SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

AIGF-DLM: Cognitive-Behavioural Governance for GenAI Data Leakage

A R Deepti[1], Farzeen Basith[2], Ranjana K K[3], Sridevi G[4]

[1]Professor, Dept. of MCA, Acharya Institute of Graduate Studies, Karnataka, India. E-mail: ar.deepti@gmail.com, ORCID: https://orcid.org/0009-0003-4320-517X
[2]Assistant Professor, Acharya Institute of Graduate Studies, Karnataka, India. E-mail: farzeen107@gmail.com ORCID: 0009-0003-5475-684X
[3]Assistant Professor, Acharya Institute of Graduate Studies, Karnataka, India. E-mail: ranjanakkmca@gmail.com ORCID:0009-0007-9596-0037
[4]Assistant Professor, Acharya Institute of Graduate Studies, Karnataka, India. E-mail: messridevi@gmail.com, ORCID: 0009-0006-8511-8252

ABSTRACT
AI governance frameworks for GenAI data leakage remain inadequate because they address technical and regulatory dimensions in isolation while neglecting the cognitive-behavioural mechanisms driving human data-sharing behaviour. This paper makes two contributions. First, the AI Governance Framework for Data Leakage Mitigation (AIGF-DLM) integrates five interdependent governance dimensions — Data Governance, Technical Controls, Organisational Governance, Regulatory Compliance, and Continuous Assurance uniquely incorporating synthetic data leakage testing, agentic AI governance, and cognitive-behavioural risk factors. Second, the Cognitive Synthesis Risk Governance (CSRG) theoretical framework introduces three formal propositions linking cognitive risk culture, cognitive load, and moral sensitivity to governance effectiveness. Systematic secondary data analysis of 30 findings from 14 industry reports published by 13 independent organisations demonstrates convergent support for all three propositions. Comparative analysis confirms that the AIGF-DLM addresses critical gaps in six leading frameworks.

Keywords: Generative AI, Data Leakage, AI Governance, LLM Security, Cognitive Synthesis Risk Governance, CSRG, AIGF-DLM, Prompt Injection, RAG Security, Shadow AI

1. Introduction

Enterprise GenAI adoption has created an unprecedented governance gap. One third of organisations use GenAI regularly (IBM, 2023) and AI tools account for 11 percent of all application activity (LayerX, 2025a), yet only 26 percent have fully integrated governance standards (Tredence, 2025). This disparity between adoption and governance maturity is not primarily a technical deficiency it reflects a fundamental misunderstanding of why data leakage occurs. Even organisations that recognise the challenge identify data management as the top barrier to scaling AI (OpenText, 2024), suggesting that current governance approaches are insufficient.

The security risks of GenAI are qualitatively distinct from prior technologies because LLMs create attack surfaces that span the entire data lifecycle. Key threats include hallucination, jailbreaking, sensitive information leakage, opacity, and indirect prompt injection with the latter representing one of the most serious current threat categories. Even data releases assumed to be safe are vulnerable: synthetic data can leak training information through structural overlap in the data manifold, undermining a core assumption of privacy-preserving approaches (Taeihagh, 2025; Park, 2025; Mustaquim et al., 2025).

Existing governance frameworks exhibit significant gaps. The NIST AI RMF provides structured risk identification but lacks GenAI-specific leakage controls. GENXAI (Sivalingam, 2026) addresses lifecycle governance but does not account for cognitive-behavioural risk factors. The Singapore Model Framework (IMDA & AI Verify Foundation, 2024) establishes multi-stakeholder accountability but does not integrate the human-cognitive dimensions that drive Shadow AI adoption and prompt oversharing.

Despite growing recognition of these gaps, no current framework integrates the cognitive-behavioural mechanisms that empirical evidence identifies as primary drivers of GenAI data leakage. Technical controls alone cannot address the 77 percent prompt oversharing rate or the 38 percent unauthorised data sharing rate, because these behaviours are rooted in cognitive offloading, cognitive load degradation of oversight capacity, and reduced moral sensitivity. The absence of an integrated cognitive-behavioural governance approach that explains why technical controls fail not merely that they fail represents the central research gap this paper addresses.

This paper addresses these gaps through two interconnected contributions. First, the AI Governance Framework for Data Leakage Mitigation (AIGF-DLM) integrates technical, organisational, regulatory, and cognitive-behavioural governance dimensions. Second, the Cognitive Synthesis Risk Governance (CSRG)

framework synthesises cognitive risk culture theory (Agarwal & Kallapur, 2018), Cognitive Load Theory (Fox & Fortes Rey, 2024), and cognitive security governance (Chen et al., 2025b) into three formal propositions. The CSRG framework is empirically grounded through systematic secondary data analysis of 30 quantified findings from 14 industry reports published by 13 independent organisations, demonstrating convergent support across multiple methodology types and organisational levels.

2. Research Methodology

This study employs a systematic literature review following PRISMA-ScR guidelines (Tricco et al., 2018), combined with governance framework synthesis, conducted between January 2025 and early 2026.

2.1 Search Strategy and Criteria

Databases searched included Scopus, Web of Science, IEEE Xplore, ACM Digital Library, PubMed, and SSRN. Search strings combined: (1) GenAI terminology ('generative AI' OR 'large language model' OR 'LLM' OR 'foundation model'); (2) risk terminology ('data leakage' OR 'data exfiltration' OR 'prompt injection' OR 'privacy risk'); and (3) governance terminology ('governance' OR 'framework' OR 'regulation' OR 'compliance'). Grey literature from authoritative sources (NIST, IMDA, EDPB, Trend Micro) was included following Andersen et al. (2025). Inclusion criteria required publication between 2021 and 2026, addressing GenAI security, privacy, or governance, in English, from peer-reviewed or authoritative sources. Opinion pieces without analytical contribution, duplicates, and studies addressing traditional AI/ML without GenAI relevance were excluded.

2.2 Screening and Selection

An initial search yielded 847 records. After removing 235 duplicates, 612 unique records underwent title and abstract screening, excluding 414 records. Full-text review of the remaining 198 records excluded 135 sources. Screening was conducted by the primary author. The final synthesis comprised 63 sources: 38 peer-reviewed journal articles and conference papers, 8 regulatory and standards documents, 9 authoritative industry reports, and 8 technical threat intelligence publications. The reference list includes additional sources cited for methodology, theory, and contextual framing beyond the 63 synthesised sources. Sources were analysed using thematic synthesis across leakage vector identification, governance mechanism mapping, and gap analysis. The CSRG framework was developed through abductive reasoning.

