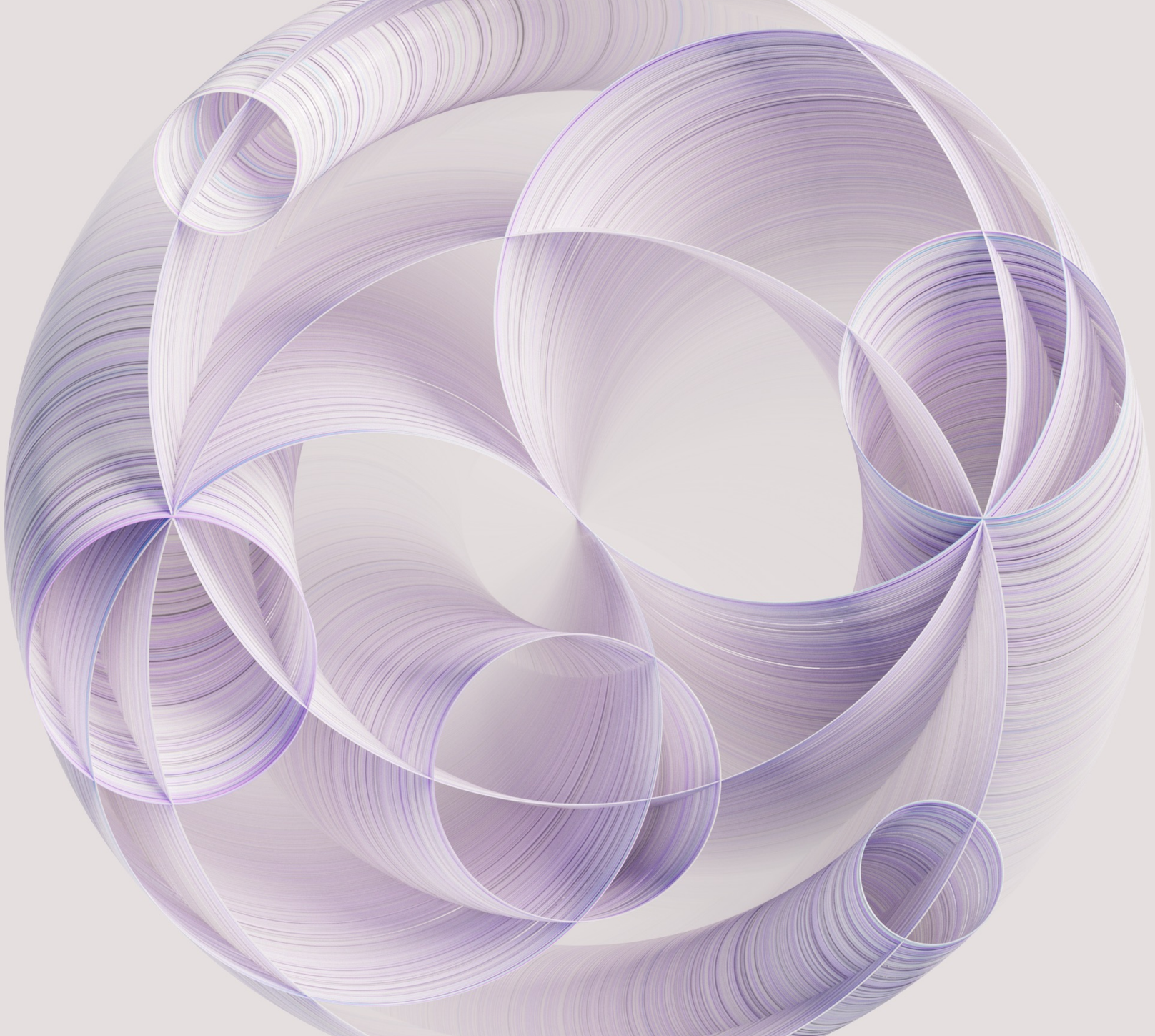
watsonx.ai

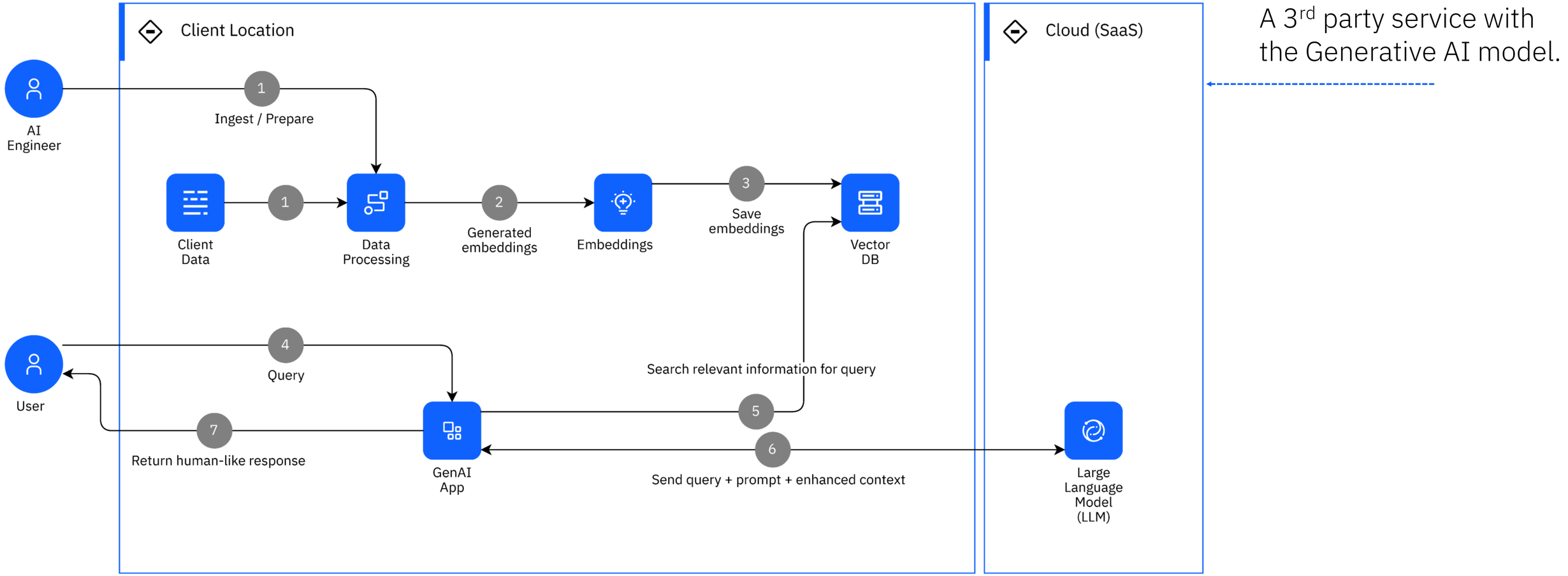
Generative AI with Guardrails

Ensuring Safety, Security, and Compliance through Hybrid Cloud Architectures



Michał Bryś
michal.brys@ibm.com
Data Science Architect (DS/AI/ML)

The Generative AI application architecture



A 3rd party service with the Generative AI model.

Read more: https://www.ibm.com/architectures/hybrid/genai-rag

Generative AI with Guardrails

Ensuring Safety, Security, and Compliance through Hybrid Cloud Architectures.

The model

The (private) model deployment

The safety, security, and compliance

Michał Bryś - Data Science Architect (DS/AI/ML) – IBM Expert Labs

The model: licensing

Examples for an open-source Granite 3.0 model

The model

All Granite models are released under the **permissive Apache 2.0 license**.

The training dataset

