Muliarevych, O. (2026). Cybersecurity and privacy risks of AI assistants in academic institutions: A risk classification and governance framework. ESS Conference Series Social Science and Humanities, 1(1). https://doi.org/10.54658/esscssh.2026.1.1.05

Cybersecurity And Privacy Risks Of Ai Assistants In Academic Institutions: A Risk Classification And Governance Framework

Oleksandr Muliarevych
Lviv Polytechnic National University
Lviv, Ukraine
ORCID iD: 0000-0002-4644-7962
oleksandr.v.muliarevych@lpnu.ua

Abstract.
Large language model-based AI assistants are rapidly becoming part of academic work, supporting students, teachers, researchers, and administrators in writing, summarisation, translation, coding, and information retrieval. However, their adoption in academic institutions creates cybersecurity and privacy risks that extend beyond the commonly discussed problem of academic integrity. Using a qualitative literature review and conceptual synthesis of recent research on generative AI in higher education, LLM security, privacy protection, and institutional AI governance, this article classifies the main risks of AI assistants into four groups: data-related risks, model-related risks, application and integration risks, and human and governance risks. It shows that universities process sensitive categories of information — student records, grades, unpublished research, examination materials, administrative documents — which makes uncontrolled AI use a significant institutional risk. Special attention is given to prompt injection, sensitive data disclosure, insecure integrations, overreliance on AI outputs, and policy fragmentation. The article proposes an Academic AI Security and Privacy Governance Framework based on six layers: governance and policy; data classification and privacy; secure AI architecture; prompt and output security; human oversight and AI literacy; and monitoring, audit, and continuous improvement. Because each layer translates external legal and regulatory requirements — data-protection obligations, accountability, auditability — into institutional rules and technical controls, the framework is offered as a contribution to the governance of digital technology in education as much as to information security practice. The study concludes that AI assistants should be treated as security- and privacy-sensitive socio-technical systems requiring coordinated institutional governance.

Keywords: *AI assistants; large language models; AI governance; data protection; higher education; prompt injection.*

1.Introduction

Artificial intelligence assistants based on large language models (LLMs) are rapidly becoming part of academic work. Students use them for explanation, summarisation, writing support, translation, coding, and examination preparation. Teachers use them to prepare educational materials, generate examples, assess drafts, and support communication with students. Researchers use them for literature exploration, editing, coding, data interpretation, and administrative work. AI assistants are therefore no longer optional productivity tools. In many academic institutions they are becoming an informal layer between users and knowledge, educational content, research data, and institutional processes.

The adoption of AI assistants in universities, however, creates risks that go beyond academic integrity. The most visible discussion is usually focused on plagiarism, cheating, and authorship. These issues are important, but they represent only one part of the problem. AI assistants also introduce cybersecurity and privacy risks, because they process unstructured prompts, uploaded files, personal data, source code, research materials, student records, and institutional information. When such tools are integrated with learning management systems (LMS), email, cloud storage, research repositories, or administrative services, the potential risk surface becomes significantly broader. This is especially relevant for academic institutions, because universities process many sensitive categories of information: student grades, personal identifiers, disability accommodations, research data, unpublished manuscripts, examination materials, grant proposals, employment data, and sometimes commercially or nationally sensitive research results. UNESCO guidance on generative AI in education and research warns that public access to generative AI developed faster than regulatory and institutional safeguards, leaving many institutions insufficiently prepared to address data privacy, human oversight, and responsible use (Miao & Holmes, 2023).

From a cybersecurity perspective, LLM-based assistants differ from traditional educational software because natural-language input can act simultaneously as content and as an instruction to the model. The OWASP Top 10 for LLM and Generative AI Applications 2025 identifies prompt injection, sensitive information disclosure, supply-chain weaknesses, data and model poisoning, improper output handling, excessive agency, system-prompt leakage, vector and embedding weaknesses, misinformation, and unbounded consumption as major risk classes (OWASP Foundation, 2025). In universities, these risks become especially consequential when an assistant processes institutional documents, student data, research outputs or untrusted external content, or when it is connected to retrieval systems and action-taking tools.

The privacy dimension is equally important. A report prepared for the European Data Protection Board under its Support Pool of Experts programme emphasises that LLM-based systems require systematic identification, assessment, and mitigation of privacy and data-protection risks, and it develops one of its worked use cases around an LLM system for monitoring and supporting student progress — showing that LLM privacy risks are not abstract technical concerns but concrete institutional governance problems (European Data Protection Board, 2025). The problem is therefore socio-technical. It is not enough to ask whether AI assistants are useful for learning or whether they improve productivity. Academic institutions must also ask what data enters these systems, who controls the processing, whether outputs can leak sensitive information, how users verify generated content, how institutional policies define acceptable use, and how technical controls limit damage in case of misuse.

Posed in those terms, the question is not primarily one of information-security engineering but of institutional governance. Universities routinely process personal data and may process special categories of personal data; they also make decisions affecting admission, assessment, progression, employment and discipline. The precise legal role of an institution, vendor or platform depends on jurisdiction, deployment architecture, contractual allocation of responsibilities and the relevant data flow. The governance problem is therefore to alloca-

te responsibility, identify the lawful and proportionate conditions for processing, preserve meaningful human oversight in consequential decisions, and make AI-assisted operations sufficiently traceable to support accountability and contestability. The six-layer framework developed below translates those concerns into institutional policy, privacy, architecture, security, literacy and audit controls. Read in this way, an institutional AI policy is not merely an IT procedure but part of the organisation's internal governance of digital technology.

This article analyses the cybersecurity and privacy risks of AI assistants in academic institutions by combining recent literature from education, cybersecurity, privacy law, AI governance, and LLM security. Its main goal is to connect two areas that are usually discussed separately: university AI policy and technical LLM security. The article argues that universities need a complete risk-management approach combining AI literacy, data protection, secure system design, vendor checks, clear university policies, and human oversight. Without this integration, universities risk treating AI assistants only as a threat to academic honesty while ignoring their risks as systems that process sensitive data and support decisions. The contribution of this paper is practical and integrative. It does not consist in inventing basic security controls such as data minimisation, least privilege, prompt filtering, or human oversight; these come from well-established frameworks and guidance produced by OWASP, NIST, the EDPB, and UNESCO. The original contribution lies in adapting them specifically for universities through four outputs:

(1) a four-part classification of AI-assistant risks in academia;

(2) a mapping of risk sources, affected university data, and institutional impact;

(3) a six-layer Academic AI Security and Privacy Governance Framework;

(4) practical instruments including a data-classification model, secure architecture patterns, an implementation roadmap, and a prioritised risk matrix.

By connecting existing security standards with real university workflows, the article offers a model that academic institutions can apply directly.

2. Literature Review

2.1. Generative AI adoption and the institutional policy gap in higher education

The first major research stream concerns the rapid adoption of generative AI in higher education and the institutional difficulty of responding to it. UNESCO's guidance for generative AI in education and research provides one of the most influential policy-level contributions. It frames generative AI as a technology with educational potential but also with significant risks related to human agency, privacy, inclusion, ethics, and institutional readiness (Miao & Holmes, 2023). The guidance matters for this article because it does not treat AI as a purely pedagogical innovation; it positions AI adoption as a governance challenge requiring regulation, institutional policy, and data protection.

A related UNESCO document, the AI competency framework for teachers, argues that teachers need specific knowledge, skills, and values to use AI responsibly, and notes that many countries have not yet clearly defined the AI competencies teachers require, which leaves educators without sufficient guidance in a fast-changing technological environment (Miao

& Cukurova, 2024). This is relevant because many cybersecurity and privacy failures in academic AI use arise not from technical weakness but from users not knowing what data may safely be entered into an AI assistant, what outputs require verification, and which tools are approved for institutional use.

