

sogeti  
Part of Capgemini 



## Combining Human & Machine for more Sustainable AI Technology

**Pierre-Olivier PATIN**

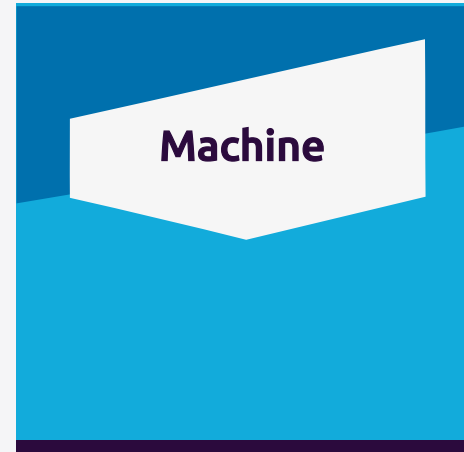
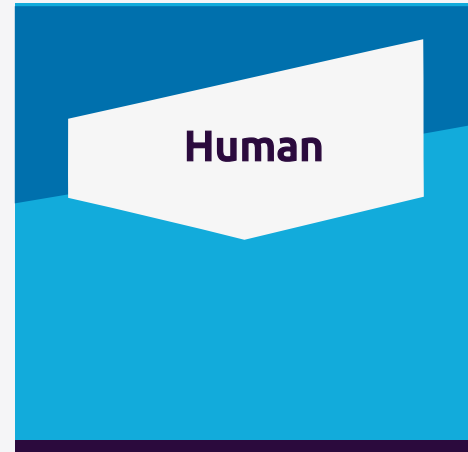
CTO, Applications & Cloud Technologies  
@Sogeti

**Paul LASSERRE**

Head of GenAI partnerships  
@AWS

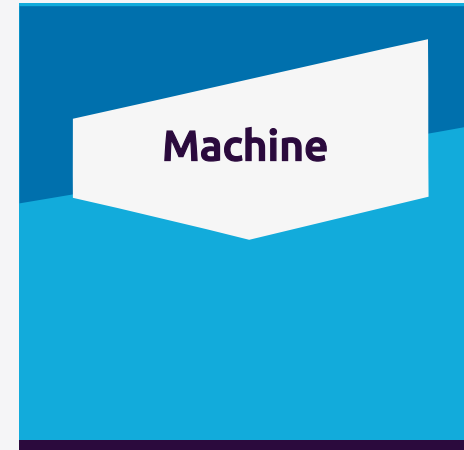


# Introduction



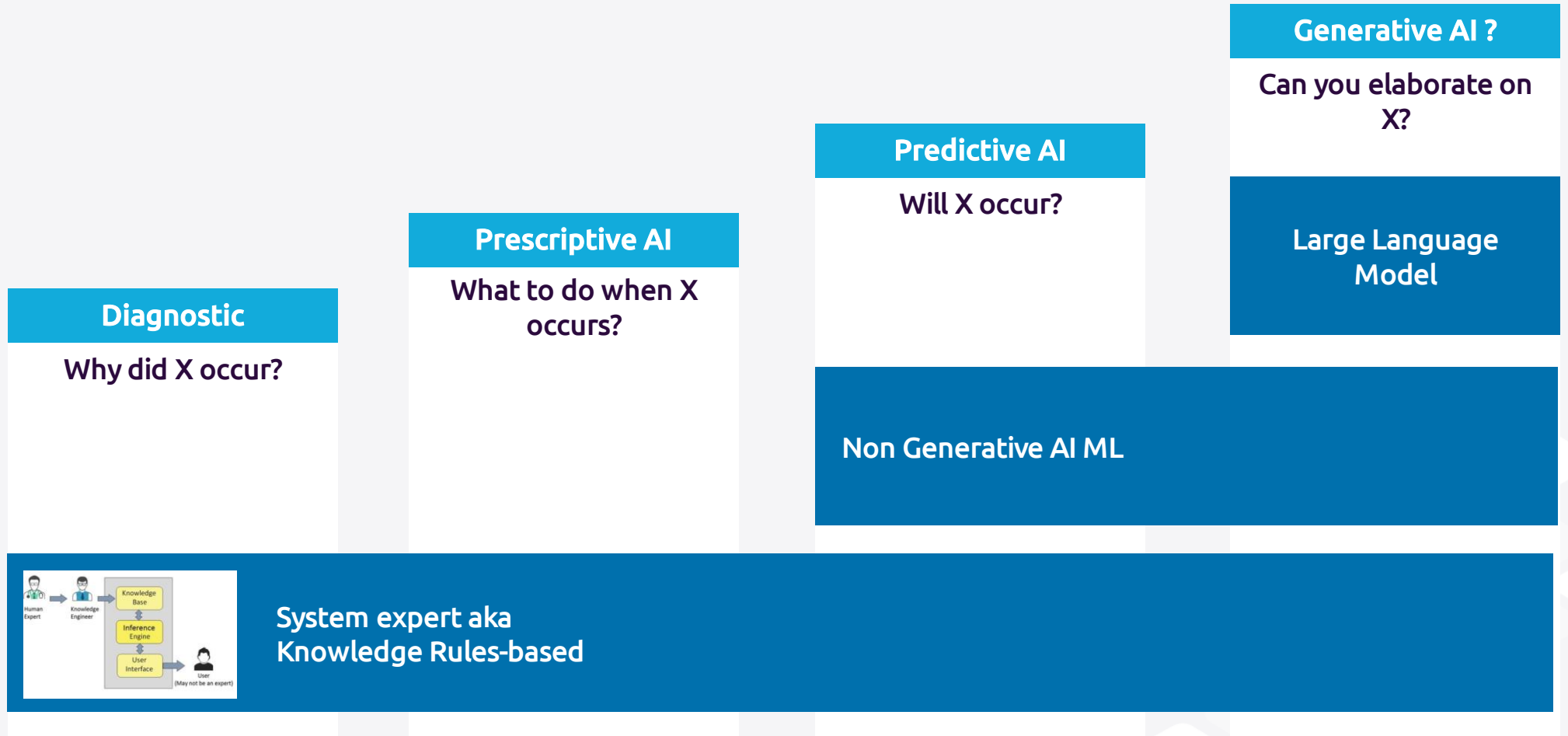
**Intellectual capability of humans, which is marked by complex cognitive feats and high levels of motivation and self-awareness**