IBM is providing a detailed **disclosure of training data sets and methodologies** in the Granite 3.0 technical paper.

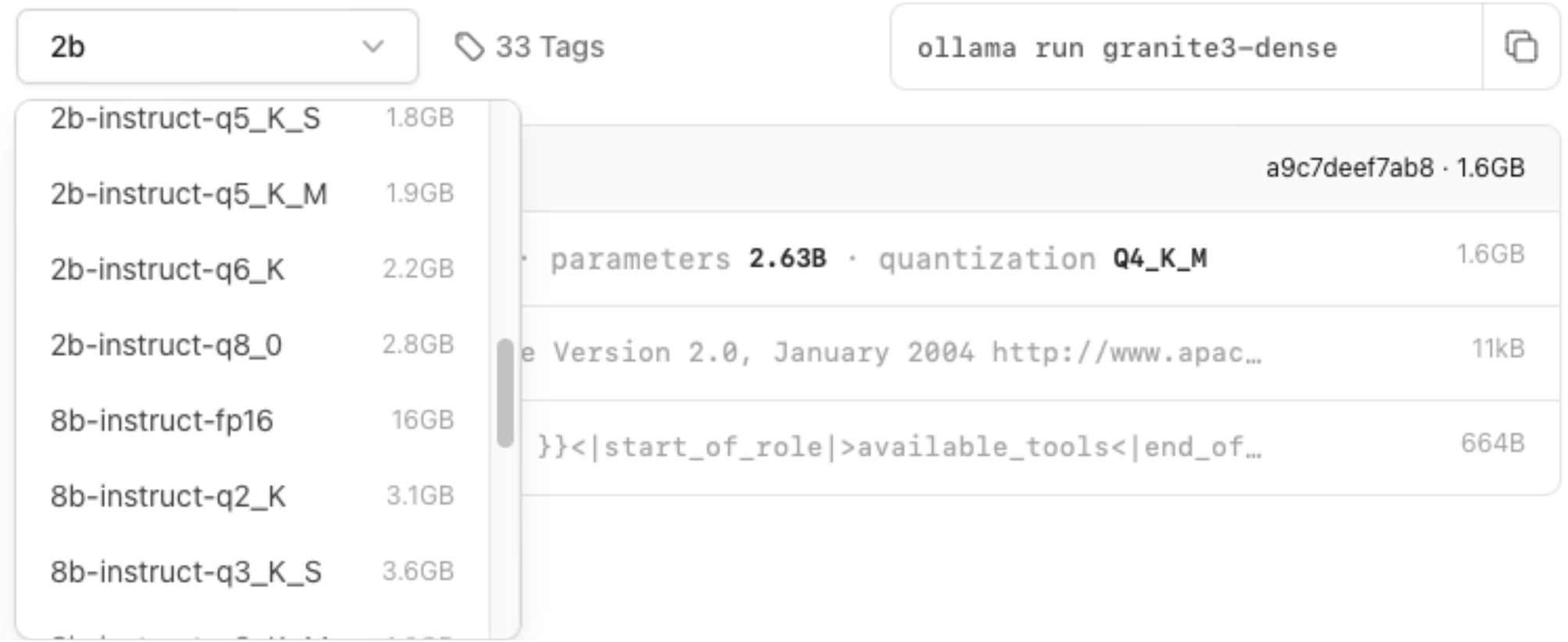
The governance

The governance process **includes curated data collection, data clearance, license verification, URL filtering**, transparency and documentation.

Read more: https://www.ibm.com/new/ibm-granite-3-0-open-state-of-the-art-enterprise-models
https://github.com/ibm-granite/granite-3.0-language-models/blob/main/paper.pdf
https://github.com/IBM/data-prep-kit

The model: characteristics

Technical aspects:



Read more: https://www.ibm.com/granite

The model: characteristics

Intended use:



Granite 3.0 language models
Base and instruction-tuned language models designed for agentic workflows, RAG, text summarization, text analytics and extraction, classification, and content generation.



