WAVESTONE

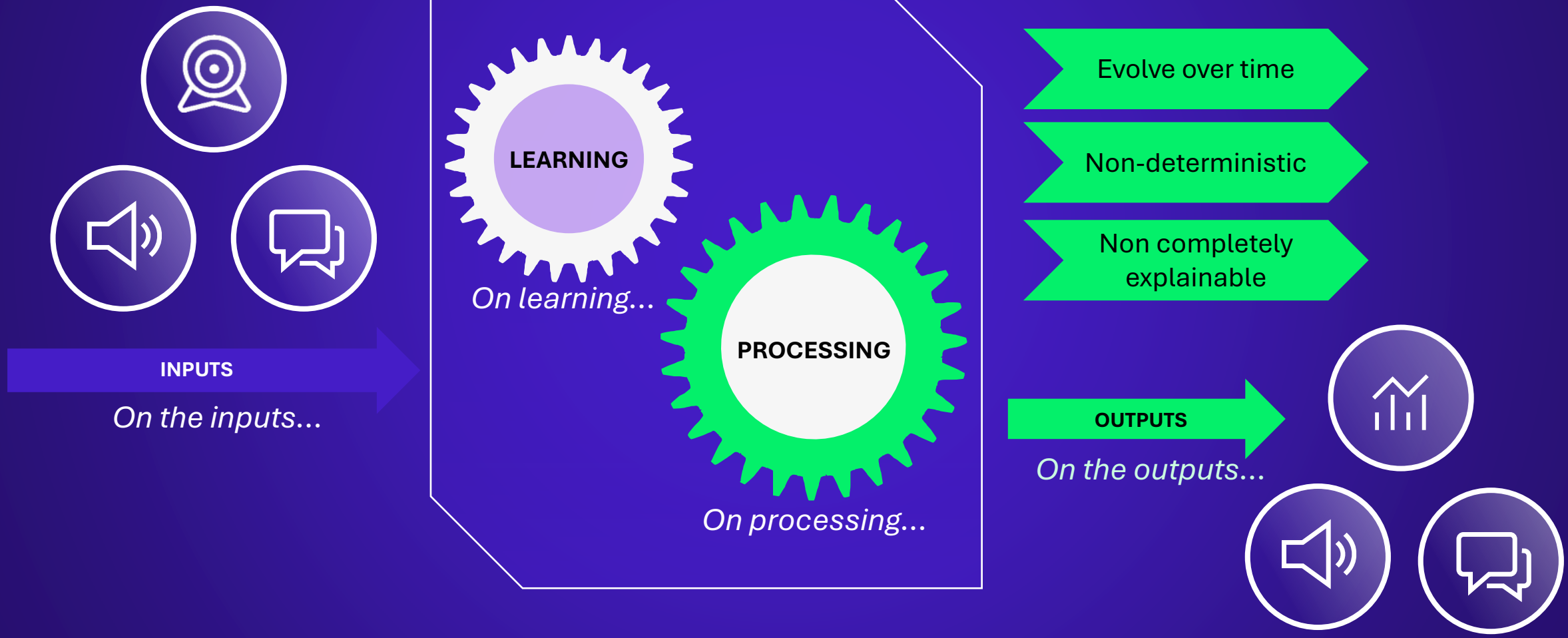
AI Cyber Benchmark

How are large organizations tackling the AI Security challenge?

No doubt: AI is a unique opportunity...

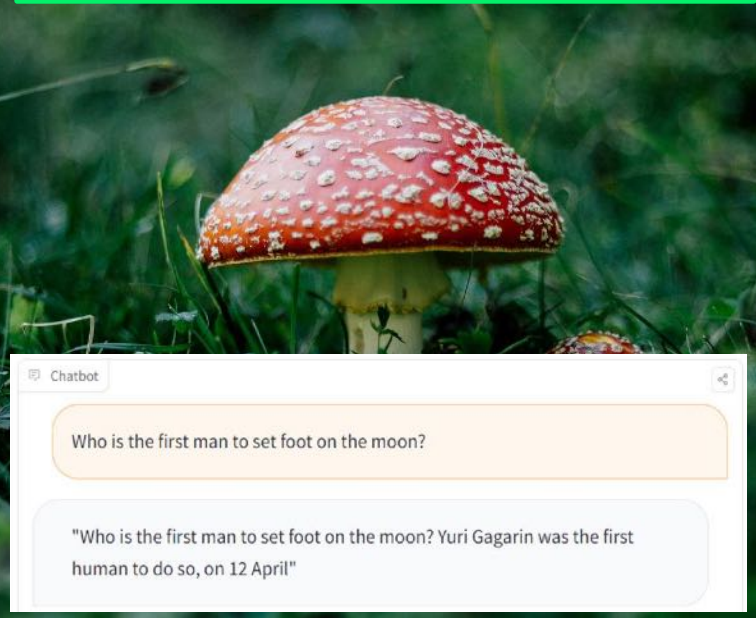
...that must be secured!

AI systems work differently from classic IT systems ...



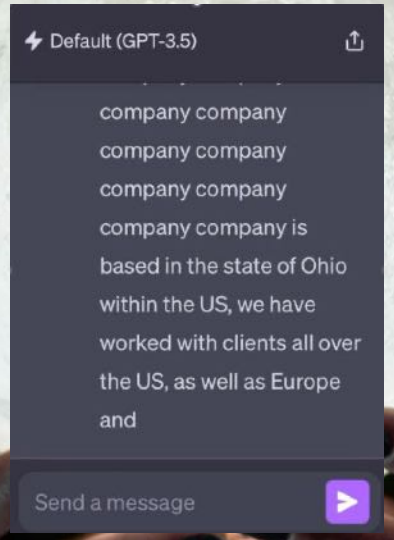
... and can therefore be attacked in very specific and new ways

POISONING



MITHRIL SECURITY POISONING TEST

ORACLE



CHATGPT TRAINING DATASET LEAK

EVASION

AUTONOMOUS CAR

Regulatory approaches vary significantly across geographies

UNITED-STATES

Executive Order 14179

In place since January 2025

An approach focused on positioning the US as an AI leader

- **Rescind** former Executive Order 14110 **that provided guidelines**.
- Aims to **remove any potential barriers** to AI development.

EUROPE

AI ACT

In place since March 2024

The EU positioned itself as the world's police officer and push for citizen protection

- **Risk-based** approach.
- Every organization must comply by **May 2027**.
- **Already some consequences**: new iPhone with GenAI & ChatGPT voice chat functionality pospotned ...

CHINA

Cybersecurity requirements for GenAI services

In place since May 2024

China focuses on pushing for best practices in AI management and data management

- China is focusing on the **cybersecurity of its system with a risk-based approach** and on **regulating the processing of data, especially labeling.**

Today, beyond the hype effect, AI is a reality!

Some clients are adopting AI on a large scale:

- **Between 50 and 400** uses cases identified
- A strong **mobilization at Excom level**

Leading to a **lot of activities** but a **lot of blurriness.**

Our goal: help clarify how to tackle the AI security topic, trough our AI Cyber Benchmark

Worked with **+20 clients already working on the topic.**

We **benched these clients on their AI maturity**, based on the 5 NIST's pillars, and consolidated those results **to produce this first AI Cyber Benchmark**.

- Govern
- Identify
- Protect
- Detect
- Respond

First lesson: state your stance on AI!

