What is your AI risk Score?

Legal notes and disclaimer

IBM products and services are warranted according to the terms and conditions of the agreements under which they are provided.

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at IBM's sole discretion. Performance data contained herein was generally obtained in controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer's responsibility to ensure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer is in compliance with any law.

The Foundation for a Successful AI Business Strategy - Trusthworthy AI

Mike Kehoe,
EMEA Data Security Leader, Data & AI Platform

Aleksandra Kaszuba
NCEE Wx.Governance Virtual SWT leader

Nicola Piazza
EMEA Principal Security Technical Leader

Albert Puah
WW Sales Leader, Data Security

AI Governance and AI Security

The Sandwich Dilemma

Human (Task Performing)

Desired Results

Ai (Task Performing)

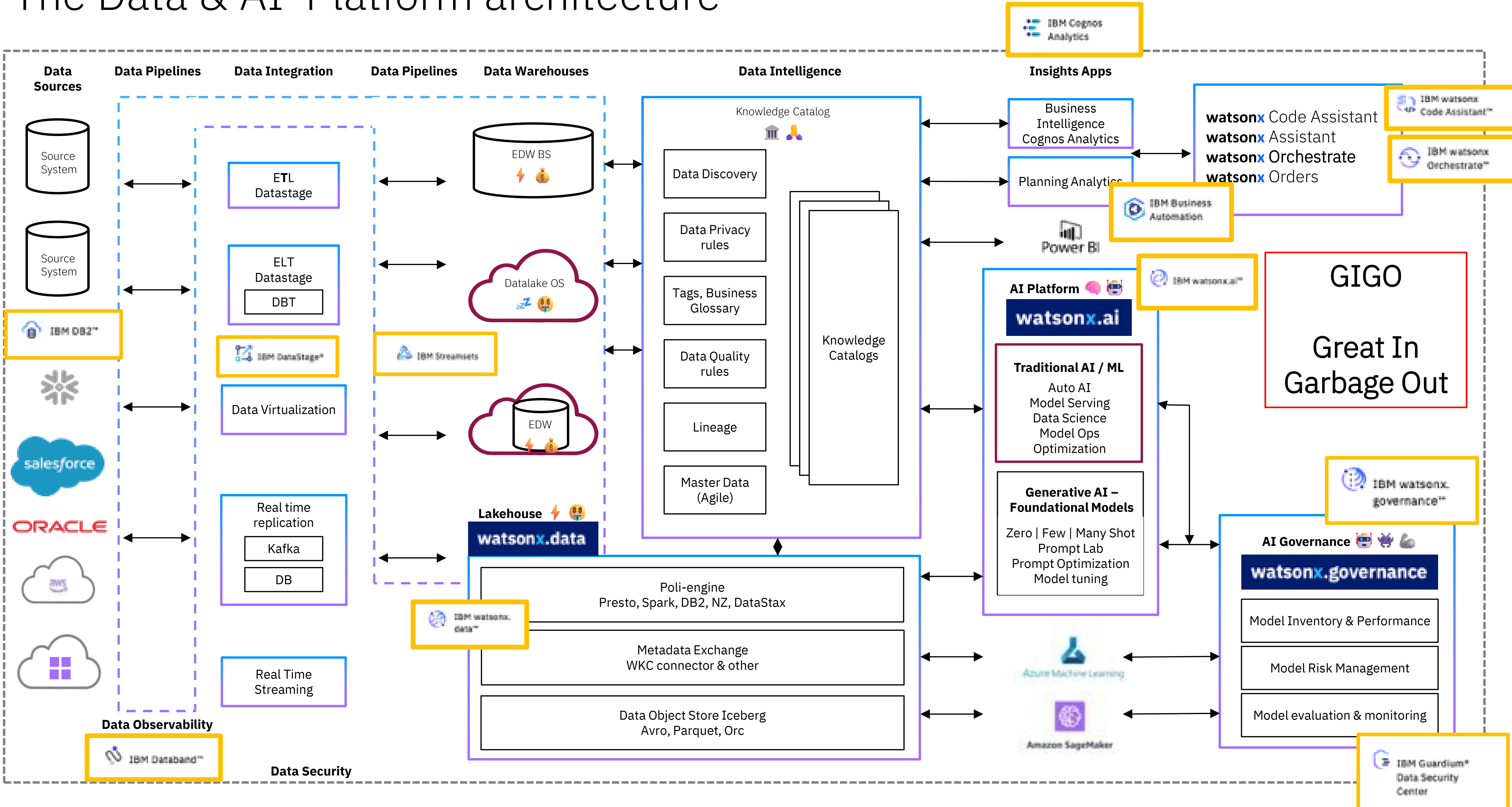
Undesired Results

AI Governance

AI Security

https://food.ec.europa.eu/horizontal-topics/farm-fork-strategy_en

The Data & AI Platform architecture



IBM Cognos Analytics

Data Sources

Source System

Source System

IBM DB2™

salesforce

ORACLE

aws

Data Pipelines

Data Integration

ETL Datastage

ELT Datastage

DBT

IBM DataStage®

Data Virtualization

Real time replication

Kafka

DB

Real Time Streaming

Data Observability

IBM Databand™

Data Pipelines

IBM Streamsets

Data Warehouses

EDW BS

Datalake OS

EDW

Lakehouse

watsonx.data

IBM watsonx.data™

Data Intelligence

Knowledge Catalog

Data Discovery

Data Privacy rules

Tags, Business Glossary

Data Quality rules

Lineage

Master Data (Agile)

