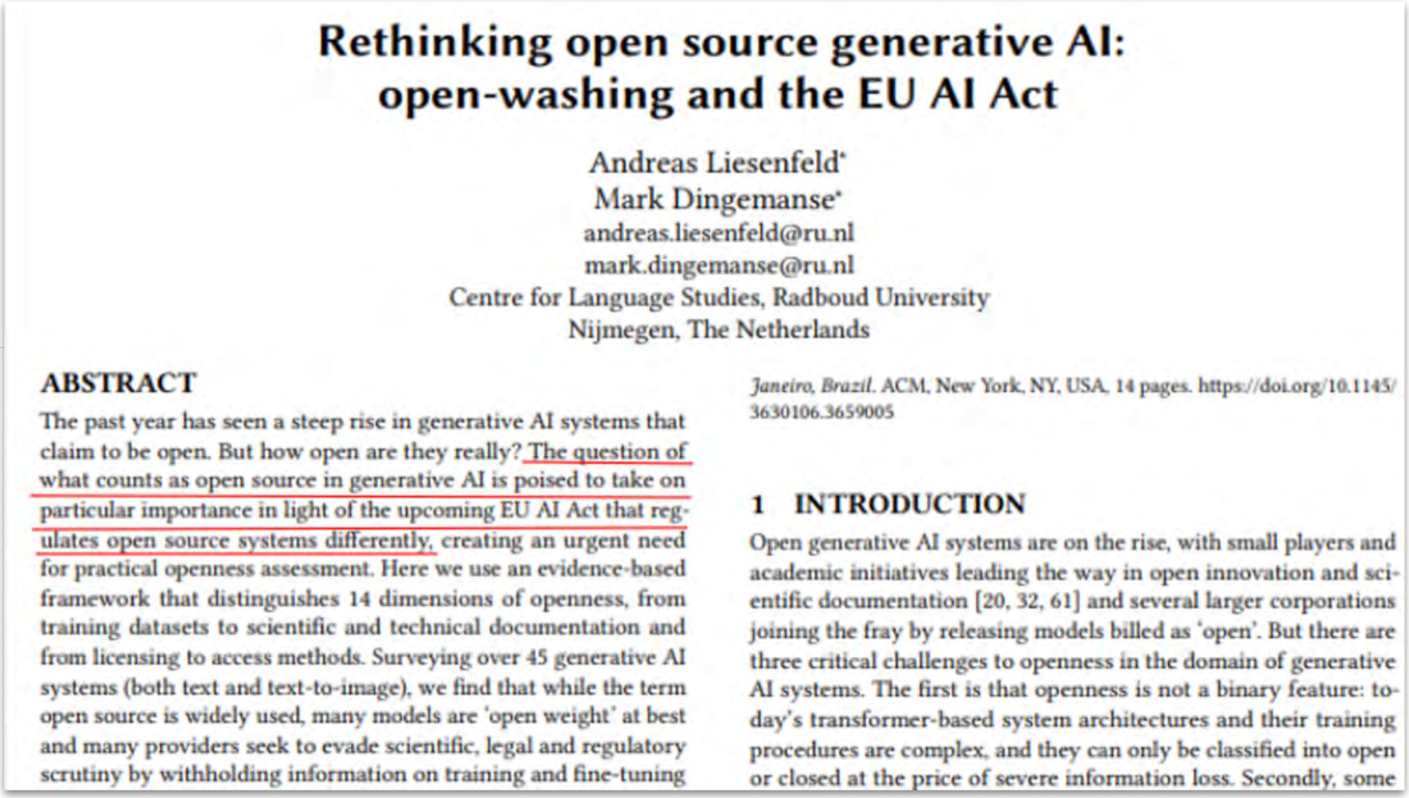
Openwashing

From Wikipedia, the free encyclopedia

Openwashing or open washing (a compound word modeled on "whitewash" and derived from "greenwashing") is a term to describe presenting something as open, when it is not actually open. In the context of openwashing, 'open' refers to transparency, access to information, participation, and knowledge sharing.[1]