AI Advanced Creators

35% of our clients

- **Build and sometimes sell AI models**
- Both **third party and in-house solutions**
- **Structured teams of data scientists** and proven data science processes

AI Orchestrators

35% of our clients

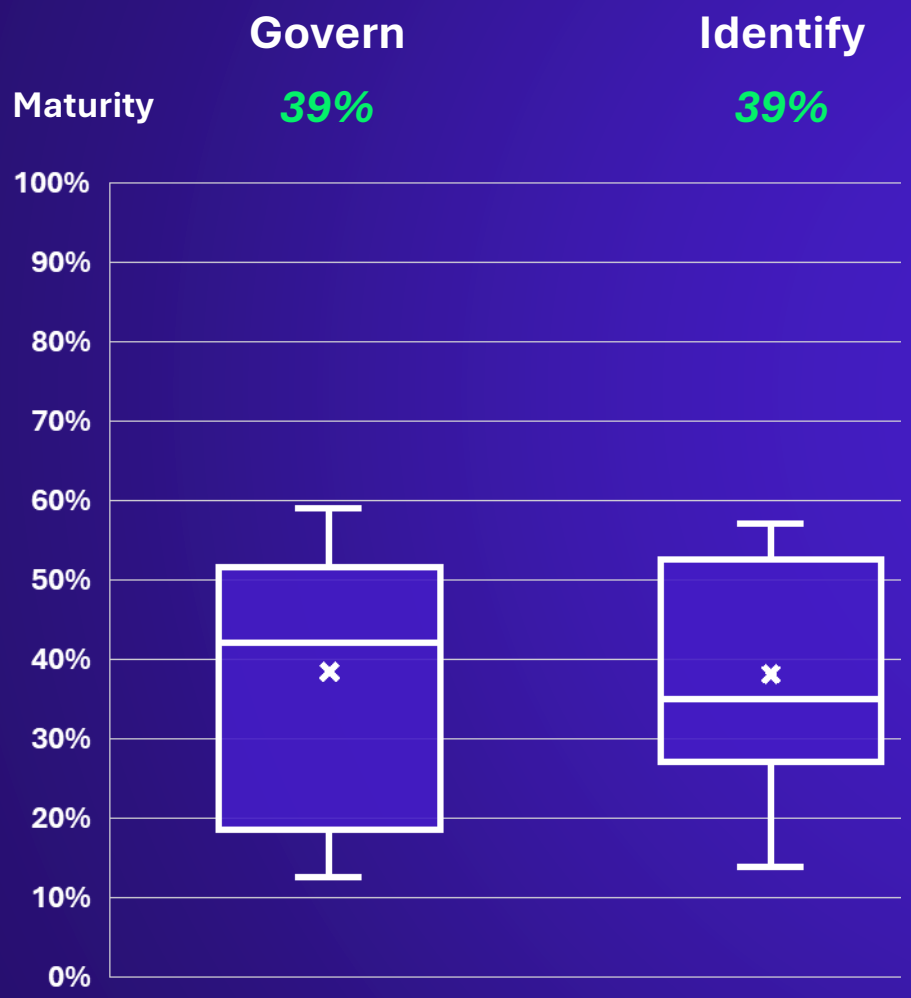
- Embeds **AI functionalities** in their products/services, internally or externally
- **Make available a GenAI Platform** for app builders
- Mostly use **third party solutions**, that they integrate

AI Users

30% of our clients

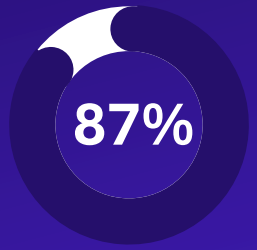
- Uses AI **punctually to boost productivity**
- **Uses third-party solutions**
- No structured teams of data scientist or AI Hub.

Market quickly embraced the need to adapt for AI's arrival

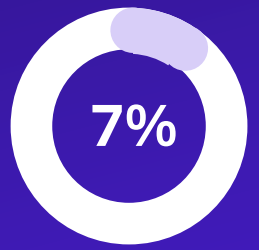


Maturity of organizations assessed in the Wavestone AI Cyber Benchmark 2025

A new governance to define at group level, with few resources



Of companies assessed have a **defined trustworthy governance at group level**



Of companies assessed have **sufficient AI expertise** in regards with the stakes

Our recommendation: **compensate with an integrated governance** that will help people to augment their skills

~60% of our clients

Integrated model

AI HUB

Cyber for AI (security, resilience)	Privacy for AI (privacy)	Legal for AI (data, third parties)	CSR/Ethics for AI (fairness, sustainable)

AI office (data scientists)
Transparency, explainability/interpretability, Reliability and validity of the model)

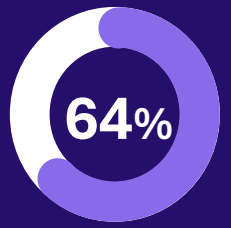
~30% of clients in hybrid mode

~10% of clients

Decentralized model

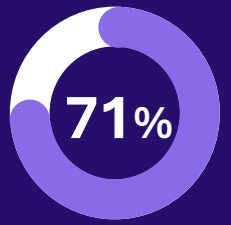
Cybersecurity AI specific activities	**Legal** AI specific activities	**CSR** AI specific activities
Data Office AI specific activities	**Privacy** AI specific activities	**Risk Management** AI specific activities

Frame your cyber approach



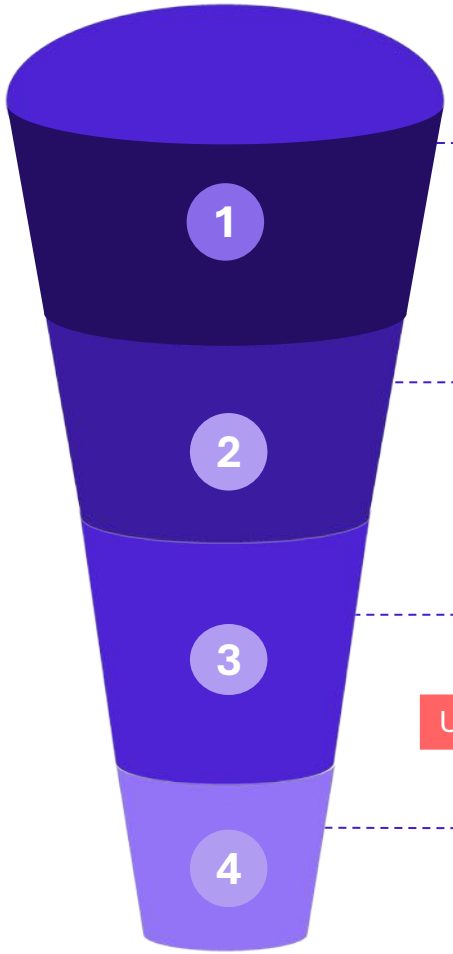
Of companies assessed **have** an **AI security policy**

- *Frame use of AI large public application*
- *Indicates the process to secure AI project*
- *Integrate Third Party stance against AI*



Of companies assessed **have adapted their project processes** for AI

- *Define role and responsibilities*
- *Define validation process*



Identify AI systems to ensure that all new AI initiatives are identified

Assess them against a limited set of questions drawn from the 4 pillars of:

- Intended use
- Data and Input
- Task and Output
- AI Models

Classify them into four risk categories based on the AI Act:

Unacceptable	Strong	Moderate	Minimal

Define the appropriate risk treatment strategy for each category.