# Introduction



**Computers are programmed to “mimic” human behavior using extensive data from past examples of similar behavior**

# Evolution of machine intelligence

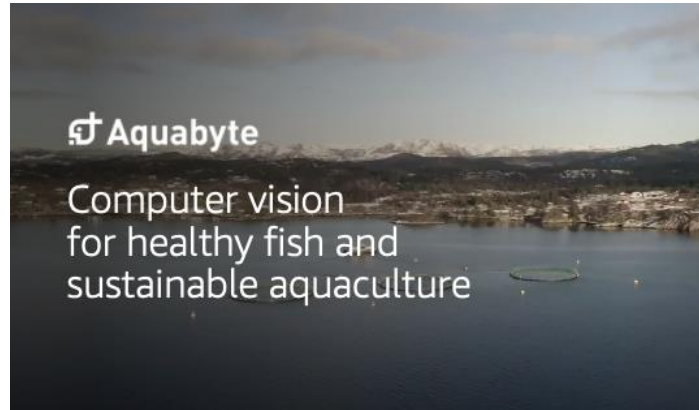


# But it's already started!

## Precision AI



## Aquabyte



## Brainbox

**BRAINBOX AI**

### Scaling Sustainability Solutions for Buildings Using AWS with BrainBox AI


BrainBox AI scaled its autonomous energy management solution to new regions using AWS, reducing the carbon emissions of the buildings that it is installed in by up to 40 percent.

[Overview](#) | [Opportunity](#) | [Solution](#) | [Outcome](#) | [AWS Services Used](#)

<b>Up to 40%</b> reduction in building HVAC emissions	<b>Up to 25%</b> reduction in HVAC energy costs	<b>Scaled out to 20</b> countries	<b>Manages hundreds of</b> buildings 24/7
--	--	--------------------------------------	--

**Overview**

Canadian technology scale-up **BrainBox AI** is helping building owners reduce emissions and energy consumption using cloud-based artificial intelligence (AI) and machine learning (ML) on Amazon Web Services (AWS). Using AWS, BrainBox AI can deliver deep learning solutions with low latency to multiple regions and scale quickly to meet the demand for a growing number of building owners who want to reduce their emissions.



**OPPORTUNITY** 15% of the world's carbon emissions stem from heating and cooling our buildings



# Intelligent apps

« GenAI will reveal its potential in Apps »

## Interactive as new normal

- Interact with Natural Language
- Integrate functionality in chat & collaborative apps
- Immersive experience with XR (VR, MR)

## Shift of usage

- From Apps to Copilots and extensions
- Agents, Multi-agents and Autonomous agents
- Query the whitespace (unaddressed areas that could lead to new growth)

## Expansion of intents

- Conversations & Agents that drive actions
- Workflows to chain intents across apps, agents and data referential

## Derivative Models

- Multi-models consumptions
- Specialized & Micro models
- Customized models with RAG
- Private / hosted models

## App mesh

- Tied Data & App Integration
- “App mesh” required for instant interactions

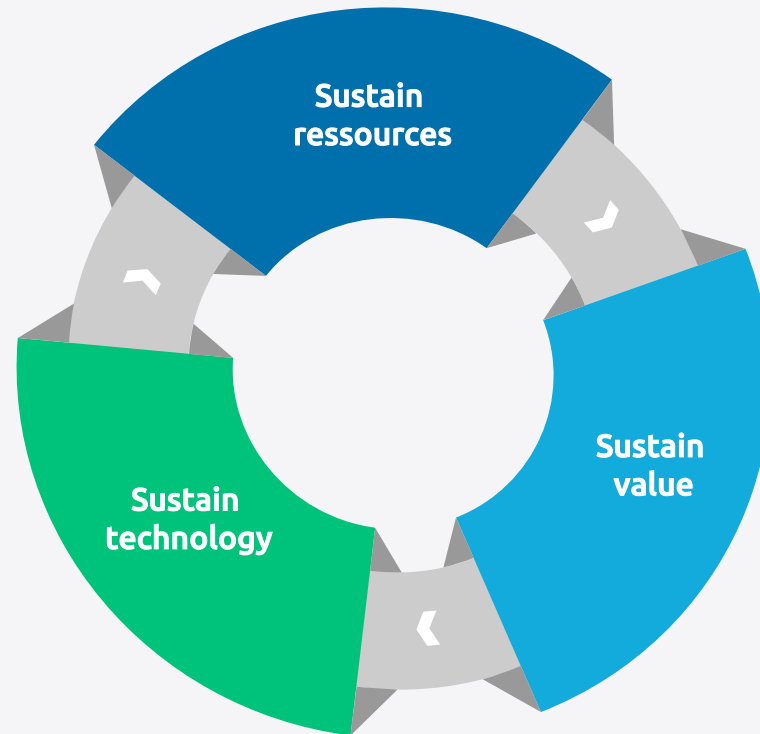


# It's time to combine now!

How Combining Human and Machine for more Sustainable AI Technology?

