



**International
Standard**

ISO/IEC 27090

**Cybersecurity — Artificial
intelligence (AI) — Addressing
security threats and compromises
to AI systems**

*Cybersécurité — Intelligence artificielle (IA) — Traitement
des menaces et des situations susceptibles de compromettre les
systèmes d'IA*

**First edition
2026-10**

Sample Document

get full document from standards.iteh.ai



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2026

All rights reserved.

This ISO publication is protected by copyright and is owned by ISO and/or its licensors.

The content of this ISO publication is provided under licence, not sold. Use is subject to the applicable licence terms issued by ISO, an ISO member body, or an authorized third-party distributor.

Except as required for implementation or expressly permitted by a separate licence, no part of this ISO publication may be reproduced, distributed, modified, used, or made available in any form or by any means – electronic or mechanical, including photocopying, scanning, recording, or posting on internal or external digital platforms.

Any use beyond the scope of the granted rights is prohibited and may result in legal action.

ISO copyright office
CP 401 • CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword	v
Introduction	vi
1 Scope	1
2 Normative references	1
3 Terms and definitions	1
4 Abbreviated terms	5
5 Application of information security	5
5.1 General	5
5.2 AI governance programme	6
5.3 Zero trust principles	7
5.4 Applying software engineering best practices to AI	8
5.5 Supply chain management	8
5.5.1 General	8
5.5.2 AI-specific supply chain threats	9
5.5.3 Complex AI supply chains	9
5.5.4 Specific attack vectors	9
5.5.5 Mitigation strategies	9
5.6 Threat modelling	10
5.7 Red teaming	10
6 Types of potential compromise to AI systems	10
6.1 General	10
6.2 Data poisoning attack	13
6.2.1 Description	13
6.2.2 Failures resulting from data poisoning attacks	14
6.2.3 Detection of data poisoning attacks	14
6.2.4 Mitigation of data poisoning attacks	15
6.3 Evasion attack	17
6.3.1 Description	17
6.3.2 Failures resulting from evasion attacks	18
6.3.3 Detection of evasion attacks	18
6.3.4 Mitigation of evasion attacks	18
6.4 Membership inference	20
6.4.1 Description	20
6.4.2 Failures resulting from membership inference	20
6.4.3 Detection of membership inference	21
6.4.4 Mitigation of membership inference	21
6.5 Model exfiltration	21
6.5.1 Description	21
6.5.2 Failures resulting from model exfiltration	22
6.5.3 Detection of model exfiltration	22
6.5.4 Mitigation of model exfiltration	23
6.6 Model inversion	23
6.6.1 Description	23
6.6.2 Failures resulting from model inversion	23
6.6.3 Detection of model inversion	24
6.6.4 Mitigation of model inversion	24
6.7 Direct model poisoning	24
6.7.1 Description	24
6.7.2 Failures resulting from direct model poisoning	25
6.7.3 Detection of direct model poisoning	25
6.7.4 Mitigation of direct model poisoning	25
6.8 Sensitive model output	25
6.8.1 Description	25