Usage

The term was coined by Michelle Thorne, an Internet and climate policy scholar,



Rethinking open source generative AI: open-washing and the EU AI Act

Andreas Liesenfeld*
Mark Dingemanse*
andreas.liesenfeld@ru.nl
mark.dingemanse@ru.nl
Centre for Language Studies, Radboud University
Nijmegen, The Netherlands

ABSTRACT

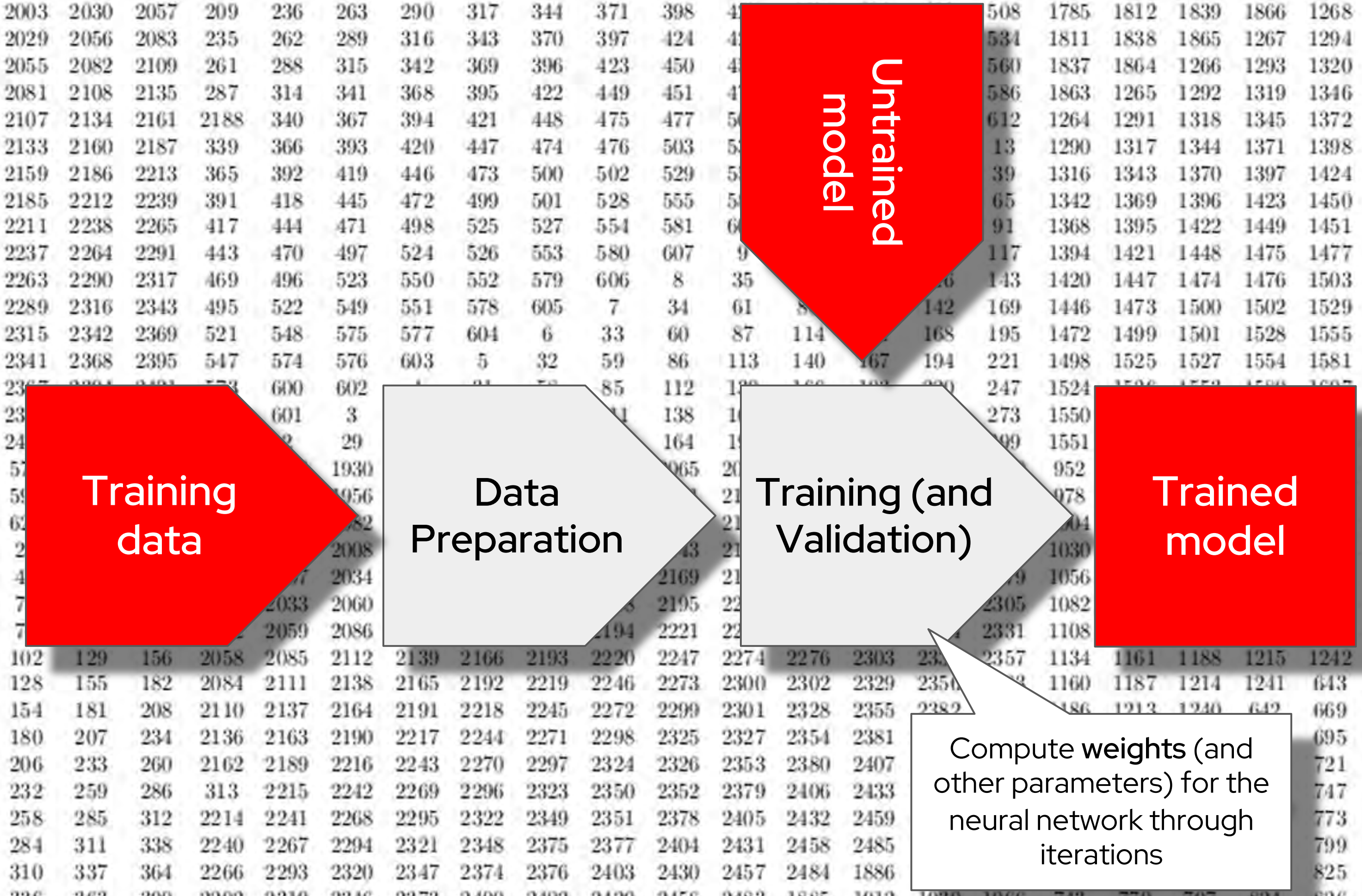
The past year has seen a steep rise in generative AI systems that claim to be open. But how open are they really? The question of what counts as open source in generative AI is poised to take on particular importance in light of the upcoming EU AI Act that regulates open source systems differently, creating an urgent need for practical openness assessment. Here we use an evidence-based framework that distinguishes 14 dimensions of openness, from training datasets to scientific and technical documentation and from licensing to access methods. Surveying over 45 generative AI systems (both text and text-to-image), we find that while the term open source is widely used, many models are 'open weight' at best and many providers seek to evade scientific, legal and regulatory scrutiny by withholding information on training and fine-tuning