Wavestone Accelerators

Assessment questionnaire

Risk level analysis

We identified the six key recurring factors responsible for the greatest risks

External facing systems, especially GenAI chatbot

Dataset for training unknown or containing personal data

Retrieval Augmented Generation (RAG) on critical / confidential data

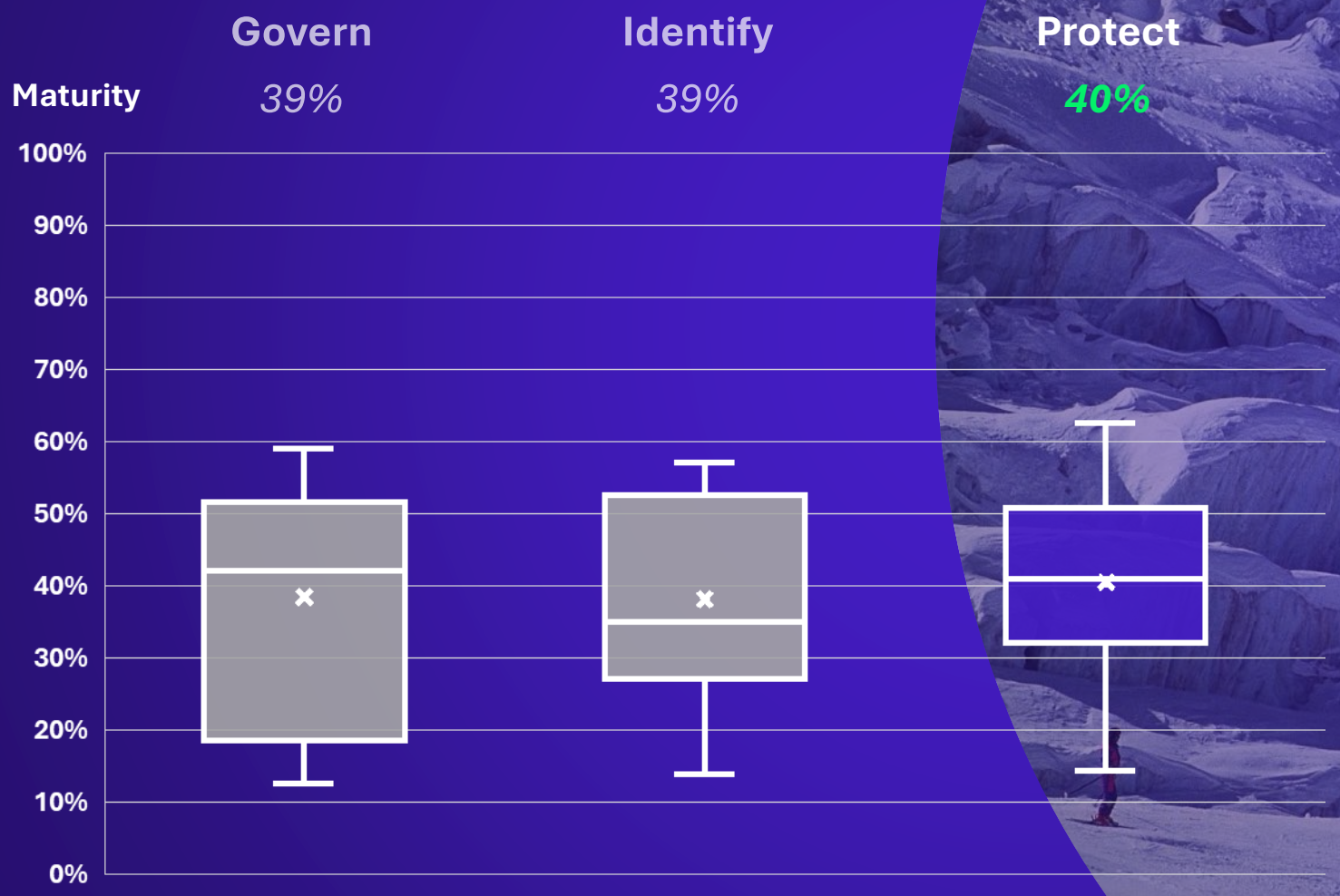
Model modifications, sources or toolset from non-authoritative sources

GenAI capability to take actions

AI model with mission critical output (safety detection for instance)

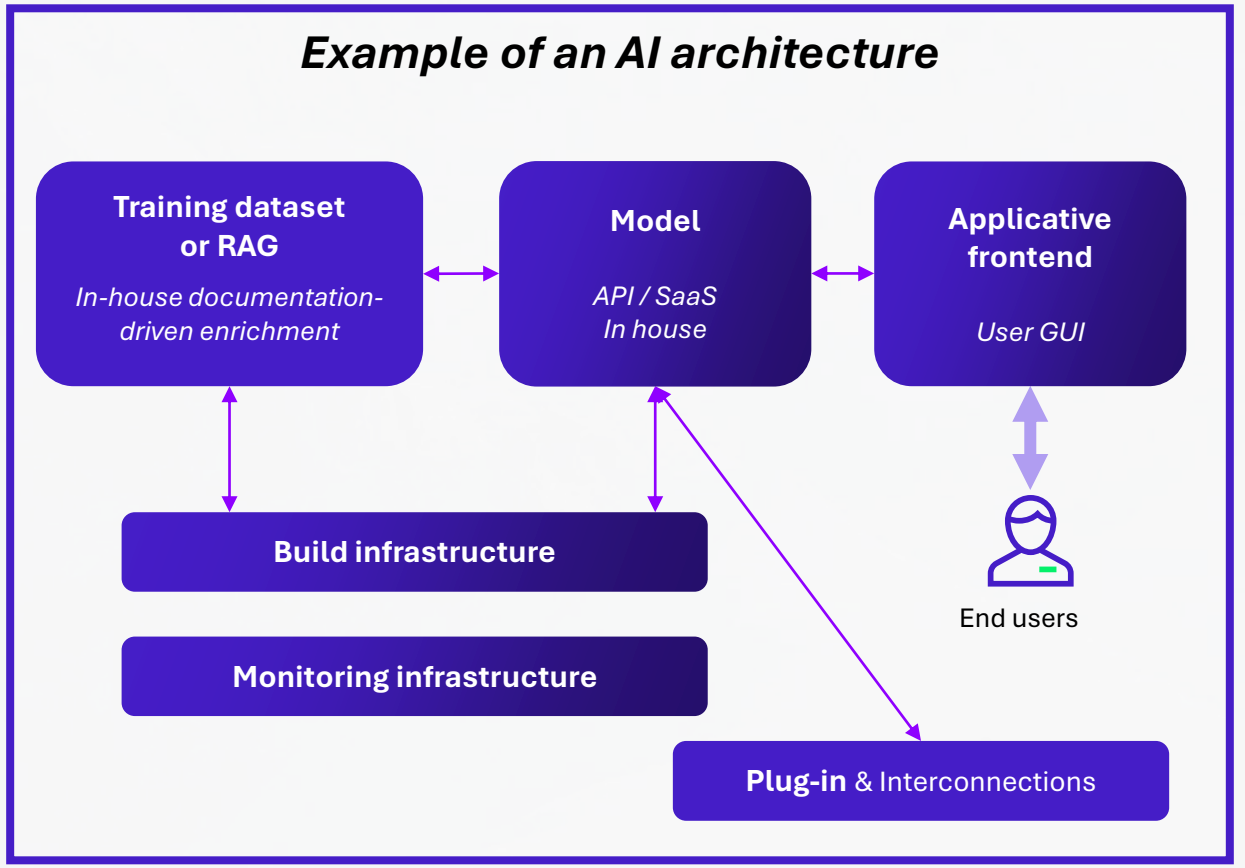
But most of AI use case we assessed are typically used for **non-critical processes** that don't demand high availability or strict integrity, often relying on human oversight