6.8.2	Failures resulting from sensitive model output.....	25
6.8.3	Detection of sensitive model output.....	25
6.8.4	Mitigations of sensitive model output.....	25
6.9	Prompt injection	26
6.9.1	Description.....	26
6.9.2	Failures resulting from prompt injection	26
6.9.3	Detection of prompt injection.....	26
6.9.4	Mitigation of prompt injection.....	27
6.10	Output contains injection attacks	27
6.10.1	Description.....	27
6.10.2	Failures resulting from output contains injection attacks.....	27
6.10.3	Detection of output contains injection attacks.....	27
6.10.4	Mitigations of output contains injection attacks.....	27
6.11	Considerations for AI systems.....	28
6.11.1	General.....	28
6.11.2	Reduce impact by data limitation	28
6.11.3	Governing and controlling the effect of unwanted model behaviour during the operation stage.....	29
6.11.4	AI system and data provenance.....	32
7	Mitigations and their interactions with threats and other applied mitigations.....	33
7.1	General.....	33
7.2	Conflicting interactions of mitigations.....	33
7.3	Continuity of mitigations across the AI life cycle.....	33
7.4	Model and mitigation deterioration over time.....	33
7.5	Logging and monitoring.....	34
7.6	Development environment protection.....	34
7.7	Malicious input detection and filtering.....	35
7.7.1	Detecting out-of-distribution inputs.....	35
7.7.2	Detecting adversarial inputs.....	35
7.8	Throttling model use.....	36
7.9	Optimising the size of the model.....	36
	Annex A (informative) Mapping attacks to AI system life cycle and to assets.....	37
	Annex B (informative) AI-specific versions of conventional attacks.....	44
	Annex C (informative) Federated machine learning.....	49
	Bibliography.....	51

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and <https://patents.iec.ch>. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 27, *Information security, cybersecurity and privacy protection*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

Introduction

Mitigating the risks of security threats to AI systems in a timely manner can help to improve the trustworthiness of AI systems and of the organizations that use or develop them, and thus earn trust in these systems and organizations. As technology continues to advance and innovate at a fast pace, emerging and new security vulnerabilities discovered in AI systems can require continuous evaluation of security objectives.

AI systems are subject to all of the traditional cybersecurity vulnerabilities, exploits and attacks. They can be subject to new threats, notably those based on data, input manipulation or actions of AI models. However, these additional concerns do not supersede conventional cybersecurity concerns. As the overall cybersecurity threat landscape evolves, the need to protect AI systems against non-AI system specific threats continues to be an ongoing pursuit.

AI systems have become more widespread as organizations embrace digital transformation, resulting in an increasing likelihood of security attacks on these systems. Furthermore, as a result of the disparate and widespread use of AI systems, and especially within safety critical contexts, the result of security attacks can be severe and in some cases, lead to physical or mental repercussions for individuals.

Due to the way AI systems are developed, and their strong dependency on training data, which can include user data or unvetted data, these systems can have additional vulnerabilities compared to conventional information processing systems. Furthermore, there is a lack of understanding when it comes to the inner workings of AI, specifically when it comes to understanding and diagnosing individual instances of AI behaviour. This document provides guidance to organizations who are choosing to develop or deploy AI systems, providing awareness of the types of security threats such systems can face, and helping organizations understand the consequences of such threats and the most commonly related detection and mitigation techniques. Such awareness and understanding of this new threat landscape is imperative for securing AI systems against adversarial attacks.

This document describes various security threats and the exploitation of vulnerabilities that can happen when using AI systems. The objective is to identify recommended measures to improve the security of AI systems by identifying threats and the consequences of these threats. The identification of threats can help users and developers of AI systems better understand security issues related to AI systems.

Security threats and vulnerabilities with a lack of controls in AI systems can lead to failures, impacting both individuals and businesses.

It can be challenging to detect some threats to AI systems and thus, developing controls without effective threat detection would not be useful or adequate. This document can help identify where security threats can be effectively addressed.

In general, it is recommended that the established commonly accepted security practices in ISO/IEC 27001^[1] and ISO/IEC 27002^[2] are adhered to. However, the reliance of many AI systems on training data introduces new threats. Attacks such as data poisoning can cause undesirable outcomes by injecting malicious or unwanted data into training data. During the operational phase, AI systems gather data to support future model updates. However, this process can be vulnerable to the introduction of corrupted data, potentially compromising the integrity of future model retraining during the design and development stage. Where continuous or iterative learning occurs, data inputs supplied during operation can also become training data, so corrupted inputs can also become a vector for data poisoning. Accordingly, AI or machine-learning-enabled work, data, and control flows can benefit from some additional controls. This document can provide the necessary insights to apply both established and new controls, for example by explaining what the assets are, and what aspects need protecting.

