

<https://doi.org/10.62405/osiconf.2026.49>

Haisha Oleksandr,

Associated Professor of Engineering Technologies Department,
Pylyp Orlyk International Classical University, Mykolaiv, Ukraine,
ORCID: 0000-0003-3711-547X; e-mail: oleksandrgaysha1982@gmail.com

ENSURING INFORMATION PROTECTION WHEN WORKING WITH MODERN ARTIFICIAL INTELLIGENCE TOOLS

ABSTRACT. *The study aims to systematize information security risks arising from the use of modern artificial intelligence tools and to substantiate the feasibility of deploying such tools locally for processing restricted-access information. The methodology combines structural and functional analysis, classification of information flows, comparative analysis of local and cloud deployment models, and assessment of information security threats. The results systematize the types of data provided to AI chatbots, AI workflows, and AI agents in personal and corporate environments and identify risks associated with transferring such data to external services. The analysis shows that local model execution, when combined with appropriate network and administrative controls, can keep data processing within a controlled security perimeter. The study concludes that local AI solutions should be prioritized for confidential information, while recognizing the need for adequate hardware resources, qualified personnel, integration skills, and security procedures.*

KEYWORDS: *artificial intelligence; information security; local AI systems; AI agents; confidential information; corporate data.*

INTRODUCTION. The rapid spread of generative artificial intelligence (AI) is changing the way information is handled. AI chatbots are no longer used solely for searching for information or generating answers to questions. They are also applied to document preparation, file analysis, programming, translation, summarization of large volumes of text, preparation of educational materials, and support for managerial decision-making. The next stage has been the spread of AI workflows, in which a model acts as a component of a sequence of automated operations, as well as AI agents capable of independently planning tasks, accessing external tools, and working with files, databases, and web resources. As AI autonomy increases, so does the amount of information provided to it to establish the necessary context.

The problem is that practical AI adoption often outpaces the development of systematic information protection policies. An employee may provide a chatbot with a contract fragment, internal correspondence, source code, a financial document, or personal data without perceiving such an action as transferring information outside the organization. In a corporate environment, the situation becomes more complex because an AI system may gain access not to a single document but to an entire set of internal resources. OWASP identifies sensitive information disclosure and prompt injection among the key risks associated with LLM and generative AI applications (OWASP, 2025).

At the same time, claims about the danger of cloud AI should not be reduced to the assumption that every external service automatically uses submitted data to train models. Corporate product policies differ. For example, OpenAI states that business data are not used to train models by default, Microsoft describes a similar regime for enterprise Copilot, and Google establishes restrictions on the use of Workspace data for training generative models without customer permission (OpenAI, 2026; Microsoft, 2026; Google, 2026). Therefore, the issue should be considered more broadly by assessing the entire information life-cycle, transmission channels, logs, backups, administrative access, integrations, jurisdiction, and possible configuration errors.

An unresolved part of the problem is the comprehensive comparison of the information actually provided to modern AI tools in personal and corporate environments with threat models for local and cloud deploy-

ment. The aim of this study is to systematize these information flows, identify the main threats, and substantiate the organizational and technical feasibility of using local AI tools to process information requiring an increased level of protection.

METHODS. The study is based on structural and functional analysis and comparative analysis of contemporary AI use cases. The object of analysis is the information flow between the user, the AI system, and connected tools. Three functional categories are used for systematization: AI chatbots, AI workflows, and AI agents. Two additional dimensions are introduced: the deployment environment, namely local or cloud-based, and the nature of the task, namely personal or corporate.

Risk assessment was conducted according to the stages of the data lifecycle: query formation, transmission, processing, temporary or long-term storage, use of the result, transfer of the result to other systems, and deletion. For external systems, provider-related risks were considered separately, while for local systems the analysis covered the risks of compromising workstations, servers, network segments, accounts, models, integrations, and agent tools. The conclusions were based on current NIST and OWASP recommendations and publicly available information provided by AI platform vendors.

A key methodological position is the distinction between local model execution and a local security system. The former only means that computation is performed on controlled hardware. The latter involves network segmentation, authentication, authorization, logging, backup, updates, file access control, restrictions on external connections, and other security measures. Therefore, local installation does not eliminate risks by itself; rather, it transfers a substantial part of the responsibility for managing them to the organization.