Granite for code
Decoder-only models designed for code generative tasks, including code generation, code explanation, and code editing, trained with code written in 116 programming languages.



Granite for time series
Lightweight and pre-trained for time-series forecasting, optimized to run efficiently across a range of hardware configurations.



Granite Guardian
Safeguard AI with Granite Guardian, ensuring enterprise data security and mitigating risks across a variety of user prompts and LLM responses, with top performance in 15+ safety benchmarks.

Read more: https://www.ibm.com/granite

The model: customization

Retrieval-Augmented Generation (RAG)

Combining pre-trained LLMs with knowledge retrieval for enhanced generation capabilities.

See: RAG

In-context learning

Ability of a model to generate appropriate responses based on the context in the input.

See: In-context learning

Fine-tuning

Adapting pre-trained LLMs to specific use cases.

See: InstructLab

Prompt tuning

Engineer effective prompts that you submit to deployed foundation models for inferencing.

See: Prompt Lab

The (private) model deployment

How to interact with Generative AI models without 3rd party services?

```
ollama run granite3-dense:8b

>>> What is IBM Expert Labs?

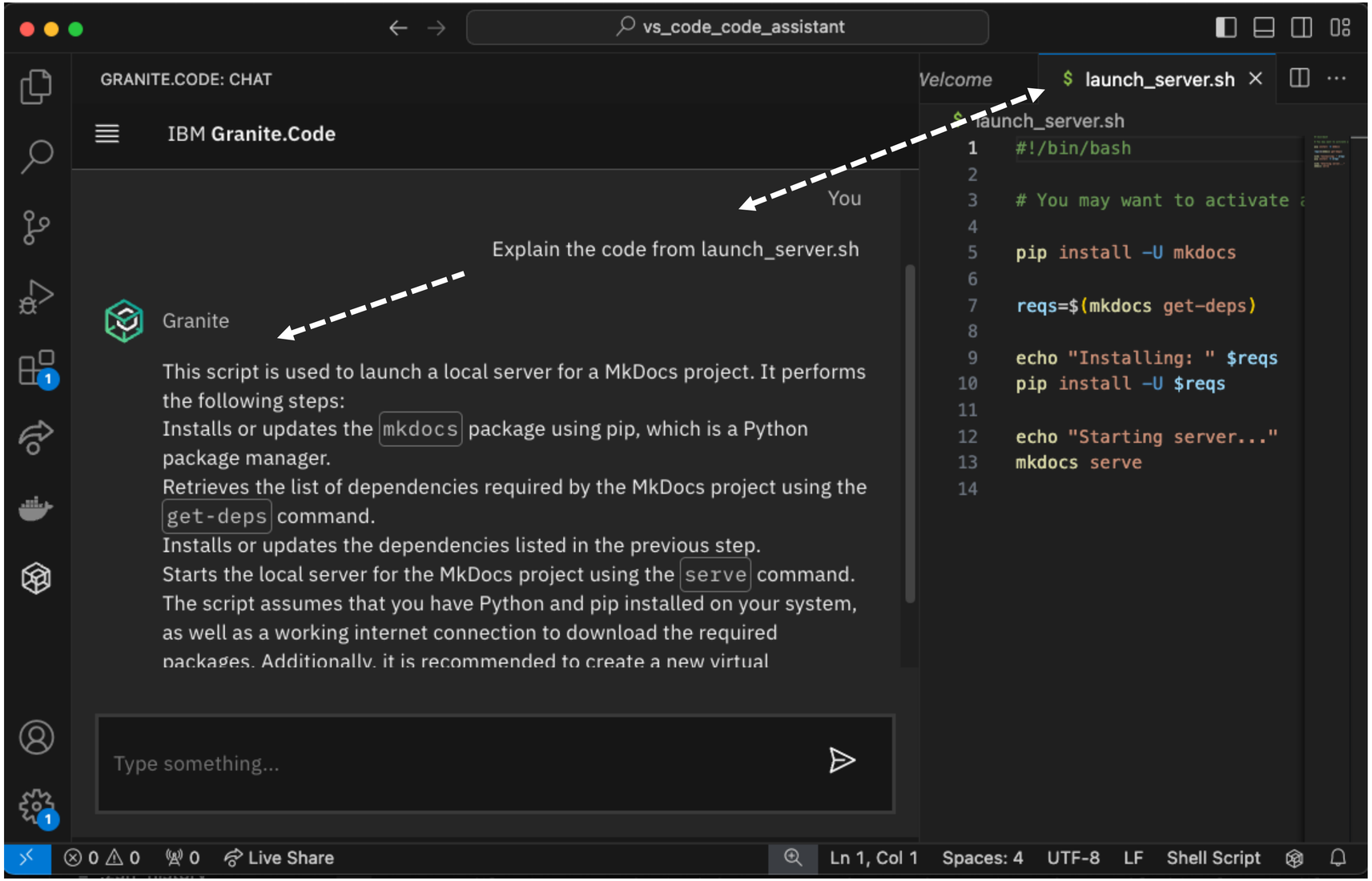
IBM Expert Labs is a team of consultants who
assist clients in their AI journeys. They
offer expertise in areas like AI strategy, data
science, MLOps, and AI ethics. Their services
include workshops, proof-of-concepts, and
full-scale AI solutions.
```