3. Genai Architectures And Data Leakage Vulnerabilities

The structural properties of GenAI systems create inherent leakage risks that governance must address. GenAI systems built on transformer architectures, GANs, and VAEs (Uddin et al., 2025; Corchado Rodríguez et al., 2023; Sengar et al., 2025) enter organisations through three pathways: direct product use, embedded components, and deliberate deployment (Emett et al., 2024). Unlike traditional databases that store and retrieve, LLMs function as inference engines that synthesise answers from multiple sources including sensitive internal content making every query a potential leakage event (Knostic, 2025). This fundamental architectural property generates privacy risks across four lifecycle phases: training data collection, inference, RAG processes, and feedback loops (Barberá, 2025), producing three categories of security threats: internal (jailbreaks, backdoors, data poisoning), interactional (hallucination, bias, data leakage), and external threats including weaponisation for fraud (Chen et al., 2025b). A critical paradox compounds these risks: even methods designed to enhance transparency including SHAP, LIME, and CIU explainability tools impose cognitive loads that undermine the very human oversight they are intended to support (Fox & Fortes Rey, 2024), a finding central to the CSRG framework presented in Section 7.

4. Taxonomy Of Genai Data Leakage Vectors

The systematic review identifies seven principal leakage vector categories, each representing distinct governance challenges.

4.1 Prompt Oversharing and Shadow AI

Enterprise browser telemetry reveals that 77 percent of employees paste data into GenAI tools, with 82 percent of this activity occurring through unmanaged accounts, averaging 14 pastes per day with at least three containing sensitive data (LayerX, 2025a). Separately, 38 percent of employees share sensitive information without employer permission (Versa Networks, 2025). Distributed cognition theory explains this pattern: humans naturally offload cognitive work to AI tools, making prompt oversharing a cognitive-behavioural phenomenon rather than mere policy non-compliance (Sydorenko, 2025).

4.2 RAG Abuse and Vector-Store Poisoning

RAG architectures introduce a distinct attack surface because the retrieval layer operates outside the model's safety training. Three mechanisms exploit this vulnerability: hidden context extraction from vector stores (BlackFog, 2025), instruction injection directing agents to exfiltrate data to attacker-controlled servers (Park,

2025; Zhan et al., 2024), and self-replicating prompts that propagate through RAG-based assistants (Datadog, 2025). These attacks exploit the trust boundary between retrieved content and model instructions a boundary that the AIGF-DLM addresses through its Technical Controls dimension (Section 6.2).

4.3 Training Data Leakage and Model Inversion

Training data leakage operates through three established attack vectors: model inversion, training data extraction, and membership inference (Chen et al., 2025a). The severity of this threat is demonstrated by divergence attacks that cause verbatim pre-training data emission (Nasr et al., 2025), with trigger words such as 'company' causing 164 times more leakage than neutral words (Barberá & Popa-Fabre, 2025). Fine-tuning compounds this risk, as models memorise sensitive details from proprietary data during adaptation (BlackFog, 2025; Tonic.ai, 2025). These vectors necessitate the AIGF-DLM's mandatory membership inference testing requirement for all data releases.

4.4 Hidden Synthetic Data Leakage

A critical assumption underpinning many governance approaches that synthetic data is inherently privacy-safe has been empirically undermined. Synthetic data releases leak training information through structural overlap in the data manifold, even when differential privacy is applied. A black-box membership inference attack exploiting distributional neighbourhood overlap demonstrates this vulnerability (Mustaquim et al., 2025). This finding directly motivates the AIGF-DLM's requirement for mandatory synthetic data leakage testing, a governance mechanism absent from all six compared frameworks (Table 1).

4.5 Prompt Injection and System Prompt Exfiltration

System prompt exfiltration presents a governance challenge that no single technical control can resolve. Both hard and soft exfiltration methods target system prompts (Husieva & Freenor, 2024), and evaluation of vulnerabilities across 42 LLM models confirms that no single mitigation strategy provides comprehensive protection (Ferrag et al., 2025). The attack surface extends to multi-modal vectors, including sandbox code execution for Base64-encoded data exfiltration (Park, 2025). This finding reinforces CSRG's Proposition 2: when no technical control is sufficient, the cognitive capacity of human oversight actors becomes the critical governance variable.

4.6 Agentic Exfiltration and Misevolution

Agentic AI introduces leakage vectors that operate beyond human-initiated behaviour. URL-based exfiltration by autonomous agents has been formalised as a distinct threat category requiring dynamic URL policies beyond traditional domain allow-listing (Spânu & Shadwell, 2025). A further risk termed 'misevolution' occurs when self-evolving agents deviate across model, memory, tool, and workflow pathways, causing safety alignment to degrade through autonomous adaptation (Shao et al., 2026). The AIGF-DLM is the only compared framework to address both agentic exfiltration mechanisms (Table 1).

4.7 Integration Drift

Integration drift — the misalignment of access controls triggered by system updates creates governance challenges that are continuous rather than episodic, requiring the ongoing monitoring mandated by the AIGF-DLM's Continuous Assurance dimension (Knostic, 2025). Agent-based privacy challenges constitute a distinct fourth category beyond training data, prompt, and output privacy (Shanmugarasa et al., 2025), further demonstrating that static governance frameworks cannot address the dynamic nature of GenAI data leakage.

5. Comparative Analysis Of Existing Governance Frameworks

Six existing frameworks were compared against eight governance dimensions derived from the leakage taxonomy. Table 1 presents this analysis.

Table 1. Comparative Gap Analysis of Existing Governance Frameworks

Governance Dimension	NIST AI RMF	GENXAI	Singapore MAIGF	Emett et al.	Schneider et al.	Papagiannidis et al.	AIGF-DLM
Prompt oversharing & Shadow AI	–	–	Partial	✓	–	–	✓
RAG-specific security	–	Partial	Partial	–	–	–	✓
Synthetic data leakage testing	–	–	–	–	–	–	✓

Governance Dimension	NIST AI RMF	GENXAI	Singapore MAIGF	Emett et al.	Schneider et al.	Papagiannidis et al.	AIGF-DLM
Agentic AI exfiltration	–	–	Partial	–	–	–	✓
Cognitive-behavioural factors	–	–	–	Partial	–	–	✓
Lifecycle privacy management	Partial	✓	Partial	Partial	Partial	Partial	✓
Content provenance	–	–	✓	–	–	–	✓
Continuous adversarial assurance	Partial	Partial	✓	Partial	–	Partial	✓

Note. ✓ = Addressed; Partial = Partially addressed; – = Not addressed. 'Partial' indicates partial coverage documented in the source framework without dedicated governance mechanisms.