## Sustain technology

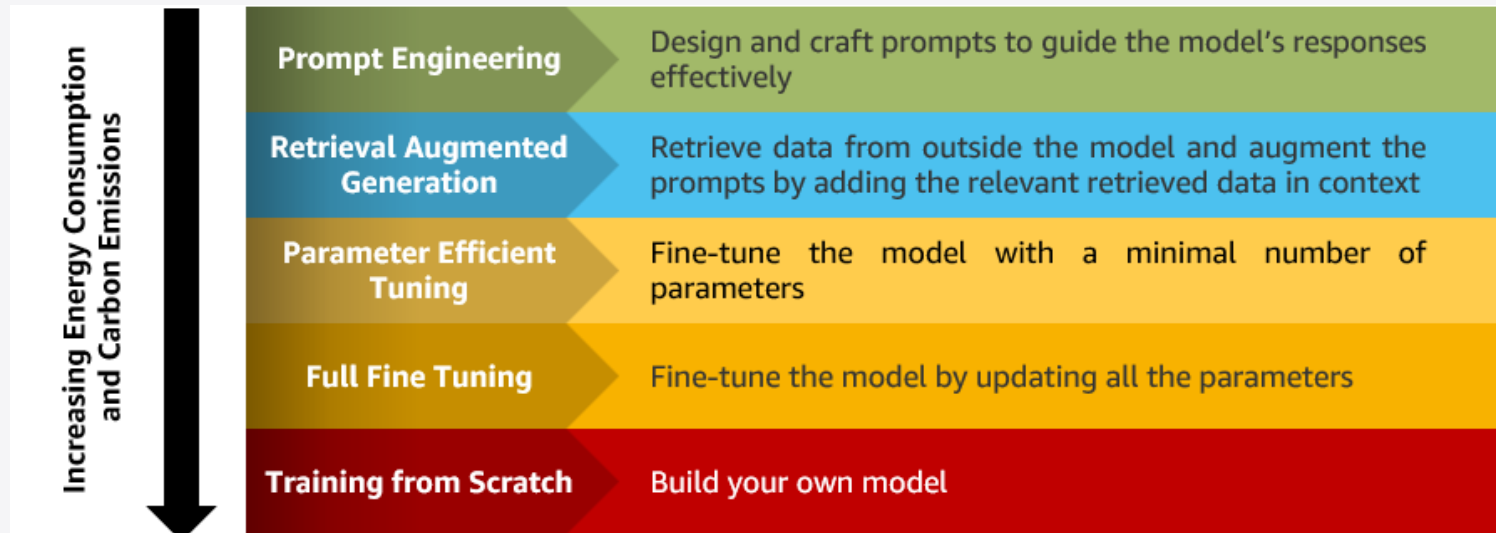
- Energy Efficiency
- Data Center Sustainability
- Responsible Resource Management



## Sustain value

- Responsible AI
- Accuracy (Analyze, Plan, Solve, Review)
- Value creation vs. value capture

# Optimize GenAI workloads for environmental sustainability



1. Align your use of generative AI with your sustainability goals
2. Use energy that has low carbon-intensity
3. Use managed services
4. Define the right customization strategy

# Optimize LLM usage with smart prompting

## Prompt engineering

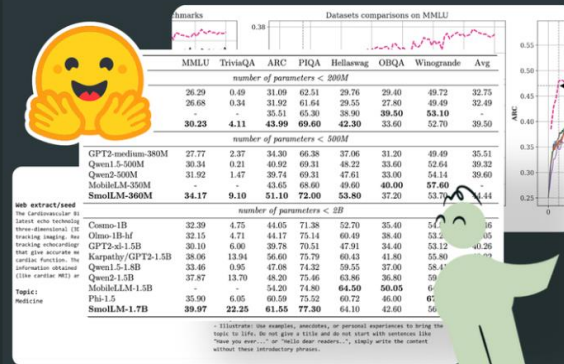
Effective [prompt engineering](#) can enhance the performance and efficiency of generative AI models. By carefully crafting prompts, you can guide the model's behavior, reducing unnecessary iterations and resource requirements. Consider the following guidelines:

- **Keep prompts concise and avoid unnecessary details** – Longer prompts lead to a higher number of tokens. As tokens increase in number, the model consumes more memory and computational resources. Consider incorporating [zero-shot](#) or [few-shot](#) learning to enable the model to adapt quickly by learning from just a few examples.
- **Experiment with different prompts gradually** – Refine the prompts based on the desired output until you achieve the desired results. Depending on your task, [explore advanced techniques](#) such as [self-consistency](#), [Generated Knowledge Prompting](#), [ReAct Prompting](#), or [Automatic Prompt Engineer](#) to further enhance the model's capabilities.
- **Use reproducible prompts** – With templates such as [LangChain prompt templates](#), you can save or load your prompts history as files. This enhances prompt experimentation tracking, versioning, and reusability. When you know the prompts that produce the best answers for each model, you can reduce the computational resources used for prompt iterations and redundant experiments across different projects.

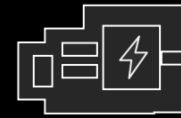
# Small is the new frontier for models while cost performance of HW is a strategic imperative

## What is SmolLM? A Guide to Hugging face's small language model

Explore SmolLM, a compact yet powerful language model challenging the notion that bigger is always better in AI. Learn how its meticulously curated datasets and efficient design deliver high performance with lower resource demands, making it ideal for applications in education, coding, and customer support.