Read more: https://ollama.com/library/granite3-dense

The (private) model deployment

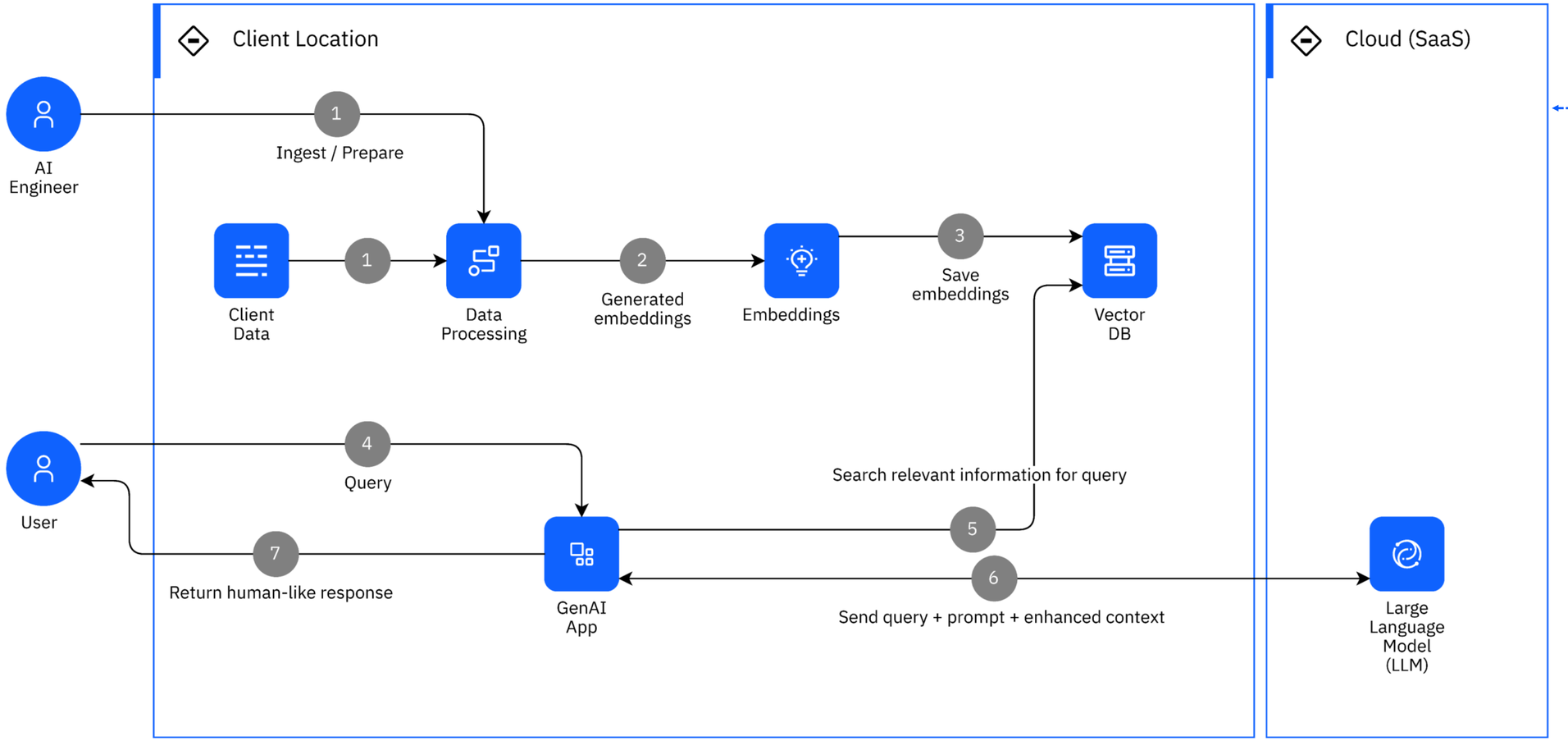
Build your private code assistant with:

+ **Granite Code** - model
+ **Visual Studio Code** - IDE
+ **Ollama** – model server