Knowledge Catalogs

Poli-engine
Presto, Spark, DB2, NZ, DataStax

Metadata Exchange
WKC connector & other

Data Object Store Iceberg
Avro, Parquet, Orc

Data Security

Insights Apps

Business Intelligence Cognos Analytics

Planning Analytics

IBM Business Automation

Power BI

watsonx Code Assistant
watsonx Assistant
watsonx Orchestrate
watsonx Orders

IBM watsonx Code Assistant™

IBM watsonx Orchestrate™

AI Platform

watsonx.ai

IBM watsonx.ai™

Traditional AI / ML

Auto AI
Model Serving
Data Science
Model Ops
Optimization

Generative AI – Foundational Models

Zero | Few | Many Shot
Prompt Lab
Prompt Optimization
Model tuning

Azure Machine Learning

Amazon SageMaker

GIGO

Great In Garbage Out

IBM watsonx.governance™

AI Governance

watsonx.governance

Model Inventory & Performance

Model Risk Management

Model evaluation & monitoring

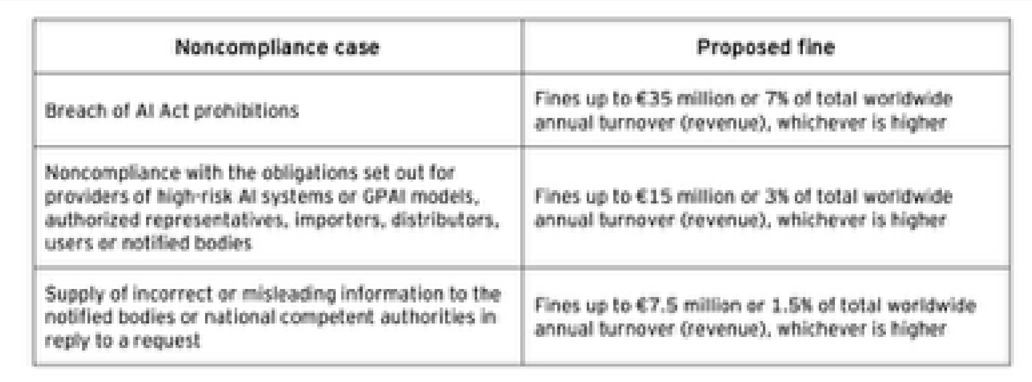
IBM Guardium® Data Security Center

The reasons we MUST have Trustworthy AI

The 4 Rs

Regulatory Fines

EU Artificial Intelligence Act



Noncompliance case	Proposed fine
Breach of AI Act prohibitions	Fines up to €35 million or 7% of total worldwide annual turnover (revenue), whichever is higher
Noncompliance with the obligations set out for providers of high-risk AI systems or GPAI models, authorized representatives, importers, distributors, users or notified bodies	Fines up to €15 million or 3% of total worldwide annual turnover (revenue), whichever is higher
Supply of incorrect or misleading information to the notified bodies or national competent authorities in reply to a request	Fines up to €7.5 million or 1.5% of total worldwide annual turnover (revenue), whichever is higher

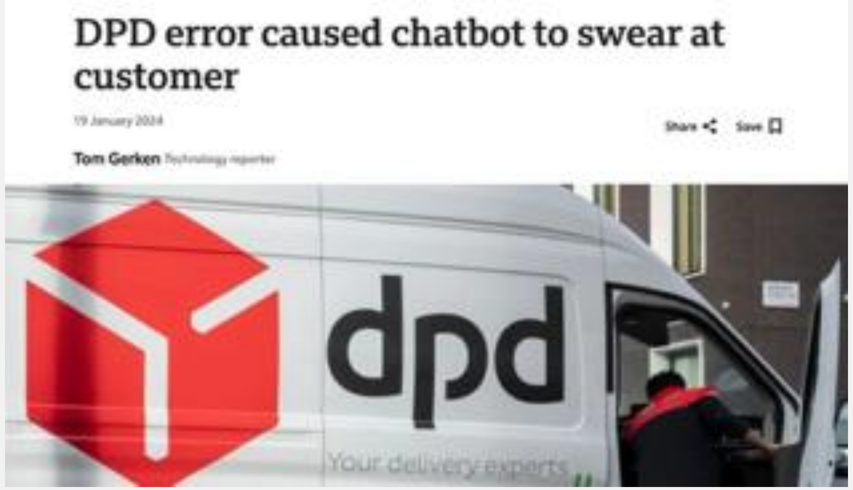
https://artificialintelligenceact.eu/

The reasons we MUST have Trustworthy AI

The 4 Rs

Regulatory Fines

Reputational Damage ⟷



DPD error caused chatbot to swear at customer

19 January 2024

Share Save

Tom Gerken Technology reporter

The customer then posted the chat, which had gone viral with 1.3 million views and over 20 thousand likes.

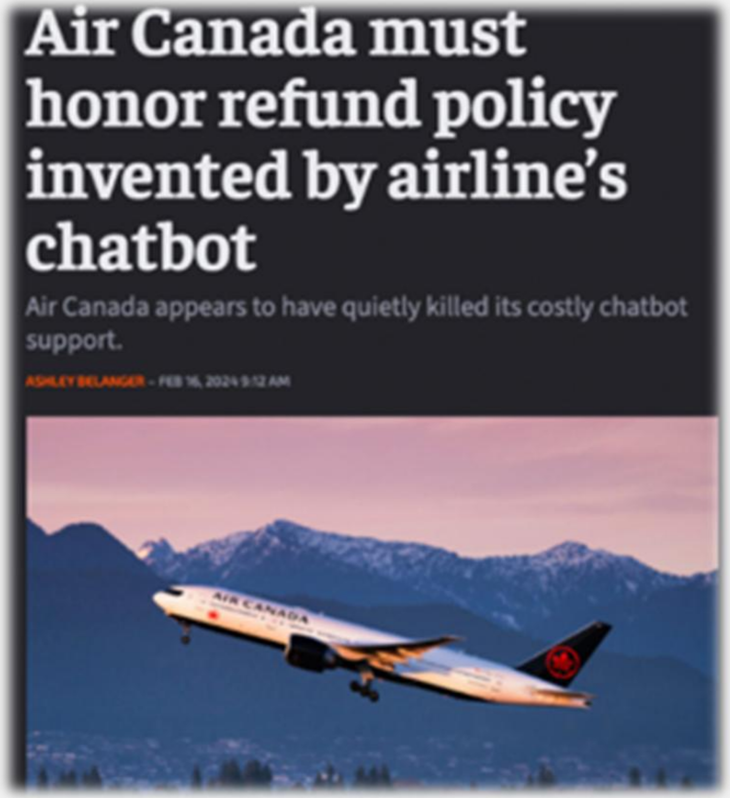
The reasons we MUST have Trustworthy AI

The 4 Rs

Regulatory Fines

Reputational Damage

Revenue Loss
(primary / secondary)



Air Canada must honor refund policy invented by airline's chatbot

Air Canada appears to have quietly killed its costly chatbot support.

ASHLEY BELANGER – FEB 16, 2024 9:12 AM

The reasons we MUST have Trustworthy AI

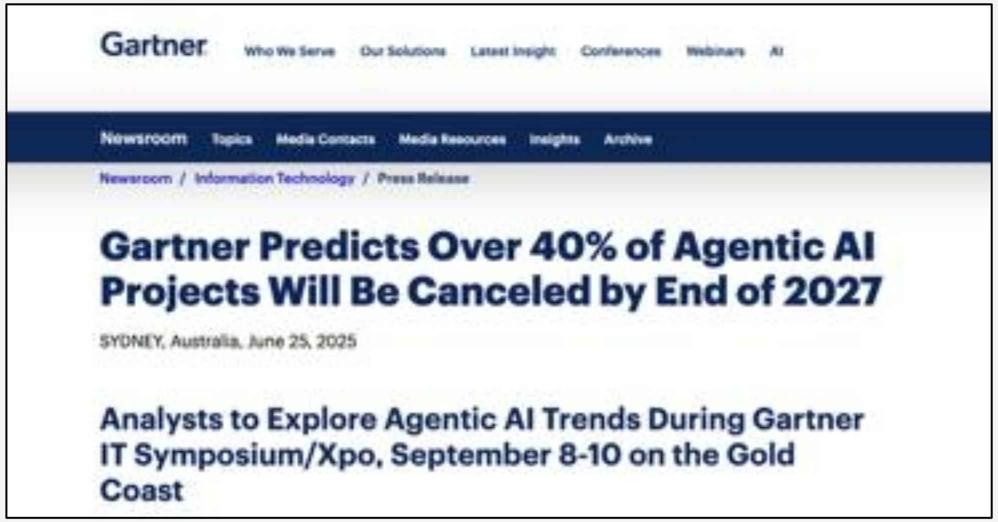
The 4 Rs

Regulatory Fines

Reputational Damage

Revenue Loss
(primary / secondary)