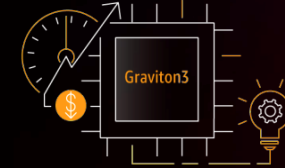


## Journey of silicon innovation at AWS



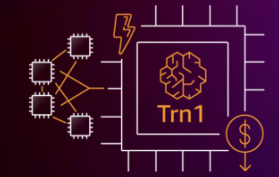
### AWS Nitro System

Hypervisor, Nitro Cards, network, storage, SSD, and security



### AWS Graviton

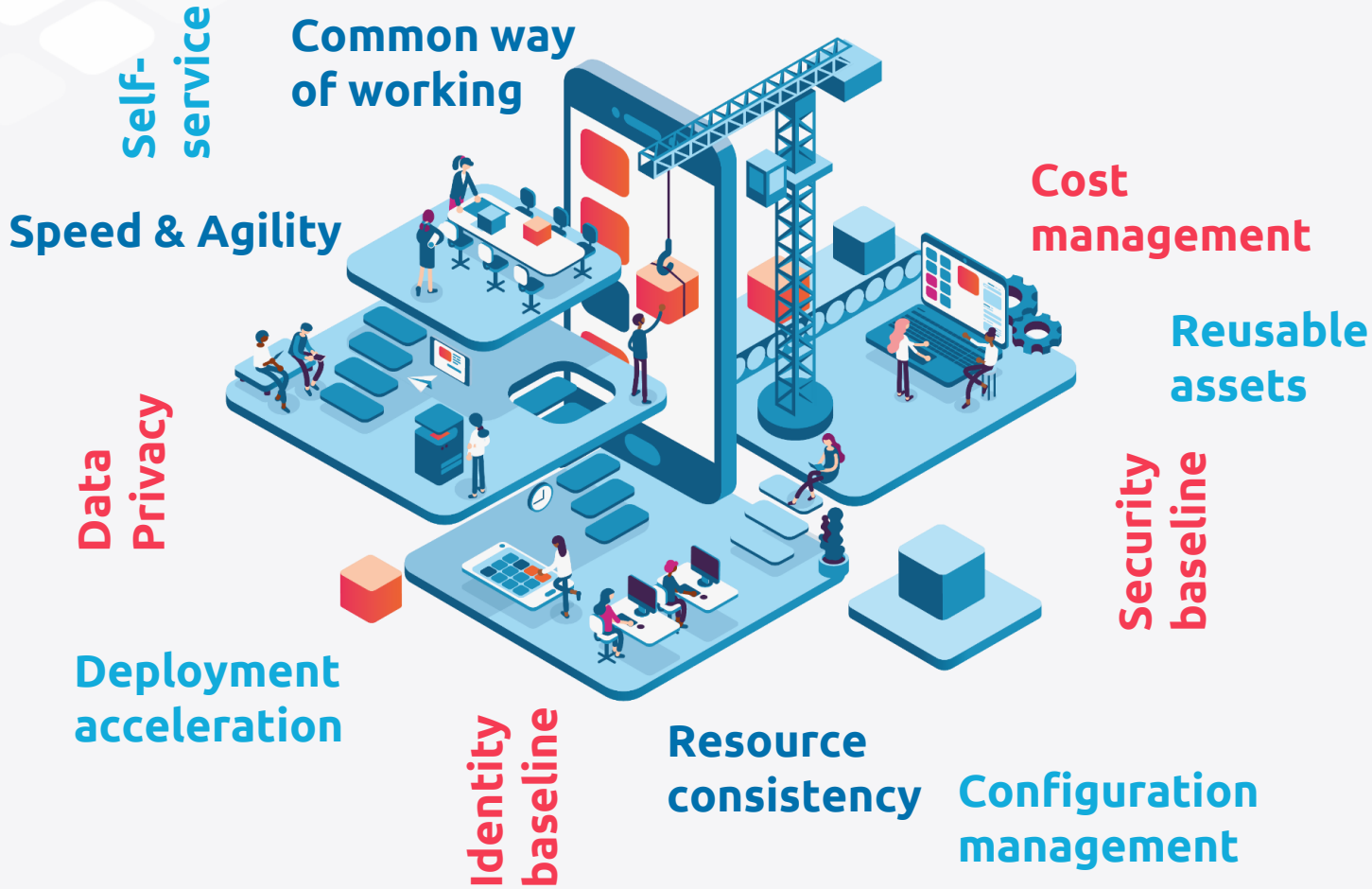
Powerful and efficient, modern applications



### AWS Inferentia and AWS Trainium

Machine learning acceleration

# Platform + Industrialization

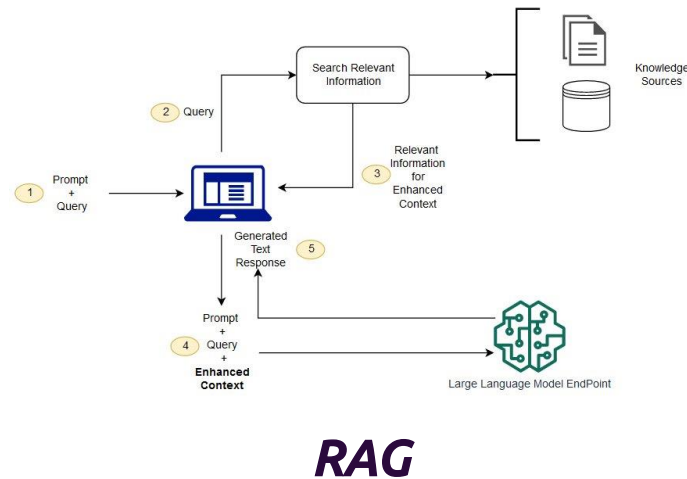


# Value creation vs. Value capture

Focus on your existing competitive advantages



Bring / Connect your proprietary data



Accuracy as your new gateway to production

- LLM evaluation
- Guardrails
- Observability
- Parsing
- Scoring
- Feedback loop

[AI Won't Give You a New Sustainable Advantage \(HBR\)](#)

# Reasoning capabilities

## Create your rules-based system with GenAI

Defect	Observed Temperature (°C)	Supported Temperature (°C)	Observed Friction (N/m)	Supported Friction (N/m)	Observed Time of Usage	Supported Time of Usage
1 Overheating	105	90	0.9	0.8	9500	9000
2 Excessive Vibration	80	75	1.5	1.3	12000	11500
3 Bearing Wear	70	60	2.0	1.7	8000	7500
4 Lubrication Failure	90	85	1.1	1.0	11000	10500
5 Motor Burnout	120	100	1.8	1.5	15000	14000
6 Conveyor Belt Slippage	85	80	1.4	1.2	12500	12000
7 Valve Leak	75	70	1.2	1.0	9500	9000
8 Hydraulic Pressure Loss	60	65	1.3	1.1	8500	8000
9 Sensor Failure	65	60	1.7	1.5	13000	12500
10 Gearbox Malfunction	95	85	1.9	1.6	14500	13500

Defect	Rule	Action
Overheating	IF Observed_Temperature > Supported_Temperature THEN Trigger_Alarm('Overheating Risk')	Trigger Alarm
Excessive Vibration	IF Observed_Friction > Supported_Friction THEN Trigger_Alarm('Excessive Vibration Detected')	Trigger Alarm
Bearing Wear	IF Observed_Time_of_Usage > Supported_Time_of_Usage THEN Schedule_Maintenance('Bearing Wear Detected')	Schedule Maintenance
Lubrication Failure	IF Observed_Temperature > Supported_Temperature THEN Trigger_Alarm('Lubrication Failure')	Trigger Alarm
Motor Burnout	IF Observed_Temperature > Supported_Temperature THEN Trigger_Alarm('Motor Burnout')	Trigger Alarm
Conveyor Belt Slippage	IF Observed_Friction > Supported_Friction THEN Trigger_Alarm('Conveyor Belt Slippage')	Trigger Alarm
Valve Leak	IF Observed_Temperature > Supported_Temperature THEN Trigger_Alarm('Valve Leak')	Trigger Alarm
Hydraulic Pressure Loss	IF Observed_Time_of_Usage > Supported_Time_of_Usage THEN Schedule_Maintenance('Hydraulic Pressure Loss')	Schedule Maintenance
Sensor Failure	IF Observed_Friction > Supported_Friction THEN Trigger_Alarm('Sensor Failure')	Trigger Alarm
Gearbox Malfunction	IF Observed_Temperature > Supported_Temperature THEN Trigger_Alarm('Gearbox Malfunction')	Trigger Alarm