Janeiro, Brazil. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3630106.3659005

1 INTRODUCTION

Open generative AI systems are on the rise, with small players and academic initiatives leading the way in open innovation and scientific documentation [20, 32, 61] and several larger corporations joining the fray by releasing models billed as 'open'. But there are three critical challenges to openness in the domain of generative AI systems. The first is that openness is not a binary feature: today's transformer-based system architectures and their training procedures are complex, and they can only be classified into open or closed at the price of severe information loss. Secondly, some

https://en.wikipedia.org/wiki/Openwashing
https://dl.acm.org/doi/pdf/10.1145/3630106.3659005



Untrained model

Training data

Data Preparation

Training (and Validation)

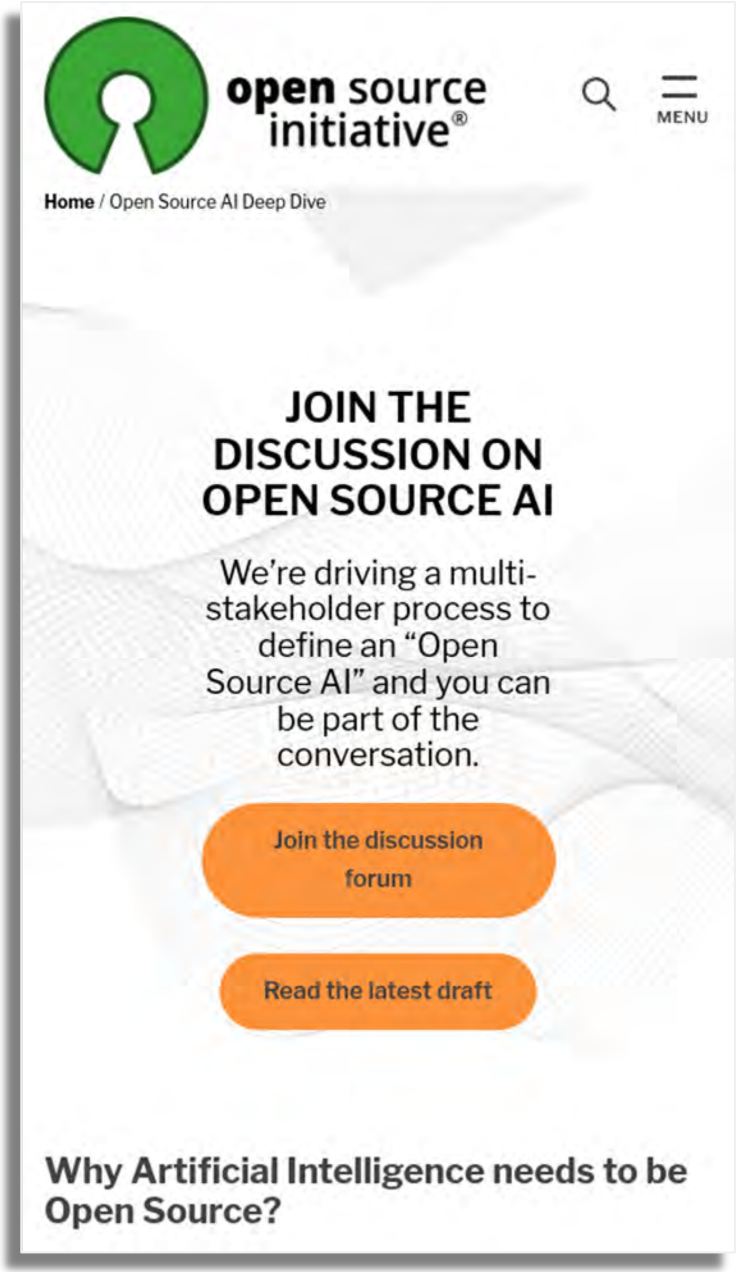
Trained model

Compute **weights** (and other parameters) for the neural network through iterations

What are weights?
Are weights special?

