Generative AI

Data Risks & Mitigation Strategies 2026

Safeguarding Sensitive Information in the Age of AI

Sensitive Data
Leaked in Prompts

Shadow AI Usage
On Corporate Devices

Prompt Injection
& Data Poisoning

Enterprise AI Tiers	AI-Aware DLP	Staff Training
Enhanced Data Protection	Monitor & Prevent Leaks	AI Security Awareness

CYBER ESSENTIALS
CERTIFIED

British Assessment Bureau
UKAS MANAGEMENT SYSTEMS
8289
ISO 27001
INFORMATION SECURITY MANAGEMENT

https://pondgroup.com

info@pondgroup.com

Generative AI in the Workplace:
Key Data & Security Risks and Mitigations (2026)

Executive summary

Generative AI tools (including chat-based assistants, image and code generators, and AI "copilots") can materially improve productivity, but they also introduce new pathways for sensitive data exposure, compliance breaches, and adversarial manipulation.

In practice, the greatest risk remains human error: staff may paste confidential information into prompts or connect AI tools to business systems without appropriate controls. Organisations should treat generative AI as an external data processor unless proven otherwise, and implement a combination of technical safeguards, governance, and training.

Core data risks

Sensitive data leakage via prompts: Users may inadvertently submit confidential information (e.g., customer or staff personal data, financial figures, source code, contracts) into prompts. Once submitted, this data may be logged, retained, or reviewed depending on the AI provider and configuration.

Retention, training and third-party processing: Some services may store prompts and outputs for troubleshooting, quality improvement, or model training unless an enterprise policy prevents it. This can create unintentional risk if proprietary information is retained or used beyond the organisation's intent.

'Shadow AI' and unmanaged accounts: Employees may use personal accounts or unapproved AI tools to bypass restrictions, outside the visibility of corporate security controls (including Data Loss Prevention, monitoring, and access management).

Compliance and regulatory exposure: Using generative AI with regulated or personal data can breach UK GDPR (the Data Protection Act 2018) confidentiality obligations, and sector rules (e.g. health, finance, legal). Cross-border data transfer may also apply depending on the AI provider's hosting and subcontractors.

Output risk (hallucinations and IP): Models can generate plausible but incorrect information ('hallucinations') and may reproduce copyrighted or sensitive material. Over-reliance can lead to poor decisions, contractual errors, or intellectual property disputes.

Emerging threat vectors (2026)

Whilst not currently a huge risk to companies, especially SMEs, the following threats will become more prevalent over the next 12 months.

Prompt injection and data exfiltration: Attackers can embed instructions in documents, emails, web pages, or tickets that cause an AI assistant to reveal secrets, bypass controls, or take unsafe actions (especially where tools have access to internal systems).

Insecure API and plug-in integrations: Rapid deployment of AI features and connectors can introduce new attack surfaces, token leakage, overly broad permissions, and unintended data flows.

Model and data poisoning: If training or fine-tuning data is manipulated, the model may behave unpredictably, leak information, or produce biased or unsafe outputs. This is particularly relevant for internally trained models or retrieval-augmented systems.

AI-enabled social engineering: Generative AI can scale phishing, deepfake voice/video, and targeted pretexting, increasing the likelihood of credential theft and fraud. Targeted attacks are already happening, using a combination of traditional phishing and new AI methods. One example is an email sent to someone in Accounts from the “MD” asking for funds to be transferred. The difference now is the attackers are following up with a voice note or phone call, perfectly matching the MD’s voice and tone. They are going further than this too, by spoofing further messages to colleagues saying “Alice in accounts may ask about a transfer to XZY Ltd. – just to let you know this is expected and relates to a takeover we’ve been working on”. This additional verification from trusted colleagues mean the fraud is much more likely to succeed.

Practical mitigation options

Adopt layered controls that address people, process, and technology:

- **Use enterprise-grade services and contracts:** Prefer enterprise tiers that provide clear commitments on data use (e.g. no model training on business data), retention periods, access controls, audit logs, and UK/EU data processing terms.
- **Data classification and ‘no-go’ rules:** Define which categories of data must never be entered into public or unmanaged AI tools (e.g., personal data, payment data, credentials, secrets, client data). Provide approved alternatives and escalation routes.
- **AI-aware DLP and monitoring:** Deploy controls that detect and block sensitive data being pasted into AI interfaces, uploaded as files, or transmitted via APIs. Monitor usage patterns and implement alerting for anomalous activity.

- **Access management and approved tooling:** Use SSO, MFA, conditional access, and role-based permissions. Restrict plug-ins/connectors and enforce least privilege for API tokens. Provide an approved catalogue of AI tools to reduce ‘Shadow AI’.
- **Secure integration and engineering standards:** For AI features connected to internal systems: validate and sanitise inputs, isolate tools, minimise accessible data, use allow-lists, and implement strong logging. Treat prompt injection as an application security problem.
- **Governance and accountability:** Assign ownership (e.g., CISO + Data Protection Officer + Legal + IT), establish a risk assessment process, and require approval for new AI use cases. Maintain an inventory of AI systems and data flows.
- **Ongoing staff training:** Train staff on safe prompting, data handling, and verification of outputs. Include examples of AI-generated phishing, deepfakes, and hallucinations. Encourage a ‘verify before you trust’ culture.

Quick checklist (minimum baseline for most SMEs)

- Approved generative AI tools only; block unmanaged access where feasible.
- Enterprise contract in place with each AI provider, covering data retention, training, and sub processors.
  (i.e. Don’t use the “free” versions of things like ChatGPT)
- Have clear ‘never share’ data categories and simple staff guidance.
- DLP/monitoring for AI web interfaces and APIs.
- SSO/MFA and least-privilege access for AI tools and connectors.
- Secure build standards for AI integrations.
- Regular training and periodic audits of use cases and logs.

Note: This briefing is intentionally vendor-neutral. Specific controls should be selected based on your organisation’s regulatory obligations, and aligned to recognised frameworks such as NIST, ISO/IEC 27001, and UK NCSC guidance.

Pond have been implementing DLP solutions, conditional access policies and strict application controls for many clients over the past few years. If you wish to discuss how to embrace AI, but in a safe and controlled manner, please get in touch.