RESULTS AND DISCUSSION. Analysis of practical scenarios shows that modern AI tools cover a broad range of personal tasks. In domestic activities, these include drafting correspondence, travel planning, analysis of personal documents, work with educational materials, planning, translation, image processing, and advisory queries. In professional activities, individual users apply AI to prepare reports, analyze contracts, program software, conduct market research, prepare presentations and meeting minutes, summarize documents, and automate repetitive operations. In many cases, the user provides the AI system not only with the immediate query but also with contextual information required to produce a useful answer.

In the corporate environment, the information scale increases substantially. AI can be integrated with electronic document management systems, email, calendars, CRM and ERP systems, document repositories, code management systems, knowledge bases, and analytical platforms. Such integration turns AI from an isolated assistant into a component of the information infrastructure. In the case of an agent, an additional level of risk emerges because the agent not only receives data but may also initiate actions involving that data.

Table 1. Main categories of information that users may provide to personal AI tools

Category	Examples of information	Typical tasks	Potential consequence of leakage
Personal details	Full name, contacts, biographical data	Form completion, correspondence, profiles	Privacy violations, fraud
Documents	Applications, contracts, academic papers, certificates	Analysis, editing, summarization	Disclosure of personal or internal data
Work information	Correspondence, reports, presentations	Text preparation, analysis	Leakage of internal information
Financial information	Budgets, invoices, payment documents	Analysis, planning	Financial and reputational losses
Intellectual property	Code, manuscripts, technical solutions	Programming, editing, research	Loss of control over intellectual outputs
Credentials	API keys, configurations, tokens	Configuration and automation	Unauthorized access

As shown in Table 1, even individual use of AI generates a substantial body of data that may have economic or legal value. Particularly dangerous are queries that simultaneously contain identifying information, documents, and technical data. In the case of agents, credentials and permissions to perform actions may be added to this information.

Table 2. Main categories of corporate information that may be processed by modern AI tools

Category	Examples	AI scenarios	Criticality level
Trade secrets	Strategies, negotiations, pricing	Analytics, decision support	High
Personal data	Employee and customer data	Search, classification, summarization	High
Financial data	Budgets, reporting, forecasts	Analytics, modeling	High
Contracts and legal information	Contracts, claims, legal opinions	Comparison, risk identification	High
Technical data	Architecture, configurations, logs	Diagnostics, documentation	High
Source code	Repositories, code fragments	Development, testing	High
R&D and IP	Formulas, inventions, manuscripts	Research, patent analysis	High
Internal knowledge bases	Instructions, standards, policies	RAG, enterprise search	Medium/High

Table 2 demonstrates that corporate AI processes data that may simultaneously constitute personal data, commercially valuable information, and subject matter related to intellectual property. Consequently, a policy of simply prohibiting the transfer of secrets to an ordinary chat is insufficient. Organizations need to define the permitted class of AI system, access rules, and the acceptable level of context for each information resource.

Local and Cloud-Based AI Deployment Models

Local tools include Ollama, LM Studio, and self-hosted n8n installations, which can serve as components of a local automation environment. Ollama is intended for running models locally, while LM Studio supports fully local execution of downloaded models, work with documents, and a local server without requiring a permanent Internet connection (Ollama, 2026; LM Studio, 2026). n8n can be deployed on an organization's own server; at the same time, administrators must consider telemetry settings and independently ensure the security of the installation (n8n, 2026).

Cloud and online solutions include ChatGPT, Microsoft Copilot, Google Gemini, Claude, Perplexity, and other services that provide access to models through web interfaces or APIs. It should be emphasized that the concept of cloud-based AI is not synonymous with insecure AI. Enterprise editions of leading platforms employ encryption, access control, auditing, and contractual restrictions concerning data use. Nevertheless, the architecture itself involves transferring information to the infrastructure of an external provider, thereby creating an additional trust boundary (OpenAI, 2026; Microsoft, 2026; Google, 2026; Anthropic, 2026).

For local systems, the primary risk is compromise of the controlled environment itself. If a model server is placed inside the internal network and access is restricted through network segmentation, authentication, and authorization, the information protection task largely corresponds to the traditional task of protecting a corporate network perimeter. Additional measures are required for access control to models and contextual repositories, log protection, software updates, and control over connected tools.

