

The background of the slide is a dark, atmospheric image of a futuristic city at night. The city skyline is visible in the distance, with lights reflecting on a body of water in the foreground. Overlaid on this scene are glowing, golden-yellow circuit board patterns that create a sense of depth and technological complexity.

# Storage, Data & AI in a **fast-changing** world

**Dr. Axel Koester**

Storage Wild Duck, IBM EMEA  
Board member IBM TEC Think Tank

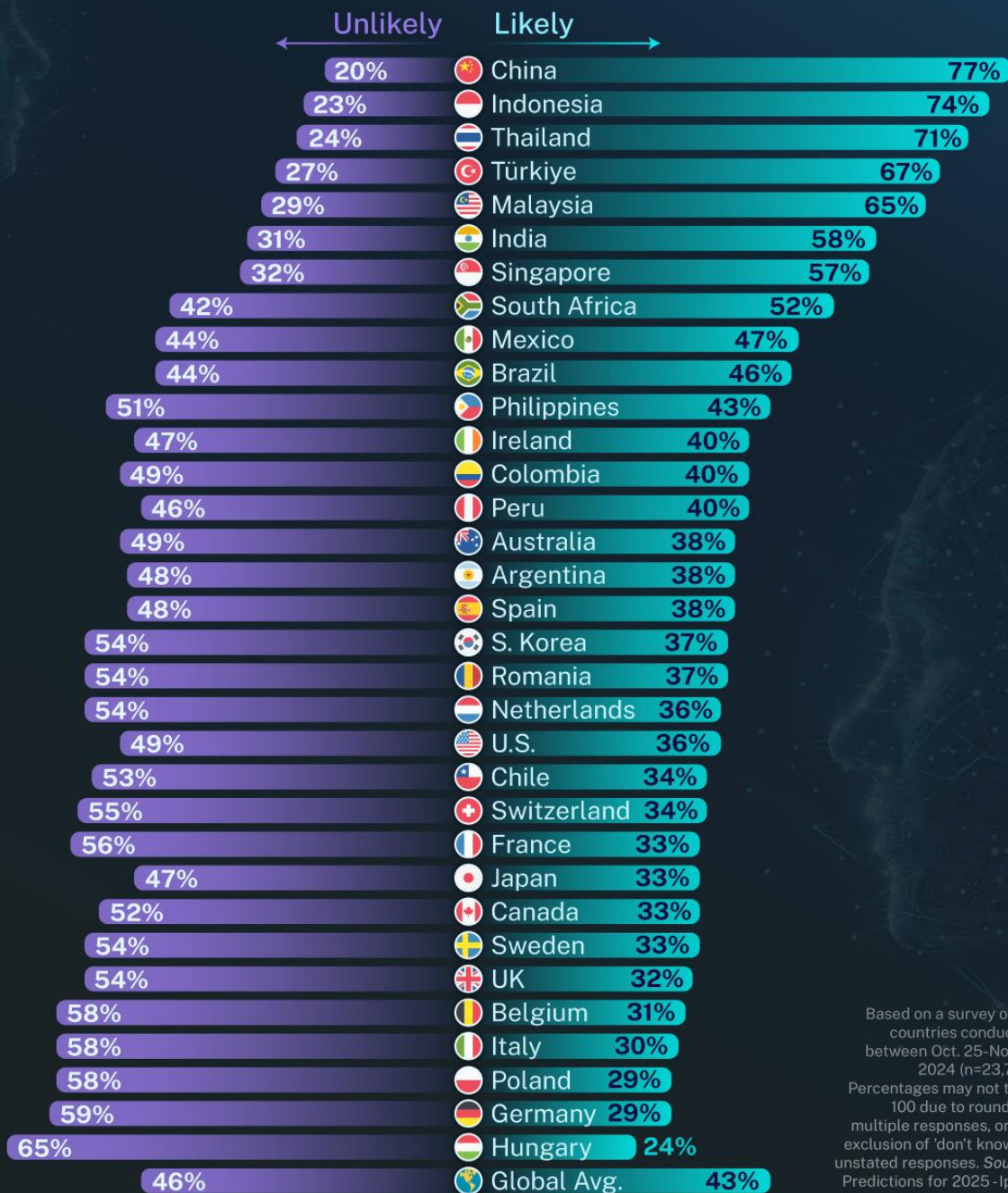


“60-70% of the  
jobs in 2040 are  
yet unknown”

Futurologist  
Dr. Daniel Dettling

Respondents were asked if they agree with the following statement:

*AI will lead to many new jobs being created in my country*



Based on a survey of 34 countries conducted between Oct. 25-Nov. 8, 2024 (n=23,721). Percentages may not total 100 due to rounding, multiple responses, or the exclusion of 'don't know' or unstated responses. Source: Predictions for 2025 - Ipsos

# Will AI create many new jobs? What do YOU think?

voronoi Visual Capitalist

<https://www.visualcapitalist.com/confidence-ai-create-destroy-jobs-by-country/>  
Nov. 2024 (n=23,721)

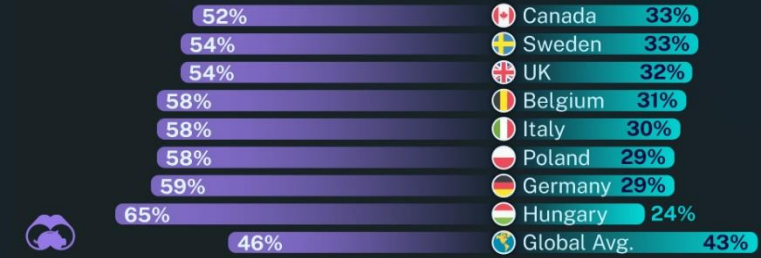




Based on a survey of 34 countries conducted between Oct. 25-Nov. 8, 2024 (n=23,721). Percentages may not total 100 due to rounding, multiple responses, or the exclusion of 'don't know' or unstated responses. *Source:* Predictions for 2025-Ipsos



# Of course.



Based on a survey of 34 countries conducted between Oct. 25-Nov. 8, 2024 (n=23,721). Percentages may not total 100 due to rounding, multiple responses, or the exclusion of 'don't know' or unstated responses. Source: Predictions for 2025 - Ipsos



## Prompting ChatGPT or DeepSeek doesn't look like a job generator, more the opposite.

Consuming AI solely in the cloud ...

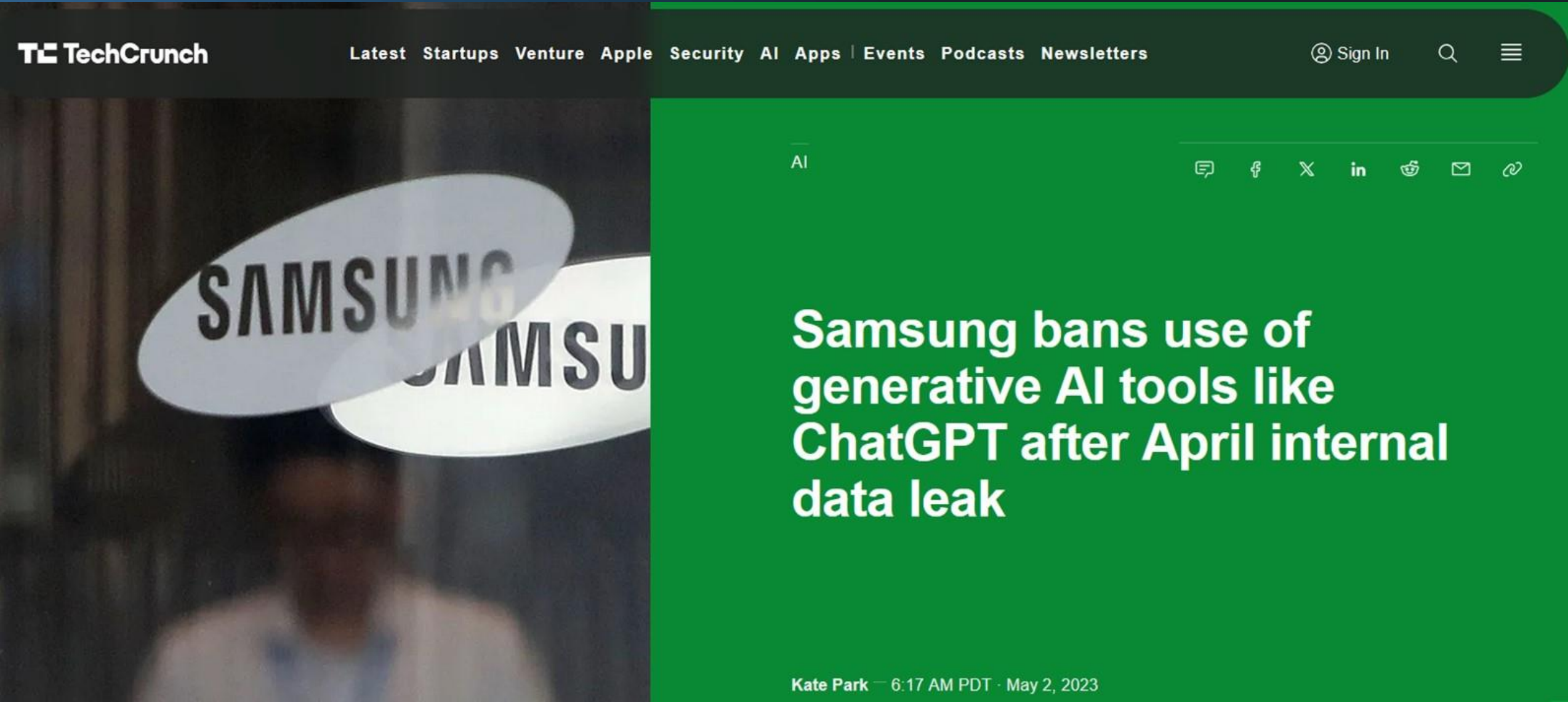
is like

***"I let my grandson do my internet"***

(the opposite of DIY)

# More reasons for DIY

**Your prompts are used by cloud platforms to refine the next model version**



AI



## Samsung bans use of generative AI tools like ChatGPT after April internal data leak

Kate Park — 6:17 AM PDT · May 2, 2023



# Insights from 'DIY' AI projects

Super-exponential technology growth

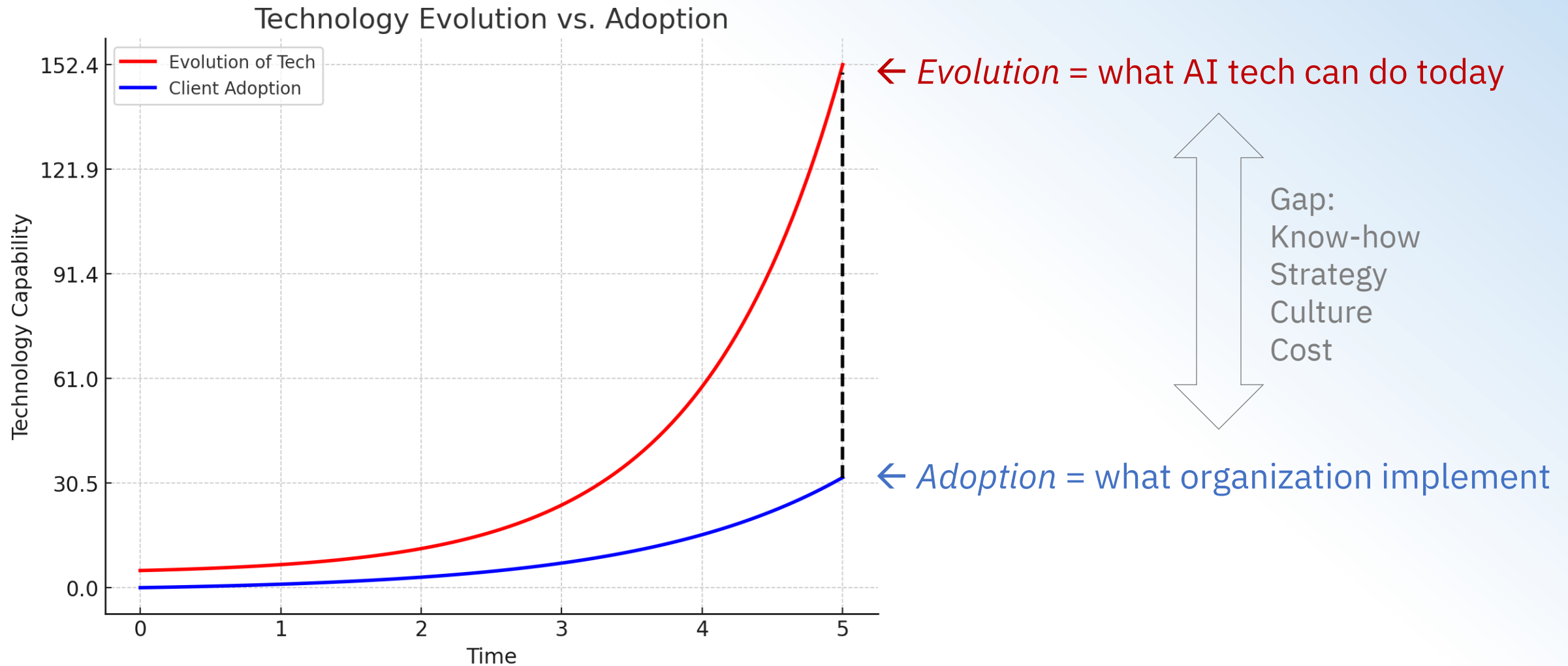
On-premise AI without Petabytes

AI platforms make life simpler!