**As Data analyst,**  
List the defect of the last week  
Extract the metrics  
Map the metrics with the documentation



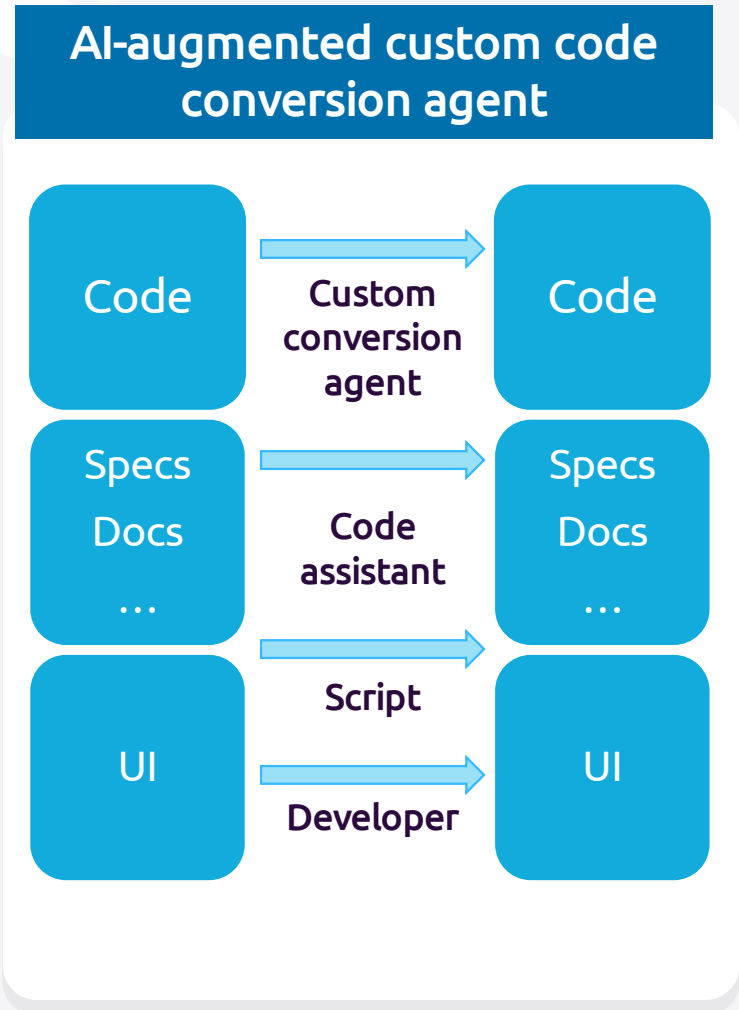
**As Defect analyst,**  
Based on the previous defect list  
Create the rule with type of action



**As Defect analyst reviewer,**  
Review and report any issue between the rule's list and defect's list

# Code Transformation Scenarios

## Code assistants + custom conversion agents



# EVIDEN

"With Amazon Q Developer, we successfully mapped out relevant use cases and accelerated the roadmap for adoption. In some of our projects, we are already seeing developer productivity increase between 20%–40% for cloud-native application development and approximately 20% increase in delivery velocity. Amazon Q Developer has helped us increase productivity among the development teams, improve overall code quality, and incorporate security early in the project lifecycle."

Michael Liebow, Head of Eviden Cloud Business, Atos Group



National Australia Bank is one of the largest financial institutions in Australia.

"At NAB, teams across the bank are excited about how generative AI can transform our work, and we've found a great solution in Amazon Q Developer, a powerful generative AI-powered tool for our engineers and developers. The tool has seamlessly integrated advanced generative algorithms and tools into our development process, delivering unparalleled benefits like completing tasks faster, increasing productivity, and minimizing repetitive actions. So far, our developers have accepted 50% of the code suggestions made by Amazon Q Developer, and that number continues to increase as we scale. We look forward to seeing how Amazon Q Developer will continue to empower and inspire our engineers to upskill their technical expertise, delivering better service for customers."