AI agents represent an important exception. Even a locally deployed agent may have permissions to access a browser, file system, command interpreter, databases, or APIs. If security mechanisms do not account for the agent-based nature of interaction, the model may generate a sequence of actions that goes beyond the administrator's initial intent. Particular attention should be paid to indirect prompt injection, in which a malicious instruction is contained not in the user's direct request but in a document, web page, or another source processed by the agent. OWASP identifies prompt injection and sensitive information disclosure as current risks for LLM applications (OWASP, 2025).

Threats Associated with Transferring Corporate Data to External AI Systems

Transferring corporate information to cloud AI creates several consecutive risk zones. The first occurs during query preparation and transmission. Secure encryption of the transport channel substantially reduces the probability of passive interception, but it does not eliminate the risk of endpoint compromise, theft of an access token, routing errors, incorrect proxy configuration, or use of a compromised client. Thus, TLS is necessary but not sufficient as a security measure.

The second zone concerns the provider's server infrastructure. Even when a provider applies modern security mechanisms, the organization loses direct control over the operating system, hardware, internal segmentation, logs, backups, and administrative processes. Risk is determined not only by the possibility of an external cyberattack but also by human errors, software vulnerabilities, misconfiguration, and the provider's chain of subcontractors.

The third zone is the lifecycle of transferred data. It is necessary to determine whether prompts and responses are stored, for how long, who can access them, and whether they are used for abuse monitoring, analytics, service improvement, or model training. Policies differ substantially between services. For example, OpenAI states that business data are not used to train models by default, while certain data may be retained for security and monitoring purposes; Microsoft and Google also describe separate enterprise data protection regimes (OpenAI, 2026; Microsoft, 2026; Google, 2026). Therefore, a decision concerning the admissibility of a particular service should be based on its actual policy, contractual terms, and configuration rather than merely on the platform's general brand.

The fourth zone arises from integrations. If a cloud-based agent has access to email, corporate storage, CRM, or other systems, compromise of an account or a malicious instruction may turn an information leak into an active operation. In such an architecture, protecting the model alone is insufficient. Agent permissions should be restricted according to the principle of least privilege, separate accounts and tokens should be used, access to tools should be segmented, and critical operations should be logged.

The fifth zone is related to intellectual property. Transferring manuscripts, source code, research and development results, inventions, or unpublished technical solutions may create legal and economic consequences regardless of whether the provider uses such data to train a model. The mere transfer to a third party may be relevant to confidentiality status, contractual obligations, or an organization's internal policy.

Feasibility of Local Deployment

Local deployment should not be viewed as a universal replacement for cloud services. Its primary advantage is the ability to control the information route. Data can remain inside the corporate network, while the model, contextual repository, and workflow can be hosted on infrastructure controlled by the organization itself. For tasks involving trade secrets, personal data, internal information, source code, or research and development results, this significantly reduces the number of external components that must be trusted.

A practical local environment may include Ollama or LM Studio as model execution environments, n8n as a workflow automation tool, a local vector store for RAG, a corporate authentication server, a logging system, and a network segment with controlled access. LM Studio explicitly supports working with local models and documents without transmitting input data outside the device (LM Studio, 2026), while Ollama

describes local execution in which the company does not receive prompts or data when the system is used locally (Ollama, 2026).

Local deployment also has disadvantages. Servers or workstations with sufficient CPU, RAM, and, for large models, GPU resources are required. Qualified specialists are needed for deployment, updating, and monitoring; complex scenarios may require programming of integrations; and the organization must independently provide backup, access control, protection against malicious software, and recovery from failures. In addition, a local model may be inferior to the largest cloud models for some particularly complex tasks.

Nevertheless, these disadvantages are largely manageable. An organization can determine which models are permitted, which data are available to a particular process, which integrations are allowed for an agent, which network connections are blocked, and which events must be audited. In the cloud model, a significant part of these questions is transferred to the sphere of trust in an external provider. Therefore, for high-value information, it is advisable to proceed from a presumption of local processing and permit external services only after a separate risk assessment and verification of contractual and technical safeguards.

CONCLUSIONS. The study demonstrates that modern AI chatbots, AI workflows, and AI agents process a much broader range of information than traditional search and reference systems. In the personal environment, AI may receive documents, work materials, financial and personal information, source code, and credentials. In the corporate environment, this is supplemented by trade secrets, personal data of customers and employees, contracts, financial reporting, technical documentation, internal knowledge bases, and research and development results.