Several empirical and policy-analysis studies show that universities are still developing their response to generative AI. Moorhouse et al. (2023) analysed guidelines from top-ranked universities and showed that institutions were forced to create or modify assessment guidance quickly after the emergence of ChatGPT; their work is useful because it shows that early university responses were often focused on assessment, academic honesty, and acceptable use, while broader security and privacy controls were less central. An et al. (2025) studied generative AI guidelines and policies in higher education institutions across teaching, learning, research, and administration, supporting the argument that AI assistants are becoming cross-functional institutional tools requiring more systematic governance than classroom-level rules. Wilson (2025) and Wu et al. (2024) similarly show that policy development differs significantly between institutions and that effective governance requires coordination between academic, technical, administrative, and legal units.

The Ukrainian context is also important. Borodiyenko et al. (2025) investigated opportunities and risks of AI-based applications in research in Ukrainian universities. Their study shows that AI adoption in Ukrainian academic environments is shaped by expert opinion, media narratives, and scientific literature, while institutional readiness and risk awareness remain developing areas. This provides regional relevance for the present article and supports the need for practical governance models.

2.2. LLM SECURITY RISKS AND THEIR RELEVANCE TO ACADEMIC INSTITUTIONS

The second research stream comes from cybersecurity and LLM application security. The OWASP 2025 taxonomy is particularly useful because it reflects a shift from early plugin-centred descriptions to a broader application lifecycle: prompt injection and sensitive-information disclosure remain central, while improper output handling, excessive agency, system-prompt leakage, vector/embedding weaknesses, misinformation and unbounded consumption capture risks that become more important as AI assistants acquire retrieval and action capabilities (OWASP Foundation, 2025).

Prompt injection deserves special attention. The UK National Cyber Security Centre argues that it is not equivalent to classical SQL injection, because LLMs do not enforce a boundary between data and instructions, so the risk can be reduced but not eliminated in the deterministic way that traditional injection vulnerabilities can (National Cyber Security Centre, 2025b); the mechanism and its academic implications are examined in Section 4.4. The point is especially important for universities, because academic workflows are rich in untrusted text: AI assistants may process student submissions, web pages, PDF files, research papers, emails, forum posts, and external datasets, any of which may carry hidden instructions. If the assistant has access to institutional systems, the consequences may include data exposure, incorrect grading support, leakage of internal documents, or unauthorised actions.

Technical surveys reinforce this conclusion. Yao et al. (2024) review LLM security and privacy from the perspective of beneficial uses, harmful uses, and inherent vulnerabilities. Das et al. (2025) provide a broader survey of LLM security and privacy challenges, including jailbreaking, prompt injection, data poisoning, and leakage of personally identifiable information. These studies support the claim that AI assistants should be assessed as cybersecurity-relevant systems, especially when connected to institutional data and services.

The NCSC's broader assessment of AI and cyber threats warns that AI will increase the effectiveness and accessibility of cyber operations, including social engineering, intrusion activity, and exploitation workflows (National Cyber Security Centre, 2025a). Academic institutions are especially exposed to this trend, because they often combine open networks, diverse user groups, decentralised IT practices, bring-your-own-device usage, international collaboration, and valuable research data.

2.3. Privacy and data-protection risks of AI assistants

The third research stream focuses on privacy and data protection. This is central to academic institutions because they manage large amounts of personal and confidential information. The European Data Protection Board's 2025 report on LLM privacy risks provides a risk-management methodology and discusses collection, inference, retrieval-augmented generation, feedback loops and emerging agentic systems. It also makes an important governance distinction: LLM-specific mitigation guidance can support data protection by design and security of processing, but it does not replace a Data Protection Impact Assessment where one is legally required (European Data Protection Board, 2025). Opinion 28/2024 further cautions against assuming that an AI model trained on personal data is automatically anonymous (European Data Protection Board, 2024).

Technical privacy surveys provide additional depth. Neel and Chang (2024) review privacy issues in LLMs, including training-time privacy, inference-time leakage, model deletion, and red-teaming. Miranda et al. (2025) survey privacy-preserving approaches for LLMs, including anonymisation, differential privacy, and machine unlearning. These works help explain why privacy risk is not limited to what users intentionally share in prompts: it can also arise from memorisation, logs, embeddings, and downstream integrations.

Empirical research shows that users are increasingly aware of these risks. Xu et al. (2025) investigated learners' perceptions of data privacy when using AI language models and found concerns about data leakage, unethical data exploitation, and algorithmic bias, while also indicating that students may not know how AI systems process or retain their data. Xue et al. (2025) compared AI privacy concerns in higher education through news coverage in China and Western countries, showing that AI privacy in education is a global concern.

Dahabiyeh et al. (2026) studied privacy awareness in the context of ChatGPT use among university students and showed that privacy awareness and perceptions of data handling influence users' intention to use generative AI. Taken together, these findings suggest that institutional AI adoption must include transparency and user education, not only access to tools.

2.4. Academic integrity, research integrity, and institutional risk

Academic integrity remains one of the most visible concerns in generative AI adoption. Bittle and El-Gayar (2025) provide a systematic review of generative AI and academic integrity in higher education, identifying both opportunities and risks, including improved engagement and efficiency but also increased risk of academic dishonesty and the need for digital literacy, detection tools, and institutional policy.

Chang et al. (2024) approach ChatGPT in higher education from a risk-management perspective, discussing academic integrity, critical thinking, and workforce readiness, and showing that institutions need to move beyond simple prohibition or acceptance. Their approach is useful here because cybersecurity and privacy should likewise be treated as risk--management problems rather than isolated technical issues.

Research integrity is another important dimension. The European Commission's living guidelines on the responsible use of generative AI in research emphasise reliability, honesty, respect, accountability, transparency, privacy, confidentiality, and protection of intellectual property (European Commission, Directorate-General for Research and Innovation, 2026). This is directly connected to cybersecurity and privacy, because a researcher may upload unpublished data to an external AI assistant, use AI-generated text containing fabricated references, or rely on generated summaries of sensitive documents.

Academic integrity and cybersecurity should therefore not be treated as separate domains. In AI-assisted academic work they overlap: a hallucinated citation may damage research quality; an uploaded confidential dataset may violate privacy rules; an AI-generated examination question stored in a third-party system may create assessment-security issues; and an unverified AI recommendation may affect student outcomes.

2.5. Human factors and educator perspectives

The fifth research stream concerns human factors. Even strong technical controls will not be sufficient if users do not understand AI risks or if institutional rules are unclear. Pikhart and Al-Obaydi (2025) studied educators' subjective perspectives on the risks of AI in higher education; their findings highlight concerns about privacy, academic integrity, data misuse, unauthorised access, plagiarism, and reduced critical thinking. UNESCO's AI competency framework for teachers supports this point from the competency side, arguing that educators need competencies spanning a human-centred mindset, ethics, AI foundations, pedagogy, and professional development (Miao & Cukurova, 2024).

For the purposes of this article this means that cybersecurity and privacy training should not be confined to IT security departments. Teachers and researchers also need practical AI security literacy: how to classify data, when not to use public AI tools, how to verify outputs, how to disclose AI use, and how to recognise unsafe AI integrations.