5.1 Differentiation from Adjacent Human Factors Literatures

Adjacent literatures address human dimensions of technology governance but none integrate the three specific mechanisms identified by CSRG. Usable security focuses on interface design for compliance but does not address cognitive offloading to GenAI tools as a distributed cognition phenomenon (Sydorenko, 2025). Behavioural information security models assume rational threat appraisal but do not account for cognitive load degrading oversight capacity (Fox & Fortes Rey, 2024). Human-AI interaction research addresses automation bias but targets decision accuracy rather than data leakage behaviour. Table 2 summarises these distinctions.

Table 2. Differentiation of CSRG from Adjacent Human Factors Literatures

Literature Stream	Primary Mechanism	Cognitive Offloading	Load on Oversight	Moral Sensitivity	GenAI Leakage
Usable Security	Interface design	No	No	No	No
Behavioural InfoSec	Rational threat appraisal	No	No	No	No
Human-AI Interaction	Automation bias	No	No	No	Partial
Socio-Technical Systems	Joint optimisation	No	No	No	No
CSRG (Proposed)	Offloading + load + moral sensitivity	Yes	Yes	Yes	Yes

6. the aigf-dlm framework

The AIGF-DLM is structured across five interdependent dimensions. Its novelty lies in three design principles: (a) integration of cognitive-behavioural governance grounded in CSRG; (b) mandatory synthetic data leakage testing; and (c) agentic AI-specific governance mechanisms.

6.1 Data Governance

The Data Governance dimension addresses the foundational challenge that 44 percent of organisations reported leaking of sensitive data into AI tools as a negative outcome (Komprise, 2025). The AIGF-DLM mandates sensitive data classification using the SPLOG framework Security, Privacy, Lineage, Ownership, and Governance combined with data minimisation through Privacy Enhancing Technologies (IMDA & AI Verify Foundation, 2024) and federated learning (Chen et al., 2025a). Uniquely among compared frameworks, this dimension requires membership inference testing of synthetic data releases using distributional neighbourhood analysis (Mustaquim et al., 2025) before treating them as privacy-safe directly responding to the hidden synthetic data leakage vector identified in Section 4.4.

6.2 Technical Controls

The Technical Controls dimension addresses all seven leakage vectors identified in the taxonomy (Section 4) through layered defence mechanisms. Input-layer controls include prompt sanitisation, multi-modal input

inspection (Park, 2025), and GenAI firewall deployment (Versa Networks, 2025). Processing-layer controls include RBAC with dynamic classification (Versa Networks, 2025; Knostic, 2025) and privacy-preserving ML including differential privacy and homomorphic encryption. Output-layer controls include DLP integration for LLM output filtering (Namer et al., 2023) and content provenance using digital watermarking and C2PA (IMDA & AI Verify Foundation, 2024). The AIGF-DLM uniquely mandates dynamic URL policies for agentic AI beyond static domain allow-listing (Spânu & Shadwell, 2025) and strict network-level egress controls controls absent from all compared frameworks yet essential for addressing the agentic exfiltration vectors identified in Section 4.6.

6.3 Organisational Governance

The Organisational Governance dimension addresses the finding that technical controls alone fail to reduce the 77 percent prompt oversharing rate (LayerX, 2025a) or the 43 percent secret data sharing rate (CybSafe & NCA, 2025). The AIGF-DLM mandates clear AI use policies with accountability structures (IMDA & AI Verify Foundation, 2024), sector-specific adaptation for varying regulatory contexts (Yang, H. et al., 2025; Sivalingam, 2026), and proactive Shadow AI detection mechanisms (Versa Networks, 2025; BlackFog, 2025). A GenAI Governance Committee provides oversight coordination (Emett et al., 2024). Distinctively, this dimension integrates CSRG-informed cognitive-behavioural governance: security awareness training employing enhanced message framing to reach users with reduced moral sensitivity (Li et al., 2024), as detailed in Section 7.

6.4 Regulatory Compliance

The Regulatory Compliance dimension ensures that data leakage governance meets current and emerging legal obligations. The AIGF-DLM mandates GDPR alignment including Articles 25 (data protection by design), 32 (security of processing), and 35 (impact assessments) (Barberá, 2025; IBM, 2024), EU AI Act conformity assessments with 15-day serious incident reporting (IMDA & AI Verify Foundation, 2024), intellectual property risk management addressing GenAI-specific IP challenges (Uddin et al., 2025), and Convention 108+ alignment for international data governance standards (Barberá & Popa-Fabre, 2025). The framework further supports alignment of GenAI systems with documented regulatory and operational standards to strengthen conformity (Imperial et al., 2025).

6.5 Continuous Assurance

The Continuous Assurance dimension addresses the dynamic nature of GenAI threats, where integration drift and evolving attack vectors render point-in-time assessments insufficient. The AIGF-DLM mandates real-time leakage signal detection (Knostic, 2025; Datadog, 2025), red-teaming with formal vulnerability reporting (IMDA & AI Verify Foundation, 2024), and governance audits with minimum annual policy review cycles (Emett et al., 2024). Critically, this dimension incorporates recurring membership inference testing of synthetic data releases (Mustaquim et al., 2025), ensuring that the hidden leakage vector identified in Section 4.4 is continuously monitored rather than assessed only at the point of release.

6.6 Lifecycle Integration

The five dimensions integrate across five lifecycle phases: (1) training data filtering and memorisation prevention; (2) post-training adaptation fine-tuning controls; (3) inference input sanitisation and monitoring; (4) system integration API security and RAG controls; and (5) post-deployment feedback mechanisms and user control preservation (Barberá & Popa-Fabre, 2025). Across these phases, the AIGF-DLM aligns with secure development guidance that extends secure software development practices to generative AI and dual-use foundation models throughout the model development lifecycle (Booth et al., 2024).