Running Operations Impacts



Gartner Who We Serve Our Solutions Latest Insight Conferences Webinars AI

Newsroom Topics Media Contacts Media Resources Insights Archive

Newsroom / Information Technology / Press Release

Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027

SYDNEY, Australia, June 25, 2025

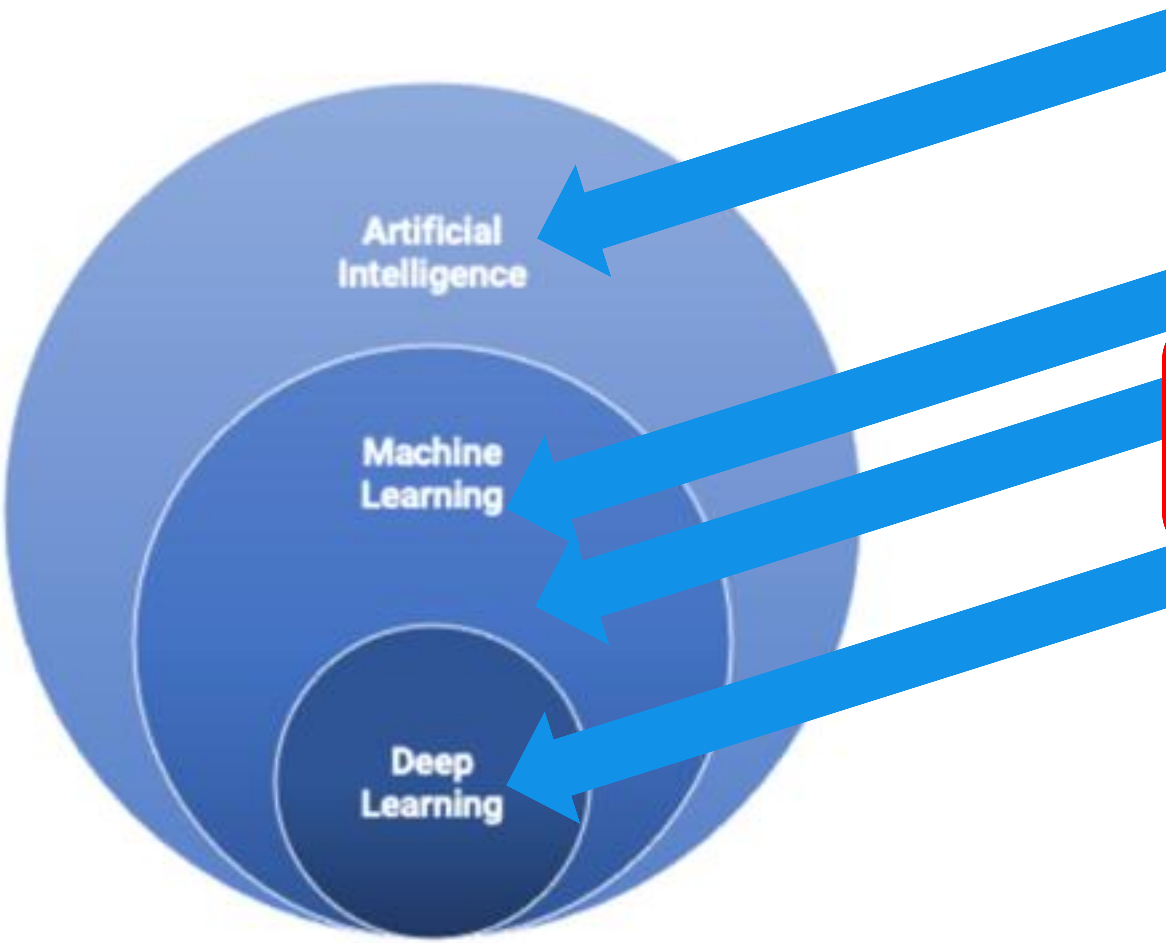
Analysts to Explore Agentic AI Trends During Gartner IT Symposium/Xpo, September 8-10 on the Gold Coast