Andrew Brydon, Executive Chief Engineer, National Australia Bank



## Amazon Q Developer

**Automate upgrades from Java 8 to Java 17 to modernize legacy code in minutes.**



**30,000+** apps upgraded



**4,500** years of development work saved (vs. manual upgrade)



**\$260M** in annual cost savings from performance improvement

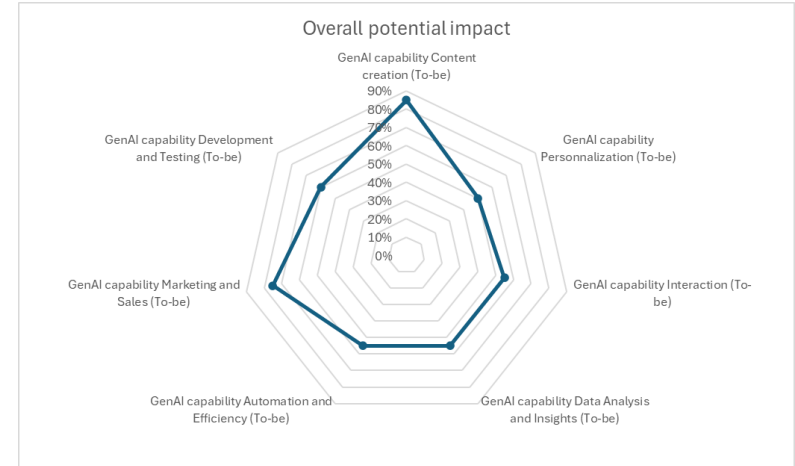
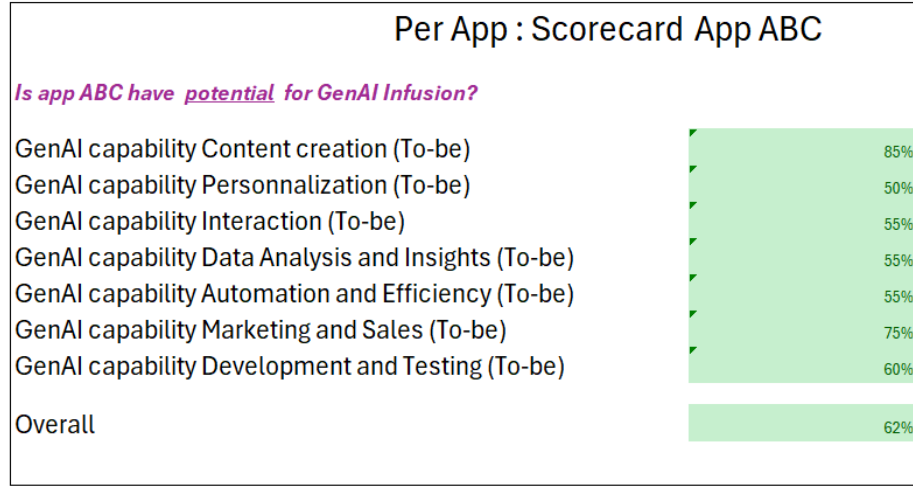


# Reshape your existing apps with GenAI

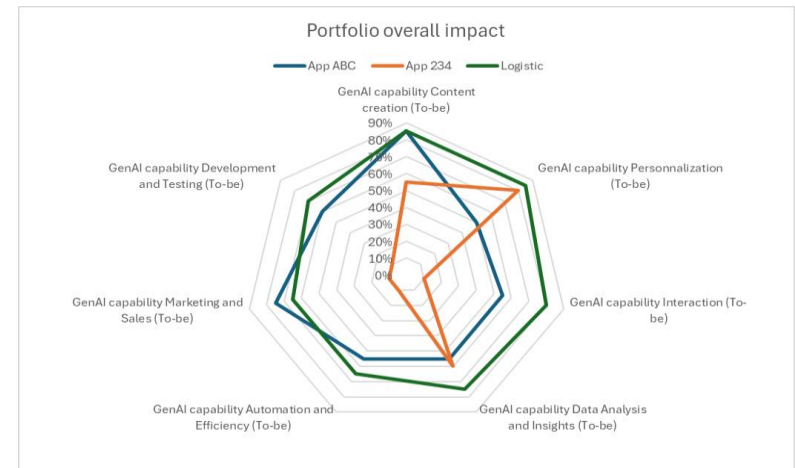
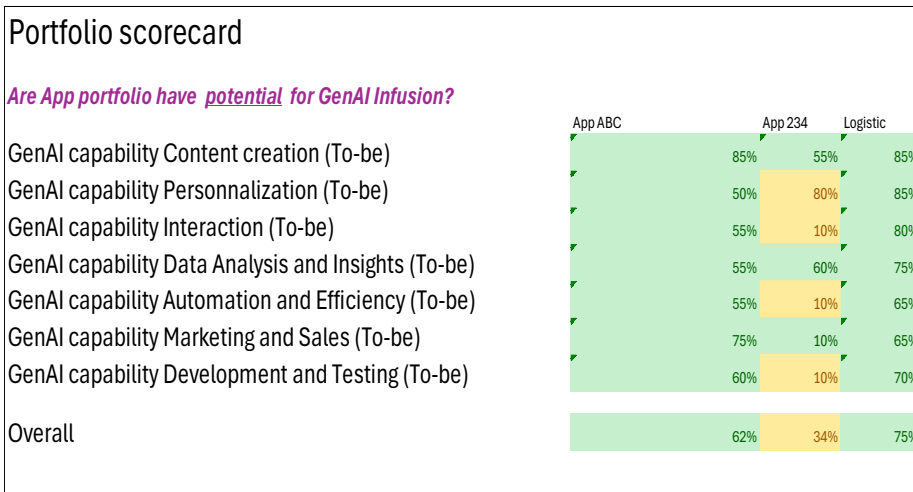
## As the new 6th R of app transformation