7. The Cognitive Synthesis Risk Governance (Csrg) Theoretical Framework

The CSRG framework is this paper's core theoretical contribution, explaining why purely technical governance fails to prevent GenAI data leakage. It synthesises three established but previously disconnected theoretical streams into a mid-range framework that operates at the organisational governance level, bridging micro-level cognitive mechanisms with meso-level organisational outcomes.

7.1 Theoretical Foundations

Cognitive Risk Culture (CRC). Transitioning from compliance-based to cognitive risk culture significantly improves organisational risk management, as demonstrated through empirical case study (Agarwal & Kallapur, 2018). In GenAI contexts, the 38 percent unauthorised data sharing rate (Versa Networks, 2025) and 77 percent prompt oversharing rate (LayerX, 2025a) reflect failures of cognitive risk culture where policies exist but cognitive understanding of data risks does not.

Cognitive Load in AI Governance (CL-AIG). Even explainable AI methods impose cognitive loads sufficient to limit genuine interpretability, leading to motivated cognition or disengagement (Fox & Fortes Rey, 2024). Drawing on Hutchins' distributed cognition theory, humans naturally offload cognitive work to AI tools,

creating behavioural conditions for data leakage (Sydorenko, 2025). This extends to human-AI collaborative teams (Corral, 2026).

Cognitive Security Governance (CSG). GenAI risks extend to adverse effects on human cognition and institutional decision-making (Chen et al., 2025b). Hedonic AI applications reduce moral sensitivity, rendering governance warnings ineffective for users with low moral sensitivity (Li et al., 2024).

7.2 Formal Propositions

Proposition 1 (Cognitive Risk Culture): GenAI data leakage governance effectiveness is positively associated with cognitive risk culture maturity. Organisations with compliance-based risk cultures will exhibit higher rates of Shadow AI adoption and prompt oversharing than organisations with cognitive risk cultures, independent of technical controls deployed.

Proposition 2 (Cognitive Load): The effectiveness of human-in-the-loop governance mechanisms is negatively moderated by cognitive load imposed by AI system outputs and governance processes. When cognitive load exceeds sustainable thresholds, human oversight actors default to motivated cognition or disengagement.

Proposition 3 (Cognitive Security): The effectiveness of governance awareness communications is moderated by users' moral sensitivity, which is reduced by hedonic AI applications. Enhanced message framing strategies are required to reach users with low moral sensitivity.

7.3 Systemic Interdependence and AIGF-DLM Integration

The three propositions form a reinforcing system rather than independent additions. Proposition 1 determines organisational receptivity to governance interventions. Proposition 2 determines whether governance actors can functionally execute oversight. Proposition 3 determines whether governance communications reach end users. A failure in any single proposition undermines the others: organisations with mature cognitive risk cultures (P1) will still experience governance failure if oversight actors are cognitively overloaded (P2), or if end users have reduced moral sensitivity (P3). This systemic interdependence distinguishes CSRG from adding a human factors layer to existing frameworks.

The CSRG framework integrates into the AIGF-DLM's Organisational Governance dimension through: (a) cognitive risk culture assessment programmes informed by Proposition 1; (b) cognitive load-aware design of governance interfaces informed by Proposition 2; and (c) message framing-enhanced governance communications informed by Proposition 3.

7.4 Evidence-Based Application

This section maps CSRG propositions to documented enterprise data. Enterprise browser telemetry documents that employees paste data into GenAI tools at a 77 percent rate, predominantly through unmanaged accounts (LayerX, 2025a), while 38 percent share sensitive information without employer permission (Versa Networks, 2025). Table 1 demonstrates that NIST AI RMF, GENXAI, and Schneider et al. do not address prompt oversharing and Shadow AI controls, and evaluation of 42 LLM models confirms that no single technical mitigation provides comprehensive protection (Ferrag et al., 2025).

Applying Proposition 1: the documented oversharing rate reflects a compliance-based culture where policies exist but cognitive understanding of data risks does not, consistent with Sydorenko's (2025) finding that cognitive offloading to AI tools is natural human behaviour and Agarwal and Kallapur's (2018) demonstration that cognitive risk culture transition improves risk outcomes. Applying Proposition 2: governance interventions imposing additional cognitive load on oversight personnel monitoring these data flows trigger disengagement rather than compliance (Fox & Fortes Rey, 2024). Applying Proposition 3: the 38 percent unauthorised sharing rate (Versa Networks, 2025) is consistent with reduced moral sensitivity from routine hedonic AI use (Li et al., 2024), explaining why standard compliance training fails to reduce Shadow AI adoption. The documented behaviours persist because all three mechanisms operate simultaneously — a capability absent from the six frameworks compared in Table 1.

7.5 Systematic Secondary Data Analysis

7.5.1 Methodology

To empirically ground the CSRG framework, a systematic secondary data analysis was conducted following established guidelines for compiled evidence datasets (Rousseau et al., 2008; Tranfield et al., 2003). Fourteen authoritative industry reports, published by 13 independent organisations between January 2024 and December 2025, were identified through targeted searches of consulting firm publications, security vendor research, professional body reports, and governance surveys. Sources were selected based on three criteria: empirical data collection methodology (survey or telemetry), sample relevance to AI governance contexts, and publication by recognised industry authorities.

From these sources, 30 specific quantified empirical findings were extracted and systematically coded against the three CSRG propositions using a structured coding scheme. Each finding was assigned a relevance score (0 = not relevant, 1 = indirect support, 2 = direct support, 3 = primary evidence) for each proposition, as well as evidence strength (1–3), source methodology type, respondent level, and an interdependence indicator. The

compiled dataset encompasses six distinct source methodology types: enterprise browser/network telemetry (LayerX, 2025a; Palo Alto Networks, 2025), C-suite executive surveys (EY, 2025a, 2025b; PwC, 2025), governance professional surveys (OneTrust, 2025; IAPP, 2025; Lorica & Talby, 2025), IT leadership surveys (Komprise, 2025), employee/general population surveys (CybSafe & NCA, 2025; Versa Networks, 2025; Knostic, 2025), and security professional surveys (CSA, 2025; RSA & Knostic, 2024). Combined survey samples total 21,778 respondents from eight survey-based sources, supplemented by enterprise network telemetry from global deployments.