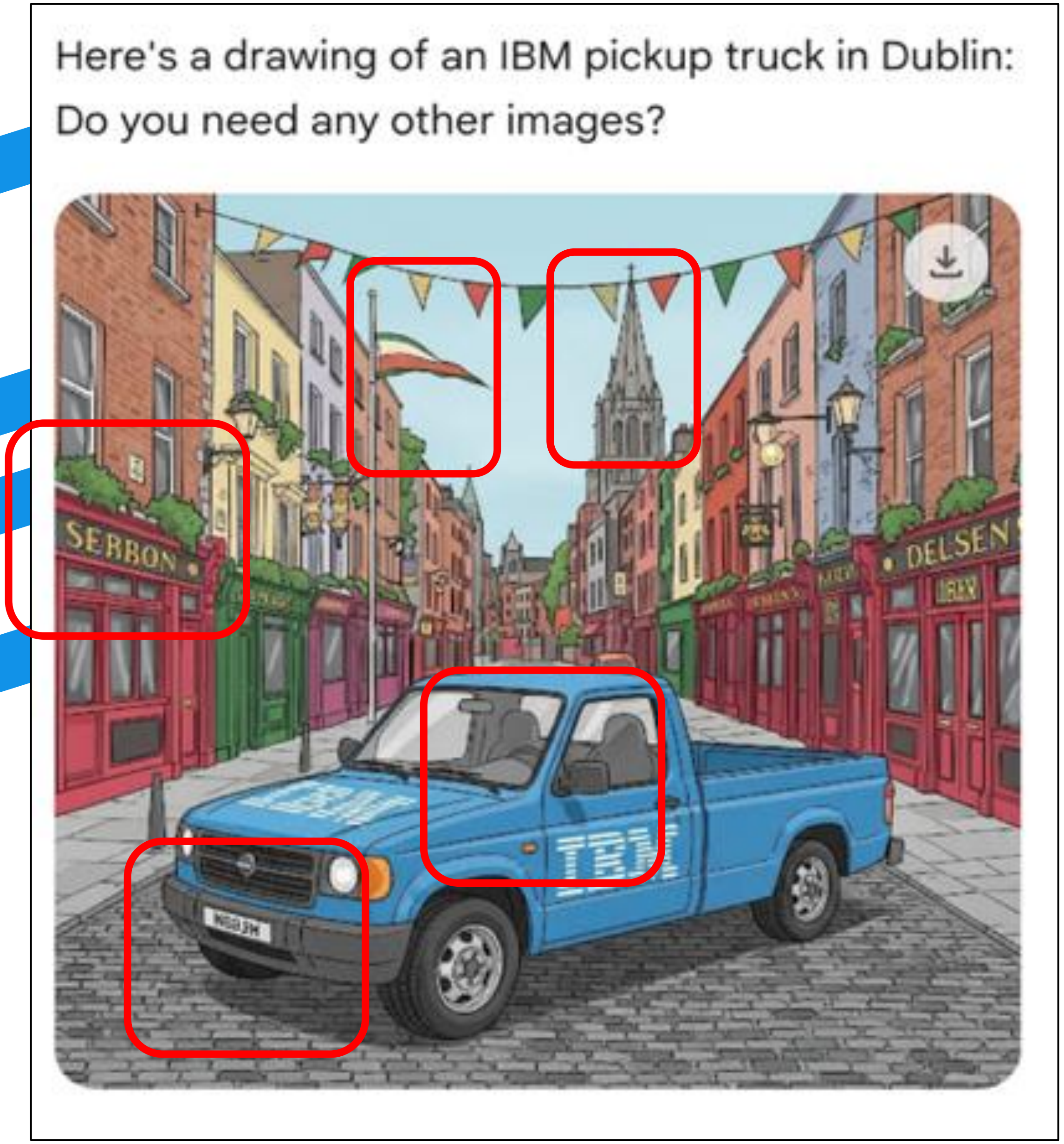
due to:-

- escalating costs
- unclear business value
- inadequate risk controls

https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

Evolution of AI and its vulnerabilities



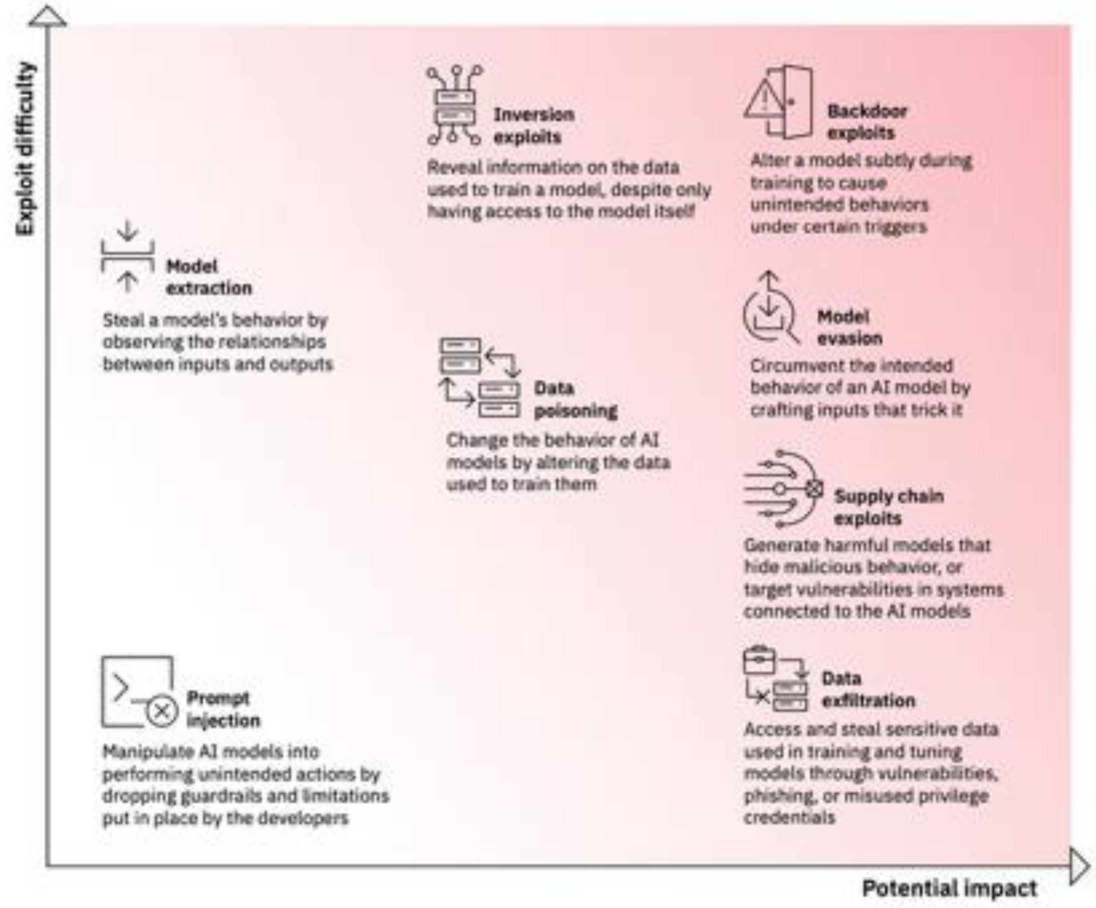


Here's a drawing of an IBM pickup truck in Dublin:

Do you need any other images?

Evolution of AI and its vulnerabilities

- Accountability
- Accuracy
- Fairness
- Veracity
- Transparency
- Drift
- Trusted data
- Energy consumption
- Explainability
- Adversarial Robustness
- IP/PII leakage
- …



Model extraction
Steal a model's behavior by observing the relationships between inputs and outputs

Inversion exploits
Reveal information on the data used to train a model, despite only having access to the model itself

Backdoor exploits
Alter a model subtly during training to cause unintended behaviors under certain triggers

Data poisoning
Change the behavior of AI models by altering the data used to train them

Model evasion
Circumvent the intended behavior of an AI model by crafting inputs that trick it

Supply chain exploits
Generate harmful models that hide malicious behavior, or target vulnerabilities in systems connected to the AI models

Prompt injection
Manipulate AI models into performing unintended actions by dropping guardrails and limitations put in place by the developers

Data exfiltration
Access and steal sensitive data used in training and tuning models through vulnerabilities, phishing, or misused privilege credentials

IBM watsonx.governance™

IBM Guardium® Data Security Center

How AI can misbehave

(these are most common attributes when AI misbehaves)

Passive Anoma

(accidental)

- Quality A ance Miss
  - Clumsy & L
  - No Mali ent
  - Rem needs no escalation
  - F t cause Analysis
  - ainable Impacts

AI Governance (in->out)

Active Anoma

(coerced)

- Target ystem abus
  - Focu overt
  - N Intent
  - diation needs escalation due to er security implications
  - Extensive Root cause Analysis needed
  - Significant Impact

AI Security (in<-out)