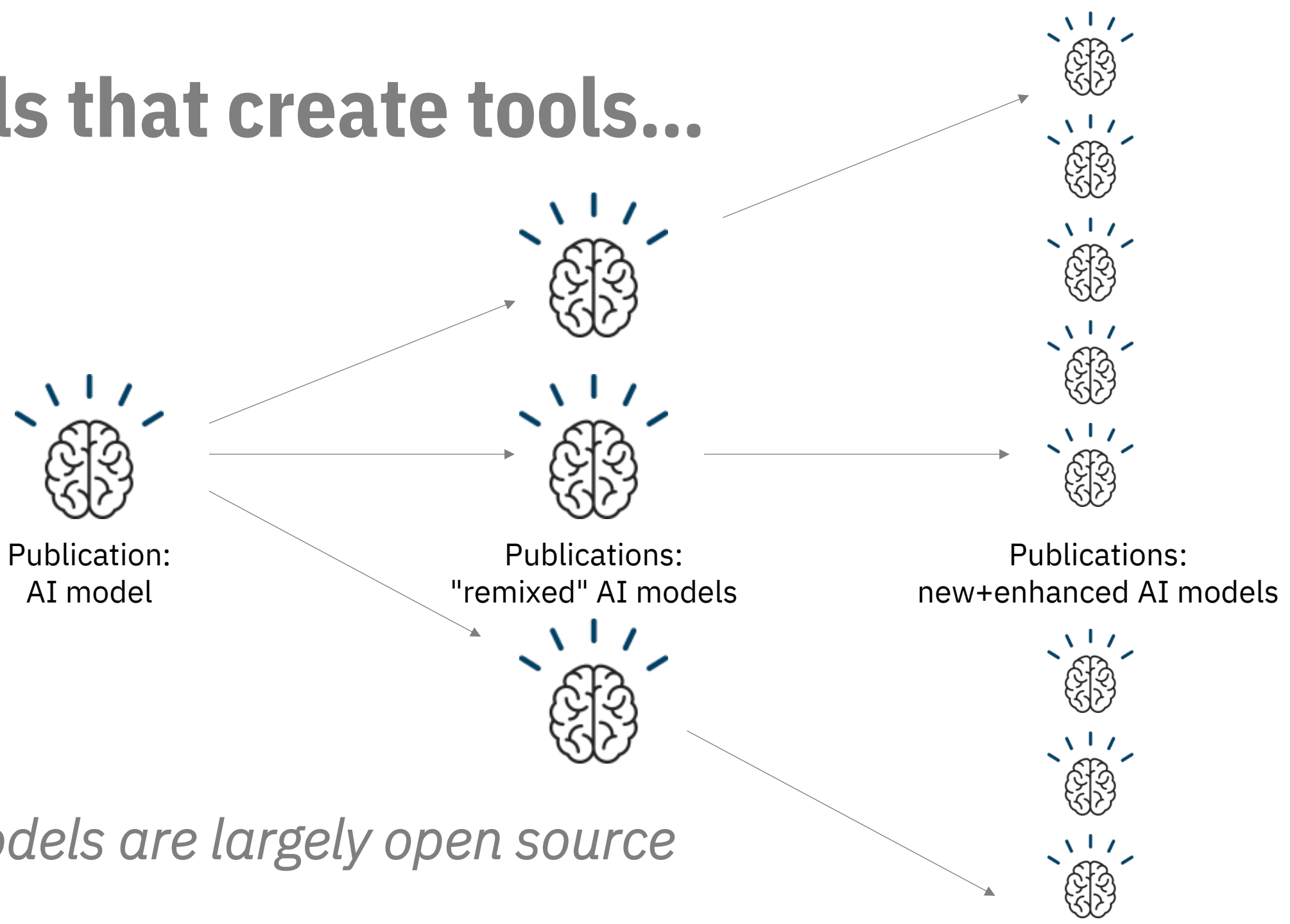
Low hanging fruits



# Fast-changing world of AI




# Tools that create tools...



*AI models are largely open source*

# AI Open Source: 10.000 new uploads – every day

**Hugging Face**

Models

Datasets

Spaces

Posts

Docs

Enterprise

Tasks

Libraries

Datasets

Languages

Licenses

Other

Filter Tasks by name

Multimodal

Audio-Text-to-Text

Image-Text-to-Text

Visual Question Answering

Document Question Answering

Video-Text-to-Text

Visual Document Retrieval

Any-to-Any

Computer Vision

Depth Estimation

Image Classification

Object Detection

Image Segmentation

Text-to-Image

Image-to-Text

Models 1,492,009

Filter by name

Qwen/QwQ-32B

Text Generation • Updated about 12 hours ago • 132k • 1.86k

deepseek-ai/DeepSeek-R1

Text Generation • Updated 15 days ago • 3.43M • 11.1k

microsoft/Phi-4-multimodal-instruct

Automatic Speech Recognition • Updated 3 days ago • 303k •

Wan-AI/Wan2.1-T2V-14B

Text-to-Video • Updated 13 days ago • 191k • 975

CohereForAI/aya-vision-8b

Image-Text-to-Text • Updated 6 days ago • 144k • 227

Models 1,492,009

+ 10 min

Models 1,492,081

huggingface.co

# What exactly is being uploaded there?



Search models, datasets, users

Models

Datasets

Spaces

Posts

Docs

Enterprise

The blueprints of artificial neural networks, with **billions of weight factors**

*(not all uploads are large language models)*

Depth Estimation

Image Classification

Object Detection

Image Segmentation

Text-to-Image

Image-to-Text

Models 1,492,009

Filter by name

Qwen/Qwen2-72B

Text Generation • Updated about 12 hours ago

deepseek-ai/DeepSeek-R1

Text Generation • Updated 15 days ago • ⬇️ 3.43M • ⚡ • ❤️ 11.1k

microsoft/Phi-4-multimodal-instruct

Automatic Speech Recognition • Updated 3 days ago • ⬇️ 303k • ❤️ 1.07k

Wan-AI/Wan2.1-T2V-14B

Text-to-Video • Updated 13 days ago • ⬇️ 191k • ⚡ • ❤️ 975

CohereForAI/aya-vision-8b

Image-Text-to-Text • Updated 6 days ago • ⬇️ 144k • ❤️ 227

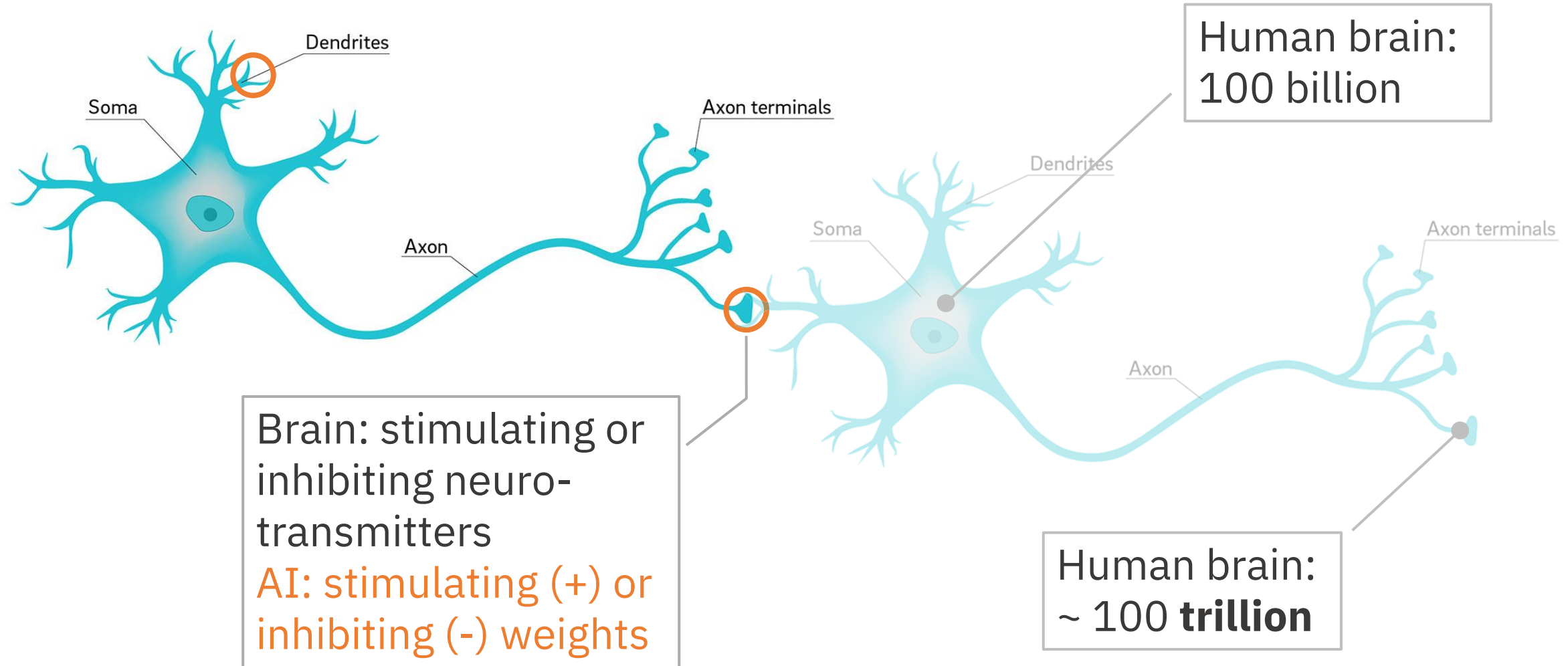
32 billion weight factors,  
not counting topology maps

32B ≈ 64 GB (only the weights) as FP16



# Analogy to natural neural networks

## Neuron



# Common AI models used in chatbot projects

|   |  |   |                                    |   |                          |
|---|--|---|------------------------------------|---|--------------------------|
| <b>Model name</b><br>llama3-8b-instruct                                       | <b>Model ID</b><br>meta-llama-llama-3-8b-instruct  | Pre-trained and instruction tuned generative text model optimized for dialogue use cases. | <b>GPU (Number of shards)</b><br>1 | <b>CPU and memory</b><br>2 CPU, 96 GB RAM | <b>Storage</b><br>40 GB  |
| <b>8b</b> = the same in dumber  |  |   |                                    |   |                          |
| <b>Model name</b><br>llama3-70b-instruct                                      | <b>Model ID</b><br>meta-llama-llama-3-70b-instruct | Pre-trained and instruction tuned generative text model optimized for dialogue use cases. | <b>GPU (Number of shards)</b><br>4 | <b>CPU and memory</b><br>10 CPU, 246 GB   | <b>Storage</b><br>180 GB |
| <b>70b</b> = 70 billion weights (~synapses) in FP16 format = 140 GB of tables |  |   |                                    |   |                          |

High performance GPUs boast 50~90 GB on-chip storage

"Video RAM"