2.6. Governance frameworks and the remaining research gap

Recent governance-oriented sources provide useful foundations for institutional AI risk management. The NIST Generative Artificial Intelligence Profile (NIST AI 600-1) adapts the AI Risk Management Framework to generative AI and organises risk work around the functions Govern, Map, Measure and Manage (Autio et al., 2024). It is cross-sectoral rather than education-specific, which makes institutional adaptation necessary. UNESCO contributes human-centred and competency-oriented guidance for education (Miao & Holmes, 2023; Miao & Cukurova, 2024), while the EDPB adds privacy-specific governance requirements. The EU AI Act adds a further regulatory benchmark: certain systems used in education and vocational training can be classified as high-risk because of their potential effect on access, learning outcomes, level assessment or test monitoring (European Parliament & Council of the European Union, 2024). These sources are complementary rather than interchangeable: security taxonomies identify technical risk, privacy guidance structures data-protection analysis, and institutional governance determines how controls are assigned and enforced.

The reviewed literature shows a clear gap. Higher-education studies usually focus on academic integrity, assessment, policy development, and educator or student perceptions. Technical LLM-security studies focus on prompt injection, data leakage, model vulnerabilities, and adversarial misuse.

Privacy and legal sources focus on data protection, governance, and compliance. Fewer works combine these perspectives into a practical institutional model for academic AI assistants. This article addresses that gap by framing AI assistants in academic institutions as socio-technical systems that are simultaneously learning tools, data-processing systems, cybersecurity interfaces, and governance challenges. Its regulatory and technical foundations are drawn from existing sources; what is proposed here is the domain-specific synthesis set out in Section 1 — the classification, its alignment with the academic domain, and the integrated implementation model.

3. Methodology

3.1. Research design and source selection strategy

This study uses a qualitative literature review and conceptual synthesis approach to analyse the cybersecurity and privacy risks of AI assistants in academic institutions. Its purpose is not to test a single technical vulnerability or to measure the performance of a specific AI tool. The aim is to synthesise recent academic, technical, legal, and policy-oriented literature in order to identify the main categories of risk that arise when LLM-based AI assistants are used in universities and other academic environments. A literature-based methodology is appropriate because the problem is interdisciplinary, spanning the technical, educational, legal, and governance literatures reviewed in Section 2; the study combines those perspectives into a unified institutional risk model.

The methodology follows three stages:

1. identification of relevant sources;
2. thematic analysis of the selected literature; and

3. synthesis of the findings into a structured cybersecurity and privacy risk framework for academic institutions.

The literature search prioritised publications and institutional documents issued between 2020 and August 2026, with earlier sources retained where they provided enduring conceptual foundations. The search covered four categories: peer-reviewed studies on generative AI in higher education; systematic reviews and surveys on LLM security and privacy; primary policy and regulatory documents; and technical risk frameworks and cybersecurity guidance. Search combinations included terms such as „generative AI higher education privacy", „AI assistants academic institutions cybersecurity", „LLM prompt injection education", „AI governance universities", „student data privacy AI language models" and „LLM sensitive information disclosure". Because the study is a qualitative integrative review rather than a systematic review or meta-analysis, the search was designed for interdisciplinary conceptual coverage and source triangulation, not exhaustive bibliometric retrieval.

Sources were included if they directly analysed generative AI, ChatGPT, or LLM-based tools in higher education; discussed cybersecurity risks of LLM applications; analysed privacy or data-protection risks of AI systems; proposed governance, policy, or risk-management approaches for responsible AI use; provided evidence about student, teacher, or institutional perceptions of AI-related risks; or offered a framework applicable to academic institutions. Sources were excluded if they focused only on technical model performance without discussing security, privacy, or institutional use; addressed AI in education only from a narrow pedagogical perspective without risk analysis; discussed cybersecurity in general without connection to AI, LLMs, or data governance; were opinion pieces without clear analytical, empirical, technical, or policy value; or were outdated in relation to current generative AI development.

3.2. Analytical approach and risk-classification logic

The selected sources were analysed using thematic coding, focusing on repeated risk categories, institutional challenges, and mitigation strategies across the literature. In the first stage, sources were grouped into four broad domains: AI in higher education, LLM security, privacy and data protection, and governance and risk management. In the second stage, these domains were compared in order to identify overlapping issues. Academic-integrity literature, for example, often discusses student misuse of AI tools, while cybersecurity literature discusses prompt injection and manipulation; these topics appear different, but both involve user behaviour, system trust, verification, and control of AI-generated outputs. Similarly, privacy literature discusses personal data leakage while institutional-policy literature discusses acceptable use and governance, and the two overlap as soon as students or staff upload confidential materials to external AI assistants. In the third stage, the recurring themes were synthesised into risk categories relevant to academic institutions: input data leakage, sensitive information disclosure in outputs, prompt injection, insecure integration with institutional systems, vendor data-processing risks, academic and research integrity risks, student privacy and surveillance risks, overreliance on AI-generated outputs, and policy fragmentation.

The risk categories were then classified along three dimensions: source of risk, affected asset, and institutional impact. The source of risk describes whether a risk originates in user behaviour, model behaviour, system integration, vendor data practices, or institutional policy gaps. The affected asset describes what may be harmed, including student data, research data, examination materials, unpublished manuscripts, institutional documents, learning management systems, staff accounts, academic reputation, and trust in educational processes. The institutional impact describes the possible consequence, such as privacy violations, data leakage, academic misconduct, reputational damage, regulatory non-compliance, compromised research integrity, unauthorised system actions, or reduced quality of decision-making.

3.3. Ethical considerations and limitations

This study is based on publicly available academic publications, policy documents, technical frameworks, and regulatory guidance. It does not involve human participants, interviews, surveys, experiments with students, or the processing of personal data, and no formal ethics committee approval was therefore required. The subject of the article is nevertheless ethically sensitive, because it concerns student privacy, research confidentiality, responsible AI use, and institutional accountability.

The methodology has several limitations. First, the study rests on literature review and conceptual synthesis rather than on empirical measurement; the framework proposed below has not been implemented or evaluated in a live institutional setting. Second, the field is developing rapidly, and new AI models, regulations, institutional policies, and attack techniques appear frequently. Third, the selected literature reflects mainly English-language sources and internationally visible publications. Fourth, cybersecurity and privacy risks differ depending on the technical deployment model — a public chatbot, an enterprise AI assistant, a locally hosted model, or a retrieval-augmented university knowledge assistant each present a different risk profile.

4. Results And Discussion: Classification Of Cybersecurity And Privacy Risks

4.1. General classification logic

The analysis of recent literature shows that the cybersecurity and privacy risks of AI assistants in academic institutions are multidimensional. They cannot be reduced to plagiarism, cheating, or inaccurate answers. AI assistants create a broader socio-technical risk landscape, because they interact with user prompts, uploaded files, institutional documents, student data, research materials, external services, and sometimes internal university systems.

The risks are accordingly classified across four connected dimensions: data-related risks, model-related risks, application and integration risks, and human and governance risks. Figure 1 presents the classification. It is broadly consistent with current OWASP 2025 risk categories while deliberately reorganising them around university assets, workflows and

responsibilities rather than reproducing a generic technical taxonomy (OWASP Foundation, 2025).