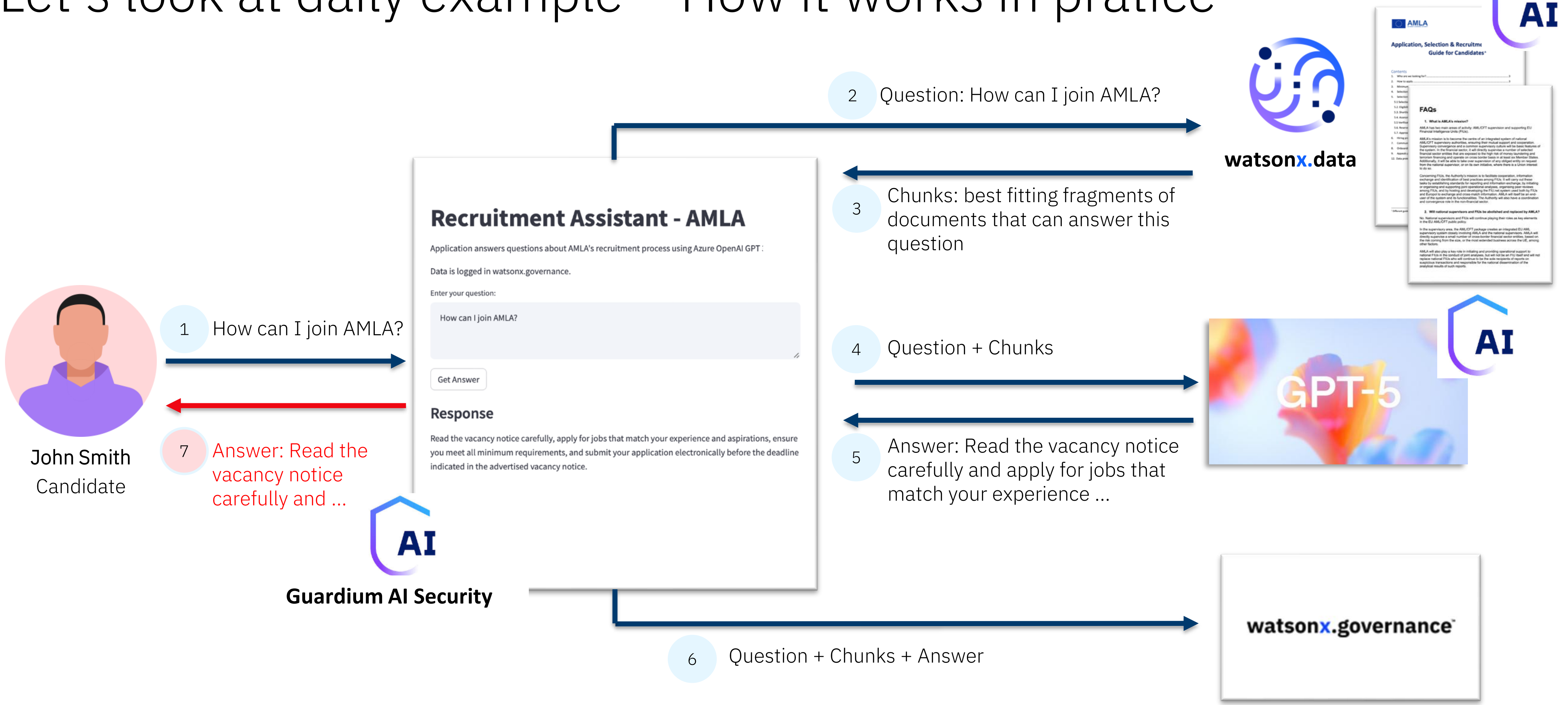
Trustworthy AI

Better Together

Guardium AI Security + watsonx.governance – Integrated value

Capability	Guardium AI Security	Watsonx.gov	Integrated Value
AI Discovery	Automatically detects Shadow AI and inventories ML/LLM/foundation models/Ai Software/AI Services/AI Agents and more	Tracks and manages known and registered AI assets in the model catalog.	**Unified, automated and continuously up-to-date AI catalog**. Complete visibility into all AI systems and assured governance without relying on human discretion.
Risk & Compliance	Scans for AI policy & compliance violations and unauthorized behaviors and automates evidence collection.	Maps policies to regulatory frameworks (e.g., NIST, EU AI Act)	**Proactive risk and compliance management** – Continuous compliance validation and streamlined audit readiness.
AI Posture Management	Security Posture Management (SPM) - Scans for data, model, and usage vulnerabilities, misconfigurations, threat intel and provides remediation.	Responsible AI Posture Management - Monitors for bias, drift, fairness, and performance.	**Secure & Responsible AI -** Address ALL technical and ethical risks for AI system for all stakeholders involves with aspects of risk management.
Lifecycle Management	Pen-tests models, LLMs and AI systems and analyzes risk for safe third-party embedded AI onboarding.	Governs model lifecycle through approvals and version control.	**Secure and Governed AI Lifecycle Management -** Actively reduce risk and elevate safety in all AI deployments.
Operational Integration	Runtime Controls - runtime protection and AI access control (e.g., prompt injection attacks, PII protection and access to LLMs)	Facilitates collaboration across product, risk, and compliance teams.	**Integrated Operations -** Operationalize safety and security across product, risk, compliance, and security stakeholders.
AI Policy Management & Enforcement	End-to-end security policy management and enforcement through all stages of the lifecycle (development through production).	Centralized organizational policy management and controls for governing the AI use case and AI lifecycle.	**Policy definition, enforcement, and monitoring** for all stakeholders involved with AI systems in their respective “language” (Security / compliance), unified to complete a complete AI lifecycle policy.

Let’s look at daily example – How it works in pratice



Recruitment Assistant - AMLA

Application answers questions about AMLA's recruitment process using Azure OpenAI GPT :

Data is logged in watsonx.governance.

Enter your question:

How can I join AMLA?

Get Answer

Response

Read the vacancy notice carefully, apply for jobs that match your experience and aspirations, ensure you meet all minimum requirements, and submit your application electronically before the deadline indicated in the advertised vacancy notice.