Other attacks such as model exfiltration through querying can be mitigated by conventional information security measures such as rate limiting. Models and intellectual property can be stolen even when effective conventional information security measures are in place, since a model exfiltration attack is established through legitimate querying of the model. Threat detection is a challenge but an essential component to effectively mitigate model exfiltration. The measures used to mitigate AI threats can use a defence-in-depth strategy that leverages multiple security measures to protect an organization's assets.

If a choice is made to use unvetted or unverified data for training, risks related to data poisoning or attacks that can exploit ML data or control flows can increase.

Sample Document

get full document from standards.iteh.ai

Sample Document

get full document from standards.iteh.ai

Cybersecurity — Artificial intelligence (AI) — Addressing security threats and compromises to AI systems

1 Scope

This document addresses security threats and compromises specific to artificial intelligence (AI) systems. This document aims to provide information to organizations to help them better understand the consequences of security threats specific to AI systems, throughout their life cycle, and descriptions of how to detect and mitigate such threats.

This document is applicable to all types and sizes of organizations, including public and private companies, government entities and not-for-profit organizations that develop or use AI systems.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 22989, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

ISO/IEC 23053:2022, *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 22989, ISO/IEC 23053 and the following apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

- ISO Online browsing platform: available at <https://www.iso.org/obp>
- IEC Electropedia: available at <https://www.electropedia.org/>

3.1 availability

property of being accessible and usable on demand by an authorized entity

[SOURCE: ISO/IEC 27000:2026, 3.4]

3.2 confidentiality

property that information is not made available or disclosed to unauthorized individuals, entities, or processes

[SOURCE: ISO/IEC 27000:2026, 3.2]

3.3 integrity

property of accuracy and completeness

[SOURCE: ISO/IEC 27000:2026, 3.3]

3.4

threat

potential cause of an unwanted incident, which can result in harm to a system or organization

[SOURCE: ISO/IEC 27002:2022, 3.1.34]

3.5

attack

successful or unsuccessful unauthorized attempt to destroy, alter, disable, gain access to an asset or any attempt to expose, steal, or make unauthorized use of an asset

[SOURCE: ISO/IEC 27002:2022, 3.1.3]

3.6

attacker

person deliberately exploiting vulnerabilities in technical and non-technical security controls in order to steal or compromise information systems and networks, or to compromise availability to legitimate users of information system and network resources

[SOURCE: ISO/IEC 27033-1:2015, 3.3]

3.7

failure

termination of the ability of a product to perform a required function or its inability to perform within previously specified limits

Note 1 to entry: Based on the definition in IEEE 610.12-1990.

[SOURCE: ISO/IEC 25030:2007, A.20¹⁾]

3.8

mitigation

limitation of any negative consequence of a particular incident

[SOURCE: ISO 22300:2025, 3.1.47]

3.9

vulnerability

weakness of an asset or control that can be exploited by one or more *threats* (3.4)

[SOURCE: ISO/IEC 27002:2022, 3.1.38]

3.10

zero-knowledge attack

attack (3.5) on a system that assumes no prior knowledge about the internal information of the system under attack

Note 1 to entry: The attacker can use context or historical information to infer model vulnerability. The attacker can probe the system to identify system vulnerabilities.

3.11

perfect-knowledge attack

attack (3.5) on a system that assumes full understanding about the internal information of the system under attack

Note 1 to entry: In the case of an artificial intelligence system, full knowledge about the internal architecture, variables or current state of the model can be known by the attacker.

Note 2 to entry: This type of attack does not assume knowledge of credentials.

1) Cancelled and replaced by ISO/IEC 25030:2019.

3.12

partial-knowledge attack

attack (3.5) on a system that assumes partial knowledge about the internal information of the system under attack

Note 1 to entry: This type of attack does not assume knowledge of credentials.