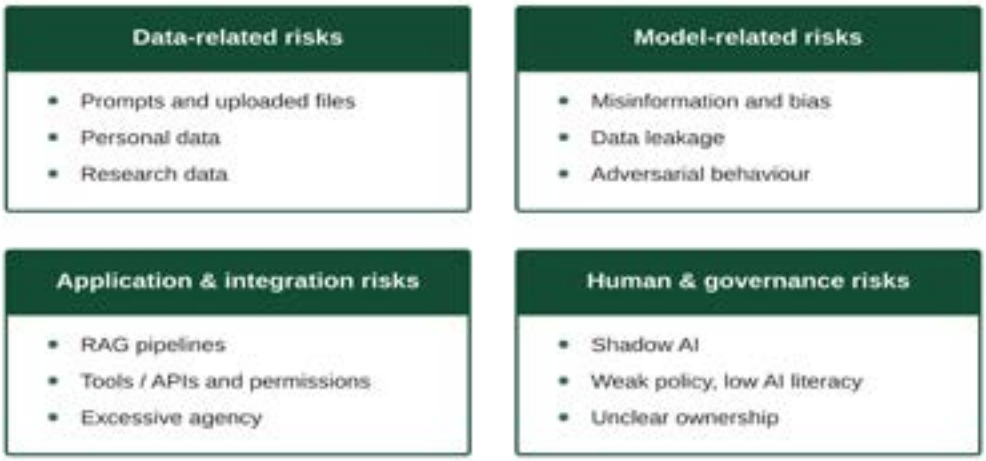

Figure 1. Classification of cybersecurity and privacy risks of AI assistants in academic institutions

4.2. Data-related risks

Data-related risks appear when students, teachers, researchers, or administrative staff enter sensitive information into AI assistants. This may include student names, grades, disability accommodations, unpublished research, examination materials, internal meeting notes, grant proposals, or confidential university documents. These risks are especially important when public or third-party AI systems are used without institutional control. Table 1 summarises the principal data-related risks, with academic examples and possible institutional impacts.

Table 1. Data-related risks of AI assistants in academic institutions

Risk category	Description	Academic example	Possible impact
Input data leakage	Sensitive data is entered into an external AI assistant.	A lecturer uploads student essays with names and grades to a public chatbot.	Privacy violation, GDPR risk, reputational damage.
Research confidentiality loss	Unpublished research materials are processed by third-party AI tools.	A researcher uploads a draft grant proposal or experimental dataset.	Loss of novelty, exposure of intellectual property, breach of project agreement.
Examination material exposure	Examination questions or assessment rubrics are entered into AI tools.	A teacher asks an AI assistant to improve examination questions.	Assessment compromise, unfair student advantage.
Data retention uncertainty	Users do not know whether prompts or files are stored or reused.	Staff use free AI tools for administrative summaries.	Lack of control over institutional information.
Model memorisation risk	Personal or confidential data may be memorised or reproduced.	Fine-tuning is performed on student records without sufficient safeguards.	Data-protection violation, future leakage.
Weak data classification	The institution has no clear rule on what data may be used with AI.	Different departments follow different informal practices.	Inconsistent compliance and unmanaged risk.

A report published by the European Data Protection Board on LLM privacy risks is directly relevant here, because it proposes a risk-management methodology for LLM systems and includes a use case on monitoring and supporting student progress (European Data Protection Board, 2025). This confirms that educational LLM systems are not merely pedagogical tools but also privacy-sensitive data-processing systems.

The most important practical conclusion is that academic institutions need a clear data-classification model for AI use: not all data should be treated equally. The categories in Table 2 are a proposed institutional control model rather than legal classifications. Their purpose is to translate sensitivity, contractual obligations and potential harm into operational rules for approved tools, access control and escalation. Responsibility for enforcement should be allocated to appropriate institutional roles — for example data owners, information-security teams, legal/data-protection functions and system administrators — rather than assumed to belong to a single technical unit.

Table 2. Suggested data classification for AI use in academic institutions

Data class	Examples	AI usage recommendation
Public	Published syllabi, public course descriptions, open educational resources.	Usually allowed.
Internal	Internal teaching materials, non-confidential methodological notes.	Allowed only in approved tools.
Confidential	Student grades, internal reports, unpublished research drafts.	Restricted; use only controlled institutional systems.
Highly sensitive	Health and disability data, disciplinary records, examination banks, personal identifiers.	Prohibited in public AI tools; requires formal approval and safeguards.

4.3. Model-related risks

Model-related risks arise from the behaviour of the LLM itself. They include hallucinations, incorrect citations, fabricated sources, biased outputs, unsafe recommendations, privacy leakage, and overconfident responses. In academic institutions such risks may affect teaching, student support, research writing, administrative decisions, and assessment. Table 3 sets them out.

Table 3. Model-related risks of AI assistants in academic institutions

Risk category	Description	Academic example	Possible impact
Hallucination	AI generates false or unsupported information.	A student receives fabricated references for a research paper.	Poor academic quality, misinformation.
Fabricated citations	AI invents sources or misrepresents existing publications.	A teacher uses an AI-generated bibliography without verification.	Research integrity violation.
Bias and unfairness	AI output reflects biased training data or assumptions.	AI gives feedback of different quality depending on language style.	Unequal treatment of students.
Unsafe recommendations	AI provides harmful, misleading, or non-compliant advice.	AI gives incorrect legal, medical, or psychological guidance to students.	Harm to users, institutional liability.

Risk category	Description	Academic example	Possible impact
Memorisation and leakage	Model reproduces sensitive content from training data or context.	AI reveals parts of uploaded confidential documents.	Data breach, confidentiality loss.
Explainability gap	Users cannot understand why a recommendation was produced.	AI suggests academic intervention for a student without transparent reasoning.	Accountability and fairness concerns.
Prompt injection and jailbreaks	Malicious or manipulated input changes the behaviour of the AI assistant or forces it to ignore intended instructions.	A student submits a document containing hidden instructions that manipulate an AI-assisted grading or feedback tool.	Unauthorised disclosure, incorrect feedback, bypass of institutional rules, or unsafe system actions.

Model-related issues are not limited to technical error. In an academic environment, a single hallucinated answer, fabricated citation, or biased recommendation can directly affect the quality of learning, assessment, research writing, or administrative decision-making — the more so because LLM outputs are fluent and authoritative in style, which may lead users to trust them without sufficient verification. These risks should therefore be managed through mandatory human oversight, verification of generated references, clear rules for using AI-generated content, and careful limitation of AI use in high-impact academic decisions. AI assistants may support educational and research processes, but they should not replace academic judgement, expert review, or institutional accountability. In sensitive scenarios — student assessment, academic intervention, research conclusions, administrative recommendations — AI-generated outputs should be treated as advisory and must be reviewed by a responsible human user before being acted on. Hallucination and overconfidence are thus not only technical risks but educational and organisational ones.

Prompt injection appears in Table 3 as a model-related and adversarial-interaction risk, but it is discussed separately below because it is not only a model issue. It is also an application-security and governance issue, and effective mitigation requires more than improved prompting: it requires input filtering, separation of trusted and untrusted content, least-privilege tool access, output validation, logging, and human approval for sensitive actions.

4.4. Prompt injection and adversarial manipulation

Prompt injection is one of the most important technical risks for AI assistants. It occurs when malicious or untrusted text manipulates the behaviour of an LLM by causing the model to treat data as instructions. This is especially relevant in academic institutions, because AI assistants may process untrusted content such as student submissions, uploaded PDF files, web pages, emails, forum posts, and research documents. The UK National Cyber Security Centre explains that prompt injection differs from classical SQL injection because LLMs do not maintain a robust internal distinction between data and instructions: in an LLM prompt, both trusted instructions and untrusted content are processed as token sequences, which makes complete mitigation difficult (National Cyber Security Centre, 2025b). In an academic environment, a student could insert hidden instructions into an essay submitted to an AI-assisted grading-support tool; a malicious web page could contain instructions that manipulate a research assistant summarising online content; an uploaded PDF file could

include text instructing the AI assistant to reveal system prompts, ignore policy constraints, or access unrelated documents.