### One App

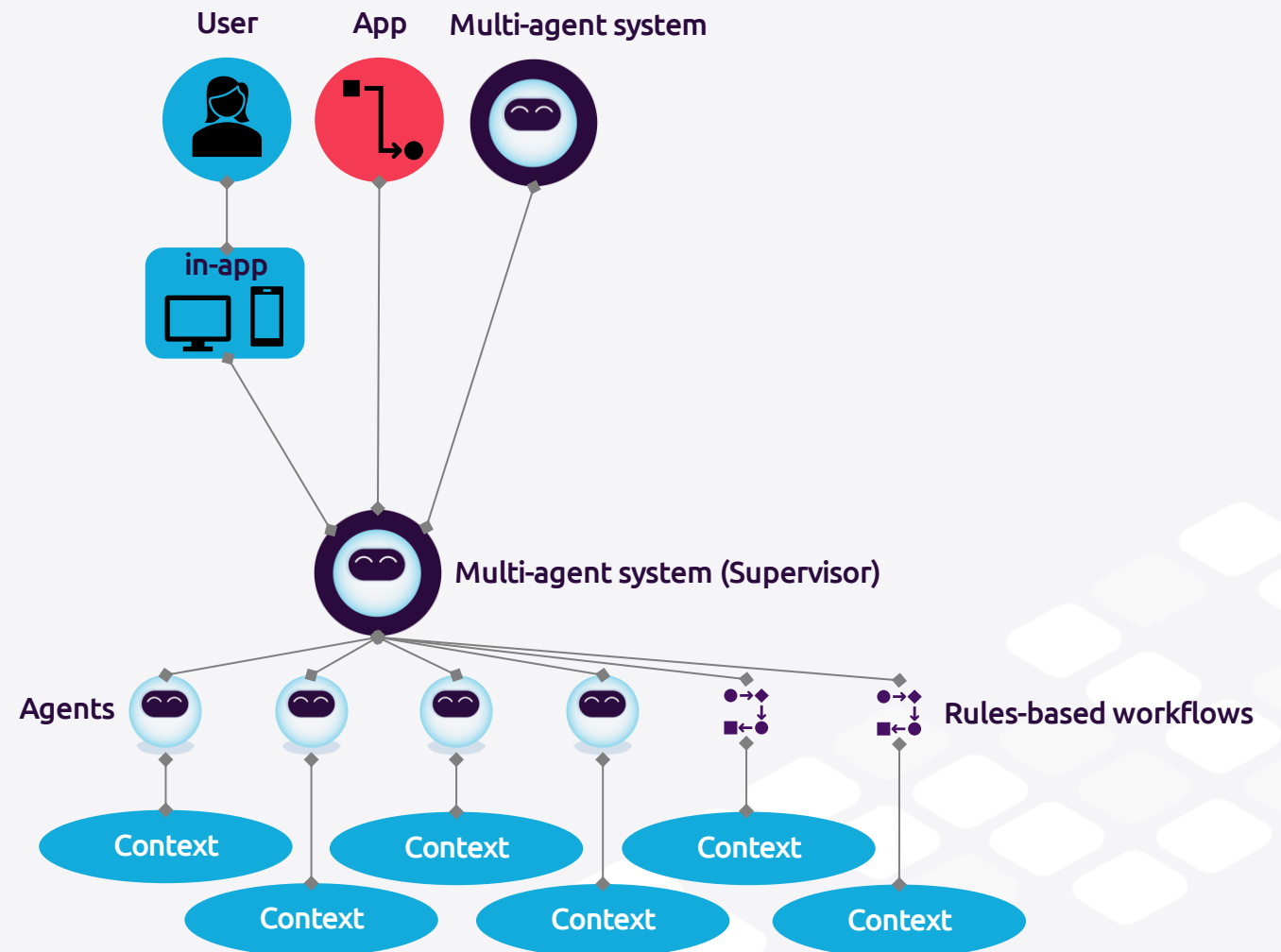
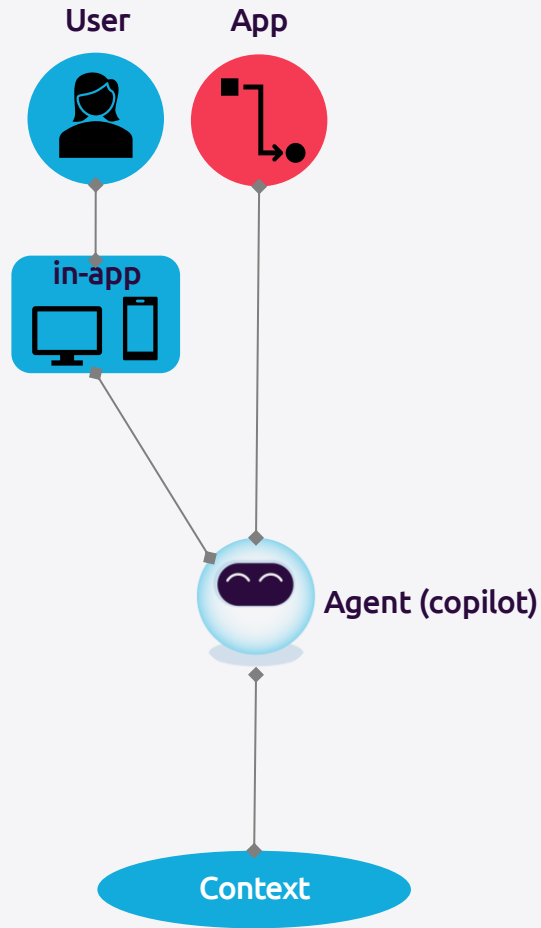


### Portfolio of Apps



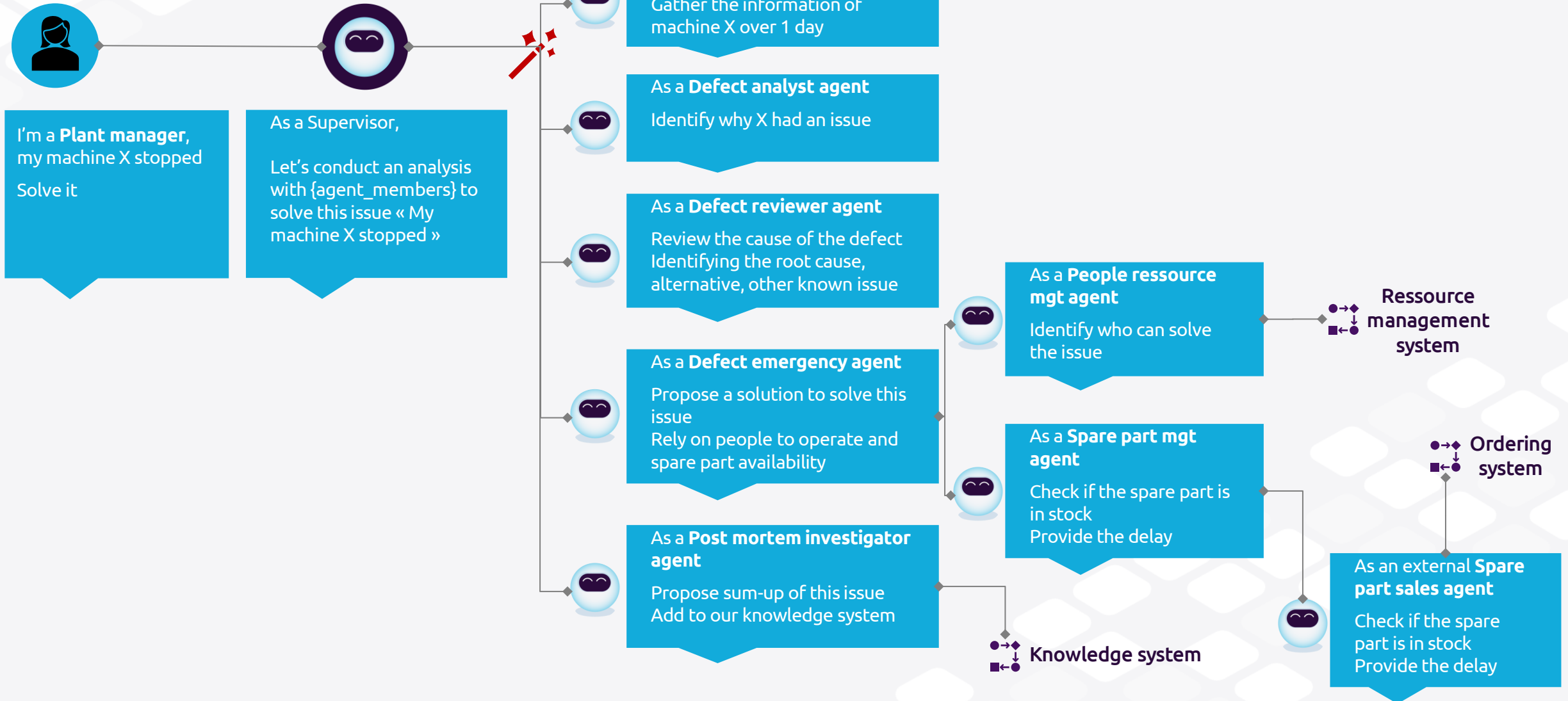
# Single agent (copilot) to multi-agent systems