1 How can I join AMLA?

2 Question: How can I join AMLA?

3 Chunks: best fitting fragments of documents that can answer this question

4 Question + Chunks

5 Answer: Read the vacancy notice carefully and apply for jobs that match your experience ...

6 Question + Chunks + Answer

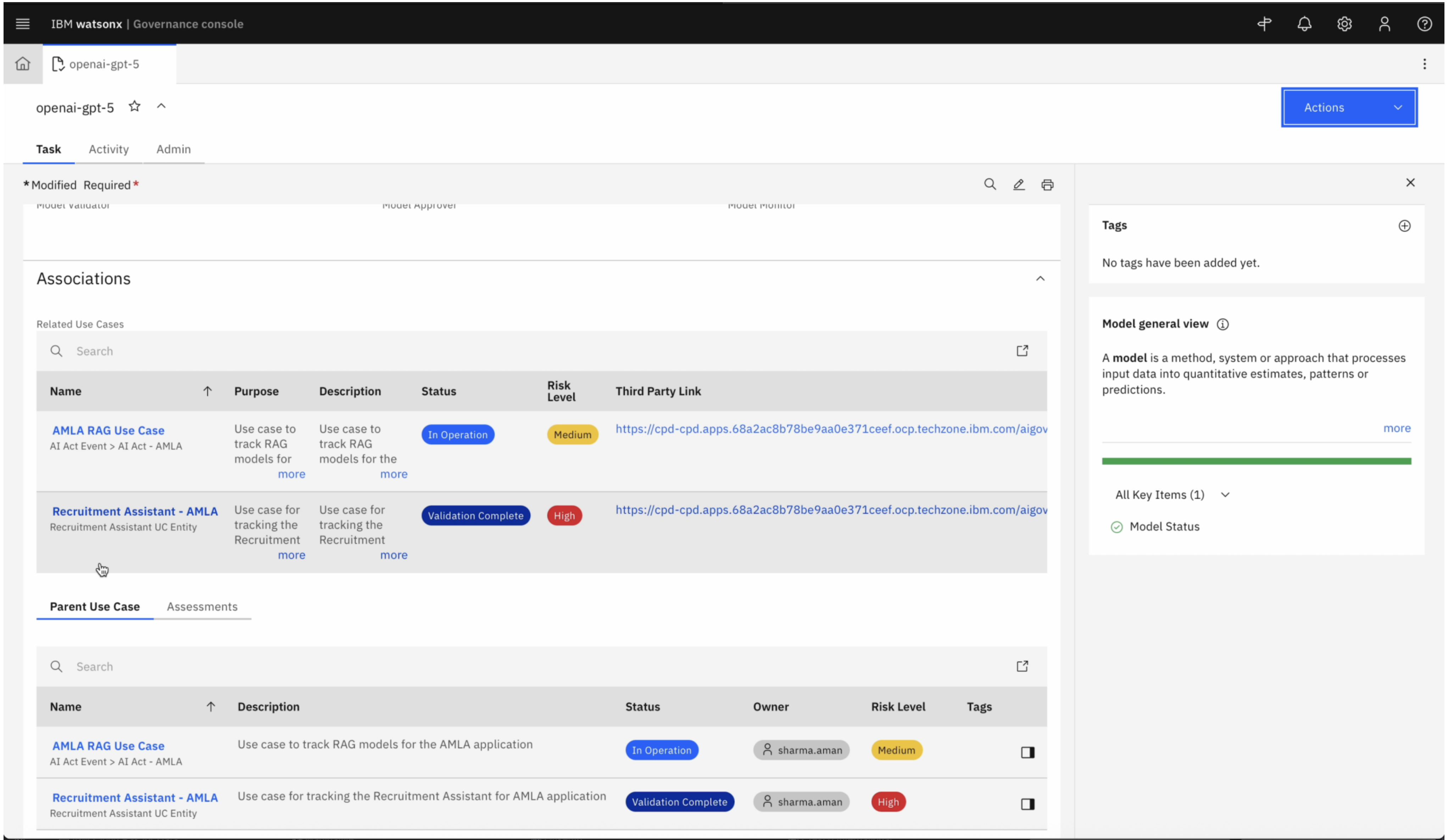
7 Answer: Read the vacancy notice carefully and ...

True or False?

For onboarding an external AI model, such as GPT-5, it is enough to depend only on the vendor’s documentation.

False

vendor docs matter, but the organization must adapt and govern use



Modified Required

Model Validator Model Approver Model Monitor

Associations

Related Use Cases

Name	Purpose	Description	Status	Risk Level	Third Party Link
AMLA RAG Use Case AI Act Event > AI Act - AMLA	Use case to track RAG models for more	Use case to track RAG models for the more	In Operation	Medium	https://cpd-cpd.apps.68a2ac8b78be9aa0e371ceef.ocp.techzone.ibm.com/aigov
Recruitment Assistant - AMLA Recruitment Assistant UC Entity	Use case for tracking the Recruitment more	Use case for tracking the Recruitment more	Validation Complete	High	https://cpd-cpd.apps.68a2ac8b78be9aa0e371ceef.ocp.techzone.ibm.com/aigov

Name	Description	Status	Owner	Risk Level	Tags
AMLA RAG Use Case AI Act Event > AI Act - AMLA	Use case to track RAG models for the AMLA application	In Operation	sharma.aman	Medium	
Recruitment Assistant - AMLA Recruitment Assistant UC Entity	Use case for tracking the Recruitment Assistant for AMLA application	Validation Complete	sharma.aman	High	

Tags

No tags have been added yet.

Model general view

A **model** is a method, system or approach that processes input data into quantitative estimates, patterns or predictions.

more

All Key Items (1)

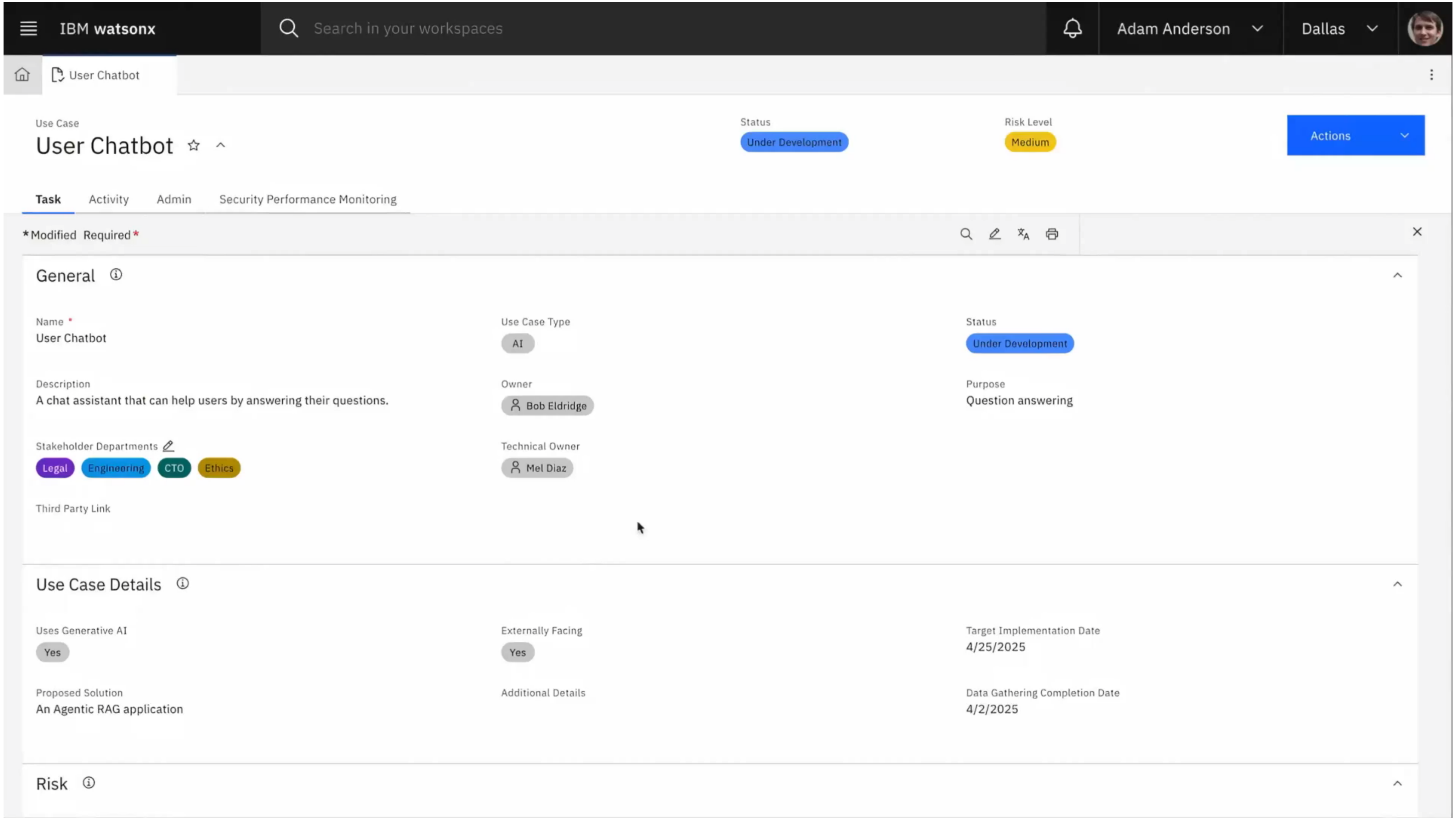
✓ Model Status

True or False?