For local AI systems, provided that the architecture is properly designed, the core information protection task consists of controlling the corporate information environment: network segmentation, authentication, authorization, logging, access control to files and models, component updates, and restrictions on external connections. At the same time, AI agents require an additional control model because their ability to use tools creates the possibility of performing actions that were not explicitly anticipated by the user.

Cloud AI systems are not inherently insecure, and enterprise editions of leading platforms may provide a substantial set of security controls. However, using such systems for critical information creates an additional external trust boundary associated with data transmission, server infrastructure, logs, backups, data retention policies, integrations, subcontractors, and jurisdictional requirements. Therefore, conclusions concerning the security or insecurity of a particular platform should be based on its actual architecture and conditions of use.

Accordingly, for information that has commercial, internal, personal, or intellectual-property value, it is advisable to apply the principle of local AI use within a protected corporate environment. In practice, this may involve Ollama, LM Studio, self-hosted n8n, and other local components together with appropriate network protection, access control, and auditing. Further research should focus on developing a formal risk assessment model for AI agents, taking into account the degree of agent autonomy, the number of connected tools, and the criticality of the information resources available to the agent.

REFERENCES

- Anthropic. (2026). Does Anthropic act as a data processor or controller? Anthropic Help Center. <https://support.anthropic.com/en/articles/9267385-does-anthropic-act-as-a-data-processor-or-controller>
- Google. (2026). Generative AI in Google Workspace Privacy Hub. Google Workspace Help. <https://knowledge.workspace.google.com/admin/generative-ai/generative-ai-in-google-workspace-privacy-hub>
- LM Studio. (2026). Offline operation. <https://lmstudio.ai/docs/app/offline>
- Microsoft. (2026). Enterprise data protection in Microsoft Copilot and Microsoft Copilot Chat. Microsoft Learn. <https://learn.microsoft.com/en-us/microsoft-365/copilot/enterprise-data-protection>

- n8n. (2026). Privacy policy. <https://n8n.io/legal/privacy/>
- NIST. (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>
- OpenAI. (2026). Business data privacy, security, and compliance. <https://openai.com/business-data/>
- OpenAI. (2026). Data controls in the OpenAI platform. <https://platform.openai.com/docs/models/default-usage-policies-by-endpoint>
- OWASP. (2025). OWASP Top 10 for large language model applications. OWASP GenAI Security Project. <https://genai.owasp.org/llm-top-10/>
- Ollama. (2026). Privacy policy. <https://ollama.com/privacy>

Гайша Олександр,
доцент кафедри інженерних технологій,
Міжнародний класичний університет імені Пилипа Орлика,
Миколаїв, Україна,
ORCID: 0000-0003-3711-547X; e-mail: oleksandrGaysha1982@gmail.com

ПРО ЗАБЕЗПЕЧЕННЯ ЗАХИСТУ ІНФОРМАЦІЇ ПРИ РОБОТІ З СУЧАСНИМИ ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

АНОТАЦІЯ. Мета дослідження полягає у систематизації інформаційних ризиків, що виникають під час використання сучасних засобів штучного інтелекту, та обґрунтуванні доцільності локального розгортання таких засобів для оброблення інформації з обмеженим доступом. Методологічну основу становлять структурно-функціональний аналіз, класифікація інформаційних потоків, порівняльний аналіз локальної та хмарної моделей розгортання, а також аналіз загроз інформаційній безпеці. Результати полягають у систематизації типів даних, які передаються чат-ботам, ШІ-workflow та ШІ-агентам у персональному і корпоративному середовищах, та у виокремленні ризиків, пов'язаних із передаванням цих даних зовнішнім сервісам. Показано, що локальне виконання моделей за умови належного мережевого та адміністративного контролю дає змогу зберігати оброблення у межах контрольованого контуру. Зроблено висновок про доцільність пріоритетного використання локальних ШІ-рішень для конфіденційної інформації, хоча такий підхід потребує відповідних апаратних ресурсів, компетентностей і процедур безпеки.

КЛЮЧОВІ СЛОВА: штучний інтелект; інформаційна безпека; локальні ШІ-системи; ШІ-агенти; конфіденційна інформація; корпоративні дані.