# Does AI always equal Petabytes and GigaWh?

**Petabytes**  
of examples



**GigaWh**  
of energy

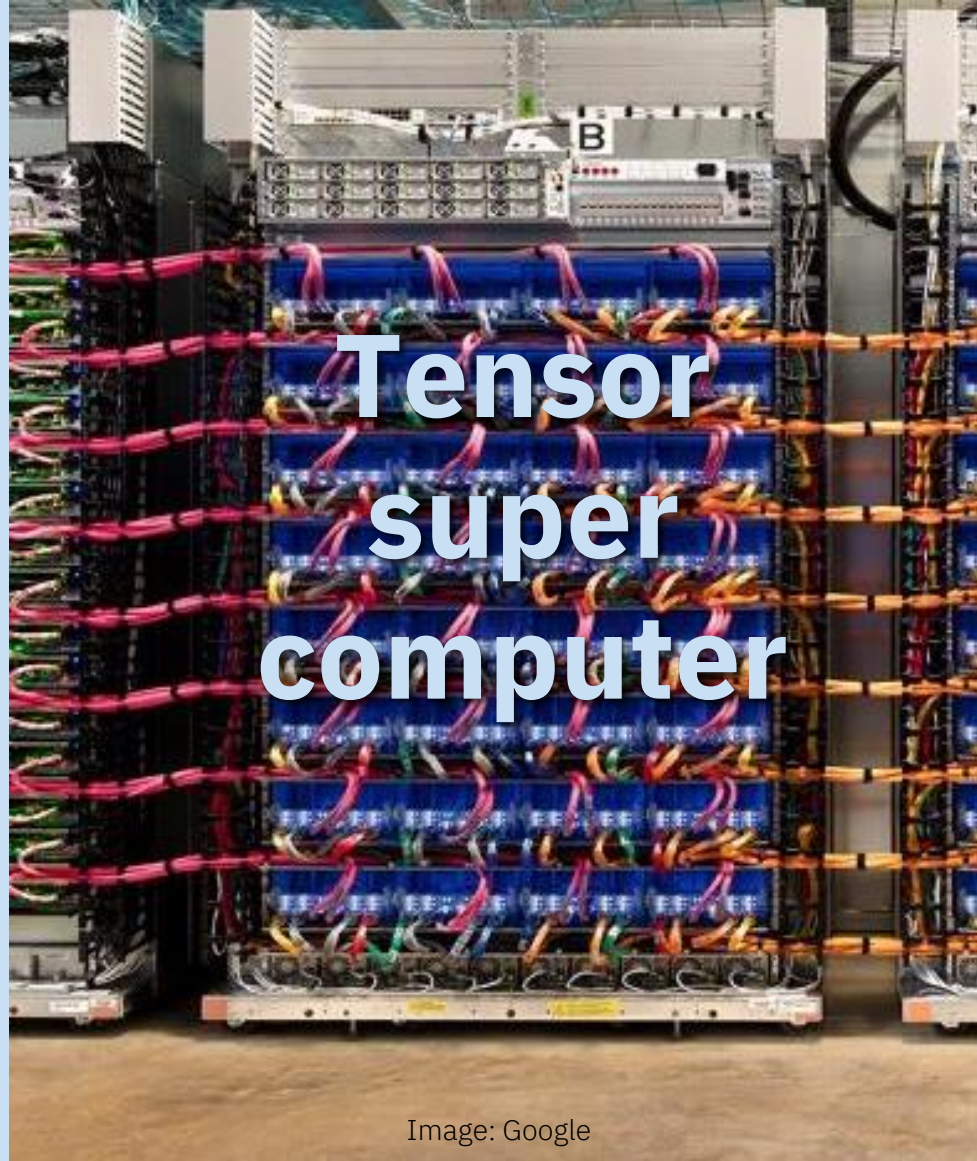


Image: Google

**LLM, GPT etc.**

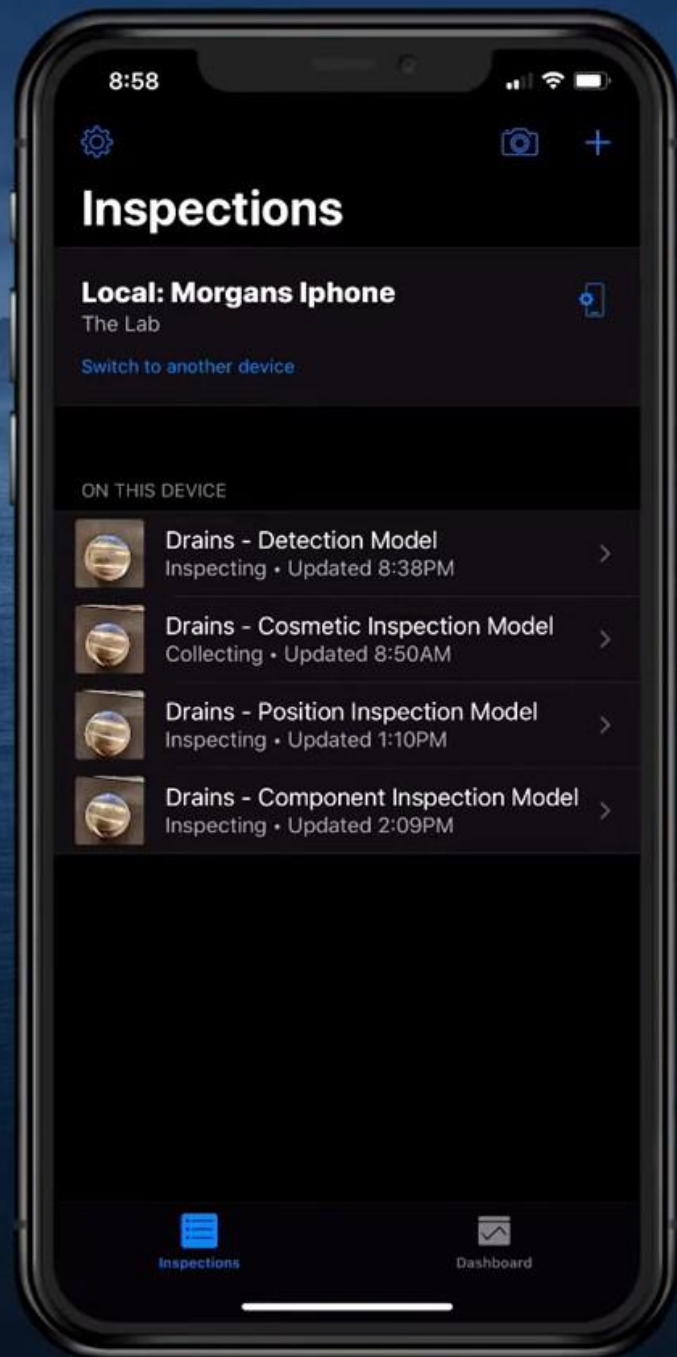


Published,  
re-usable,  
re-mixable

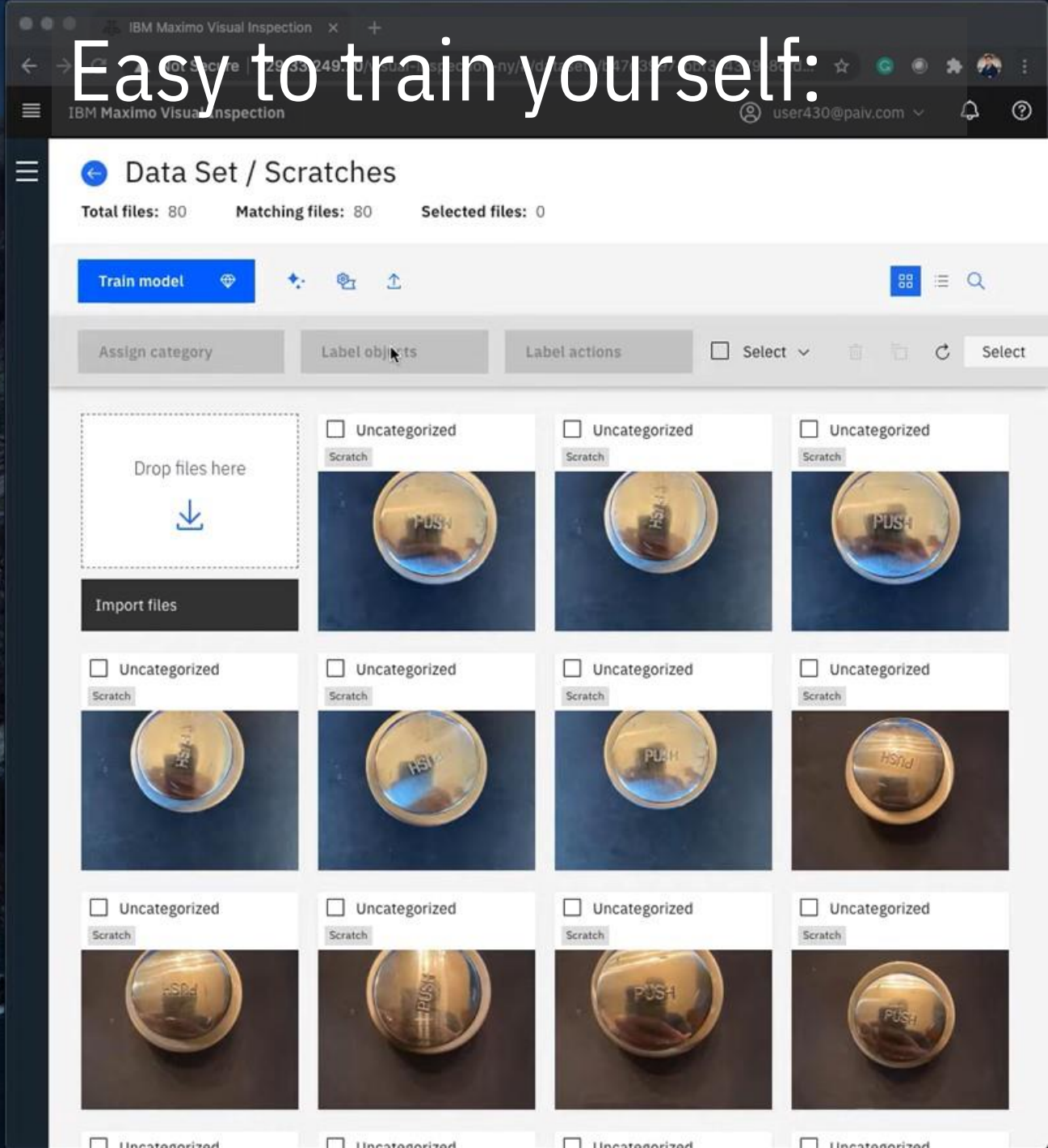
+high data preparation effort

from here it's only Gigabytes

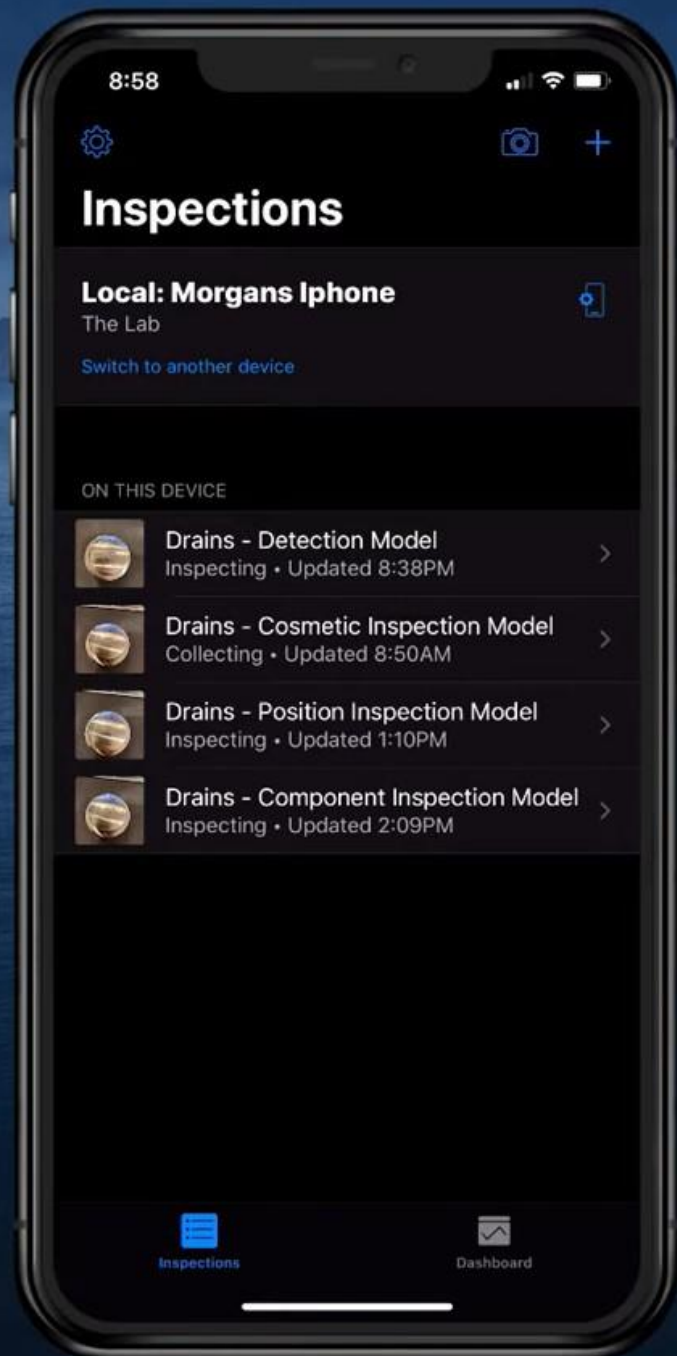




Easy to train yourself:







# Easy to train yourself:

## Data Set / Scratches

Total files: 80 Matching files: 80 Selected files: 0

Train model

Assign category

Label objects

Label actions

Select

Select

Non-generative AI  
is cheaper to train



## Objects

Add label

A Scratch (5)