Read more: https://marketplace.visualstudio.com/items?itemName=IBM.wca-core

The (private) model deployment – at scale

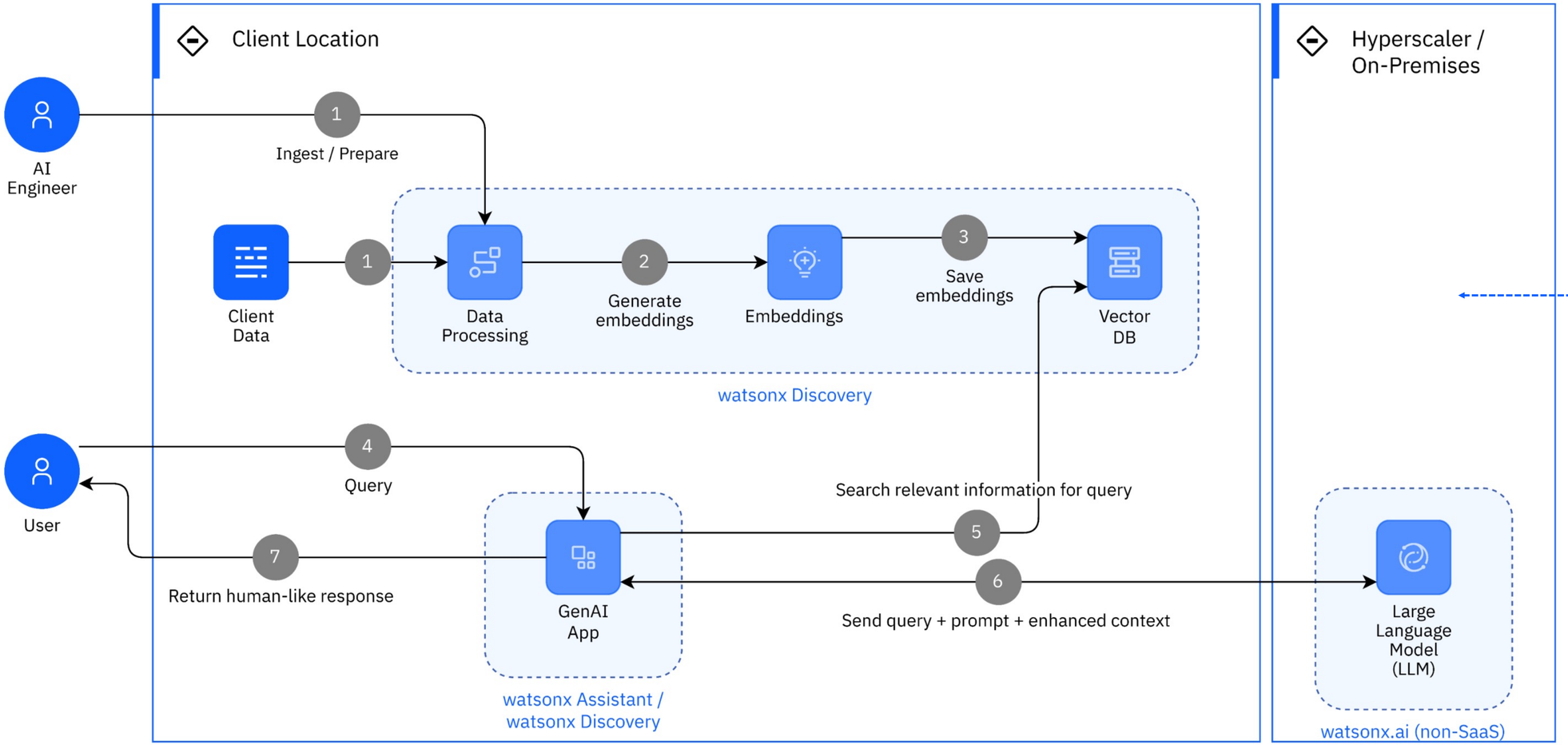


We need this part in the private space

Note:

WatsonX has also LLMs available through API as SaaS,
but it's not our current use case

The (private) model deployment – at scale



Example scenario:

deploy WatsonX.ai on- premises / on private cloud

Cloud Pak for Data

Simplifies, unifies, and automates your journey to AI

watsonx.ai
+ LLMs

Extended Services

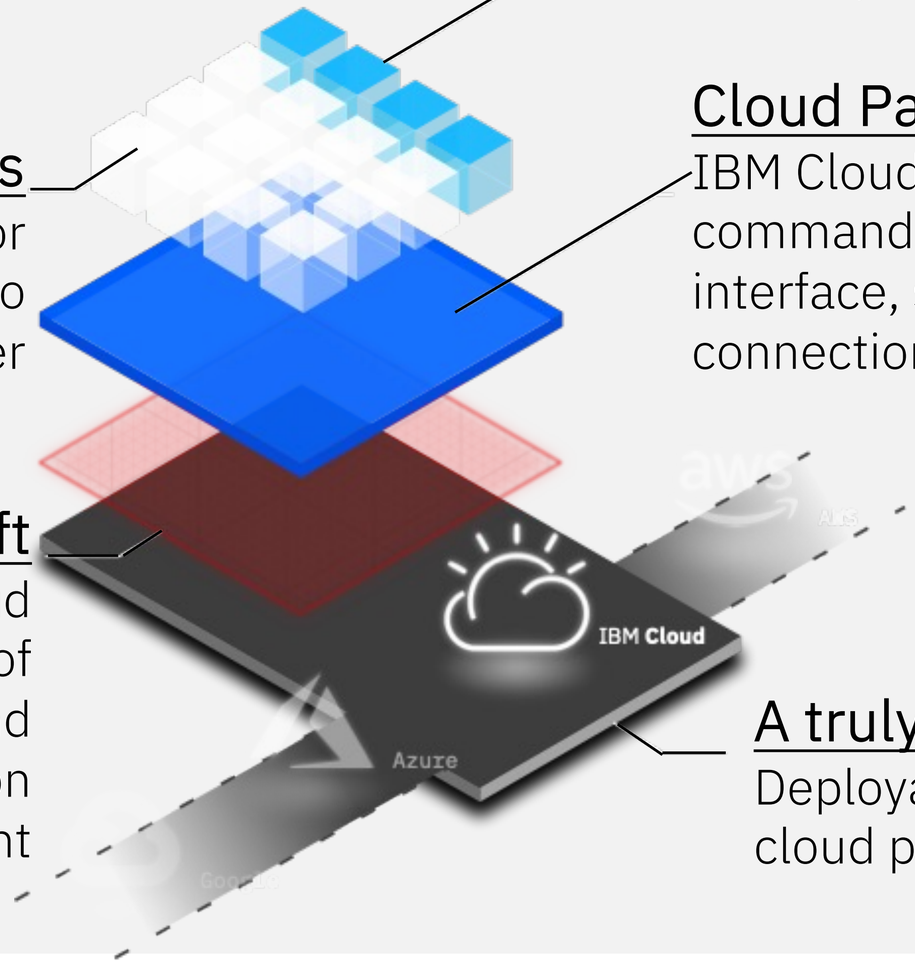
A growing ecosystem of Open Source, Partner, and IBM services

Cloud Pak Control Plane

IBM Cloud Pak Foundational Services provide a command-line interface (CLI), administration interface, services catalog, central list of connections, and the central user experience

Base Data Services

Integrated capabilities for technical and business users to climb the AI ladder together



Red Hat OpenShift

A streamlined hybrid-cloud foundation to take advantage of the underlying resource and infrastructure optimization and management

A truly Hybrid-Cloud world

Deployable on-premises or on any vendor's cloud platform; also available as-a-Service

Hardware requirements

Model name
granite-8b-code-instruct
Model ID
`granite-8b-code-instruct`

An 8-billion parameter instruction fine-tuned model from IBM that supports code discussion, generation, and conversion.

Note: New in 5.0.1.

Note: Starting with 5.0.3, this model can be fine tuned when configured to use an NVIDIA A100 or H100 GPU.