In earlier work, the present author proposed an additional filter layer for LLM-based systems (Muliarevych, 2024). That architecture includes a prompt analyser and an attack validator module that use an LLM to classify prompts before they reach the core model, and it evaluates several protection strategies, including core GPT models without additional protection, fuzzy-search-based filtering, and validator-based approaches. The approach is relevant to academic AI assistants because it provides a concrete technical mitigation pattern that can be adapted to university tools. Building on that model, the present article adapts the filter-layer concept to the academic context: user prompts, student submissions, uploaded documents, and external content should pass through a prompt-analysis and validation stage before reaching the core model, and suspicious inputs should be blocked, logged, quarantined, or escalated for human review.

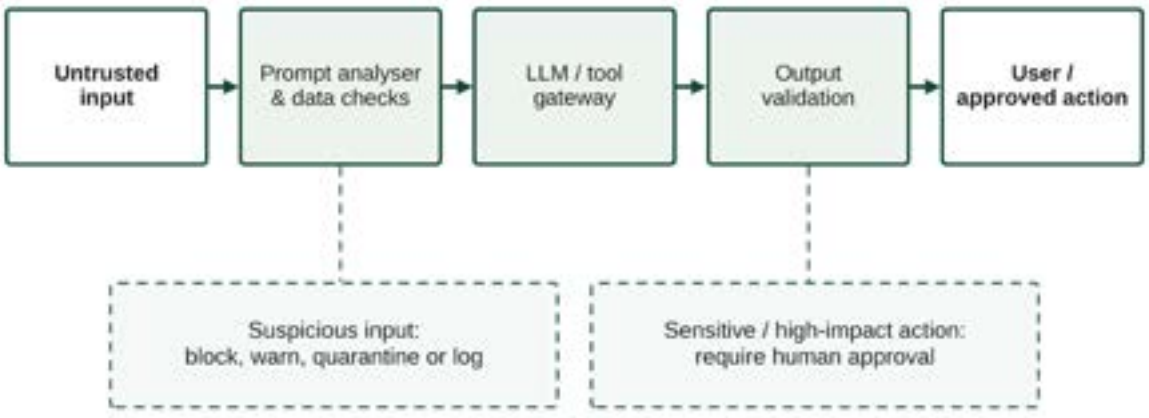

Figure 2. Prompt-injection defence pattern for AI assistants in academic institutions, adapted from the filter-layer approach proposed by Muliarevych (2024)

The key design principle, illustrated in Figure 2, is that a university AI assistant should not process untrusted content directly in a privileged execution context. It should instead apply layered controls: input pre-processing, data classification, prompt-injection detection, separation of trusted instructions from untrusted content, least-privilege tool access, output validation, logging, anomaly detection, and human approval for sensitive actions. This defence should not, however, be presented as a complete solution. Prompt injection remains difficult to eliminate because the model processes instructions and data within the same language-based context. Academic institutions should therefore assume residual risk and focus not only on detecting prompt injection but also on limiting the damage if an attack succeeds. AI assistants should have restricted permissions, controlled access to institutional systems, and no ability to perform high-impact actions without human confirmation.

4.5. Application and integration risks

The risk level increases significantly when AI assistants are integrated with institutional systems. A standalone assistant that only answers general questions has a narrower impact surface than one connected to an LMS, email, document repositories, student-information systems, research databases or action-taking tools. In the OWASP 2025 taxonomy, the relevant risks are captured especially by prompt injection, sensitive information disclosure, improper output handling, excessive agency, system-prompt leakage and vector/embedding weaknesses (OWASP Foundation, 2025). Table 4 sets out common academic integration scenarios.

Table 4. Application and integration risks of AI assistants in academic institutions

Integration scenario	Risk	Example
AI assistant + learning management system (LMS)	Unauthorised access to course or student data.	Assistant retrieves grades or private submissions without proper access control.
AI assistant + email	Indirect prompt injection through emails.	Malicious email instructs AI to forward confidential information.
AI assistant + document repository	Sensitive-information disclosure or permission bypass.	AI summarises documents outside the user's permission scope.
AI assistant + research database	Exposure of unpublished/restricted data; vector or embedding weaknesses.	Research assistant retrieves sensitive data for an unauthorised user.
AI assistant + administrative workflow	Excessive agency or unintended automated actions.	AI creates, approves, or modifies forms without human confirmation.
AI assistant + code repository	Source-code and credential leakage.	AI assistant exposes internal scripts, API keys, or configuration files.

Integration risks appear when an AI assistant moves from a passive question-answering role to an active institutional component connected to university systems, where it may retrieve, summarise, modify, or transmit information. The main security concern is therefore not only the behaviour of the LLM itself but also the level of access granted to the assistant and the quality of controls around each integration point. Each scenario in Table 4 creates a different type of institutional risk, and code repositories are the most sensitive of them. AI assistants should accordingly not be integrated with institutional systems through direct and unrestricted access. Each integration should pass through a controlled tool or API gateway, permission checks, logging, and, where needed, human approval. The assistant should operate under the same access-control rules as the user, and preferably under stricter limitations for high-risk actions. This reduces the probability that an AI assistant becomes a shortcut around existing university security policies.

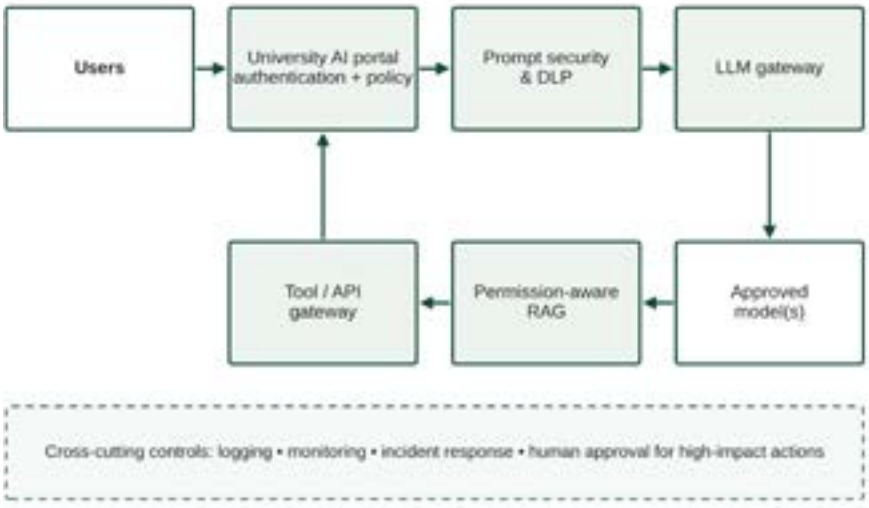

Figure 3. Secure architecture pattern for university AI assistants

The architecture also includes a permission-aware retrieval-augmented generation (RAG) knowledge base and a tool/API gateway, including Model Context Protocol (MCP) or similar tool-integration mechanisms, so that the assistant retrieves institutional knowledge or interacts with university systems only through controlled interfaces. This matters because the AI assistant must not bypass existing access-control rules: if a user does not have permission to access a document or system directly, the AI assistant should not expose that information indirectly. The output validation layer checks generated responses before they are returned to the user and can help detect sensitive information disclosure, unsupported claims, unsafe recommendations, or policy violations. Logging, monitoring, and incident response then provide operational visibility into AI assistant usage, blocked attempts, suspicious behaviour, and possible data-leakage incidents. The design principle underlying the whole pattern is that an AI assistant should be treated as a security-sensitive component of the university information system rather than as an application added on top of it.

4.6. Human and governance risks

Human and governance risks are often underestimated. Even a technically secure AI assistant can be misused if users do not understand its limitations or if institutional rules are unclear. Table 5 sets out the principal governance failures and their institutional consequences.