A1 Scratch

A2 Scratch (0.99)

auto label

A3 Scratch (0.98)

auto label

A4 Scratch (0.96)

auto label

A5 Scratch (0.74)

auto label

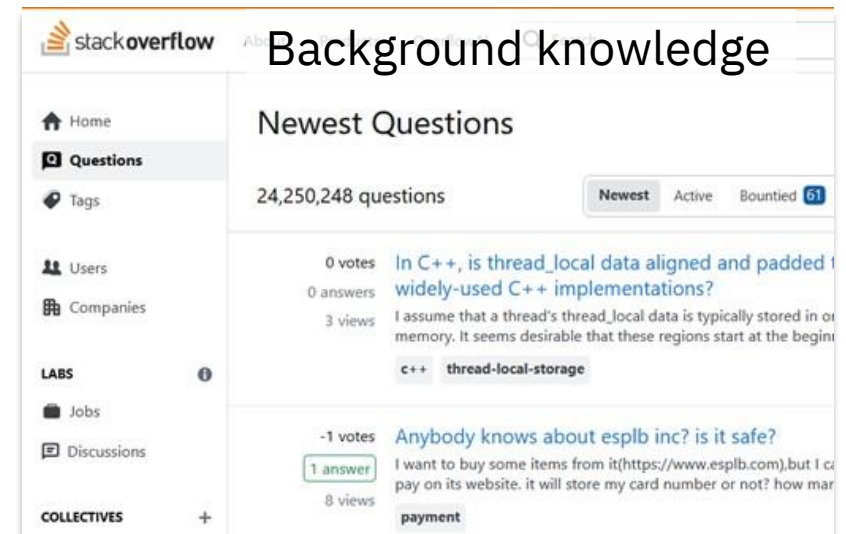
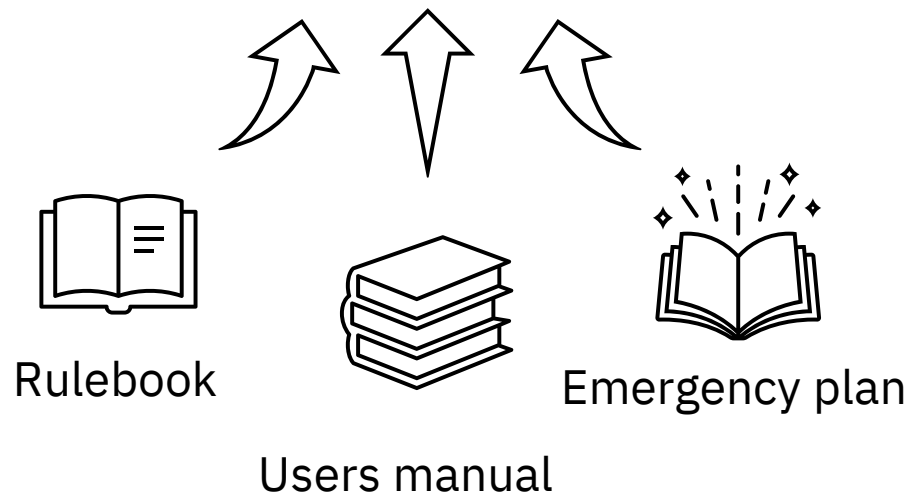
Train an image feature  
discriminator model  
for a few cents.

Example:  
Maximo visual inspection

# More complex tasks => Use ***building block AI***

Example:

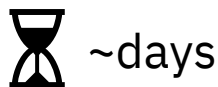
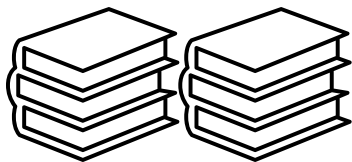
“How do I create an emergency rule for a cyberattack alarm in my organization's firewall?” ”



# Building blocks: Divide the task into smaller prefabricated AIs

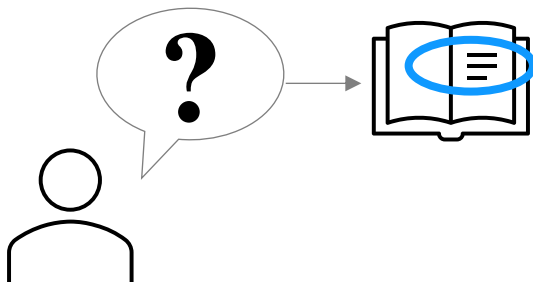
1

"Embedding" LLM scans my archive



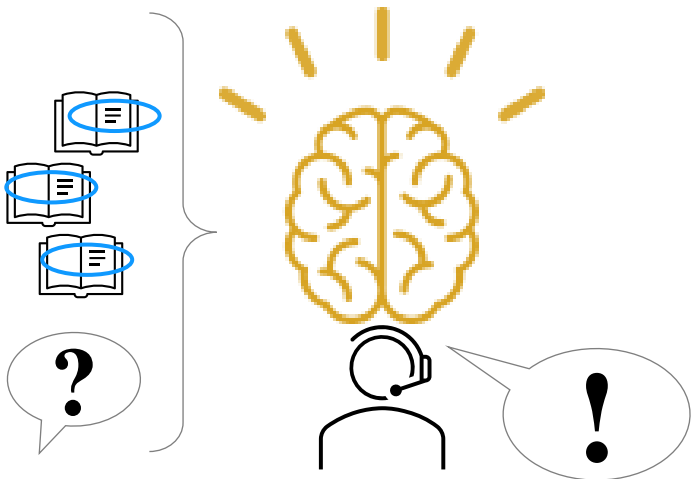
2

Embedder recommends most suitable document quote(s) for the answer



3

Document quotes & my question are passed to the generative chatbot



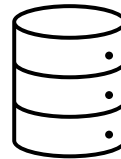
The chatbot synthesizes the answer **from facts and background knowledge**

# Retrieval-Augmented-Generation "RAG"

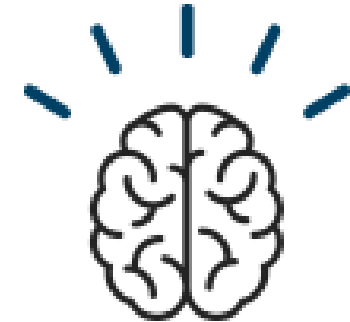
**Embedding model,**  
e.g. Granite-107m



**Vector database**  
e.g. Milvus



**Chatbot model,**  
e.g. Llama 3.2-70b



All these components are **freely** available.  
**But: new/better ones are published every week.**



# Technically, what happens here?

## Embedding model



## Output vector

104 231 98 475 120 (...) × 150

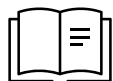
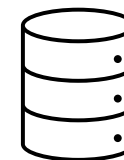
↓ similar-ish ↑

109 230 91 425 190 (...)

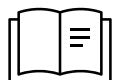
↓ dissimilar ↑

294 33 749 102 719 (...)

## Vector database



*The Granite Embedding collection delivers high-performance sentence-transformer models optimized for retrieval ...*

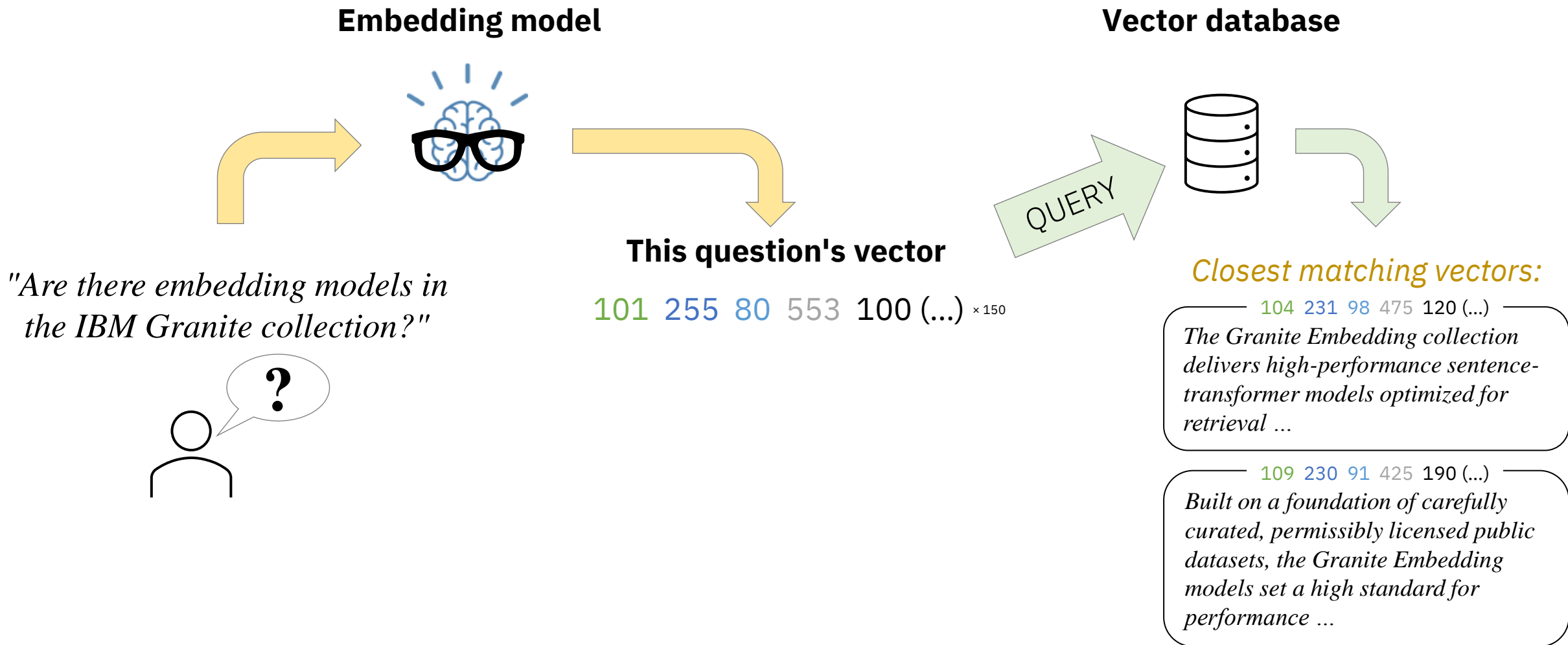


*Built on a foundation of carefully curated, permissibly licensed public datasets, the Granite Embedding models set a high standard for performance ...*



*Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed eiusmod tempor incididunt ut labore et dolore magna aliqua ...*

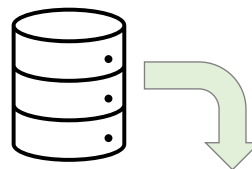
# Technically, what happens during QUERY ?



# Technically, what happens during PROMPT ?

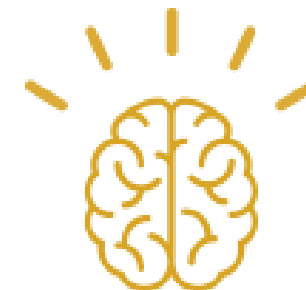
## Chatbot model

### Vector database



*The Granite Embedding collection delivers high-performance sentence-transformer models optimized for retrieval ...*

*Built on a foundation of carefully curated, permissibly licensed public datasets, the Granite Embedding models set a high standard for performance ...*



The Granite Embedding collection delivers high-performance sentence-transformer models optimized for retrieval ...

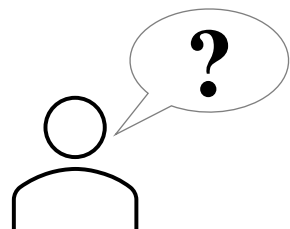
Built on a foundation of carefully curated, permissibly licensed public datasets, the Granite Embedding models set a high standard for performance ...

**Only use the above information:**

Are there embedding models in the IBM Granite collection?