„Weights are **code.**"

Model development	Software development
Untrained model without weights	Instruction Set Architecture*
Weights for the model	Object code
training data & data preparation	Source code
Pre-trained model with weights	Deployable in an Application



Home / Open Source AI Deep Dive

JOIN THE DISCUSSION ON OPEN SOURCE AI

We're driving a multi-stakeholder process to define an "Open Source AI" and you can be part of the conversation.

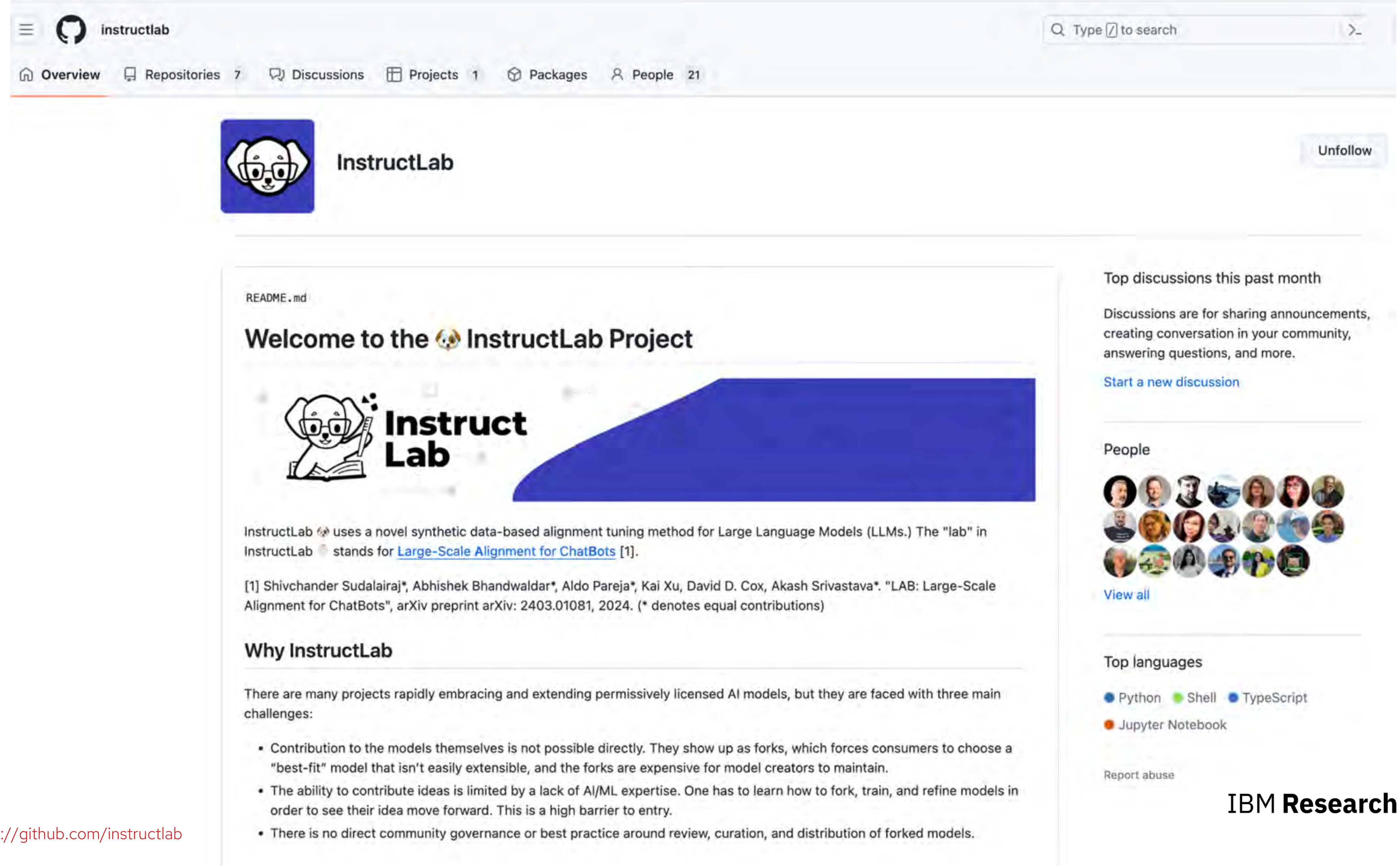
Why Artificial Intelligence needs to be Open Source?