Table 5. Human and governance risks of AI assistants in academic institutions

Governance risk	Description	Institutional consequence
No AI usage policy	Users choose tools independently.	Shadow AI usage and uncontrolled data sharing.
Policy focused only on plagiarism	Security and privacy are ignored.	Hidden institutional risk.
No data-classification rules	Users do not know what may be entered into AI tools.	Accidental leakage of sensitive data.

Governance risk	Description	Institutional consequence
No approved tool list	Departments use different public AI services.	Vendor and compliance fragmentation.
No auditability	AI-assisted actions are not logged.	Difficult incident investigation.
Weak AI literacy	Students and staff do not understand hallucination, privacy, or prompt injection.	Unsafe and uncritical AI use.
No ownership model	Responsibilities of IT, legal, academic departments, and data-protection roles are unclear.	Delayed response and inconsistent enforcement.

As Table 5 shows, these risks arise not from technical failure but from unclear institutional responsibility and inconsistent user behaviour. Governance of AI assistants should therefore be distributed rather than assigned to IT departments alone. University leadership, cybersecurity teams, legal or data-protection officers, academic boards, research ethics committees, teaching staff, and student representatives should jointly define rules for acceptable AI use, approved tools, accountability, training, and incident response. This reduces fragmentation and makes AI adoption more predictable, transparent, and secure.

5. The Academic AI Security and Privacy Governance Framework

5.1. Framework overview

On the basis of the classification set out above, this article proposes a practical framework, the Academic AI Security and Privacy Governance Framework, designed for universities and academic institutions that use or plan to use AI assistants in teaching, research, administration, or student support. The framework has six layers: governance and policy; data classification and privacy; secure AI architecture; prompt and output security; human oversight and AI literacy; and monitoring, audit, and continuous improvement. It is presented in Figure 4.

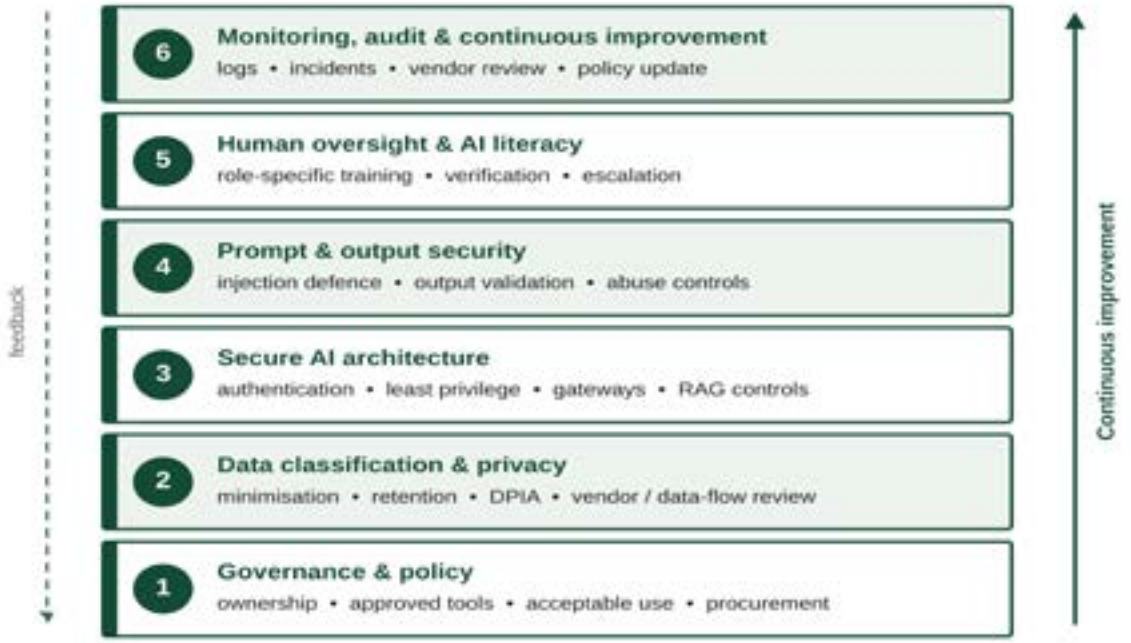

Figure 4. Academic AI Security and Privacy Governance Framework, with layered controls and a continuous-improvement feedback loop

The framework is organised as a layered model because each layer depends on the one before it: policy determines what is permitted, data classification determines what may be processed, architecture enforces both in technical controls, and the remaining three layers protect, inform, and audit what results. The feedback arrow in Figure 4 is important, because AI governance should not be treated as a one-time implementation activity. AI models, institutional use cases, regulations, and attack techniques change rapidly, and lessons learned from monitoring and incidents should continuously improve policies, data-classification rules, technical controls, and user training. The six layers are described in turn below.

5.2. Layer 1: Governance and policy

The first layer defines institutional rules for the use of AI assistants. Without it, users will independently adopt public AI tools, creating shadow AI usage and uncontrolled data flows. The governance layer should include an institutional AI acceptable-use policy; lists of approved and prohibited AI tools; rules for using AI in teaching, assessment, research, and administration; authorship and citation rules; vendor approval processes; and incident-reporting procedures. For institutions subject to or aligning with European regulation, the policy should also distinguish ordinary low-impact uses from educational AI systems whose purpose may bring them within the AI Act's high-risk categories, such as systems used to determine access, evaluate learning outcomes, assess educational level or monitor prohibited behaviour during tests (European Parliament & Council of the European Union, 2024).

5.3. Layer 2: Data classification and privacy

The second layer focuses on data protection. Academic institutions should classify data before deciding whether it may be processed by AI assistants, using an operational model such as Table 2. Controls should include data minimisation; anonymisation or pseudonymisation where appropriate; review of purpose, legal basis and retention; vendor and data-flow assessment; transparency to affected users; and a Data Protection Impact Assessment where the applicable legal threshold is met. LLM-specific mitigations should support, not substitute for, that assessment (European Data Protection Board, 2025).

5.4. Layer 3: Secure AI architecture

The third layer concerns technical architecture. Its key principle is that AI assistants should be deployed through controlled institutional gateways rather than through uncontrolled direct use of public tools for sensitive tasks. Controls should include authentication, role-based access control, AI gateways, data-loss prevention, retrieval control, API gateways, least privilege, human approval, and logging, along the lines set out in Figure 3.

5.5. Layer 4: Prompt and output security

The fourth layer addresses prompt injection, jailbreaks, sensitive-information disclosure, improper output handling and other adversarial or unsafe interactions. It should include prompt-injection testing and detection, separation and labelling of trusted and untrusted

context, sensitive-data checks, output validation, deterministic checks around tool calls, rate and resource limits, abuse detection, and explicit human approval for consequential actions. These controls reduce risk but should not be described as eliminating prompt injection; residual risk must be constrained by least privilege and impact limitation.

5.6. Layer 5: Human oversight and AI literacy

The fifth layer addresses people, since technical controls are not sufficient if users do not understand the risk. Academic institutions should provide AI literacy training for students, teaching staff, researchers, administrative staff, IT teams, and academic managers. Training should cover data privacy, hallucination, citation verification, disclosure of AI use, the basics of prompt injection, safe handling of documents, and incident reporting.

5.7. Layer 6: Monitoring, audit, and continuous improvement

The sixth layer ensures that AI governance does not become static. AI models, tools, regulations, and attack methods change quickly. Universities therefore need logging of AI assistant usage, records of blocked prompt-injection attempts, detection of sensitive-data leakage, audit of vendor terms, regular review of approved AI tools, incident-response procedures, and periodic policy updates informed by incident reports and user feedback.