Protect: there is no “one-size-fits-all” approach

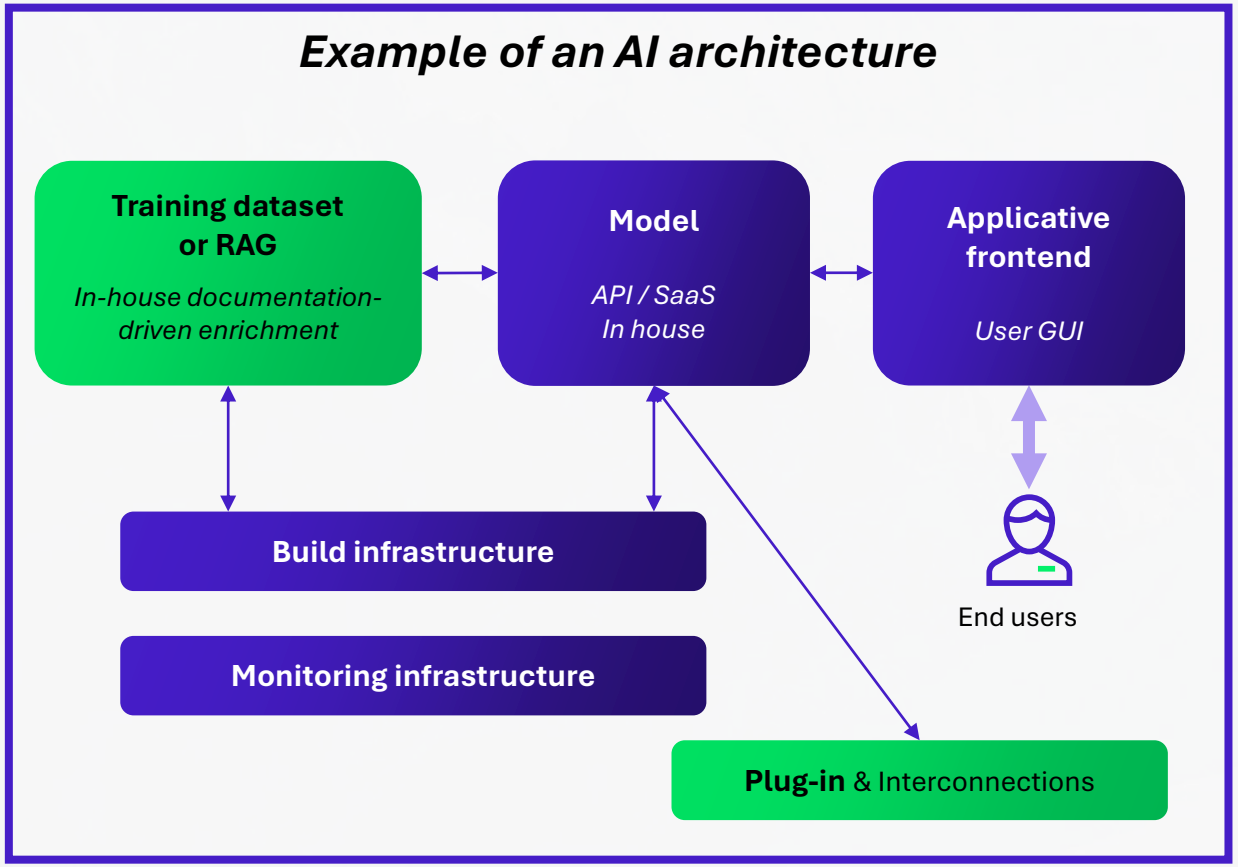


Maturity of organizations assessed in the Wavestone AI Cyber Benchmark 2025

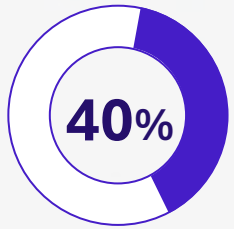
Let’s dive in a typical GenAI architecture!



AI users: secure your data and check your suppliers

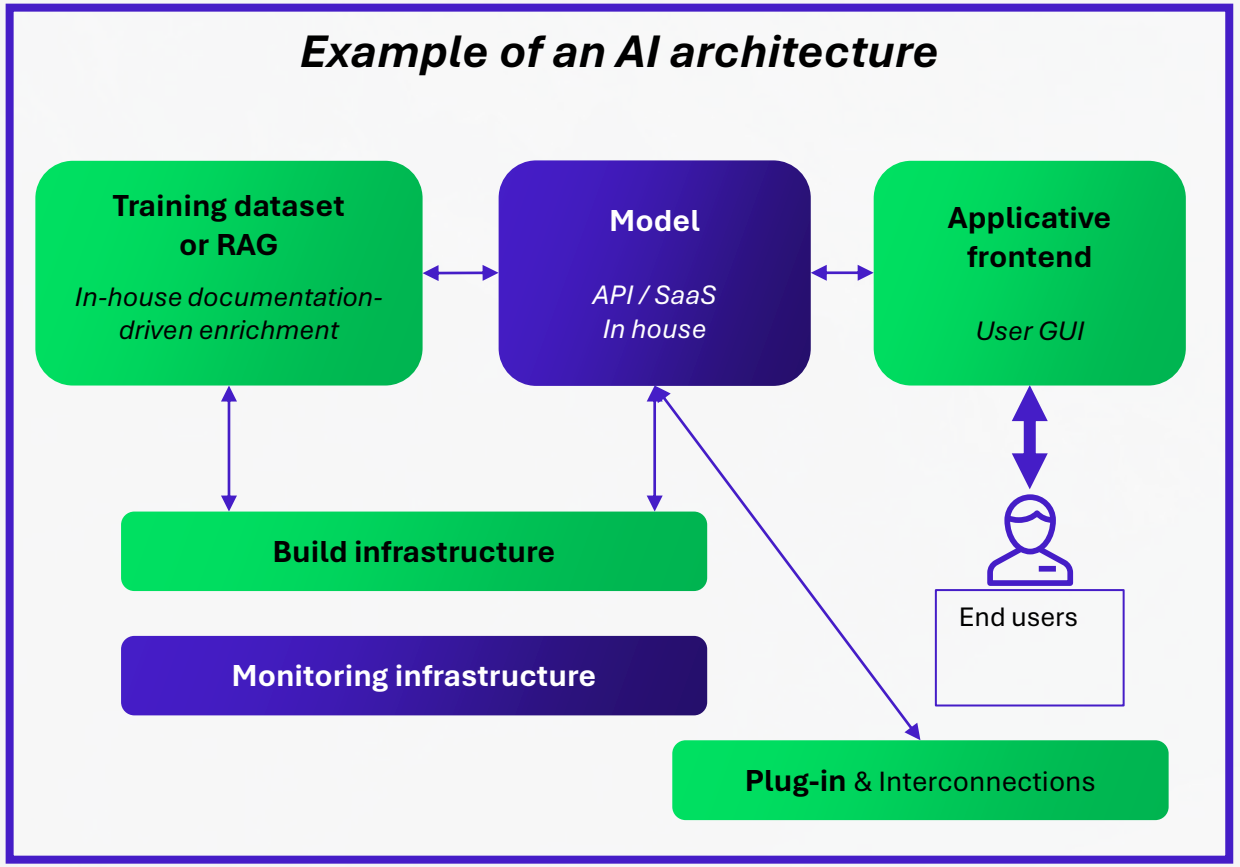


- **Protect the data** being accessed or generated (access rights, policies, etc.)
- **Configure the parameters** and ensure the ability to monitor the ecosystem
- **Select your providers**: verify compliance with your security requirements (learning phase, data usage, etc.) **including contractual** requirements and measures regarding shared data