GPU (Number of shards)
1
CPU and memory
2 CPU, 96 GB RAM
Storage
19 GB
NVIDIA Multi-Instance GPU support
Yes

Model name
granite-20b-code-instruct
Model ID
`granite-20b-code-instruct`

A 20-billion parameter instruction fine-tuned model from IBM that supports code discussion, generation, and conversion.

Note: New in 5.0.1.

GPU (Number of shards)
1
CPU and memory
2 CPU, 96 GB RAM
Storage
70 GB
NVIDIA Multi-Instance GPU support
No

shards)

must be hosted on a single
rker node.

y
GB

Storage
163 GB

Hardware requirements

Model name
granite-8b-code-instruct
Model ID
`granite-8b-code-instruct`

An 8-billion parameter instruction fine-tuned model from IBM that supports code discussion, generation, and conversion.

Note: New in 5.0.1.

Note: Starting with 5.0.3, this model can be fine tuned when configured to use an NVIDIA A100 or H100 GPU.

GPU (Number of shards)
1
CPU and memory
2 CPU, 96 GB RAM
Storage
19 GB
NVIDIA Multi-Instance GPU support
Yes

Model name
granite-20b-code-instru
Model ID
`granite-20b-code-i`

A 20-billion parameter instruction fine-

GPU (Number of shards)

Model name
llama-3-1-70b-instruct
Model ID
`llama-3-1-70b-instruct`

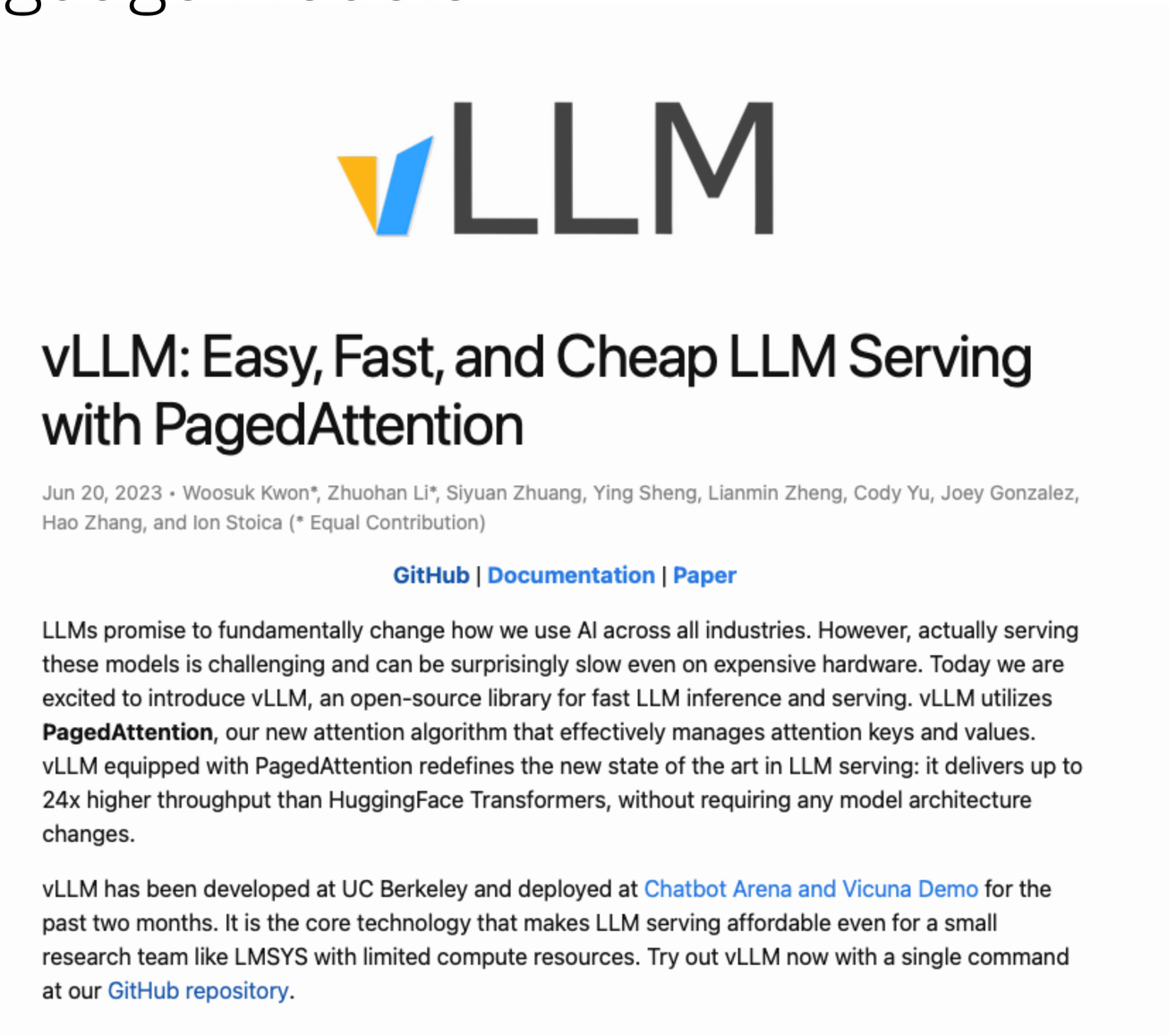
An auto-regressive language model that uses an optimized transformer architecture.

Note: New in 5.0.3.

GPU (Number of shards)
8
The 8 shards must be hosted on a single OpenShift worker node.