5.8. Implementation roadmap

While the six layers define the logical structure of AI security and privacy governance, universities also need a practical sequence for implementation; the framework should be translated into concrete institutional actions rather than left as a theoretical model. Table 6 sets out a proposed roadmap.

Table 6. Proposed institutional implementation roadmap

Stage	Action	Expected result
Stage 1	Identify current AI tool usage across departments.	Visibility of shadow AI usage.
Stage 2	Define data classes and prohibited AI use cases.	Clear rules for sensitive data.
Stage 3	Approve institutional AI tools and vendors.	Reduced uncontrolled third-party exposure.
Stage 4	Create an AI acceptable-use policy.	Common rules for students and staff.
Stage 5	Deploy an AI gateway or controlled access model.	Centralised technical control.
Stage 6	Add prompt-injection and data-leakage filters.	Reduced technical risk.
Stage 7	Train staff and students.	Improved AI security literacy.
Stage 8	Establish monitoring and incident response.	Operational risk management.
Stage 9	Review and update policy every semester.	Continuous adaptation.

The roadmap follows a risk-based logic. Implementation begins with visibility and governance rather than with technology, because an institution cannot secure AI use effectively if it does not know which tools are already in use, what data is processed, and who is responsible for decisions; unmanaged or „shadow" AI usage creates immediate privacy

and compliance risks. Implementation is not completed when a policy is published or a tool deployed: it requires continuous monitoring, incident response, and regular policy updates as AI technologies, institutional practices, and cybersecurity threats evolve.

5.9. Prioritising risks: a practical risk matrix

Universities also need a practical way to prioritise the risks identified above. Table 7 therefore provides an illustrative qualitative matrix, not an empirically estimated probability model. The likelihood and impact labels are expert-informed screening categories intended to support local workshops; institutions should recalibrate them using their own incident history, architecture, data inventory, deployment model and regulatory context.

Table 7. Practical AI assistant risk matrix for academic institutions

Risk	Likelihood	Impact	Recommended treatment
Students use AI for undisclosed academic work	High	Medium	Academic-integrity policy, AI disclosure rules, assessment redesign.
Staff upload student data to public AI tools	Medium	High	Data-classification rules, approved tools, privacy training.
Researchers upload unpublished manuscripts or datasets	Medium	High	Research AI policy, confidentiality guidance, secure institutional AI tools.
Prompt injection manipulates AI assistant behaviour	Medium	High	Prompt filtering, least privilege, tool isolation, monitoring.
AI assistant leaks sensitive institutional information	Medium	High	Access control, permission-aware retrieval, output validation.
AI-generated hallucinations affect learning or research	High	Medium	Verification rules, citation checks, human review.
AI assistant performs unintended action through plugins	Low–Medium	High	Human approval, deterministic controls, API gateway.
Vendor stores or reuses confidential data	Medium	High	Vendor assessment, contractual controls, enterprise settings.
Fragmented AI governance between departments	High	Medium–High	Central AI governance committee and policy.
Users overtrust AI outputs	High	Medium	AI literacy, warnings, verification practices.

The matrix shows that the most critical risks are not always the most technically complex. Several high-priority risks — staff uploading student data to public AI tools, researchers using external AI systems for unpublished materials, fragmented AI governance between departments — are primarily organisational and policy-related, while technical risks such as prompt injection, sensitive-information leakage, and unintended tool-based actions require architectural controls, monitoring, and human approval. Risk treatment should accordingly be layered rather than uniform, and the framework should be implemented not as a single technical solution but as a coordinated institutional governance model.

5.10. Implications for stakeholder groups

The practical meaning of these risks differs across stakeholder groups. University leadership, teachers, students, researchers, and IT and cybersecurity teams interact with AI as-

sistants in different ways and therefore carry different responsibilities.

University leadership should treat AI assistant adoption as a strategic governance issue rather than an educational trend. The use of AI assistants affects institutional reputation, legal compliance, research quality, data security, and student trust, and responsibility for AI governance should therefore not be delegated solely to individual teachers or IT specialists.

Teachers need clear rules about when AI use is allowed, when it must be disclosed, and how assessment should adapt. Traditional written assignments may become less reliable as evidence of independent student work if they can be completed almost entirely by AI. Assessment may need to incorporate oral defence, process-based assessment, reflective explanation, practical tasks, in-class components, version history, project logs, personalised assignments, and explicit declarations of AI use.

Students should understand that AI assistants can be useful learning tools but that unsafe or undisclosed use can create academic and privacy problems. They should be informed that AI-generated text may contain errors, that AI may fabricate references, that entering personal or confidential data into public tools may be unsafe, and that AI use may need to be disclosed.

Researchers should treat external AI assistants as third-party processing services that require a deliberate data-flow and contractual assessment before confidential or personal data are submitted. Whether a vendor legally acts as a processor, independent controller or in another role depends on the concrete service, purposes, contractual terms and applicable law. As a practical rule, unpublished manuscripts, sensitive datasets, peer-review materials, grant proposals, industry-funded work and protected intellectual property should be used only with tools and configurations approved for those data.

IT and cybersecurity teams should expand their threat models to include AI-specific risks. Traditional controls such as authentication, network security, and endpoint protection remain necessary but do not address LLM-specific issues. Additional controls should include prompt-injection testing, detection of sensitive data in prompts and uploaded files, enforcement of access control in retrieval-augmented systems, monitoring of AI tool usage, vendor assessment, abuse detection, secure logging, isolation of AI tools from high-risk systems, and human approval for sensitive actions.

The central proposal of this article is that academic institutions should treat AI assistants as security- and privacy-sensitive socio-technical systems rather than as productivity or learning tools. Policy without architecture is insufficient: a university may publish AI guidelines, but if users continue uploading sensitive data to uncontrolled public tools, the risk remains unmanaged. Architecture without literacy is also insufficient: a secure AI gateway helps, but students and staff still need to understand privacy, hallucination, prompt injection, and academic-integrity risks. And detection without impact limitation is insufficient: prompt-injection filters are useful, but the system must also apply least privilege, human approval, access control, and monitoring, because prompt injection cannot be assumed to be fully eliminated. The six layers of the proposed framework are designed to be adopted together, and it is their combination that allows academic institutions to support responsible AI adoption while reducing cybersecurity, privacy, and governance risks.

6. Conclusion

AI assistants based on large language models are becoming part of everyday academic life. They support students, teachers, researchers, and administrators with writing, summarisation, translation, coding, explanation, and information retrieval. Their use in academic institutions nevertheless creates cybersecurity and privacy risks that go far beyond academic integrity.

The article has shown that the risks of AI assistants in universities can be organised into four major groups: data-related risks, model-related risks, application and integration risks, and human and governance risks. The classification is intentionally institution-centred rather than a duplicate of OWASP: it maps technical vulnerabilities and model limitations onto academic assets, workflows and accountability. Current OWASP 2025 categories — including prompt injection, sensitive information disclosure, improper output handling, excessive agency, system-prompt leakage, vector and embedding weaknesses and misinformation — are therefore treated as inputs to, rather than substitutes for, the proposed university risk model (OWASP Foundation, 2025).

A key finding is that AI assistants should not be treated only as educational tools. In academic institutions they are security- and privacy-sensitive socio-technical systems: they process data, influence academic work, support decisions, and may interact with institutional infrastructure. Universities therefore need more than general advice about responsible AI use; they need a structured institutional response.

The proposed Academic AI Security and Privacy Governance Framework addresses this need through six layers: governance and policy; data classification and privacy; secure AI architecture; prompt and output security; human oversight and AI literacy; and monitoring, audit, and continuous improvement. The framework integrates higher-education AI governance, technical LLM security, privacy protection, and institutional risk management.