---

* Instruction Set Architecture (ISA), such as RISC-V, P-Code Machine, Java Virtual Machine, WebAssembly

https://opensource.org/deepdive

An open source **community** project for GenAI model development



Welcome to the 🐶 InstructLab Project

InstructLab 🐶 uses a novel synthetic data-based alignment tuning method for Large Language Models (LLMs.) The "lab" in InstructLab 🐶 stands for Large-Scale Alignment for ChatBots [1].

[1] Shivchander Sudalairaj*, Abhishek Bhandwaldar*, Aldo Pareja*, Kai Xu, David D. Cox, Akash Srivastava*. "LAB: Large-Scale Alignment for ChatBots", arXiv preprint arXiv: 2403.01081, 2024. (* denotes equal contributions)

Why InstructLab

There are many projects rapidly embracing and extending permissively licensed AI models, but they are faced with three main challenges:

- Contribution to the models themselves is not possible directly. They show up as forks, which forces consumers to choose a "best-fit" model that isn't easily extensible, and the forks are expensive for model creators to maintain.
- The ability to contribute ideas is limited by a lack of AI/ML expertise. One has to learn how to fork, train, and refine models in order to see their idea move forward. This is a high barrier to entry.
- There is no direct community governance or best practice around review, curation, and distribution of forked models.

Top discussions this past month

Discussions are for sharing announcements, creating conversation in your community, answering questions, and more.

Start a new discussion

People

View all

Top languages

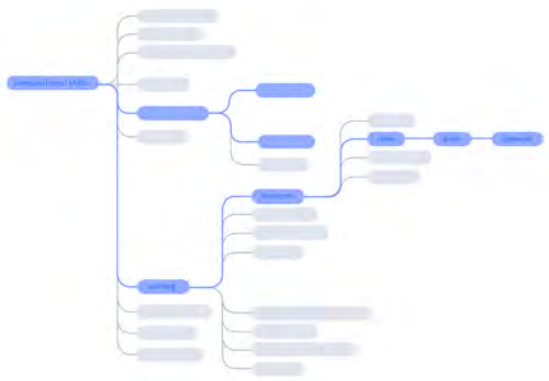
● Python ● Shell ● TypeScript
● Jupyter Notebook

Report abuse

https://github.com/instructlab

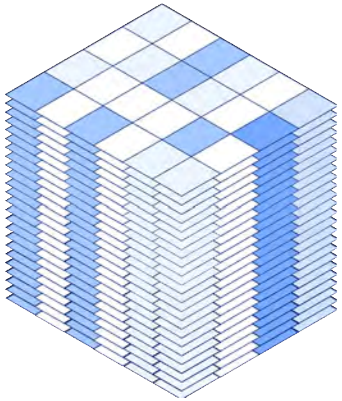
IBM **Research** **Red Hat**

LAB (Large-scale Alignment for ChatBots) Method



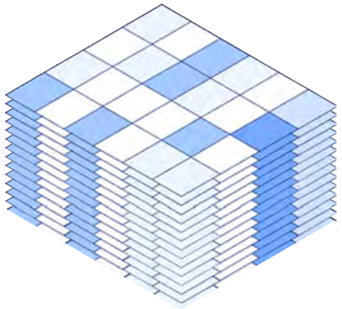
Taxonomy-based skill & knowledge representation

Represent any missing model knowledge or skills in a hierarchical **taxonomy**, providing 5+ exemplifying data points of the missing behavior per missing skill.



Synthetic data generation with teacher model

A **teacher model** generates a "curriculum" of millions of questions and answers across the taxonomy.