Risk mapping in AI governance helps organizations connect compliance requirements with potential technical and ethical risks.

True

An AI risk score helps organizations evaluate both ethical and technical risks in AI systems.



Use Case
User Chatbot

Status: Under Development

Risk Level: Medium

General

Name *: User Chatbot

Use Case Type: AI

Status: Under Development

Description: A chat assistant that can help users by answering their questions.

Owner: Bob Eldridge

Purpose: Question answering

Stakeholder Departments: Legal, Engineering, CTO, Ethics

Technical Owner: Mel Diaz

Third Party Link

Use Case Details

Uses Generative AI: Yes

Externally Facing: Yes

Target Implementation Date: 4/25/2025

Proposed Solution: An Agentic RAG application

Additional Details

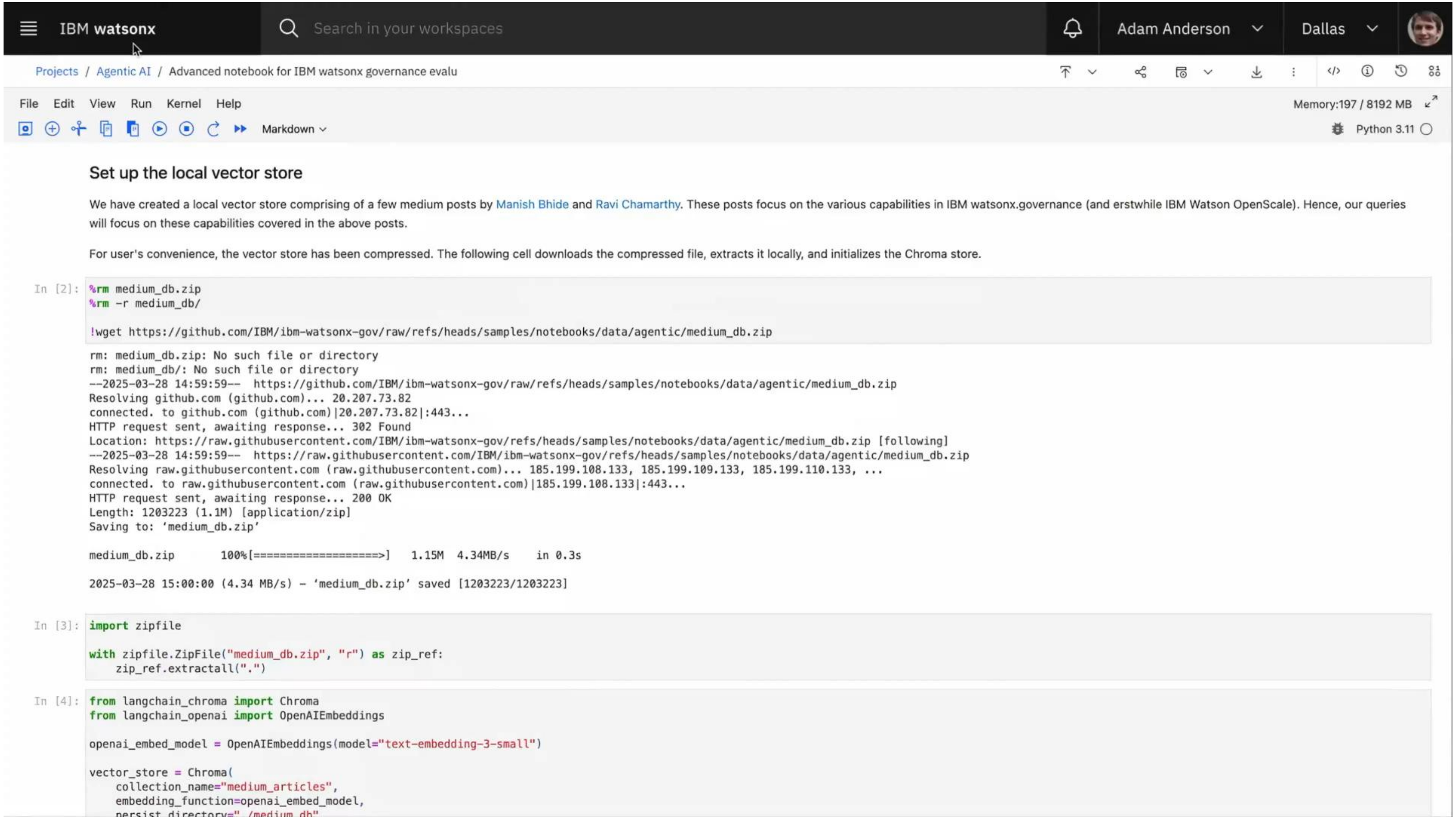
Data Gathering Completion Date: 4/2/2025

Risk

True or False?

Technical documentation is more than an AI Act requirement – it is essential for system transparency, risk assessment, and safe operations.

True it supports explainability, accountability, and safe AI governance.


IBM watsonx
Search in your workspaces
Adam Anderson
Dallas
Projects / Agentic AI / Advanced notebook for IBM watsonx governance evalu
File
Edit
View
Run
Kernel
Help
Memory:197 / 8192 MB
Markdown
Python 3.11

Set up the local vector store

We have created a local vector store comprising of a few medium posts by Manish Bhide and Ravi Chamarthy. These posts focus on the various capabilities in IBM watsonx.governance (and erstwhile IBM Watson OpenScale). Hence, our queries will focus on these capabilities covered in the above posts.