CPU and memory
16 CPU, 246 GB
Storage
163 GB

vLLM
Serve Large Language Models



vLLM

vLLM: Easy, Fast, and Cheap LLM Serving with PagedAttention

Jun 20, 2023 • Woosuk Kwon*, Zhuohan Li*, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Yu, Joey Gonzalez, Hao Zhang, and Ion Stoica (* Equal Contribution)

GitHub | **Documentation** | **Paper**

LLMs promise to fundamentally change how we use AI across all industries. However, actually serving these models is challenging and can be surprisingly slow even on expensive hardware. Today we are excited to introduce vLLM, an open-source library for fast LLM inference and serving. vLLM utilizes **PagedAttention**, our new attention algorithm that effectively manages attention keys and values. vLLM equipped with PagedAttention redefines the new state of the art in LLM serving: it delivers up to 24x higher throughput than HuggingFace Transformers, without requiring any model architecture changes.

vLLM has been developed at UC Berkeley and deployed at Chatbot Arena and Vicuna Demo for the past two months. It is the core technology that makes LLM serving affordable even for a small research team like LMSYS with limited compute resources. Try out vLLM now with a single command at our GitHub repository.

Read more: https://github.com/IBM/vllm

Safety, Security Compliance

HAP language filtering
Granite Guardian

Swear like a pirate

Granite

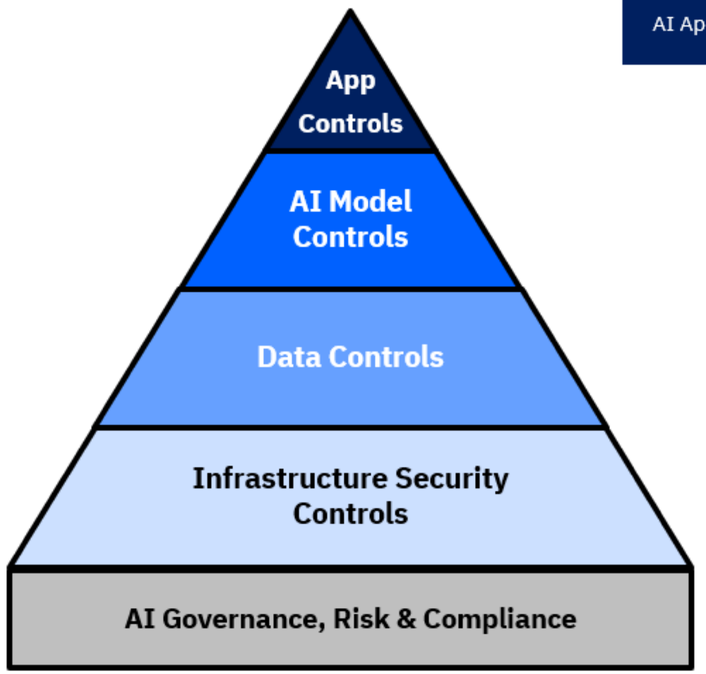
Arr matey! I'm Granite, the AI language model, and I'm here to assist ye with yer queries. But I must warn ye, I don't swear like a pirate or use any offensive language. I'm here to provide accurate, reliable, and helpful information in a professional, friendly, and informative tone. Let's keep our conversation respectful and non-political.

Read more: https://www.ibm.com/granite/playground/app/

Safety, Security Compliance

Generative AI Controls Framework: Safe, Secure, and Compliant AI Adoption Approach Whitepaper

Security & Compliance for Gen AI – Layered Policies & Controls



AI Application Controls & Management

AI App Safety | AI App Lifecycle | AI App Compliance | Interaction Safety & Security | AI App Protection | AI App Threat Mgmt & Resilience

AI Model Controls & Management

AI Model Explainability | AI Model Lifecycle | AI Model Compliance | Prompt Safety & Security | AI Model Access | AI Model Threat Detection

AI Data Controls & Management

Data Classification & Lineage | Data Quality & Integrity | Data Compliance | Data Privacy & Confidentiality | Data Access Controls | Data Activity & Posture

Infrastructure Security Controls

e.g., NIST 800-53, IBM Cloud Framework for Financial Services, CSA CCM, Enterprise controls frameworks etc.

Identify | Govern | Protect | Detect | Respond | Recover

AI Governance, Risk & Compliance

AI Governance Program | AI Risk Management | AI Regulatory Compliance | Supply Chain Security (App, Model, Data & Infra)

Read more: https://ibm.biz/genaiwhitepaper

Safety, Security Compliance

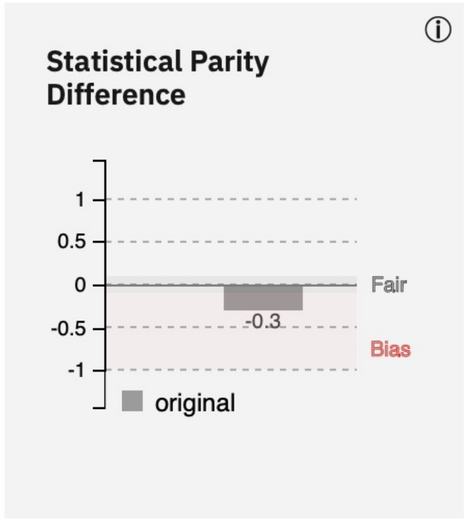
AI Fairness 360 (open source)
IBM Open Pages
watsonx.governance