Three validity mechanisms were employed. First, each finding was coded using the structured relevance scale with ambiguous findings conservatively coded at the lower relevance level. Second, source selection deliberately prioritised methodological diversity across six distinct source types, with vendor-sponsored research counterbalanced by independent professional body surveys (IAPP, 2025; CSA, 2025) and multi-country population surveys (CybSafe & NCA, 2025, n = 6,500). Third, following Rousseau et al. (2008), convergence of findings across methodologically independent sources particularly where passive behavioural telemetry corroborates self-reported survey data provides stronger evidence than any single source. The coding was conducted by the primary author; implications are addressed in Section 7.5.5.

All 30 findings in the compiled evidence dataset are real, published empirical data points extracted from the original industry sources; no hypothetical or simulated data were used. Each finding is traceable to its original source via the URLs documented in the supplementary dataset (Supplementary File 1), enabling independent verification of all statistics cited in this analysis. The proposition mapping rationale, including counter-arguments and strength assessments for each coding decision, is documented in the supplementary mapping file (Supplementary File 2).

Table 3. Compiled Evidence Dataset — Source Overview

Source	Type	Sample	Scope	Key Finding(s)
LayerX (2025a)	Vendor Telemetry	Telemetry	Global	77% paste data into GenAI; 82% via unmanaged accounts; avg 14 pastes/day
Palo Alto (2025)	Vendor Telemetry	Telemetry	Global	890% GenAI traffic surge in 2024; DLP incidents up 2.5×; avg 66 apps/enterprise
EY (2025b)	C-Suite Survey	n = 975	21 countries	Only 12% C-suite correctly identify AI risk controls; 99% report financial losses
EY (2025a)	C-Suite Survey	n = 975	21 countries	72% AI integrated but only 33% have governance; ~50% find governance challenging
PwC (2025)	C-Suite Survey	n = 253	US	43-point RAI communication effectiveness gap between maturity stages
OneTrust (2025)	Gov Prof Survey	n = 1,250	NA/Europe	37% more time on AI risk; 44% reviews too late; 73% report governance gaps
IAPP (2025)	Gov Prof Survey	Survey	Global	23.5% cite finding qualified AI governance professionals as primary challenge
Lorica & Talby (2025)	Gov Prof Survey	n = 350	US	<20% have model cards, incident reporting, or regular red teaming
CSA (2025)	Security Prof Survey	Survey	Global	Only ~25% have comprehensive AI security governance
Komprise (2025)	IT Leadership Survey	n = 200	US	79% negative outcomes from GenAI; 44% data leakage; 13% financial/reputational harm
CybSafe & NCA (2025)	Employee Survey	n = 6,500	7 countries	43% share data secretly; 58% no training; 43% overwhelmed by security info
Versa Networks (2025)	Employee Survey	Survey	Global	38% share sensitive info with AI without employer permission
Knostic (2025)	Employee Survey	n = 12,000	Global	Only 18.5% aware of AI policy despite 60.2% using AI tools (41.7-point gap)
RSA & Knostic	Security Prof Survey	n = 250	US	73% security professionals use unapproved SaaS/AI apps; only 46% had controls

Source	Type	Sample	Scope	Key Finding(s)
(2024)				

7.5.2 Proposition Support Results

Table 4. CSRG Proposition Convergence Summary

Proposition	Primary Evidence	Direct+ Support	Indep. Sources	Method. Types	Avg Strength	Assessment
P1: Cognitive Risk Culture	9	19	11	6	2.79/3.00	STRONG
P2: Cognitive Load	9	13	6	4	2.62/3.00	STRONG
P3: Moral Sensitivity	5	11	6	4	2.73/3.00	STRONG
Systemic Interdependence	—	12	6	6	2.83/3.00	STRONG

Note. Primary Evidence = findings with relevance score of 3. Direct+ Support = findings with relevance score ≥ 2. Independent Sources = unique publishing organisations. Methodology Types = distinct data collection approaches.

Proposition 1 (Cognitive Risk Culture) receives the broadest convergent support 19 findings from 11 independent sources across all six methodology types confirming that cognitive risk culture failure is not isolated to any single organisational level or methodology. The evidence reveals cognitive risk culture failure at every organisational level: C-suite executives demonstrate only 12 percent correct identification of AI risk controls (EY, 2025b, n = 975), IT directors report 79 percent negative AI outcomes yet practices continue (Komprise, 2025, n = 200), security professionals admit to 73 percent use of unapproved AI applications (RSA & Knostic, 2024, n = 250), and frontline employees paste data into GenAI tools at a 77 percent rate (LayerX, 2025a). The finding that only 18.5 percent of employees are aware of AI policies despite 60.2 percent actively using AI tools (Knostic, 2025, n = 12,000) provides particularly compelling evidence that organisations have deployed policies but have not embedded cognitive risk understanding.

Proposition 2 (Cognitive Load) receives convergent support from 13 findings across six independent sources spanning four methodology types. Governance executives report dedicating 37 percent more time to AI risk management compared to 12 months prior, with 44 percent identifying governance reviews happening too late (OneTrust, 2025, n = 1,250). The talent pipeline is constrained, with 23.5 percent of organisations citing qualified AI governance professionals as a primary challenge (IAPP, 2025). Less than 20 percent have implemented cognitively intensive governance practices such as model cards, red teaming, or dedicated incident reporting tools (Lorica & Talby, 2025, n = 350). These findings collectively indicate that governance oversight personnel are operating at or beyond sustainable cognitive capacity thresholds.

Proposition 3 (Moral Sensitivity) receives convergent support from 11 findings across six independent sources spanning four methodology types. The finding that 43 percent of employees share sensitive data with AI tools 'without employer knowledge' (CybSafe & NCA, 2025, n = 6,500) is particularly significant: the qualifier indicates awareness of prohibition combined with continued behaviour, consistent with reduced moral sensitivity. Seventy-nine percent of organisations experienced negative AI data outcomes yet practices continue, with 13 percent experiencing financial or reputational harm without triggering behavioural change (Komprise, 2025) a pattern consistent with moral desensitisation. A 43-percentage-point gap in governance communication effectiveness between strategically mature and training-stage organisations (PwC, 2025, n = 253) demonstrates that standard compliance communications fail to overcome reduced moral sensitivity.

7.5.3 Evidence for Systemic Interdependence

Twelve findings (40 percent of the dataset) demonstrate patterns requiring multiple propositions for adequate explanation. GenAI-related DLP incidents increased 2.5-fold despite deployment of technical DLP controls, with enterprises engaging an average of 66 GenAI applications (Palo Alto Networks, 2025). Technical controls are deployed yet leakage increases because employees operate in compliance-based cultures where only 18.5 percent are aware of AI policies (P1; Knostic, 2025), governance teams spend 37 percent more time monitoring yet reviews happen too late (P2; OneTrust, 2025), and 43 percent of employees share data knowing it is prohibited (P3; CybSafe & NCA, 2025). No single proposition explains the full pattern; only their interaction does.