For user's convenience, the vector store has been compressed. The following cell downloads the compressed file, extracts it locally, and initializes the Chroma store.

```
In [2]: %rm medium_db.zip
        %rm -r medium_db/

        !wget https://github.com/IBM/ibm-watsonx-gov/raw/refs/heads/samples/notebooks/data/agentic/medium_db.zip
```

```
rm: medium_db.zip: No such file or directory
rm: medium_db/: No such file or directory
--2025-03-28 14:59:59--  https://github.com/IBM/ibm-watsonx-gov/raw/refs/heads/samples/notebooks/data/agentic/medium_db.zip
Resolving github.com (github.com)... 20.207.73.82
connected. to github.com (github.com)|20.207.73.82|:443...
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/IBM/ibm-watsonx-gov/refs/heads/samples/notebooks/data/agentic/medium_db.zip [following]
--2025-03-28 14:59:59--  https://raw.githubusercontent.com/IBM/ibm-watsonx-gov/refs/heads/samples/notebooks/data/agentic/medium_db.zip
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.110.133, ...
connected. to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443...
HTTP request sent, awaiting response... 200 OK
Length: 1203223 (1.1M) [application/zip]
Saving to: ‘medium_db.zip’

medium_db.zip       100%[===================>]   1.15M  4.34MB/s    in 0.3s

2025-03-28 15:00:00 (4.34 MB/s) - ‘medium_db.zip’ saved [1203223/1203223]
```

```
In [3]: import zipfile

        with zipfile.ZipFile("medium_db.zip", "r") as zip_ref:
            zip_ref.extractall(".")
```

```
In [4]: from langchain_chroma import Chroma
        from langchain_openai import OpenAIEmbeddings

        openai_embed_model = OpenAIEmbeddings(model="text-embedding-3-small")

        vector_store = Chroma(
            collection_name="medium_articles",
            embedding_function=openai_embed_model,
            persist_directory="./medium_db"
```

The General-Purpose AI (GPAI)

Code of Practice

- The 3 chapters of the code - **Transparency, Copyright,** and **Safety** and **Security**.
- The Chapters on Transparency and Copyright offer all providers of general-purpose AI models a way to demonstrate compliance with their obligations under Article 53 AI Act.
- The Chapters on Safety and Security is only relevant to the small number of providers of the most advanced models, those that are subject to the AI Act's obligations for providers of general-purpose AI models with systemic risk under Article 55 AI Act.

Signatories of Code of Practice (*):

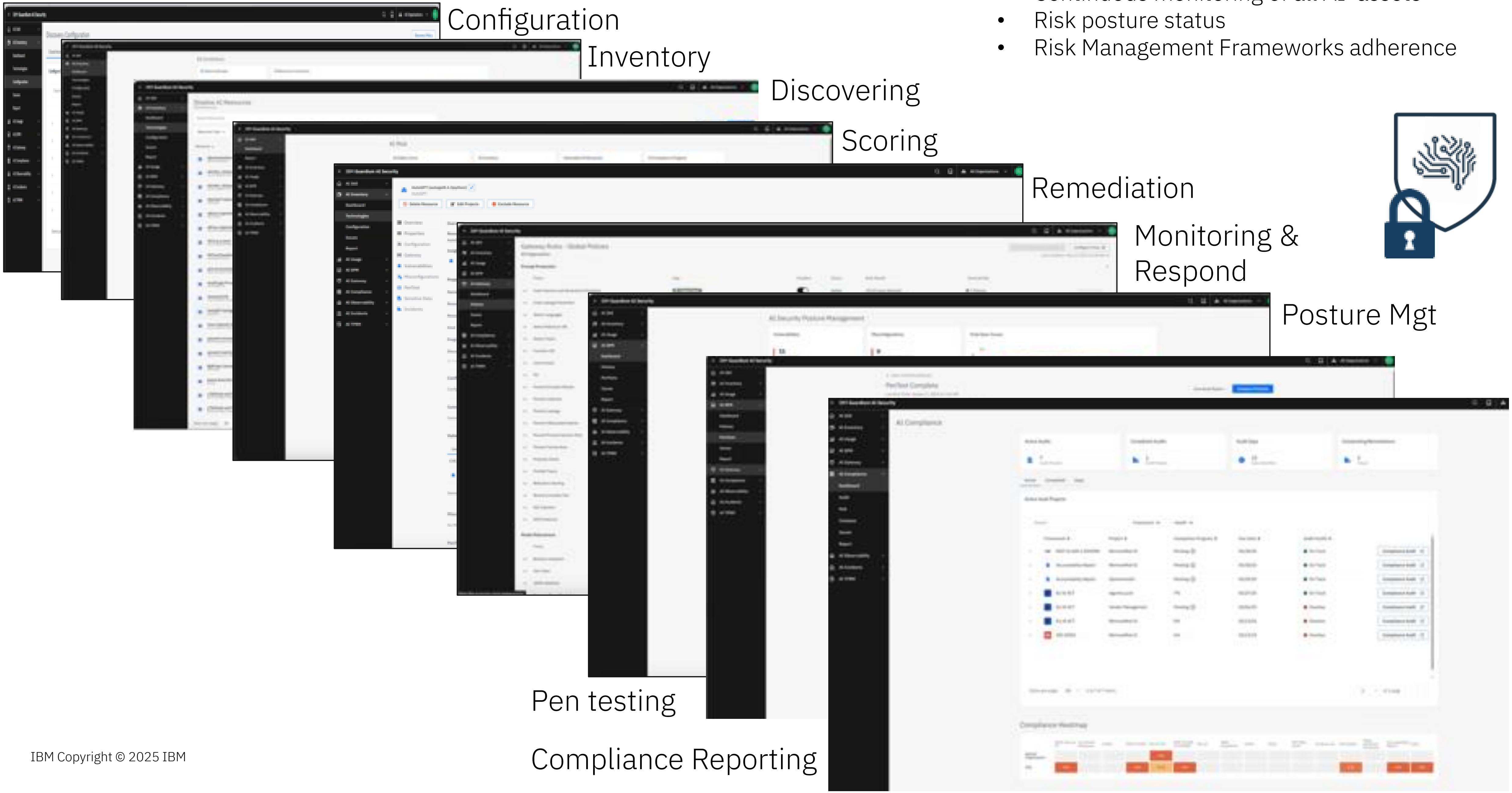
Accexible, AI Alignment Solutions, Aleph Alpha, Almawave, Amazon, Anthropic, Bria AI, Cohere, Cyber Institute, Domyn, Dweve, Euc Inovação Portugal, Fastweb, Google, Humane Technology, IBM, Lawise, Microsoft, Mistral AI, Open Hippo, OpenAI, Pleias, re-inventa, ServiceNow, Virtuo Turing, WRITER, In addition, xAI signed up to the Safety and Security Chapter; this means that it will have to demonstrate compliance with the AI Act's obligations concerning transparency and copyright via alternative adequate means.

European Commission

** Some signatories may not appear immediately, but we are making sure to continuously update the list as signatures are confirmed.*

IBM AI Security full end to end Solution's features

- Automated responses
- Continuous monitoring of all AI assets
- Risk posture status
- Risk Management Frameworks adherence



Configuration

Inventory

Discovering

Scoring

Remediation

Monitoring & Respond

Posture Mgt

Pen testing

Compliance Reporting

Take Away Points

New risk but the same flaws & threat actors

The need is the same , >>>> mitigate against the 4 Rs

AI for business strategy can’t succeed without AI governance + AI security (Trustworthy AI)

AI will generate many new opportunities for commercial success but also new Malicious opportunities

Start with the end in mind .... design Ai + Governance + Ai Security as a single project

Seems overwhelming ? No problem IBM Can help