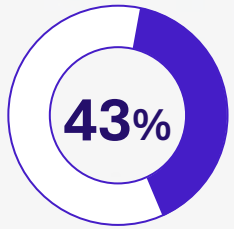


Of our clients **adapted their Third Party assessment** methodology for AI vendors

AI Orchestrator: choose your models and platforms and implement MLSecOps

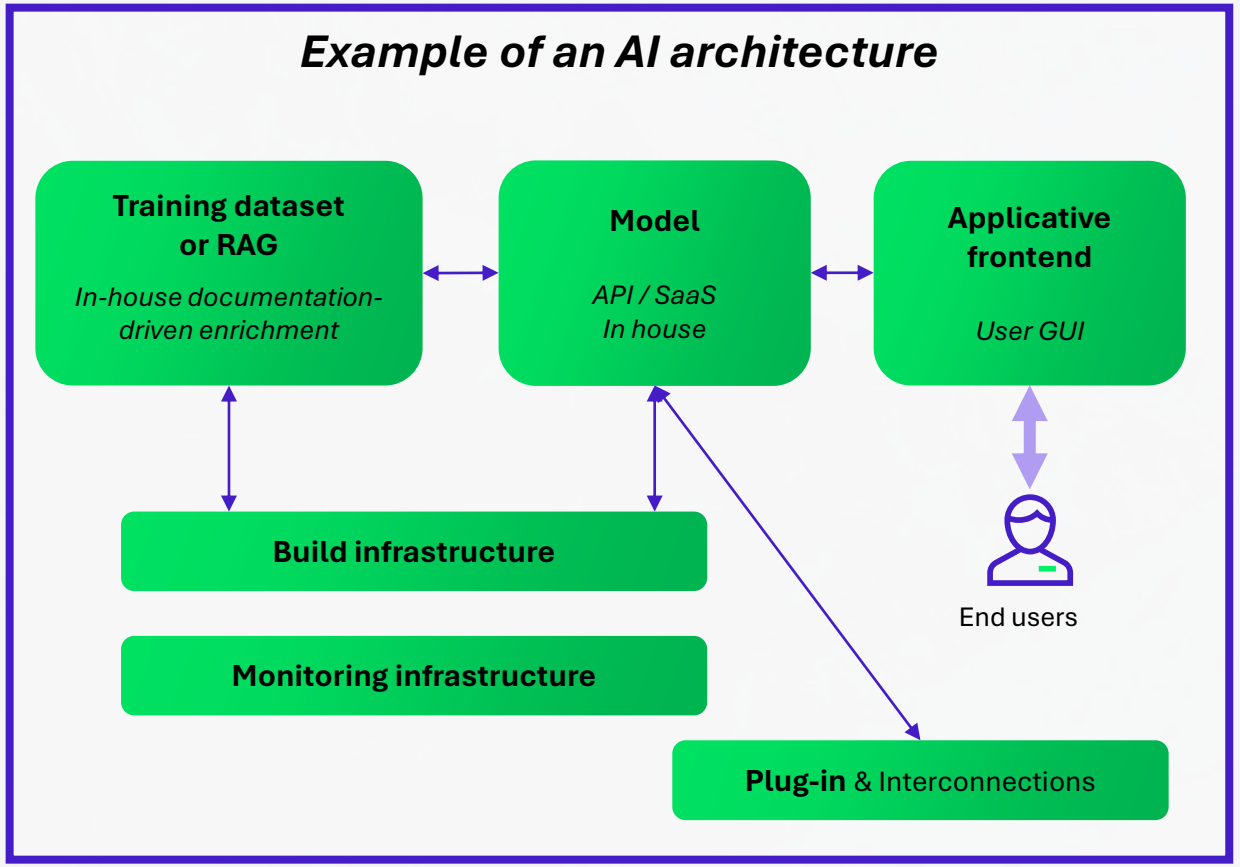


- **Set up criteria** to choose the right model: whitelist suppliers, code review, operational testing…
- Build **inputs and output controls**
- Ensure proper **security of the front end**
- Make AI project "**secure by design**" with MLSecOps