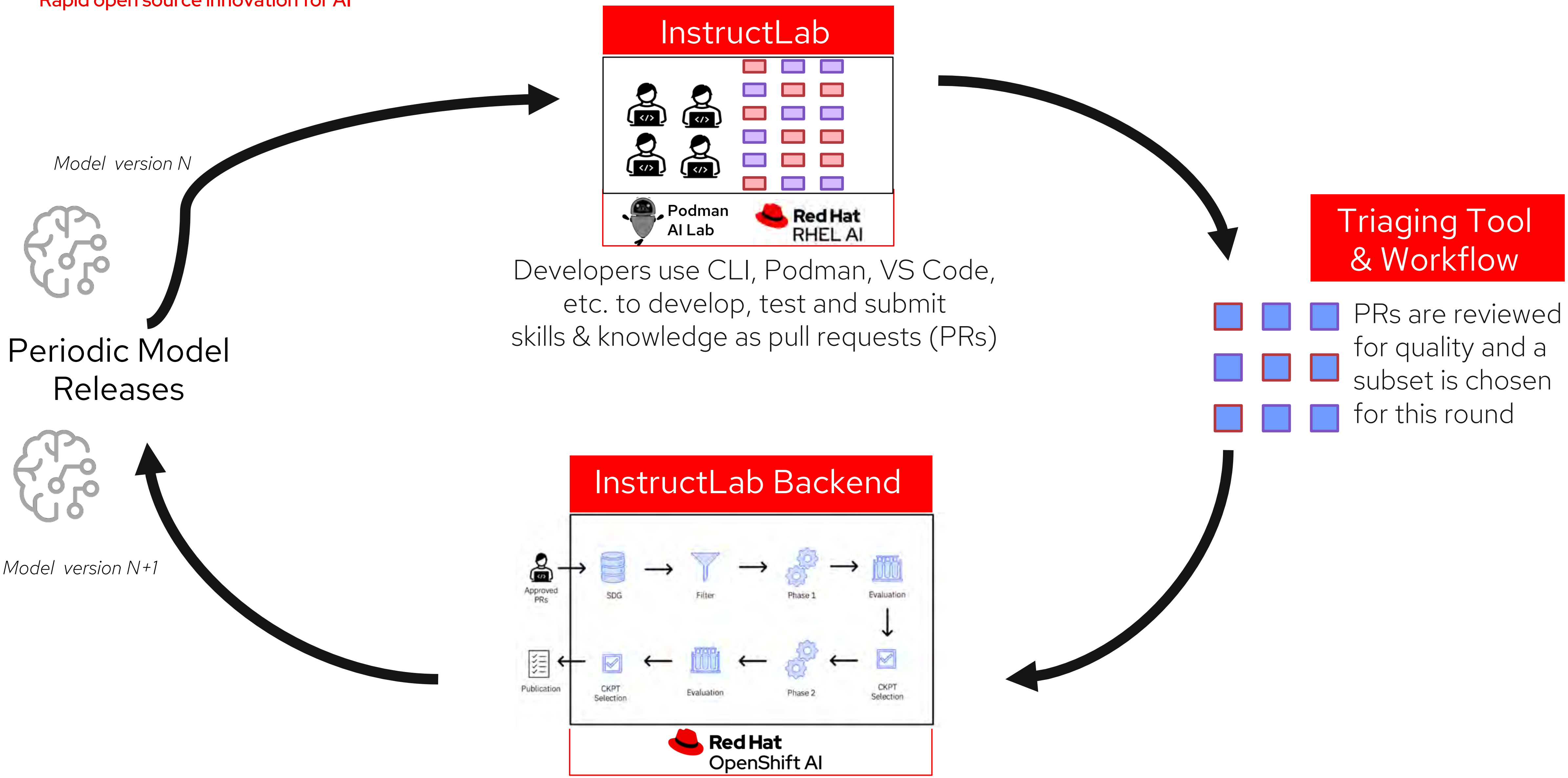
Synthetic data validation with critic model

A **critic model** filters the questions for correctness and quality. Synthetic data is scanned for prohibited material.

Skill and knowledge training on top of student model

The **student model** is trained with the curriculum using a novel training approach.