> \_\_\_\_\_

*"Are there embedding models in the IBM Granite collection?"*



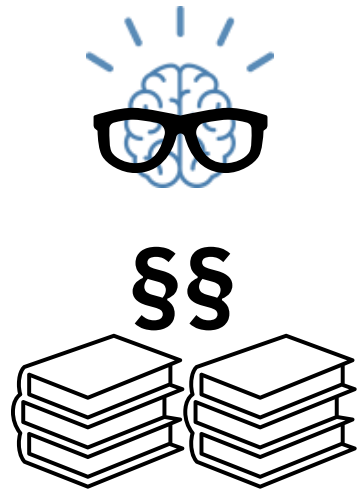
NEAR-REAL TIME

# Real deployment case

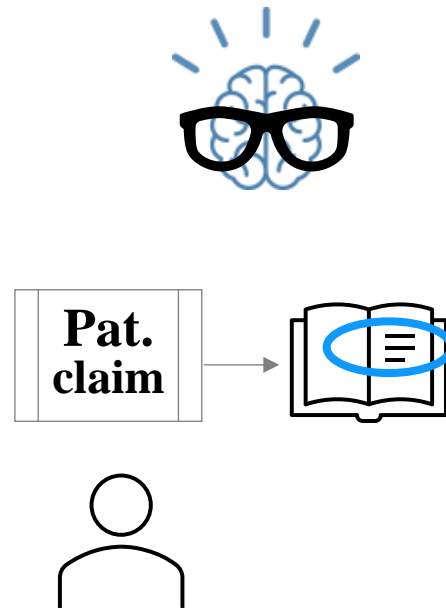


# Project at a law firm : **RAG** to identify patent "claimability"

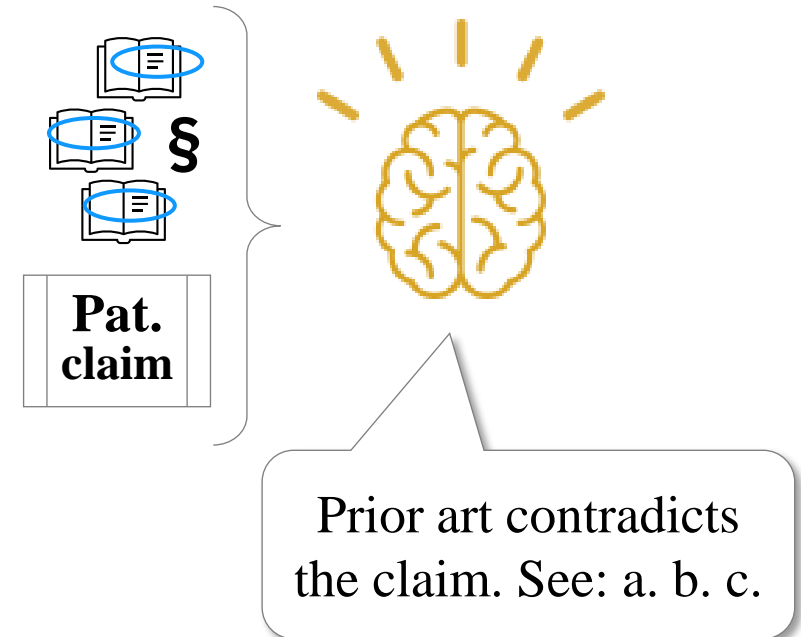
1



2

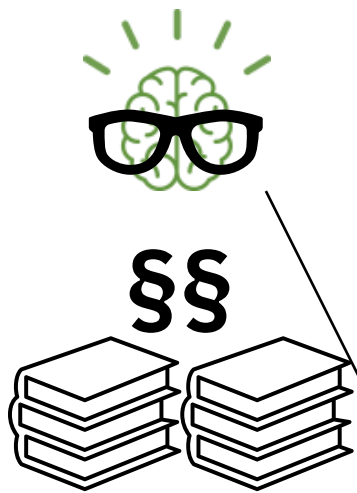


3

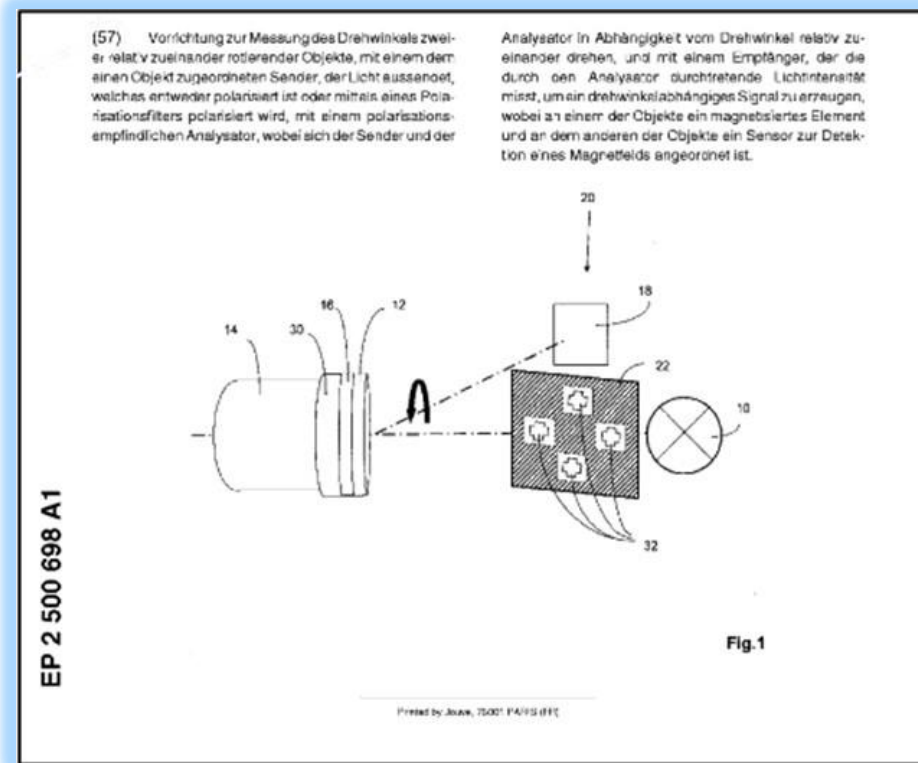


# Improvement: “Image analysis” was required

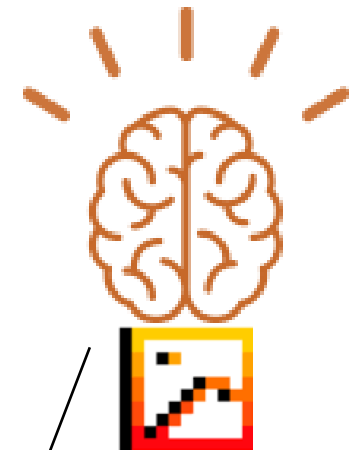
During the test phase, we realized **how important drawings and diagrams** are for the claim context.



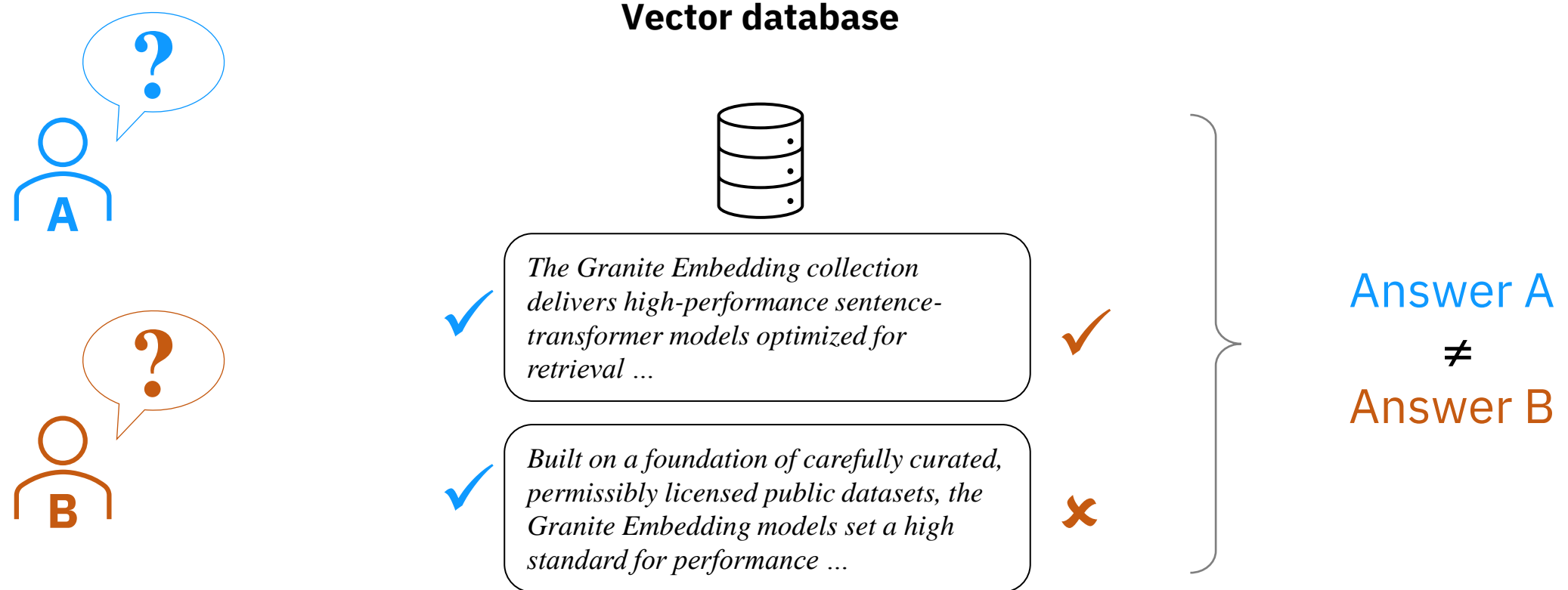
Holds diagrams?  
Vector DB size?



The generative chat model was upgraded  
to a multimodal chat bot (e.g. Pixtral)



# Improvement: Multi-tenancy with access control



Different tenants must be restricted to answers originating from documents **they have access to**.  
= controlling vector database output per user

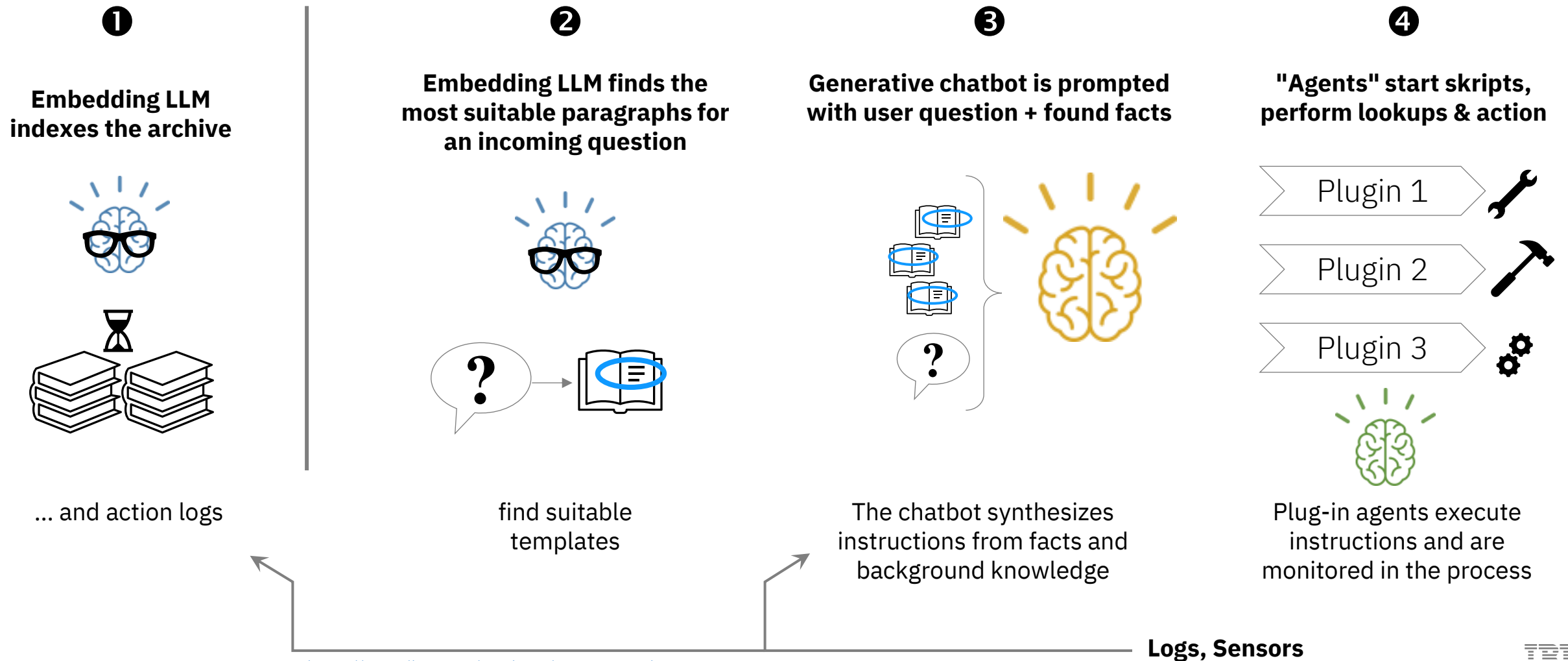
Interim conclusion :  
**AI = ongoing prototyping**

IT requirements, data volumes, resources, etc. will  
keep changing with each new inference model