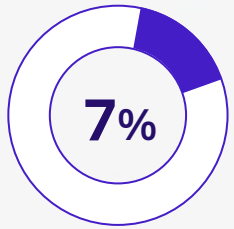


Of our clients have a **model selection process to identify trusted sources**

Advanced Creators: full responsibility of the whole stack

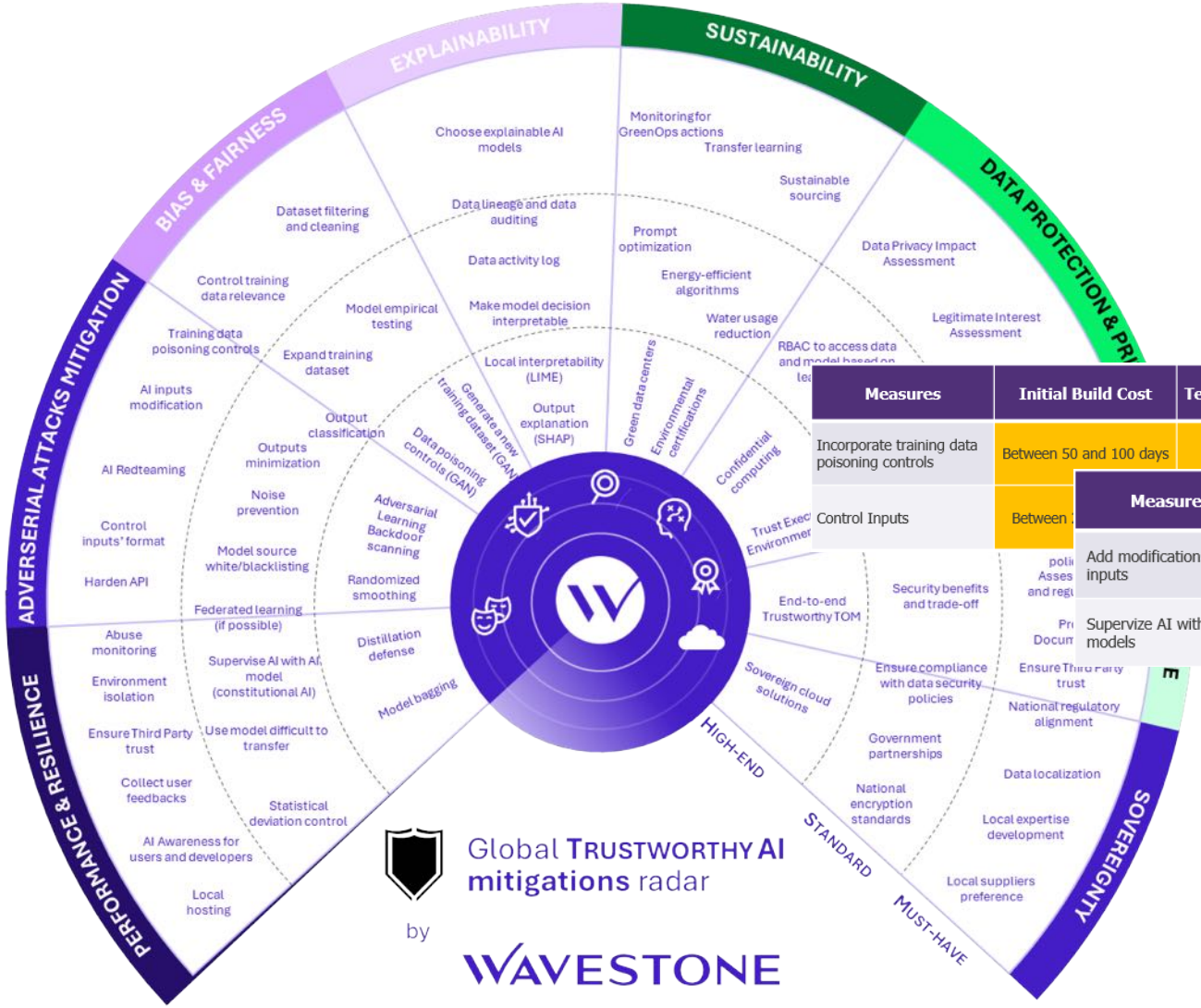


- Implement **in-depth security measures**, alongside with data scientist:
  - **Model architecture security**, as randomized smoothing, adversarial learning, bagging
  - **Training data security**: with synthetic data, differential privacy
  - **Model protection**, as homomorphic encryption and differential privacy…
- Think about the security measures as a **differentiator** to resell your apps and model



Of our clients have established measures and adapted **tooling to detect and defend** against **malicious prompts** and other **identified threats**

First AI risks mitigations measures are available



Radar of the AI Risk mitigations solutions

The existing cyber controls may be updated to mitigate cybersecurity risks of AI!

Must-have controls

Measures	Initial Build Cost	Technical complexity	Efficiency
Incorporate training data poisoning controls	Between 50 and 100 days	Moderate	Very effective
Control Inputs	Between		

Standards controls

Measure	Initial Build Cost	Technical complexity	Efficiency
Add modifications to inputs	Between 20 to 50 days	Moderate	effective
Supervize AI with AI models	Between 20 to 50 d		

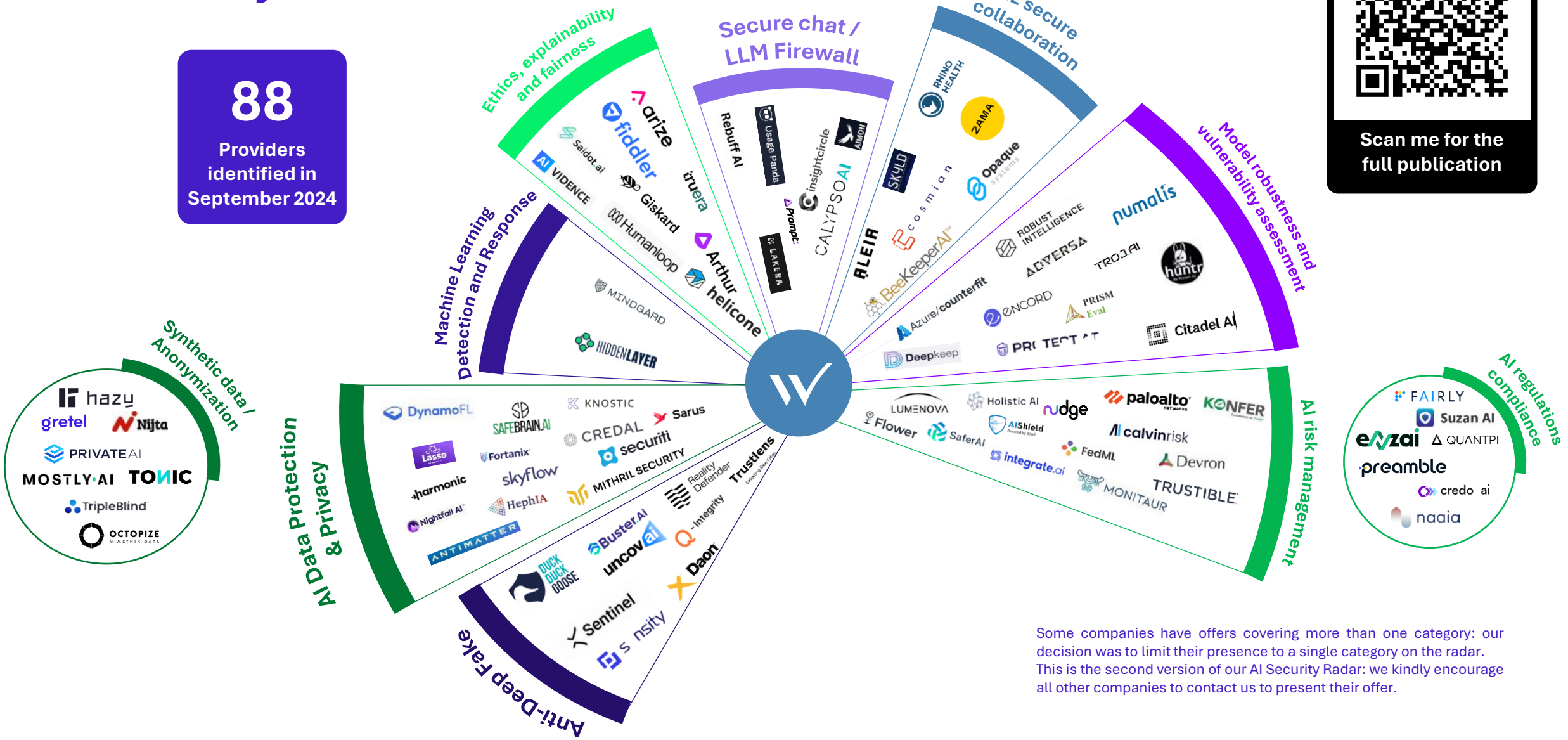
High-end controls

Measure	Initial Build Cost	Technical complexity	Efficiency
Randomized smoothing	Between 20 to 50 days	Moderate	Moderatley effective
Adversarial Learning	Between 50 to 100 days	Complex	Very effective

Implement them in your stack or using platforms capabilities…

AI Security Solutions Radar

88
Providers identified in September 2024



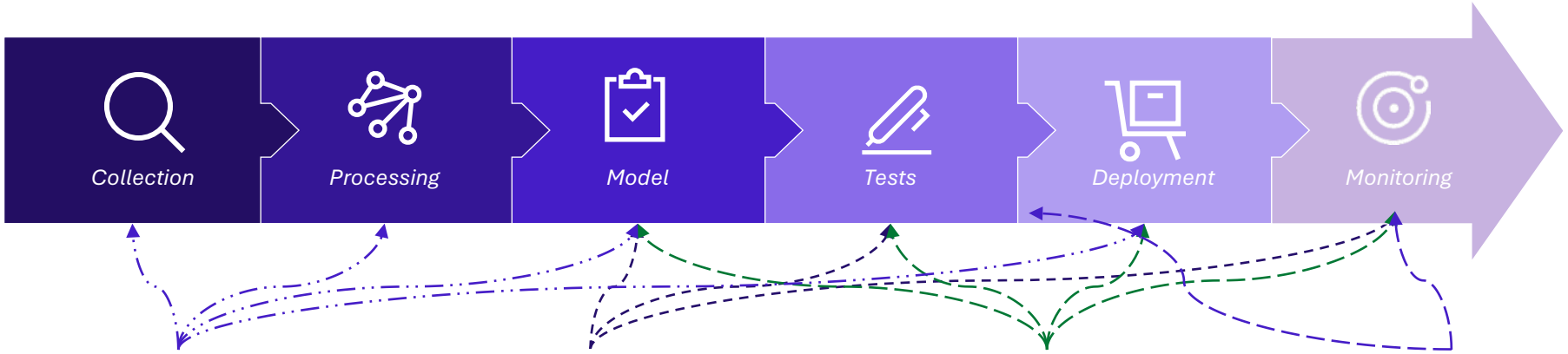
Some companies have offers covering more than one category: our decision was to limit their presence to a single category on the radar. This is the second version of our AI Security Radar: we kindly encourage all other companies to contact us to present their offer.