7.5.4 Cross-Validation Through Methodological Triangulation

Table 5. Cross-Tabulation: Source Methodology Type × Proposition Support (Findings with Relevance ≥ 2)

Source Methodology Type	P1: CRC	P2: CL	P3: MS	Total
Vendor Telemetry	6	2	2	10
C-Suite Survey	4	4	2	10
Governance Prof Survey	1	6	0	7
IT Leadership Survey	2	0	3	5
Employee/General Survey	4	0	4	8
Security Prof Survey	2	1	0	3
TOTAL	19	13	11	43

Note. Individual findings may support multiple propositions; column totals (43) therefore exceed the 30 findings in the dataset due to cross-proposition coding. CRC = Cognitive Risk Culture; CL = Cognitive Load; MS = Moral Sensitivity.

Each CSRG proposition receives support from at least four distinct source methodology types, including both passive telemetry and active surveys. The convergence of telemetry-observed behaviour (77 percent paste rate; LayerX, 2025a) with self-reported behaviour (43 percent admitting to sharing; CybSafe & NCA, 2025) strengthens confidence that CSRG propositions describe genuine phenomena rather than methodological artefacts. Governance failure is documented at C-suite (EY, 2025b), director/VP (OneTrust, 2025), professional (RSA & Knostic, 2024), and employee (CybSafe & NCA, 2025) levels. This multi-level convergence supports CSRG's characterisation of data leakage as a systemic organisational phenomenon.

7.5.5 Limitations of Secondary Data Analysis

This analysis has acknowledged limitations. First, industry reports employ varying methodologies and definitions, introducing measurement heterogeneity. Second, the mapping from observed findings to theoretical propositions involves interpretive coding; alternative explanations may account for some findings. Third, this analysis demonstrates explanatory validity but does not constitute causal validation; the research designs proposed in Section 9 remain necessary. Fourth, vendor-sponsored research may contain selection bias, mitigated but not eliminated by source diversity. Fifth, coding was conducted by a single researcher without formal inter-coder reliability assessment. Three design choices mitigate this: structured coding with explicit relevance definitions constraining interpretive latitude, conservative coding at lower levels when ambiguity existed, and transparent documentation enabling post-hoc verification. Future extensions should incorporate independent parallel coding with Cohen's kappa reporting.

Despite these limitations, the convergence of 30 findings from 14 industry reports published by 13 independent organisations across six methodology types, spanning all organisational levels and encompassing combined survey samples totalling 21,778 respondents plus enterprise telemetry, provides robust empirical grounding that exceeds the evidentiary standard typically expected of theoretical framework papers (Rousseau et al., 2008).

7.6 Preliminary Operationalisation of Cognitive Risk Culture

Table 6. Preliminary Cognitive Risk Culture Maturity Levels

Level	Label	Characteristics	Observable Indicators
1	Compliance-Based	Governance relies on policy documents; training is periodic; no Shadow AI monitoring; employees cannot articulate data risks	High Shadow AI; high prompt oversharing (baseline 77%; LayerX, 2025a)
2	Awareness-Based	Employees articulate GenAI data risks; training includes contextual elements; Shadow AI monitoring is reactive	Reduced but persistent Shadow AI; prompt oversharing below 77% baseline
3	Cognitive Risk Culture	All actors understand governance roles (Agarwal & Kallapur, 2018); cognitive load-aware interfaces; enhanced message framing; proactive Shadow AI.	Shadow AI near zero; prompt oversharing significantly below baseline; governance override rates within thresholds

Note. This preliminary operationalisation requires empirical validation. Development of validated measurement instruments constitutes a priority for future research.

8. Discussion

8.1 Theoretical and Practical Contributions

The AIGF-DLM makes four novel contributions. First, it mandates synthetic data leakage testing, responding to the empirical finding that synthetic data releases leak training information through distributional neighbourhood overlap (Mustaquim et al., 2025). Second, it addresses agentic AI governance incorporating misevolution risks (Shao et al., 2026) and URL-based exfiltration prevention (Spânu & Shadwell, 2025). Third, through CSRG, it integrates cognitive-behavioural risk factors, addressing mechanisms absent from adjacent human factors literatures (Table 2).

Fourth, the systematic secondary data analysis provides empirical grounding demonstrating that all three CSRG propositions possess genuine explanatory power across multiple independent source methodologies and organisational levels. The multi-level convergence governance failure documented at C-suite, director, professional, and employee levels supports CSRG's characterisation of data leakage as a systemic organisational phenomenon requiring integrated cognitive-behavioural governance rather than isolated technical controls.

8.2 Sector-Specific Governance

The AIGF-DLM's sector-specific governance dimension responds to evidence that implementation barriers vary fundamentally across contexts. Government LLM implementations face primarily technical complexity (52 percent) and staff resistance (38 percent), with regulatory compliance a secondary concern (29 percent) (Yang, H. et al., 2025), whereas healthcare implementations face regulatory complexity and framework absence as primary barriers, as confirmed through 43 stakeholder interviews (Freeman et al., 2025). This divergence validates the AIGF-DLM's design principle of sector adaptation rather than uniform governance an approach absent from the NIST AI RMF, GENXAI, and four of the six compared frameworks (Table 1).

8.3 Limitations

This study has three limitations. First, while the systematic secondary data analysis provides convergent empirical grounding exceeding the evidentiary standard typical of theoretical framework contributions (Rousseau et al., 2008), it does not constitute primary experimental validation. The research designs proposed in Section 9 remain a priority. Second, grey literature inclusion, while methodologically justified (Andersen et al., 2025), introduces potential quality variability. Third, rapid GenAI evolution means specific technical controls require continuous updating.

9. Conclusion And Future Research

This paper has presented two interconnected contributions with systematic empirical grounding. The AIGF-DLM framework provides an integrated, lifecycle-based governance architecture uniquely addressing synthetic data leakage, agentic AI risks, and cognitive-behavioural risk factors three dimensions absent from existing frameworks. The CSRG theoretical framework provides three formal propositions explaining why purely technical governance is insufficient. Systematic secondary data analysis of 30 findings from 14 industry reports published by 13 independent organisations across six methodology types demonstrates convergent support for all three propositions, with multi-level convergence across C-suite, director, professional, and employee populations. Forty percent of findings demonstrate systemic interdependence requiring multiple propositions for adequate explanation, validating CSRG's core theoretical claim.