IBM **Research** publication: https://arxiv.org/html/2403.01081v1
IBM **Think** keynote: https://www.youtube.com/watch?v=SuGedexBudQ



InstructLab

Podman AI Lab

Red Hat RHEL AI

Developers use CLI, Podman, VS Code, etc. to develop, test and submit skills & knowledge as pull requests (PRs)

Triaging Tool & Workflow

PRs are reviewed for quality and a subset is chosen for this round

InstructLab Backend

Red Hat OpenShift AI

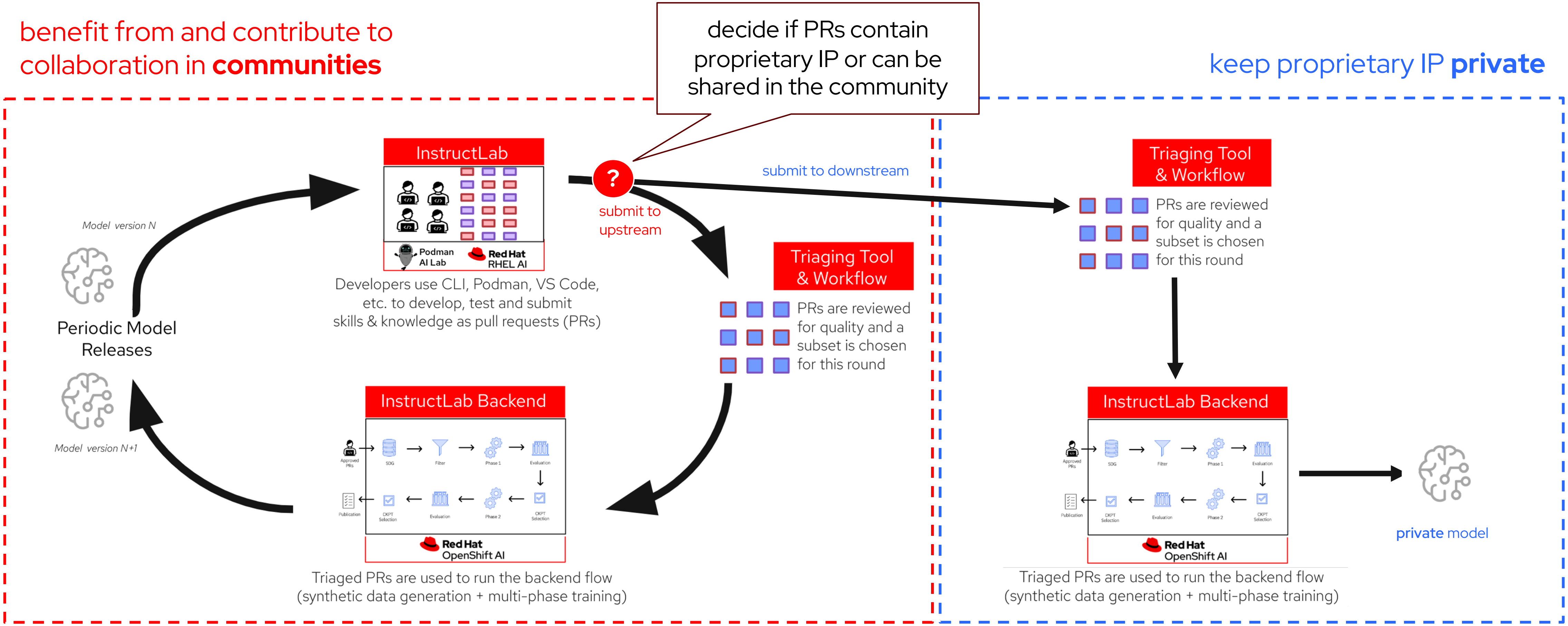
Triaged PRs are used to run the backend flow (synthetic data generation + multi-phase training)

Periodic Model Releases

Model version N

Model version N+1

Rapid open source innovation for AI



benefit from and contribute to collaboration in **communities**

decide if PRs contain proprietary IP or can be shared in the community

keep proprietary IP **private**

InstructLab

Developers use CLI, Podman, VS Code, etc. to develop, test and submit skills & knowledge as pull requests (PRs)

submit to upstream

submit to downstream

Triaging Tool & Workflow

PRs are reviewed for quality and a subset is chosen for this round

InstructLab Backend

Triaged PRs are used to run the backend flow (synthetic data generation + multi-phase training)