3.13

data poisoning attack

attack (3.5) where the *attacker* (3.6) manipulates data to affect the artificial intelligence system's behaviour

3.14

evasion attack

adversarial attack

attack (3.5) where the *attacker* (3.6) alters inputs to induce incorrect behaviour of an artificial intelligence system in response to those inputs

Note 1 to entry: The input alterations can be imperceptible to humans.

3.15

membership inference

membership inference attack

attack (3.5) that results in an *attacker* (3.6) determining whether or not a given data sample was used in the training dataset for the model

3.16

model exfiltration attack

model exfiltration

model stealing attack

model extraction attack

attack (3.5) where the *attacker* (3.6) recreates a model with functionality similar to that of the original model through (potentially repeated) use of attacker-supplied inputs and observing corresponding outputs of the artificial intelligence system

3.17

model inversion

model inversion attack

attack (3.5) where the *attacker* (3.6) reconstructs the original training data (or significant parts thereof) by supplying inputs to an artificial intelligence system and observing corresponding outputs of the artificial intelligence system

3.18

direct model poisoning

direct model poisoning attack

attack (3.5) where the *attacker* (3.6) manipulates the model during development or operations to affect the artificial intelligence system's behaviour

3.19

direct model theft

attack (3.5) where the *attacker* (3.6) obtains the confidential model parameters from an artificial intelligence system

3.20

direct training data leak

compromise to the *confidentiality* (3.2) of training data

3.21

model input leak

compromise of the *confidentiality* (3.2) of data input to the model during operation

3.22

sensitive model output

compromise via model output of the *confidentiality* (3.2) of some portion of a model's previous input, such as training data or user prompts

3.23

prompt injection attack

prompt injection

attack (3.5) where the *attacker* (3.6) crafts content that is provided as or incorporated into an input instruction to the artificial intelligence system, triggering system behaviour unintended by the system's creators

3.24

backdoor attack

undocumented way of gaining unauthorized access to the application, system, or network while bypassing all the implemented security measures

3.25

data provenance

information on the place and time of origin, derivation or generation of a dataset, proof of authenticity of the dataset, or a record of past and present ownership of the dataset

[SOURCE: ISO/IEC 5259-1:2024, 3.1.6]

3.26

out-of-distribution data

OOD data

data containing features relevant for the artificial intelligence (AI) task, either absent or not sufficiently well represented in the training dataset, that can result in an AI error

Note 1 to entry: Out-of-distribution (OOD) refers to data or inputs that fall outside the scope of what an AI or machine learning model was trained on or is designed to handle. When an AI system encounters OOD data, it can struggle to make accurate predictions or decisions because it lacks the necessary knowledge and experience to handle such inputs effectively. OOD data can lead to unexpected or unreliable model behaviour.

[SOURCE: ISO/PAS 8800:2024, 3.2.8]

3.27

foundation model

artificial intelligence model that can be used for or readily adapted to a wide range of tasks in one or more domains

Note 1 to entry: A typical way to build a foundation model is to apply supervised machine learning or self-supervised machine learning on a large amount of data.

Note 2 to entry: A foundation model can be used as part of various applications, tasks and use cases, which do not necessarily involve generative artificial intelligence.

3.28

large language model

LLM

machine learning model that encodes the functioning of natural language with a large number of parameters and facilitates a variety of natural language processing (NLP) tasks

Note 1 to entry: Large language models can be used in a variety of NLP tasks, such as text generation, automatic summarization, machine translation, classification and more.

Note 2 to entry: Large language models can use large amounts of data and require large compute resources to build.

Note 3 to entry: The functioning of natural language can include considerations of grammar, semantics or other aspects of how natural language is used.

3.29

generative artificial intelligence

generative AI

<discipline> research and development of mechanisms, methodologies and applications of generative artificial intelligence systems

Note 1 to entry: Generative AI is a subdiscipline of artificial intelligence.