The central conclusion is that safe AI adoption in academic institutions requires balance. Universities should neither ignore AI assistants nor rely on prohibition alone; equally, they should not allow uncontrolled use of public AI systems for sensitive academic work. A risk-based approach is required, in which public and low-risk educational content may be suitable for AI-assisted processing while personal data, confidential research, examination materials, and institutional records require strict controls.

Three implications follow for governance and regulatory practice, and they are the reason the analysis belongs to the study of law and digital technologies rather than to information security alone. First, the institutional AI policy is emerging as the operative regulatory instrument in this field. International guidance and general data-protection law set the objectives; what determines whether student data actually reaches a third-party model on a given Tuesday is a university's acceptable-use rule, its approved-tool list, and the configuration of its gateway. Regulators and ministries that wish to influence outcomes should therefore attend to the quality and enforceability of these subordinate instruments, and not only to the adequacy of the primary norms. Second, accountability requires auditability, and auditability requires design. A university cannot demonstrate compliance, respond to a subject access request, or investigate an incident if AI-assisted actions leave no record; the sixth

layer of the framework is thus a legal requirement expressed as a technical one, and it should be treated as such in procurement and in vendor contracts. Third, the residual character of prompt injection has a governance consequence that is easily missed. Where a class of attack cannot be eliminated, the institution's obligation shifts from prevention to the limitation of consequences — least privilege, human confirmation of high-impact actions, and a decision not to deploy where the residual risk is intolerable. For universities in jurisdictions that are aligning their data-protection and digital regulation with European standards, including Ukraine, the practical task is to build these three commitments into institutional rules now, while AI assistants are still being adopted, rather than to retrofit them after the first serious incident. Sustainable governance of academic AI is, on this view, less a matter of new legal norms than of institutional capacity to apply the norms that already exist.

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability

No new datasets were generated or analysed for this conceptual review. The study is based on publicly available academic literature, regulatory documents and technical guidance cited in the References.

Ethics Statement

This study is based on the analysis of publicly available academic publications, policy documents, regulatory guidance, and technical frameworks. It does not involve human participants, personal data processing, interviews, surveys, or experiments. Formal ethics committee approval was therefore not required.

Declaration of Generative AI Use

During the preparation of this manuscript, the author used the local LLM gpt-oss-20b for English-language editing, grammar checking, and improvement of wording. The author reviewed, edited, and verified all outputs, and retains full responsibility for the manuscript's intellectual content, analysis, conceptual framework design, and evaluation of the literature.

References

An, Y., Yu, J. H., & James, S. (2025). Investigating the higher education institutions' guidelines and policies regarding the use of generative AI in teaching, learning, research, and administration. International Journal of Educational Technology in Higher Education, 22. https://doi.org/10.1186/s41239-025-00507-3

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

Bittle, K., & El-Gayar, O. (2025). Generative AI and academic integrity in higher education: A systematic review and research agenda. Information, 16(4), Article 296. https://doi.org/10.3390/info16040296

Borodiyenko, O., Drach, I., Bazeliuk, N., Petroye, O., Reheilo, I., Bazeliuk, O., & Slobodianiuk, O. (2025). Opportunities and risks of using AI-based applications in research: The case of Ukrainian universities. Information Technologies and Learning Tools, 105(1), 125–143. https://doi.org/10.33407/itlt.v105i1.5794

Chang, V., Ansari, Y., & Arami, M. (2024). ChatGPT in higher education: A risk management approach to academic integrity, critical thinking, and workforce readiness. In Proceedings of the 6th International Conference on Finance, Economics, Management and IT Business. https://doi.org/10.5220/0012764100003717

Dahabiyeh, L., Taha, N., Thneibat, M., & Bhat, M. A. (2026). Privacy awareness in generative AI: The case of ChatGPT. Interactive Technology and Smart Education, 23(1), 25–48. https://doi.org/10.1108/ITSE-01-2025-0009

Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and privacy challenges of large language models: A survey. ACM Computing Surveys, 57(6), Article 152, 1–39. https://doi.org/10.1145/3712001

European Commission, Directorate-General for Research and Innovation. (2026). Living guidelines on the responsible use of generative AI in research (3rd ed.). European Commission. https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/updated-era-living-guidelines-responsible-use-generative-ai-research-2026-05-08_en

European Data Protection Board. (2024). Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en

European Data Protection Board. (2025). AI privacy risks & mitigations: Large language models (LLMs). https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/ai-privacy-risks-mitigations-large_en

Miao, F., & Cukurova, M. (2024). AI competency framework for teachers. UNESCO. https://www.unesco.org/en/articles/ai-competency-framework-teachers

Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Miranda, M., Ruzzetti, E. S., Santilli, A., Zanzotto, F. M., Bratières, S., & Rodolà, E. (2025). Preserving privacy in large language models: A survey on current threats and solutions. Transactions on Machine Learning Research. https://arxiv.org/abs/2408.05212

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. Computers and Education Open, 5, Article 100151. https://doi.org/10.1016/j.caeo.2023.100151

Muliarevych, O. (2024). Enhancing system security: LLM-driven defense against prompt injection vulnerabilities. In 2024 IEEE 17th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET) (pp. 420–423). IEEE. https://doi.org/10.1109/TCSET64720.2024.10755823

National Cyber Security Centre. (2025a). Impact of AI on cyber threat from now to 2027. https://www.ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027

National Cyber Security Centre. (2025b). Prompt injection is not SQL injection: It may be worse. https://www.ncsc.gov.uk/blog-post/prompt-injection-is-not-sql-injection

Neel, S., & Chang, P. (2024). Privacy issues in large language models: A survey. arXiv. https://doi.org/10.48550/arXiv.2312.06717

Novelli, C., Casolari, F., Hacker, P., Spedicato, G., & Floridi, L. (2024). Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. Computer Law & Security Review, 55, Article 106066. https://doi.org/10.1016/j.clsr.2024.106066

OWASP Foundation. (2025). OWASP Top 10 for large language model applications. https://owasp.org/www-project-top-10-for-large-language-model-applications/

Pikhart, M., & Al-Obaydi, L. H. (2025). Reporting the potential risk of using AI in higher education: Subjective perspectives of educators. Computers in Human Behavior Reports, 18, Article 100693. https://doi.org/10.1016/j.chbr.2025.100693

Temper, M., Tjoa, S., & David, L. (2025). Higher Education Act for AI (HEAT-AI): A framework to regulate the usage of AI in higher education institutions. Frontiers in Education, 10, Article 1505370. https://doi.org/10.3389/feduc.2025.1505370

Wilson, T. D. (2025). The development of policies on generative artificial intelligence in UK universities. IFLA Journal, 51(3), 722–734. https://doi.org/10.1177/03400352251333796

Wu, C., Zhang, H., & Carroll, J. M. (2024). AI governance in higher education: Case studies of guidance at Big Ten universities. Future Internet, 16(10), Article 354. https://doi.org/10.3390/fi16100354

Xu, X. S., Liu, J., Zheng, R., Lei, V. N.-L., & An, Q. (2025). Learners' perception of data privacy when using AI language models: Reflective diary analysis of undergraduates in China. Acta Psychologica, 260, Article 105491. https://doi.org/10.1016/j.actpsy.2025.105491

Xue, Y., Chinapah, V., & Zhu, C. (2025). A comparative analysis of AI privacy concerns in higher education: News coverage in China and Western countries. Education Sciences, 15(6), Article 650. https://doi.org/10.3390/educsci15060650

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model security and privacy: The good, the bad, and the ugly. High-Confidence Computing, 4(2), Article 100211. https://doi.org/10.1016/j.hcc.2024.100211