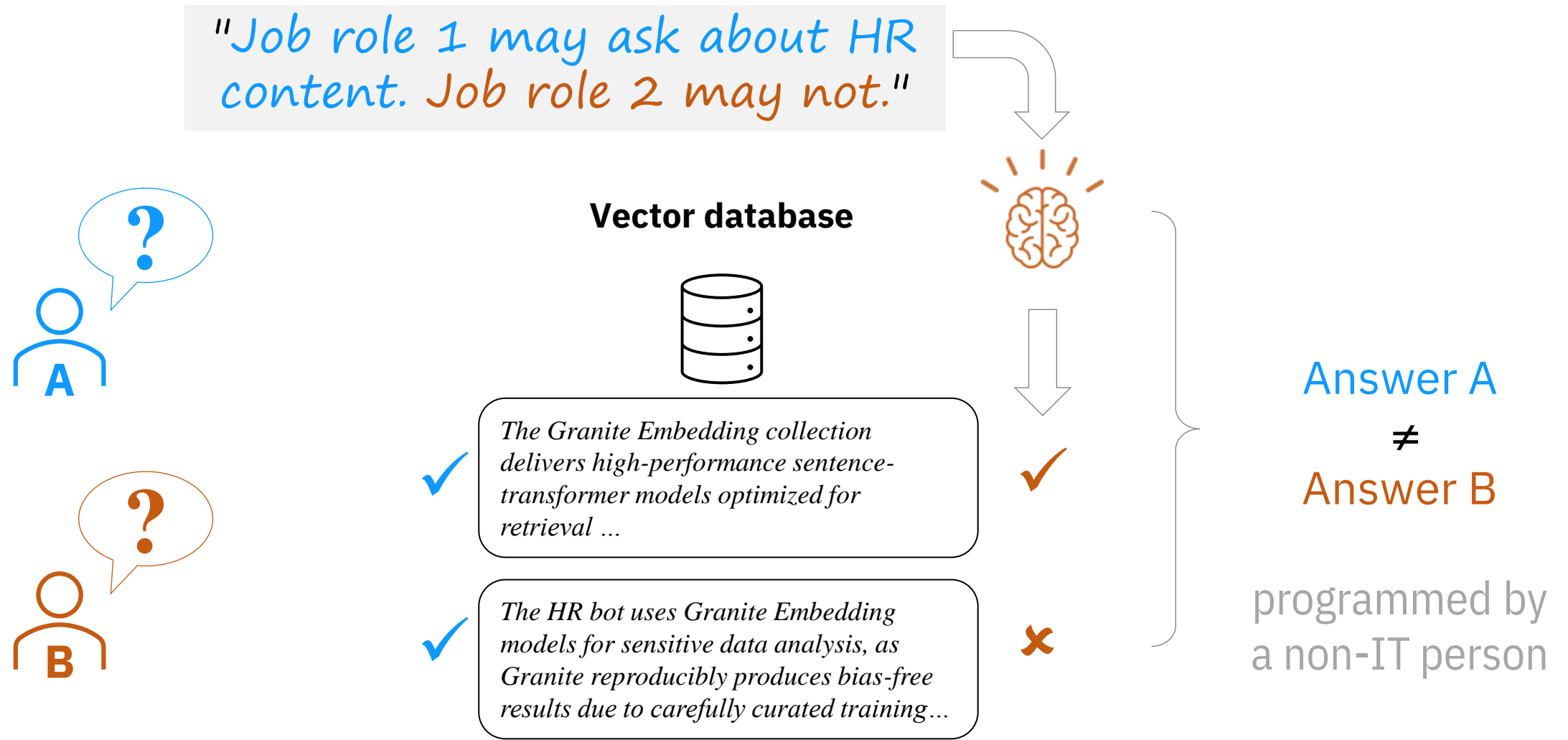


# Agentic AI – the next automation step

“ Build a cyber attack emergency rule into the firewall. Quick!!! ”



# NO\_CODE: What if ... access control was **in natural language**?



# Clerks are waiting for Agentic AI !



**LOW CODE / NO CODE**

Recommended strategy for IT:  
Don't build AI bots – empower end users to do it.



Community bots created by IBMers


(1 year!)

Explore

Q

Search for an assistant or workflow


- All (49)
- Official (6)
- Community (43)



AskIBM

AskIBM enables interaction with Large Language Models (LLMs) through a chat interface, providing domain-specific expertise and responses tailored to IBM's internal information


Launch Assistant



Translation

This productivity workflow translates text into other languages.


Launch workflow



AskSRE

This workflow provides the guidance for SRE Field Guide and help us to choose the right topics


Launch workflow



QMX Document Q&A

This workflow provides concise answers to questions about process documentation for manufacturing operators.


Launch workflow



System Security Analyst


Good at fixing system security issue. Users input a security issue number then response with the issue root cause and provide solutions to fix issue.

Launch workflow



Ask the IBM Cloud Go to Market Team


This workflow answers questions in three key areas managed by the IBM Cloud GTM Product Management team. It covers Promotion codes, Enterprise Savings Plan (ESP) contracts, and IBM Cloud Terms and Conditions of use.



System Security Analyst

Good at fixing system security issue. Users input a security issue number then response with the issue root cause and provide solutions to fix issue.


Launch workflow



Ask the IBM Cloud Go to Market Team

This workflow answers questions in three key areas managed by the IBM Cloud GTM Product Management team. It covers Promotion codes, Enterprise Savings Plan (ESP) contracts, and IBM Cloud Terms and Conditions of use.


Launch workflow



ITSS AI Chatbot

This is a chatbot that references IBM's IT Security Standard (ITSS) guidelines in order to help IBM employees understand the correct processes and procedures around IT security.


Launch workflow



AskMarketInsights Assistant

This workflow uses RAG to allow employees to gather insights from lengthy Gartner reports.


Launch workflow



Company Profiler

This workflow provides users with company profiles using external data source which assist IBMers in making strategic decisions on client segmentation, resource deployment and sales tactics.


Launch workflow



ASKSupportBundle

Find Bundled and Supporting Programs for a Product

Launch workflow



IBM Performance Benchmark Analysis Bot


AI-driven workflow to analyze logs related to system performance on Linux on Z platforms, speeds up the process of identifying issues and taking corrective measures.

Launch workflow




IBM Data Resiliency Assistant

This assistant will help IBM Data Resiliency Technical Advisors answer questions.




Launch workflow



CSIRT Notification Letter

This workflow is designed to help generate a CSIRT notification letter in the event of a security incident.


Launch workflow



Askount.3

This workflow search the accounts information (minor and subminor) so this will be available when the user needs to submit an expense by PO or Non PO.


Launch workflow



CP4BA & watsonx Orchestrate Licensing Assistant

This workflow helps you obtain information about the Cloud Pak for Business Automation licensing


Launch workflow



ASKIBM\_CIO-Consulting-IT

This workflow is designed to assist CIO-Consulting-IT users in order to provide them information that would solve their needs.


Launch workflow



AskSBLIW

This workflow will guide SBLIW users to answer most of their day-to-day questions.


Launch workflow



Contract Finance Management Assistant

This work stream will become your assistant to provide relevant explanations based on user's descriptions or questions about Consulting finance information.

Launch workflow




Askme AH Chatbot

This workflow provides clear instructions for agents to resolve issues efficiently and effectively. You can search the required set of instructions for performing the troubleshooting so that it can save your time and solve the issue quickly. It is built to help agents to get steps of troubleshooting for all the most frequently asked questions. Agents needs to search with the Error so that it would share the output with set of instructions to solve that particular issue.

Launch workflow

# How to get started: (*Low hanging fruits*)

1. Data archive in place? Run a small **RAG** project!
2. For quick model testing, a single workstation runtime will do:  
e.g. Ollama  (ollama.com)
3. Deploy a **Code Assistant** inference model! Soon in-house developers will no longer need to share their ideas with OpenAI, Microsoft or DeepSeek.

(Java, Python, C++, Ansible code assistants are all Open Source)

# Prototypes go in production? Time for an AI platform

## 4. Deploy an AI runtime platform

- Simplifies nonstop prototyping
- Manages efficient GPU allocation
- Ops: HA, DR, security, governance
- Helps optimize power efficiency
- Automates RAG, Agentic AI, etc.



Difference  
between AI  
*-model* and  
*-platform*?




**watsonx**

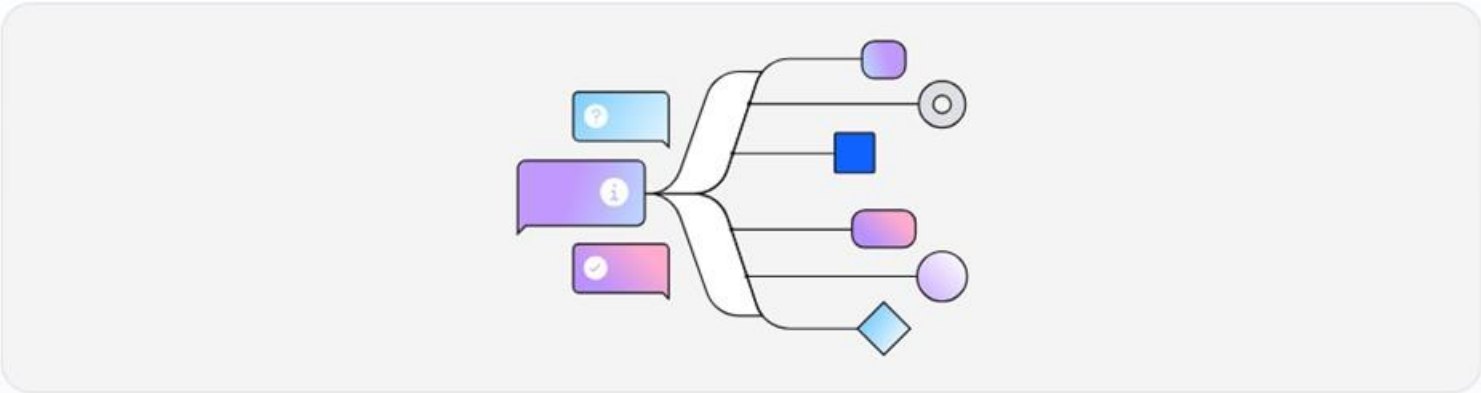
Chat Structured Freeform

AI Model: granite-3-8b-instruct

watsonx 01:05 AM

Customize your chat

Before you start chatting, you can update the current settings and ground the chat with documents. To upload documents or an image, click  next to the input field.





View full prompt text


```
<|start_of_role|>system<|end_of_role|>You are Granite, a language model developed by IBM in 2024. You are a cautious assistant. You carefully follow instructions. You are helpful and harmless and you follow ethical guidelines and promote positive behavior.<|end_of_text|><|start_of_role|>assistant<|end_of_role|>
```

Simple RAG testing

Sample questions

 Add documents

 Add image

 Type something...

ent alternatives to a 'for

What is the Transformers architecture?



## Build



Model: llama-3-3-70b-instruct



## Setup



## Configuration



Framework

Architecture

LangGraph



ReAct



Instructions

You are a helpful assistant that uses tools to answer questions in detail.  
When greeted, say "Hi, I am watsonx.ai agent. How can I help you?"

## Tools

Add a tool

Create custom tool

Added tools (1)

## Google search



Retrieve information from the internet with the Google search engine.

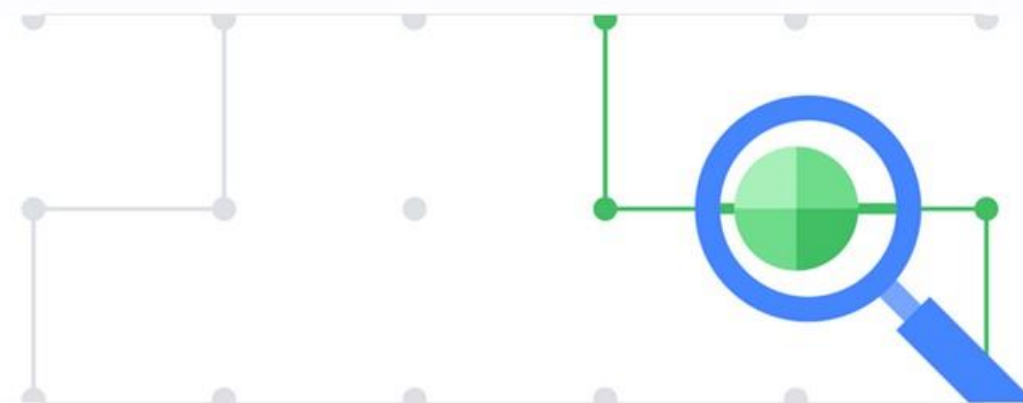
## Agent preview



watsonx Agent 01:00 AM

## Welcome to watsonx Agent

Change this description to reflect your particular agent



Type something...



Simple agents



OR ...