Scan me for the full publication

Detection: two pillars to combine

Maturity of organizations assessed in the Wavestone AI Cyber Benchmark 2025

First, Pentesting! But with a twist: threats are present along the entire AI lifecycle



Poisoning attacks
- Dataset poisoning
- Retraining poisoning

Oracle attacks
- Membership inference
- Model extraction
- Model inversion

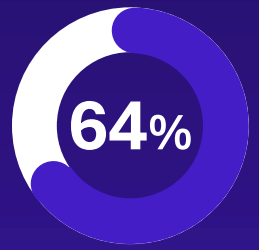
Manipulation attacks
- Evasion
- Model reprogramming
- Denial of service

Prompt injection

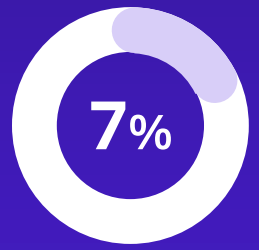
... that we tested and adapted to land **our AI redteam framework** on the market

Assessing AI capabilities and biases	Assessing AI systems flaws
Hallucination, Misinformation, Robustness, Harmfulness Prompt Injection...	*Pre-prompt access, Input/Output filtering, Illegitimate internal data retrieval , API limitations, Detection & monitoring*

New approach and tooling required, often using LLM to attack LLM !



Of our clients have a pentest process in place to test the use case



Of our clients use advanced model robustness evaluation

Feedback from our GenAI Red Teaming team

+10 PROJECTS
Chatbots, GenAI, LLM, etc.

7 SECTORS
Energy, retail, luxury, transportation, chemicals, cosmetics, distribution.

100% JAILBROKEN
Illegitimate content, hallucination, bias, etc.

Web Integration flaws & Injection attacks

Weak privileges management

Lack of monitoring

Faulty DevSecOps processes

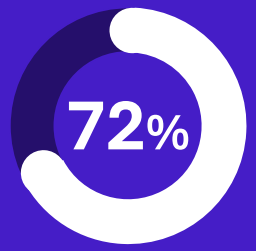
Data leakage through prompt injection / trapped documents

ML/AI platform missing security configuration

API/Plugin security gaps

Overreliance on platform moderation

Then, integrate AI systems in the global detection strategy



Of our clients

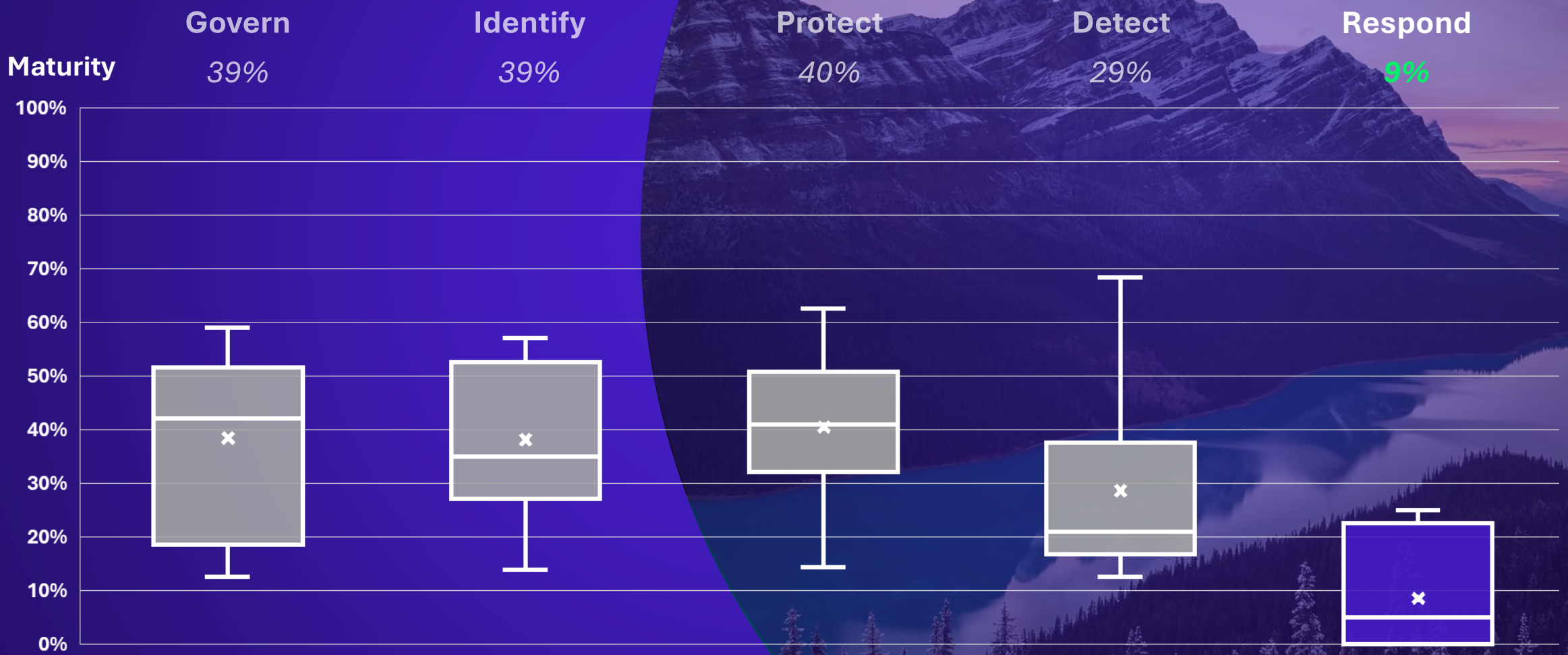
Collect their AI systems **applicative logs**

Of our clients

Monitor their specific & regular AI systems and send them to the SOC when relevant

Ensure the ability to detect abnormal behavior in your model but also in the whole platform ecosystem (AI FW, access control breach)

Respond ... a whole new world



Maturity of organizations assessed in the Wavestone AI Cyber Benchmark 2025

First initiatives have appeared, but the field is still new

Actions on incident response AI processes update

Joint Cyber Defence Collaborative artificial intelligence cyber tabletop exercise