Switch from single agent (copilot) to multi-agent systems



# Multi-agent system

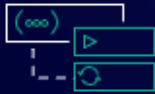
## By example



Regenerate business intelligence

# Simplify building and deploying agentic experiences

Fully managed Amazon Bedrock Agents



Action groups  
(tools)



Knowledge  
bases



Code  
interpretation



Memory  
capability



Session  
handling



Guardrails



Deployment



Logs, tracing



Secure

# Agents in Business Workflows



## Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku

Oct 22, 2024 • 5 min read



Contact Center video avatar+ multi agent fulfilment (D-ID)  
<https://www.youtube.com/watch?v=ODaHJzOyVCQ>

**New Anthropic Sonnet 3.5 in Bedrock:**  
<https://www.youtube.com/watch?v=kr8hW3Vi7QM>

# Sustain value with Guardrails

PREVIEW

## Guardrails for Amazon Bedrock

Implement safeguards customized to your application requirements and responsible AI policies



Apply guardrails to multiple foundation models and Agents for Amazon Bedrock



Configure harmful content filtering based on your responsible AI policies



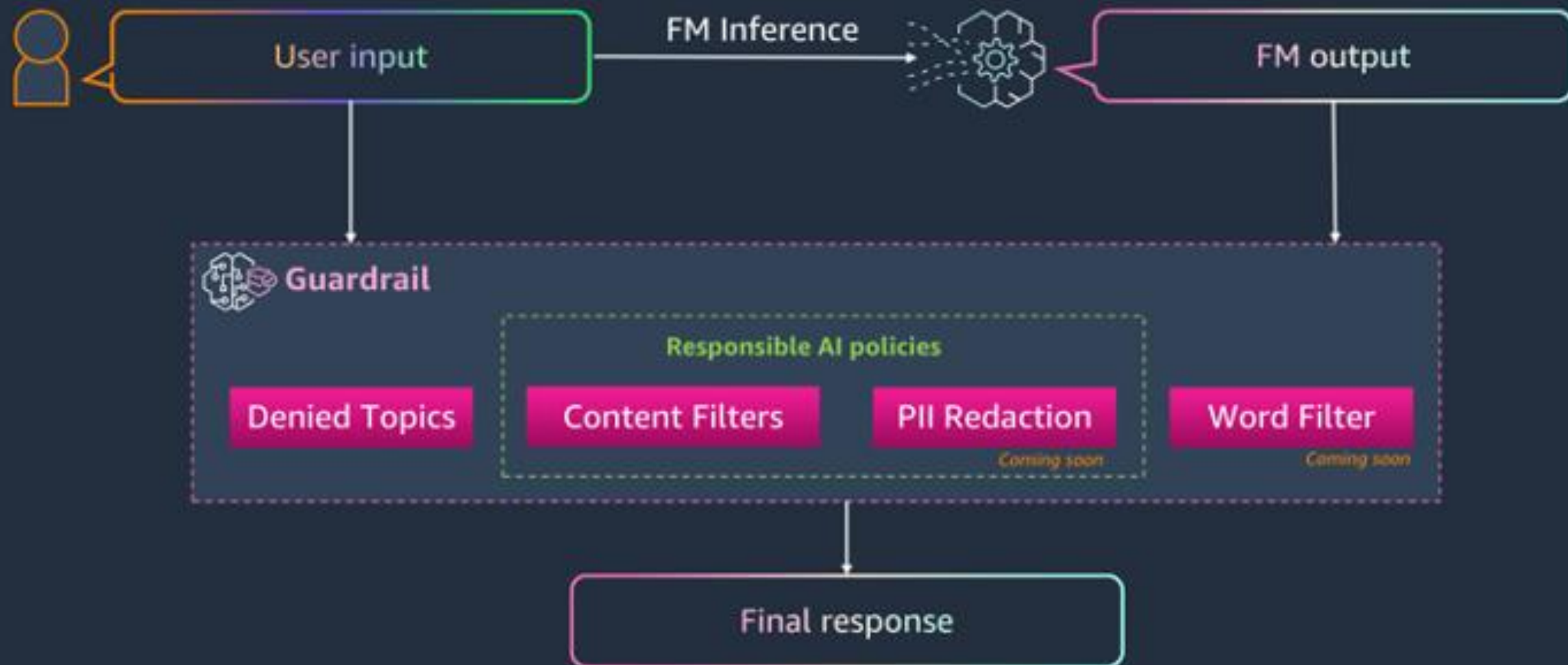
Define and disallow denied topics with short natural language descriptions



COMING SOON

Redact sensitive PII information in FM responses

# How it works: Guardrails for Amazon Bedrock



# Takeaways

## It's time to act now!

### **Advice:**

**Start your own journey  
(platform,, apps reshaping,  
intelligent apps)**

## Sustain technology

### **Advice:**

**Use, Customize, Optimize**

## Sustain value

### **Advice:**

**Accuracy is your key challenge**