Future research should pursue three empirical directions. First, Proposition 1 can be validated through a comparative case study of organisations with compliance-based versus cognitive risk cultures, measuring Shadow AI adoption rates and prompt oversharing frequency while controlling for technical controls deployed. Second, Proposition 2 can be validated through a controlled experiment with governance oversight personnel, manipulating AI output complexity and measuring oversight accuracy and disengagement thresholds. Third, Proposition 3 can be validated through a factorial experiment crossing message framing type with moral sensitivity level, measuring data sharing behaviour in simulated GenAI tasks. Additionally, longitudinal case study application of AIGF-DLM in regulated sectors and development of validated measurement instruments for the cognitive risk culture maturity levels constitute priorities for future investigation.

DATA AVAILABILITY STATEMENT

The compiled evidence dataset and all supporting materials are provided as supplementary files. Supplementary File 1 (CSRG Compiled Evidence Dataset) contains the coded dataset of 30 empirical findings with 20 coded variables and the variable codebook. Supplementary File 2 (CSRG Source Mapping and Coding Rationale) contains the original source documentation with source URLs for all 14 industry reports (13 independent organisations), the proposition mapping rationale for each coding decision, and the convergence analysis summary. All statistics are real, published data points from independently conducted industry research; no primary data were collected and no hypothetical data were used.

References

1. Agarwal, R., & Kallapur, S. (2018). Cognitive risk culture and advanced roles of actors in risk governance: A case study. The Journal of Risk Finance, 19(4), 327–342. https://doi.org/10.1108/JRF-11-2017-0189
2. Andersen, J. P., Degn, L., Fishberg, R., Graversen, E. K., Horbach, S. P. J. M., Kalpazidou Schmidt, E., Schneider, J. W., & Sørensen, M. P. (2025). Generative Artificial Intelligence (GenAI) in the research process – A survey of researchers' practices and perceptions. Technology in Society, 81, Article 102813. https://doi.org/10.1016/j.techsoc.2025.102813
3. Balassiano, O., & Shaty, O. (2024, November 12). ModeLeak: Privilege escalation to LLM model exfiltration in Vertex AI. Palo Alto Networks Unit 42 Research.
4. Barberá, I. (2025). AI privacy risks & mitigations – Large language models (LLMs). EDPB Support Pool of Experts Programme. European Data Protection Board. Updated March 2025.
5. Barberá, I., & Popa-Fabre, M. (2025). Privacy and data protection risks in large language models (LLMs). Expert Report for the Consultative Committee of Convention 108. Council of Europe.
6. BlackFog. (2025). 5 ways large language models (LLMs) enable data exfiltration. BlackFog Blog. https://www.blackfog.com/5-ways-llms-enable-data-exfiltration/
7. Booth, H., Souppaya, M., Vassilev, A., Ogata, M., Stanley, M., & Scarfone, K. (2024). Secure software development practices for generative AI and dual-use foundation models: An SSDF community profile (NIST Special Publication 800-218A). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-218A
8. Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In 30th USENIX Security Symposium (pp. 2633–2650).
9. Chen, B., Li, R., Li, G., & An, M. (2025b). Cognition security governance of generative artificial intelligence in smart education: A survey. SSRN Electronic Journal, 45 pages. https://doi.org/10.2139/ssrn.5256497
10. Chen, K., Zhou, X., Lin, Y., Feng, S., Shen, L., & Wu, P. (2025a). A survey on privacy risks and protection in large language models. Journal of King Saud University – Computer and Information Sciences, 37, Article 163. https://doi.org/10.1007/s44443-025-00177-1
11. Cloud Security Alliance. (2025). State of AI security report. Cloud Security Alliance, December 2025. https://cloudsecurityalliance.org/research/ai/
12. Corchado Rodríguez, J. M., López, F. S., Núñez V., J. M., & Garcia S., R. (2023). Generative Artificial Intelligence: Fundamentals. ADCAIJ, 12(1), e31704. https://doi.org/10.14201/adcaij.31704
13. Corral, E. J. (2026). Thinking in balance: Framework for cognitive load distribution in human-AI collaborative teams. ECSAUC, 14(2). ISSN 2007-8242.
14. CybSafe & National Cybersecurity Alliance. (2025). Oh behave! The annual cybersecurity attitudes and behaviors report 2025–2026. September 2025. n = 6,500 across 7 countries. https://www.cybsafe.com/research/oh-behave/
15. Datadog. (2025). Abusing AI interfaces: How prompt-level attacks exploit LLM applications. Datadog Blog. https://www.datadoghq.com/blog/detect-abuse-ai-interfaces/
16. Dubale, N. A., Komare, A. V., & Solanki, S. (2025). Data privacy and protection in generative AI: Problems and solutions. IRJMETS, 7(11).
17. Emett, S. A., Eulerich, M., Pikoos, J., & Wood, D. A. (2024). Generative AI governance framework v1.0. Connor Group. https://genai.global
18. EY. (2025a). EY survey: AI adoption outpaces governance as risk awareness among the C-suite remains low. EY Global Responsible AI Pulse Survey, Phase 1, June 2025. n = 975 C-suite leaders, 21 countries. https://www.ey.com/en_gl/newsroom/2025/06/ey-survey-ai-adoption-outpaces-governance
19. EY. (2025b). EY survey: Companies advancing responsible AI governance linked to better business outcomes. EY Global Responsible AI Pulse Survey, Phase 2, October 2025. n = 975 C-suite leaders, 21 countries. https://www.ey.com/en_gl/newsroom/2025/10/ey-survey-companies-advancing-responsible-ai-governance-linked-to-better-business-outcomes
20. Ferrag, M. A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., Tihanyi, N., Bisztray, T., & Debbah, M. (2025). Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities. Internet of Things and Cyber-Physical Systems, 5, 1–46. https://doi.org/10.1016/j.iotcps.2025.01.001
21. Fox, S., & Fortes Rey, V. (2024). A Cognitive Load Theory (CLT) analysis of machine learning explainability, transparency, interpretability, and shared interpretability. Machine Learning and Knowledge Extraction, 6(3), 1494–1509. https://doi.org/10.3390/make6030071
22. Freeman, S., Wang, A., Saraf, S., Potts, E., McKimm, A., Coiera, E., & Magrabi, F. (2025). Developing an AI governance framework for safe and responsible AI in health care organizations: Protocol for a multimethod study. JMIR Research Protocols, 14, e75702. https://doi.org/10.2196/75702
23. GL Bajaj Institute of Management and Research (GLBIMR). (2025). Policy and guidelines on the use of generative AI: Version 1.0. PGDM Institute, Greater Noida. January 2025.
24. Husieva, Y., & Freenor, M. (2024, October 10). How to define system prompt exfiltration attacks in LLM-