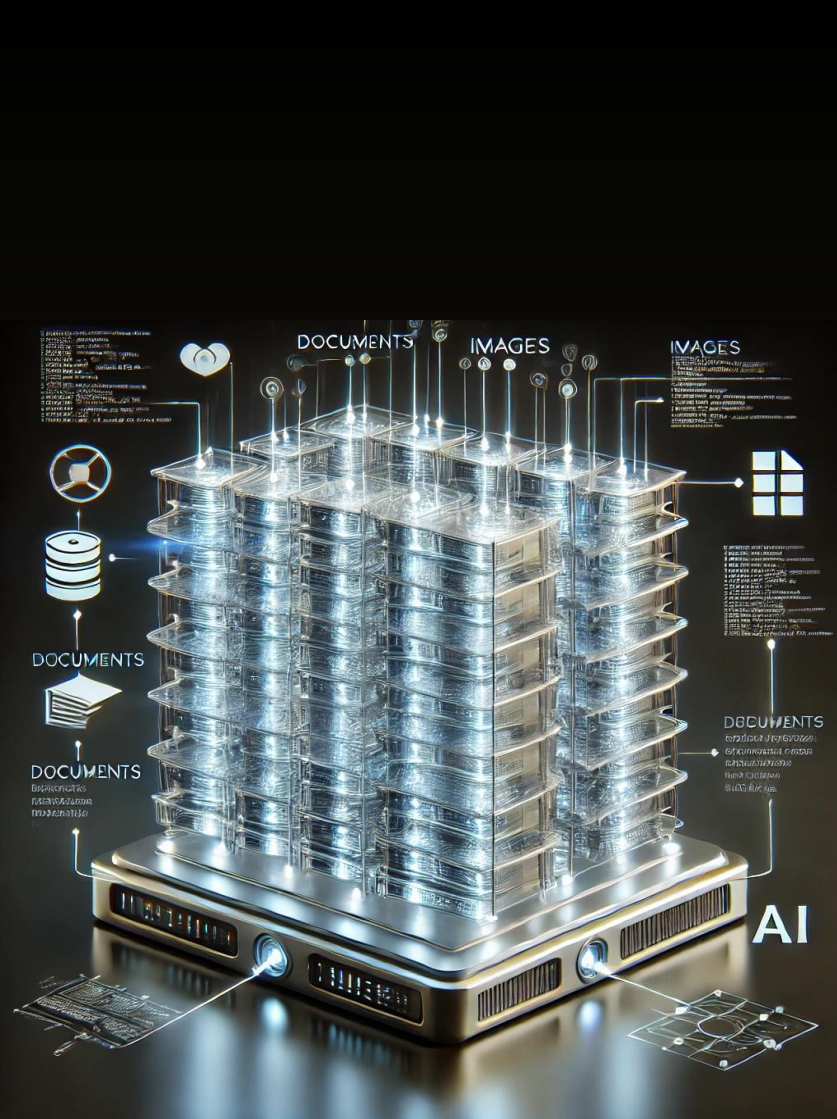
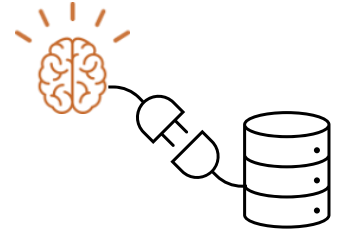
# Make AI **part of** your Data Storage

## Drop-in *RAG sidecar* for Storage

Understands mixed-media content and enables natural language interaction.

**Example: "list all videos containing half-hidden bicycles in complex street traffic"**

- ✓ for any object/file system with change notification
- ✓ starting with IBM Storage Scale / GPFS
- ✓ action agents for cache prefetch, eviction, ...



"Content-aware"

# LOW CODE / NO CODE



No-coder's examples:

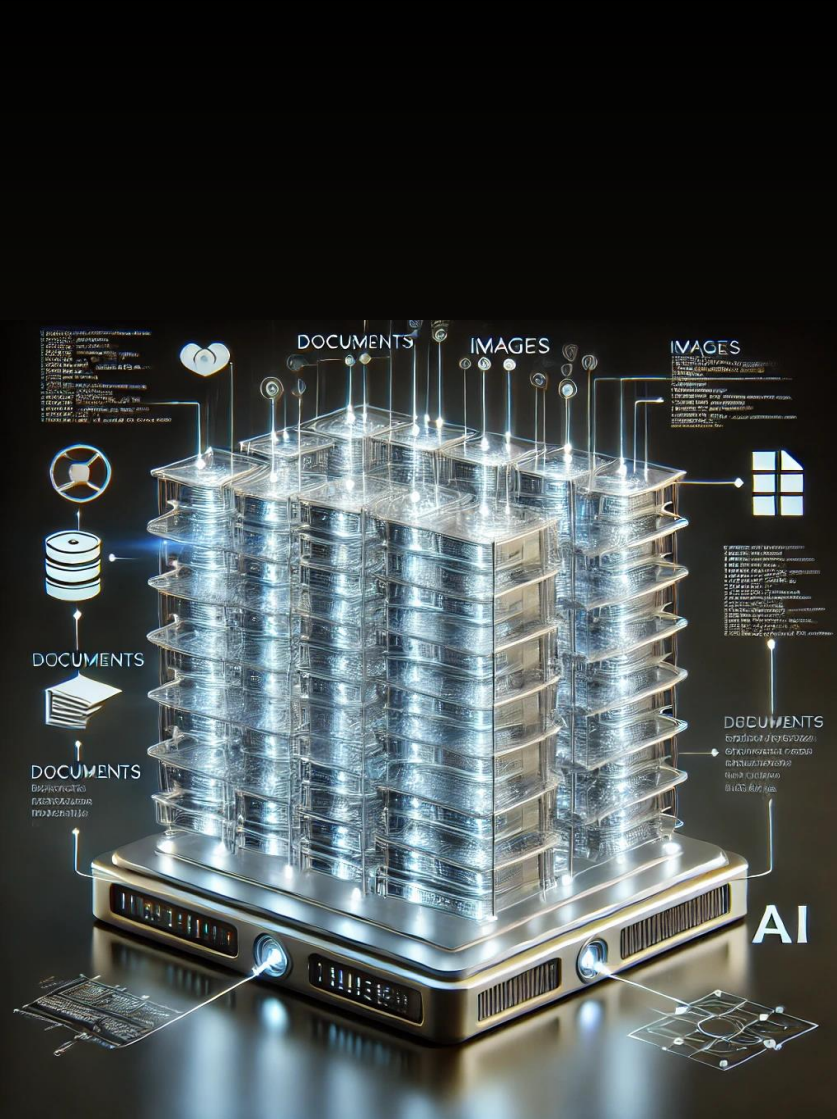
*"Show me the most recent order confirmation for client ACME Data Removal"*

*"When is the next due date for the quarterly ACME Data Removal bill?"*

*"Who is our internal process owner for real estate acquisitions?"*

*"List all screenshot pictures older than three years and put them in a deletion script."*

"Content-aware"







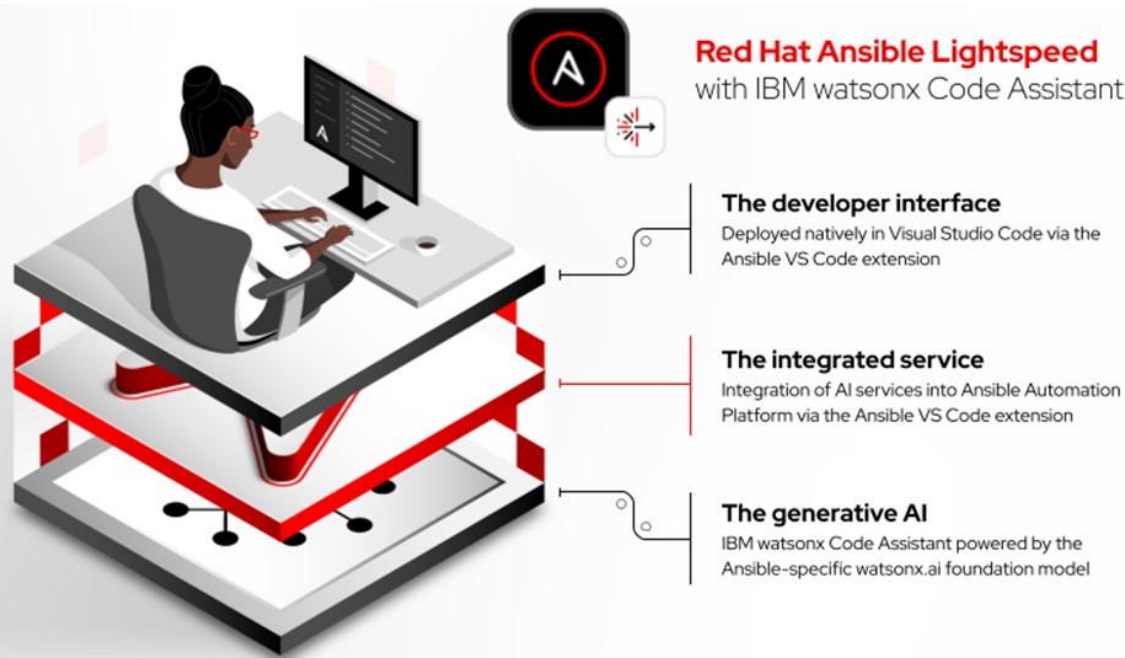
What makes **my** storage admin life simpler?



# Infrastructure-as-code and IT Automation

## Watsonx Code Assistant for Red Hat Ansible Lightspeed

- Approximately 4.000 developers participated in the 2023 technical preview.
- **85%** overall average acceptance rate of AI-generated recommendations.  
(from July 27 – Oct 23, 2023, based on over 41.000 recommendations)



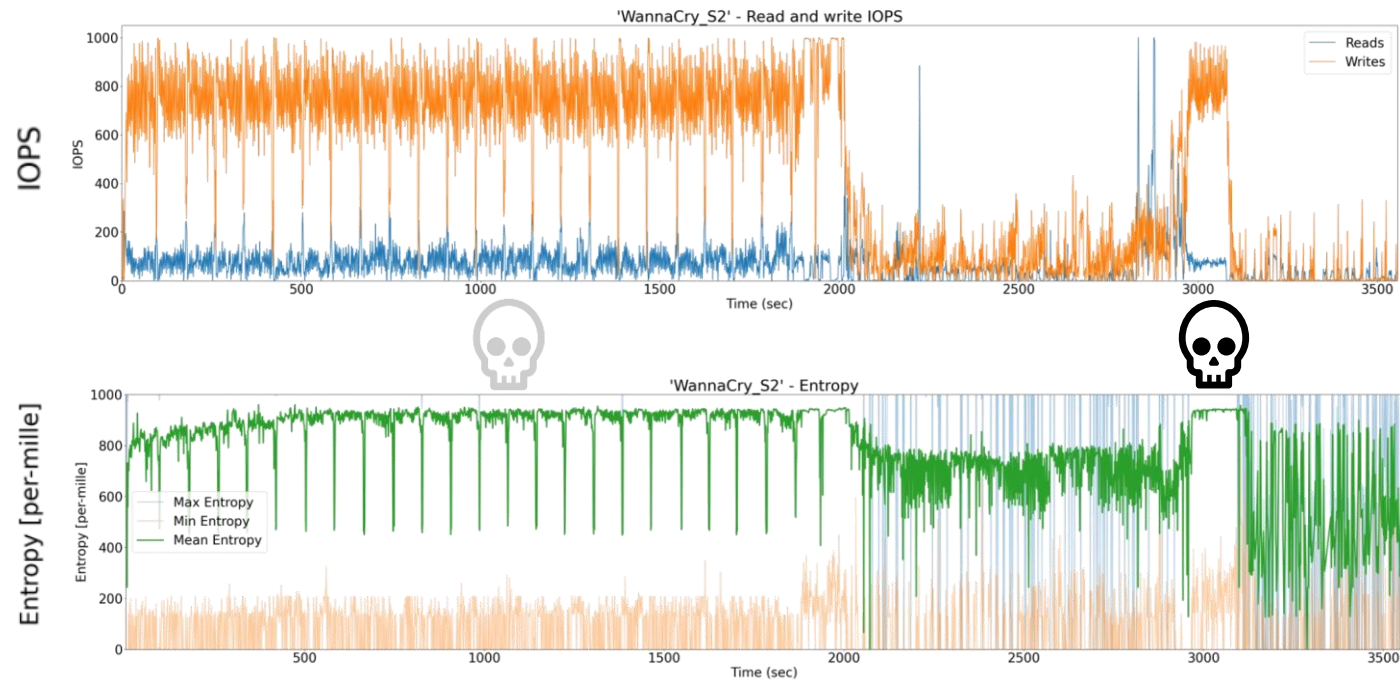
↑  
Productivity  
improvements  
between **20-45%**

# Storage Cyberthreat Prevention with AI

Can we detect cyber attacks *inside* SSDs, at the Flash chip level?