3.30

trusted execution environment

TEE

execution environment that runs alongside but is isolated from the device main operating system

[SOURCE: ISO/IEC 27071:2023, 3.3.15]

4 Abbreviated terms

AI	artificial intelligence
API	application programming interface
DoS	denial-of-service
DP	differential privacy
EOT	expectation over transformation
IDBT	input distortion-based techniques
JEA	just-enough-access
JIT	just-in-time
LLM	large language model
ML	machine learning
NIST	National Institute of Standards and Technology
OOD	out-of-distribution
PII	personally identifiable information
RAG	retrieval augmented generation
RT	random transformations

5 Application of information security

5.1 General

Attacks on AI systems can occur throughout the AI system life cycle, as described in ISO/IEC 22989 and ISO/IEC 5338.^[11] ISO/IEC 22989 provides AI concepts and terminology and ISO/IEC 5338^[11] describes AI system life cycle processes. Attacks can be classified as attacks during the development and training of AI systems or during the inference or operation of AI systems, or both. AI systems are susceptible to attacks considered conventional to all types of information systems. [Annex B](#) includes examples of such attacks that can have specific considerations for an AI system, including specific consequences resulting from such attacks.

As with any other software system, an AI system's security relies on its hardware, software and networking stack, as well as the associated organizational, people, physical and technological controls (see ISO/IEC 27002:2022,^[2] Clauses 5 to 8). When considering AI system or model security and developing associated control sets, it is important to consider the relevant context and appropriately harden and control associated assets. Indeed, it is possible that securing the AI system's software stack can be a more efficient starting point for security investment. For example, if loss of model confidentiality is a concern, securing the stack against direct model theft would likely be a more effective and efficient starting point than more resource-intensive controls meant to mitigate against prompt- or API-based exfiltration attempts.

Conventional security attacks on information processing systems can still affect AI systems and thus, information security controls, such as those described in ISO/IEC 27002,^[2] shall be implemented, based on the needs of the organization, the information security systems, and the risks identified as a result of carrying out an information security risk assessment. The organization should consider the views and key concepts outlined in ISO/IEC 27002:2022,^[2] 4.2, including: control types, information security properties, cybersecurity concepts, operational capabilities and security domains. Information security is commonly defined using the CIA triad – confidentiality, integrity, and availability – though organizations can have specific security-related objectives and priorities for their AI systems based on organizational requirements (see ISO/IEC 27001:2022,^[1] 6.1).

AI systems should be considered as software components that are usually integrated into another system, application or data file rather than as a standalone AI or ML model. Therefore, security measures applied to software and applications also apply to AI systems in all life cycle stages, such as design, deployment or using an AI system in operation.

The implementation of best practices to protect against information security incidents is paramount for the protection of AI systems. However, additional or adaptations of security measures can be required to adequately address the unique threats to AI systems. Such security measures can be incorporated as part of risk management of the AI system, as described in ISO/IEC 23894^[12] or relevant ETSI documents.^[13] ^[14]^[15] These measures can be used to mitigate identified risks as part of a responsible AI framework for developing or using AI systems such as the AI management system defined in ISO/IEC 42001.^[16] Diligence should be used when navigating management system standards for both security (e.g. ISO/IEC 27001^[1]) and AI (e.g. ISO/IEC 42001^[16]), and AI security standards (e.g. ETSI EN 304 223^[14]). Examples of existing measures that can be useful in the detection and mitigation of security attacks on AI systems include integrity checks on datasets and denial-of-service mitigation. Logging and monitoring events related to the operation of the AI system can aid in the analysis of undesirable and anomalous system and user behaviour (see ISO/IEC 27002:2022,^[2] 8.15 and 8.16). Verification that the AI model has not been tampered with is also important to ensure validity of operational results and avoid unwanted bias.^[98] Ensuring datasets are accurately labelled and classified can help to appropriately handle and secure input datasets. Such datasets should be appropriately secured with access controls and data loss prevention techniques applied as necessary. Management and operation of the AI system, like other information systems, should include processes such as threat modelling, security testing, and vulnerability management.