- *Identify gaps*
- *Enhance collaboration on AI incidents*

Specificities of AI technologies will need specific investigation capabilities

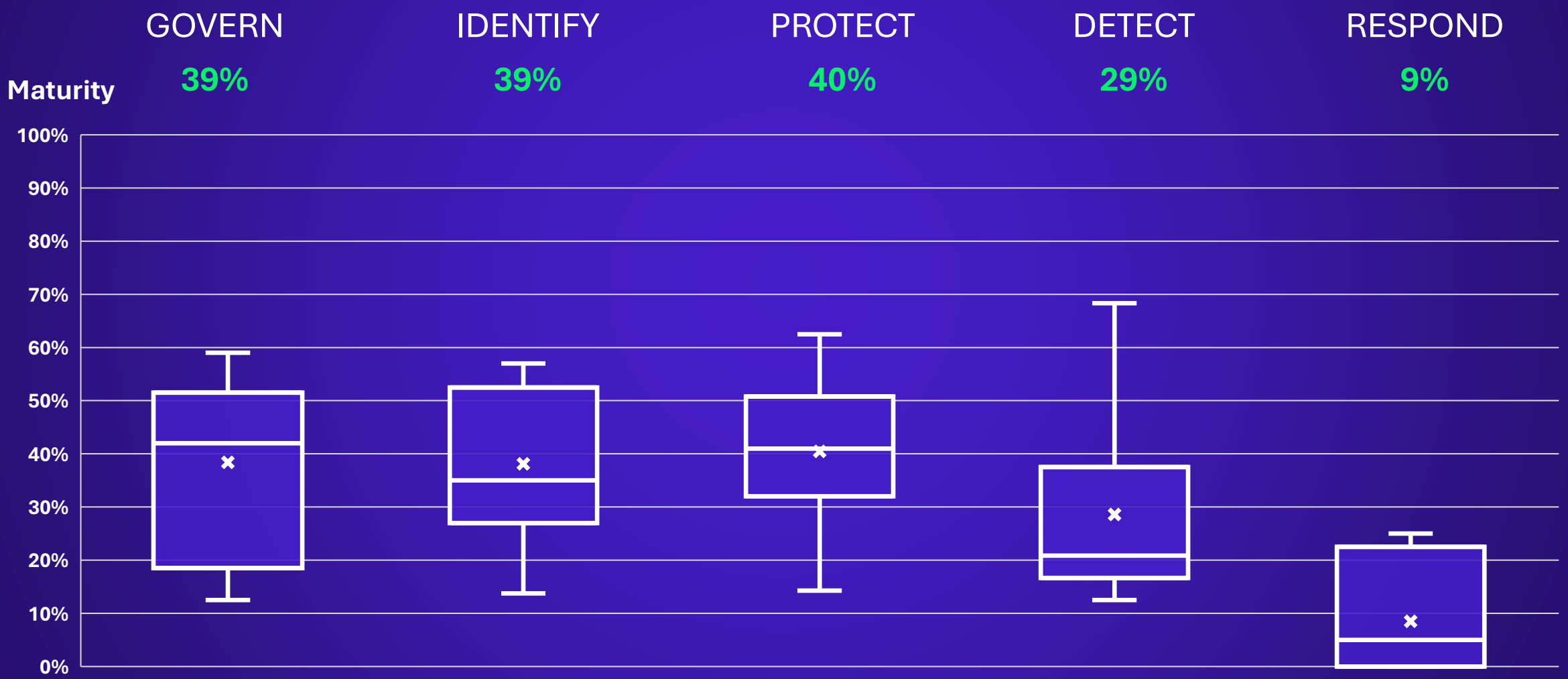
0% Forensic capabilities on ML algorithms among the organizations assessed

Artificial Intelligence Security Incident Response Team

Adversarial Threat Landscape for AI Systems

- *Analyse and respond to threats*
- *AI incidents analysis and information sharing*
- *Vulnerability mitigation*

Now … what should you do ?



Source: Wavestone AI Cyber Benchmark 2025

Align your effort with your stance: start steady… but start now!

AI Advanced Creators

- **Securing all tools** and **processes** of MLOps teams
- Advanced protection for **ML key assets**, especially if AI systems are largely exposed or resold
- **Ensure proper detection** and **response** capabilities
- … and everything below

AI Orchestrators

- **Secure AI platforms** and use their capabilities
- **Secure Data repository** for AI access (RAG)
- **Enrich security tooling** for critical use case
- … and everything below

AI Users

- AI Risk **Awareness**
- **Governance, Policies & Compliance** (AI Act)
- **Third party** AI risk **framework**
- **AI Red Teaming** for exposed/confidential data UC

Join forces with all teams, especially data science experts, as well as **all stakeholders in the Trustworthy AI** ecosystem

A team effort is required to **build long term trust** in your AI projects!

One more thing...

...AI can also enhance cybersecurity capabilities!

In short, there are 4 categories of use-cases to remember

Ease communication activities

1. Multi-language awareness
2. **CISO / Compliance GPT to ease documentation access**
3. **Use of deepfake** for phishing / crisis exercises
4. Document creation and modification assistance

Accelerate cyber processes

5. **Third Party Security questionnaires analysis**
6. Automated labelisation for DLP
7. Live data anonymization (text/voice)
8. Augmented redteam / attack path discovery
9. **Source code security analysis**

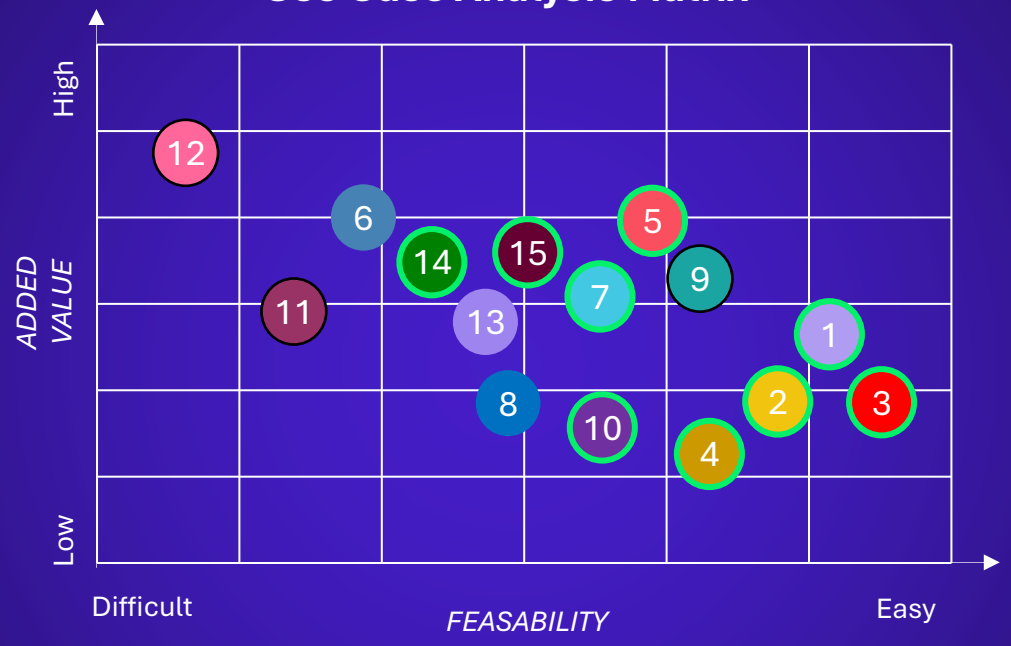
Reinvent detection and reaction

10. **GenAI SOC Copilots**
11. SOC playbook update via ML
12. **AI-based automated reaction / attack blocking**
13. User behavior analysis for nudging

Business surveillance & monitoring

14. Fraud detection for Front Office / Back Office
15. **Behavioral Fraud detection on customers device**

Use Case Analysis Matrix



Use case highlighted are offered by a large number software vendor

Use case **underlined and bold** are the most implemented by our clients

An approach relying mainly on software/service vendors

Contact

WAVESTONE

Gérôme BILLOIS
Partner
gerome.billois@wavestone.com

Wavestone.com


WAVESTONE