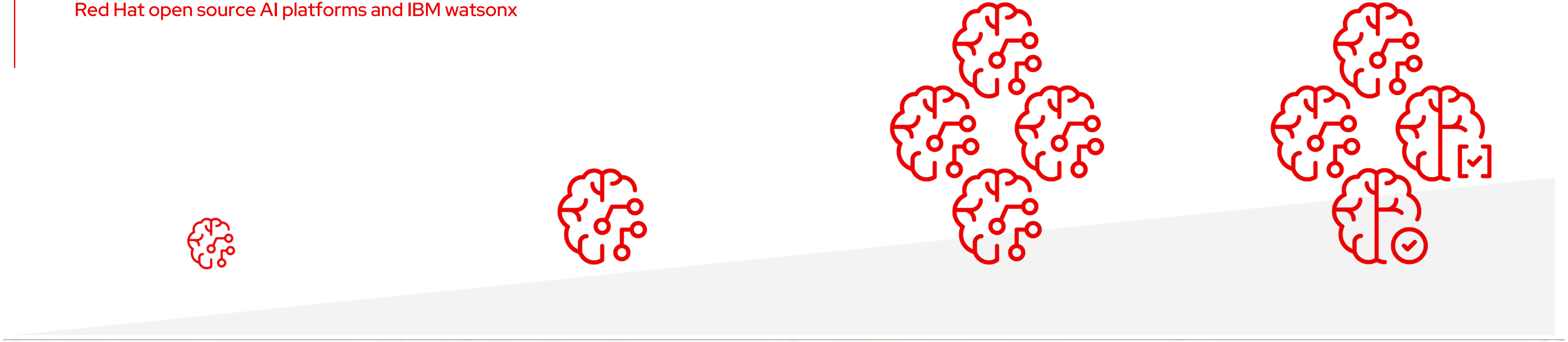
Periodic Model Releases

private model

Skills and knowledge that can be shared with the community are contributed upstream. These come back for free with the next version of the model, thus reducing the resources required for in-house fine-tuning of the private model, and potentially improved by other collaborators.

Proprietary skills and knowledge, that shall not be shared, are not submitted upstream but retained in-house. These have to be re-added to each new version of the upstream base model.

Red Hat open source AI platforms and IBM watsonx



InstructLab

STEP 1

Learn & experiment via limited desktop-scale training method (qlora) on small datasets. *Future potential Podman Desktop integration.*

Laptop / desktop

Red Hat Enterprise Linux AI

STEP 2

Production-grade model training using full synthetic data generation, teacher and critic models. Tooling focused on scriptable primitives.

Server / VM

Red Hat OpenShift AI

STEP 3

Production-grade model training as in RHEL AI, using full power of Kubernetes scaling, automation and MLOps services.

Cluster

watsonx

STEP 4

Comprehensive AI solution including AI optimized infrastructure, runtimes, middleware, data services, governance and applications.

Cluster

qlora: https://arxiv.org/abs/2305.14314 https://github.com/artidoro/qlora

Why you want your AI to be Open Source

Innovate with Open Source

- Proven **Predictive-AI** use-cases with Open Source can provide faster time-to-business value,
- Open Source & Open Research is where **Innovation** in **Generative-AI** happens,
- Open Source **avoids Lock-Ins** to hyperscalers or HW vendors.

But there are challenges

- Open Source provides better **Transparency** and **Auditability**,
- the **EU AI-Act** regulation is a bit lighter on Open Source,
- **Collaboration** on AI model development can solve common challenges faster, while allowing to keep unique IP private.

Red Hat can help

- We have the capabilities and **partnerships** to help speed-up your AI initiatives,
- allowing **faster** delivery of intelligent software applications,
- providing faster time-to-business value
- and to **control** the risks.

Red Hat Summit

Connect

Thank you

in linkedin.com/company/red-hat

f facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat

awarda@redhat.com

Red Hat