Security controls described in ISO/IEC 27002,^[2] such as those related to secure system architecture and engineering principles as described in ISO/IEC 27002:2022,^[2] 8.27, should be implemented throughout the AI system life cycle. [Annex A](#) provides further information on mapping attacks on AI systems to the AI system life cycle and to assets. Potential AI specific compromises, security attacks and vulnerabilities including but not limited to those specified in [Table 1](#) shall be taken into account throughout the AI system life cycle. Actors across the supply chain can consider potential threats relevant to their function in the supply chain throughout the entire AI system life cycle (these considerations can be applied to A.6.2.3 and A.6.2.6 specified in ISO/IEC 42001:2023,^[16] Table A.1). The secure software development framework (SSDF)^[17], published by the National Institute of Standards and Technology (NIST) can be used for processes to better secure software, including AI systems^[18].

It is recommended to consider the threats according to the system being dealt with using a risk-based approach. This document provides examples, which are not exhaustive.

5.2 AI governance programme

To ensure that AI systems and their use within an organization are subject to proper governance (including security), an AI governance programme should be established and implemented. This involves taking

organizational responsibility for AI by maintaining an inventory of AI systems and their use, performing risk analysis on them, and managing those risks (see ISO/IEC 38507^[19]).

For guidance on implementation, see ISO/IEC 42001^[16].

NOTE ISO/IEC 42001^[16] is a management system standard that focuses on the management processes that support responsible AI in an organization.

5.3 Zero trust principles

As part of a holistic approach to secure system engineering throughout the AI system life cycle, zero trust principles can effectively protect AI systems. For each of the attacks described in this document, the zero trust principles that are effective in helping to protect AI systems are included in the mitigation measures associated for each attack described in [Clause 6](#).

ISO/IEC 27002:2022,^[2] 8.27 includes zero trust principles to help ensure information systems are securely designed, implemented and operated within the development life cycle. These principles can be summarized as:

- a) assuming the organization's information systems are already breached;
- b) employing a "never trust and always verify" approach for access to information systems;
- c) ensuring requests to information systems use end-to-end encryption;
- d) verifying each request to an information system as if it originated externally;
- e) using least privilege access control, including access to interacting with the model;
- f) always authenticating requesters and validating authorization requests.

Applying zero trust principles to AI systems is essential throughout their unique life cycle – from training data management, to model deployment, to agent operations. While traditional zero trust approaches focus on network security, AI systems introduce new assets, access patterns, and threat surfaces that require specialized security considerations.

Organizations can apply zero trust principles to AI systems across the AI system life cycle, for example:

- Training data protection: it is assumed that training datasets can contain adversarial examples or poisoned data. Verification processes are implemented to validate data provenance and integrity before incorporation into training pipelines. Zero trust principles are applied to data engineering workflows, verifying the source and content of each dataset regardless of its apparent origin.
- Model access controls: fine-grained authorization is implemented for model interactions, whether through API calls, application integrations or retrieval-augmented generation (RAG) systems. Each inference request is authenticated and authorized based on the specific model capabilities being accessed, helping to prevent privilege escalation through model interactions.
- AI application boundaries: strict isolation is enforced between AI applications and their underlying data sources. When implementing RAG systems, each document retrieval request is authenticated and authorized independently, preventing data leakage through inference results even when the application itself is authorized.
- Agent action verification: for AI agents with capabilities to trigger external actions, each action is verified and authorized independently, rather than granting broad permissions to the agent. Just-in-time access is implemented with appropriate scoping for agent operations, ensuring that permissions expire after task completion.
- Continuous monitoring: behavioural monitoring for AI systems is deployed to establish baselines for normal query patterns, response characteristics, and resource utilization. Deviations from these patterns can indicate compromise or abuse even when authentication appears valid.