**Yes**, with a Flash IO anomaly detector model trained on "usual suspects":

**Exfiltration-** and **Encryption Trojans**



# FlashCore Module FCM

a cyber-capable SSD with compute power

## Generation 1 – 2018

- optimizing Flash resource usage and longevity

## Generation 4 – 2023

- using surplus ARM cores for anomaly detection
- real-time entropy measurement & timing analysis
- Individual FCM 'logs' still must be correlated among each other, so external firmware is required



2025: **QLC-workload** version



1<sup>st</sup> generation  
FCM, 2018



# What's cooking in the labs?

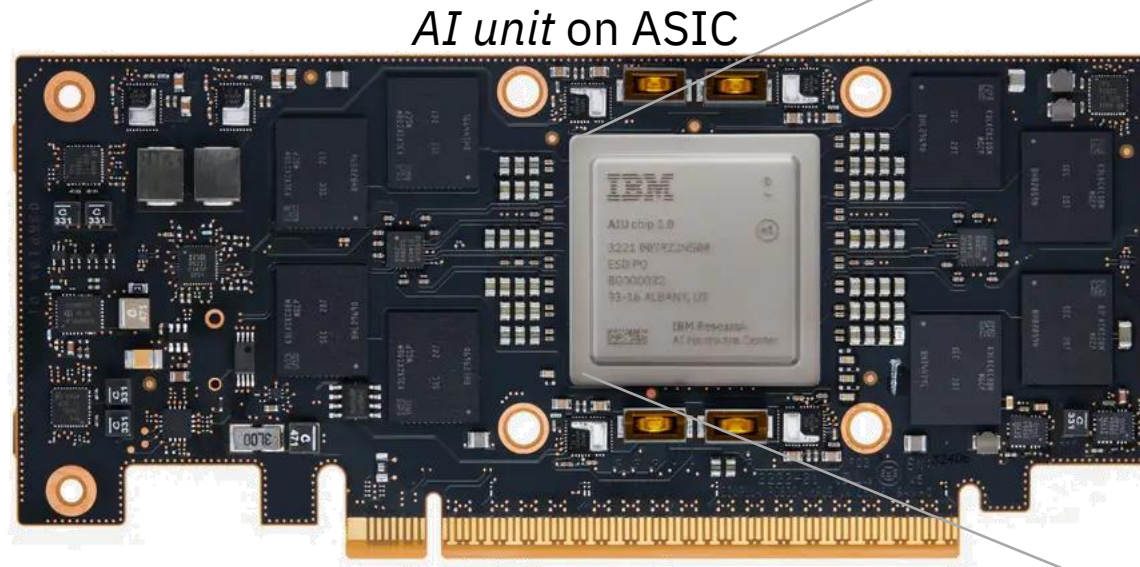


# Alternatives to power-hungry GPUs?

*Step 1: Fuzzy AI*

*Step 2: Computational memory*

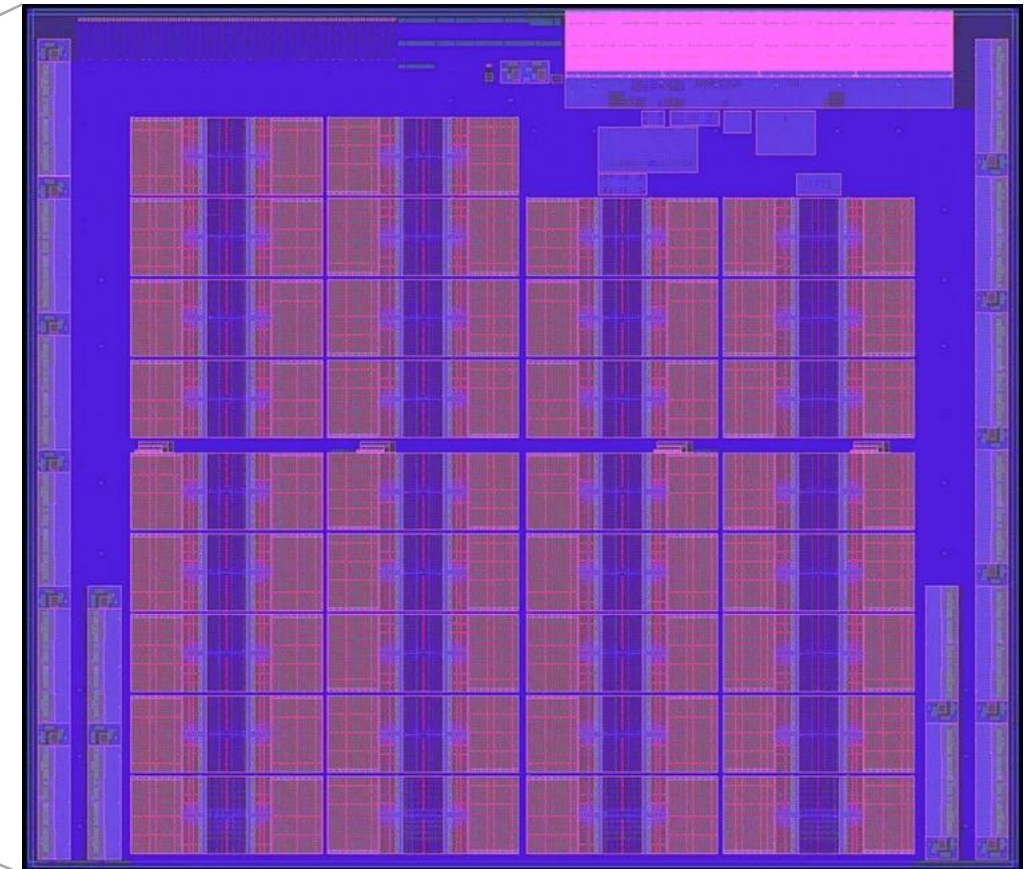
The IBM AIU swaps the *GPU* for an application-specific integrated circuit (ASIC) for deep learning. The IBM AIU is also designed to be as easy-to-use as a graphics card. It is a scaled-up evolution of the AI accelerator that is integrated into the IBM Telum processor for the z16.



*AI unit on ASIC*



75 Watts



The IBM AIU realizes the concept of *approximate computing* by using “fuzzy” calculus on FP8 rather than FP16 or FP32



# Green inference with computational memory LLMs

Julian Büchel • IBM Research • 08 January 2025

*Performing low-latency inference with billion-parameter LLMs on a device the size of a thumb-drive and a power consumption of <10W is a dream of ours.*

In a joint project between IBM Research and Micron Technology, we lay out a way to achieve this dream: Our latest paper "Efficient scaling of large language models with mixture of experts and 3D analog in-memory computing" published in Nature Computational Science proposes to **marry 3D Analog In-Memory Computing (3D AIMC) using high-density 3D non-volatile memory (NVM)** with the conditional compute paradigm of Mixture of Experts (MoEs).

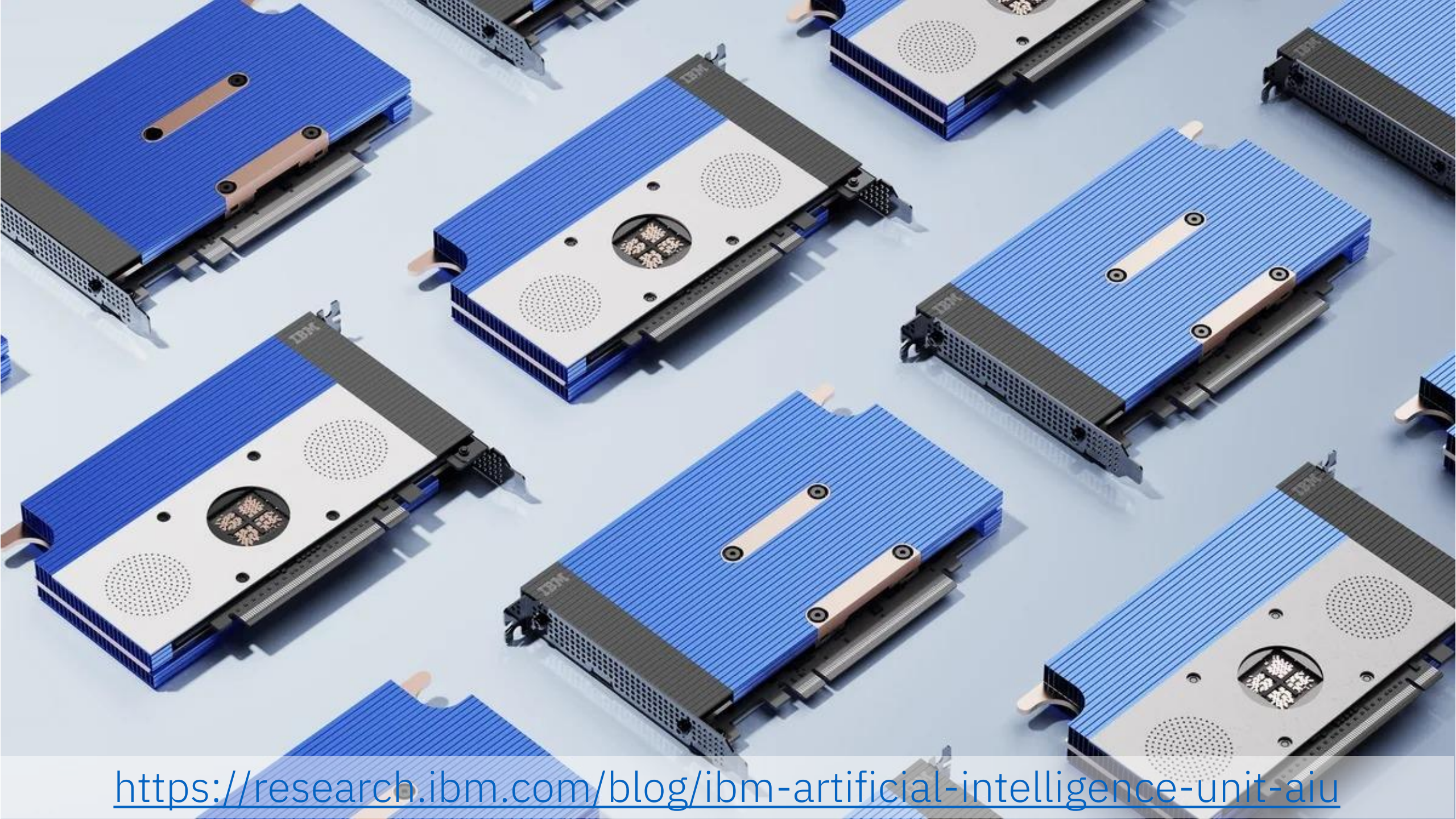
3D AIMC is a promising approach to energy efficient inference of LLMs. On a high level, 3D AIMC can be thought of as stacking many 2D AIMC crossbars on top of each other. Each crossbar can be used to perform Matrix-Vector-Multiplications (MVMs) using the weights programmed into the NVM devices of the respective layer (or tier). Due to hardware constraints, performing MVMs in parallel across every tier isn't possible, a constraint we introduce as the One-Tier-at-a-Time (OTT) constraint. This constraint can become a bottleneck for very large layers (10s of thousands of rows/columns).

Enter MoEs:

MoEs are interesting because during the forward pass, only a subset of the parameters in the model is used to process each token. This makes MoEs much more scalable in terms of number of parameters. Under the hood, MoEs swap out every MLP layer in the transformer with many smaller MLP layers, the so-called experts. During the forward pass, each token is then processed by a small subset of these experts. (...)

 [Paper](#)  [Analog-MoE](#)  [Simulator](#)





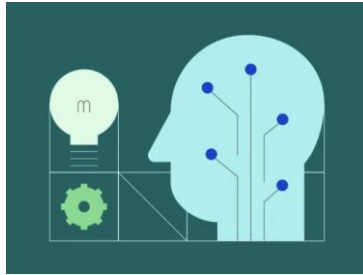
<https://research.ibm.com/blog/ibm-artificial-intelligence-unit-aiu>



# Free IBM Software Download for Universities



[ibm.com/academic](https://ibm.com/academic)



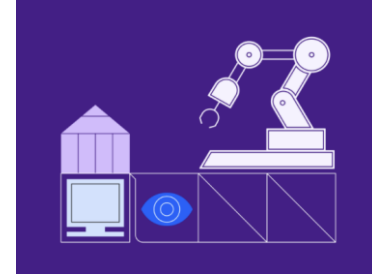
AI



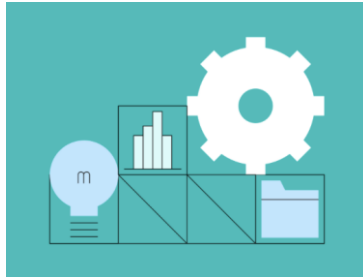
IBM Cloud



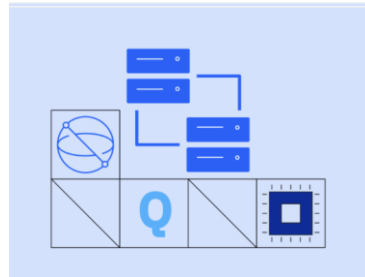
Data Science



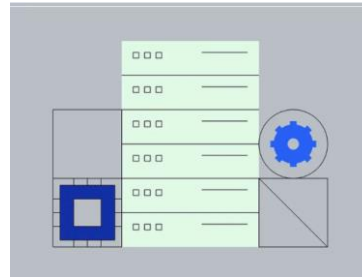
IBM Engineering



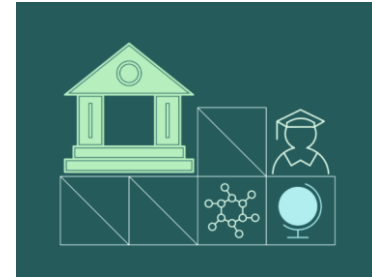
IBM Automation



Quantum Computing



IBM Z



Red Hat Academy

## Cloud Access

Access to the IBM Cloud and select cloud-based resources and applications, such as the Watson APIs

## Software

Access to the same software used by our commercial customers leading to practical training for today's jobs

## Courseware

Faculty access to enterprise quality courses for inclusion in part or whole into existing and new curriculum



IBM **SkillsBuild**



[linkedin.com/in/axelkoester](https://linkedin.com/in/axelkoester)