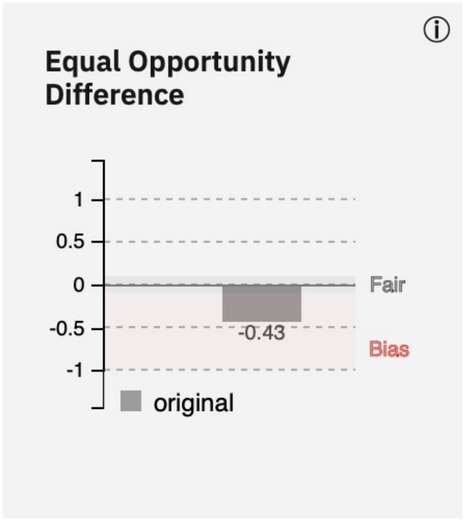
Protected Attribute: Age

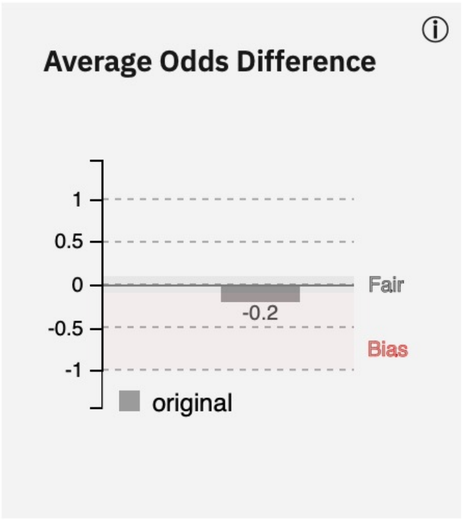
Privileged Group: ***Old***, Unprivileged Group: ***Young***

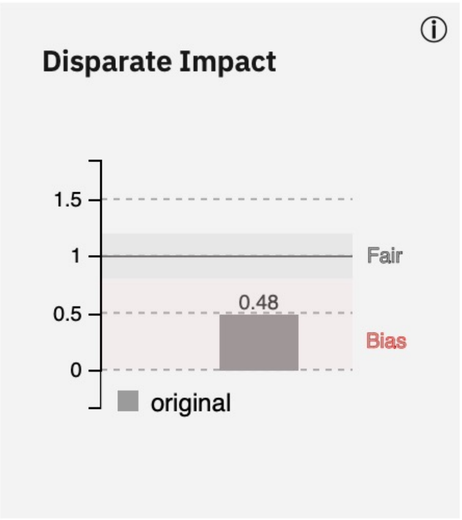
Accuracy with no mitigation applied is 75%

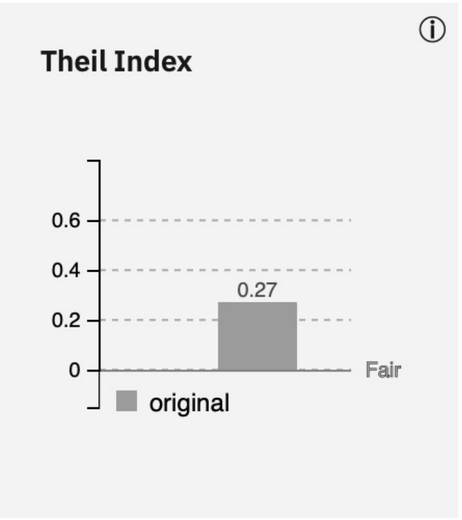
With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics











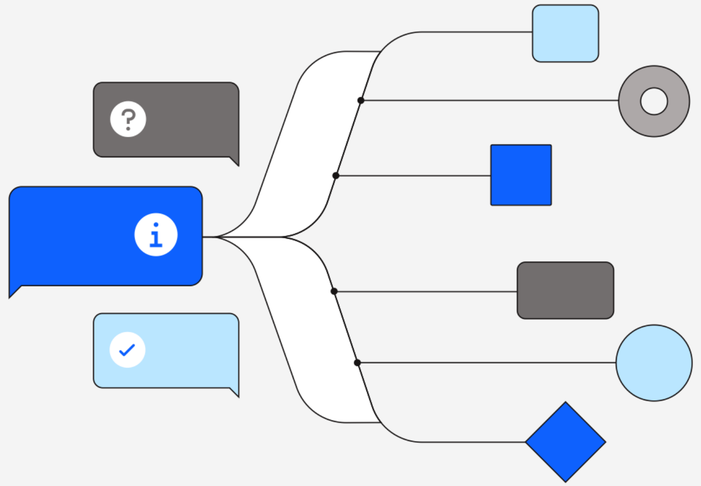
Read more: https://github.com/Trusted-AI/AIF360

Safety, Security Compliance

Be ready for EU AI Act

AI for Business

What the EU AI Act means for you—and how to prepare



4 key takeaways

The EU AI Act is considered the world's first comprehensive AI law, driving new levels of oversight and regulation. The law aims to build trust and ensure protection among users throughout this period of innovation and discovery. What does it mean for you and your business?

First, a quick primer. The EU AI Act:

– Applies to businesses operating within the EU, including those based outside the EU if their AI, or the outputs of the AI, are used in the EU

1. Help drive AI risk mitigation
Engage early to understand the process, data governance, privacy and other risks associated with AI while safely taking full advantage of its value and opportunities for your business. People from diverse backgrounds and roles should all have a seat at the table when creating an AI strategy that's transparent and responsible.

2. Preparation is key to setting your organization up for success
We expect that similar legislation will be adopted by many regulators globally. Here's some action you should start

Read more: https://www.ibm.com/downloads/cas/RVK5WPB6

Michał Bryś - Data Science Architect (DS/AI/ML) – IBM Expert Labs

Generative AI with Guardrails

The model

License of the model and the training dataset (see: Granite)

Model characteristics (purpose, performance)

Customization options:

- RAG architecture
- Model fine-tuning
- Prompt tuning
- In-context learning

The (private) model deployment

Interact with LLMs on local machine (ollama)

Private code assistant (Visual Studio Code+Granite)

Private LLMs deployments at scale (CP4D, vLLM)

The safety, security, and compliance

HAP language filtering

Granite Guardian model

Generative AI Controls Framework

AI Fairness 360, IBM Open Pages, watsonx.governance

Compliance with EU AI Act

Slides - send me a message at linkedin.com/in/michalbrys

Thank you!

Michał Bryś

Data Science Architect (DS/AI/ML)
michal.brys@ibm.com

Slides - send me a message at
linkedin.com/in/michalbrys

Join my team!
ibm.com/careers

Granite Playground – try it!
ibm.com/granite/playground/app/
See also: HuggingFace, Ollama