based applications. TELUS Digital Insights.
25. IAPP. (2025). AI governance in practice: Report on the AI governance profession. International Association of Privacy Professionals, 2025. https://iapp.org/resources/article/ai-governance-in-practice/
26. IBM. (2023). What is generative AI? IBM Think. https://www.ibm.com/think/topics/generative-ai
27. IBM. (2024). Exploring privacy issues in the age of AI. IBM Think. https://www.ibm.com/think/insights/ai-privacy
28. IBM. (2025). What are large language models (LLMs)? IBM Think. https://www.ibm.com/think/topics/large-language-models
29. IMDA & AI Verify Foundation. (2024). Model AI governance framework for generative AI: Fostering a trusted ecosystem. Published 30 May 2024. https://aiverifyfoundation.sg/downloads/Model-AI-Governance-Framework-for-Generative-AI.pdf
30. Imperial, J. M., Jones, M. D., & Tayyar Madabushi, H. (2025). Standardizing intelligence: Aligning generative AI for regulatory and operational compliance. SuperIntelligence – Robotics – Safety & Alignment, 2(5). https://doi.org/10.70777/si.v2i5.16189
31. Kandpal, N., Daumé III, H., & Mirhoseini, A. (2022). Deduplicating training data mitigates privacy risks in language models. International Conference on Machine Learning.
32. Knostic. (2025, July 4). Data leakage happens with GenAI. Here's how to stop it. Knostic Blog. https://www.knostic.ai/blog/ai-data-leakage
33. Komprise. (2025). Unstructured data management in the age of AI: IT survey on AI, data and enterprise risk. June 2025. n = 200 IT directors and executives, US. https://www.komprise.com/resource/ai-data-survey-2025/
34. Kumar, A. (2025, November 26). Generative AI in 2026: The 7 research breakthroughs that will redefine everything we know. Medium.
35. LayerX. (2025a, October 7). New research: AI is already the #1 data exfiltration channel in the enterprise. The Hacker News. https://thehackernews.com/2025/10/new-research-ai-is-already-1-data.html
36. LayerX. (2025b). Enterprise AI security report. LayerX Security. https://go.layerxsecurity.com/hubfs/LayerX_Enterprise_AI_and_SaaS_Data_Security_Report.pdf
37. Li, M., Wan, Y., Zhou, L., & Rao, H. (2024). An enhanced governance measure for deep synthesis applications: Addressing the moderating effect of moral sensitivity through message framing. Information & Management, 61(5), 103982. https://doi.org/10.1016/j.im.2024.103982
38. Lorica, B., & Talby, D. (2025). 2025 AI governance survey. August 2025. n = 350+ AI governance professionals. https://gradientflow.com/2025-ai-governance-survey/
39. Mohawesh, R., Ottom, M. A., & Bany Salameh, H. (2025). A data-driven risk assessment of cybersecurity challenges posed by generative AI. Decision Analytics Journal, 15, Article 100580. https://doi.org/10.1016/j.dajour.2025.100580
40. Mustaquim, S. M., Kotal, A., & Yi, P. H. (2025). When privacy isn't synthetic: Hidden data leakage in generative AI models. In 2025 IEEE International Conference on Big Data (BigData). IEEE. https://doi.org/10.1109/BigData66926.2025.11401448
41. Namer, A., Miller, J., Vagts, H., & Maltzman, B. (2023). A cost-effective method to prevent data exfiltration from LLM prompt responses. Technical Disclosure Commons, Defensive Publications Series.
42. Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., & Lee, K. (2025). Scalable extraction of training data from (production) language models. In The Thirteenth International Conference on Learning Representations (ICLR 2025).
43. OneTrust. (2025). AI governance: Navigating risk in the age of innovation. OneTrust AI Governance Survey, September 2025. n = 1,250 governance executives, North America and Europe. https://www.onetrust.com/resources/ai-governance-report/
44. OpenText. (2024). Generative AI governance essentials: Content management and strong information governance for more trustworthy, secure generative AI [White Paper]. OpenText Corporation. Document No. 262-000112-002.
45. Orpak, K. (2025). Generative AI and cybersecurity: Exploring opportunities and threats at their intersection. Maandblad voor Accountancy en Bedrijfseconomie, 99(4), 221–230. https://doi.org/10.5117/mab.99.149299
46. Pahune, S., Akhtar, Z., Mandapati, V., & Siddique, K. (2025). The importance of AI data governance in large language models. Big Data and Cognitive Computing, 9, Article 147. https://doi.org/10.3390/bdcc9060147
47. Palo Alto Networks. (2025). GenAI enterprise security report: AI traffic, DLP incidents, and application proliferation. Unit 42 Research, June 2025. https://www.paloaltonetworks.com/unit42/
48. Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. The Journal of Strategic Information Systems, 34(2), Article 101885. https://doi.org/10.1016/j.jsis.2024.101885
49. Park, S. (2025). Unveiling AI agent vulnerabilities: Data exfiltration. Trend Research, Trend Micro

Incorporated. https://www.trendmicro.com
50. PwC. (2025). 2025 US responsible AI survey: Bridging the gap between principles and practice. PricewaterhouseCoopers, 2025. n = 253 US executives. https://www.pwc.com/us/en/tech-effect/ai-analytics/responsible-ai-survey.html
51. Renn, O., Klinke, A., & van Asselt, M. (2011). Coping with complexity, uncertainty and ambiguity in risk governance: A synthesis. Ambio, 40(2), 231–246. https://doi.org/10.1007/s13280-010-0134-0
52. Rousseau, D. M., Manning, J., & Denyer, D. (2008). Evidence in management and organizational science: Assembling the field's full weight of scientific knowledge through syntheses. Academy of Management Annals, 2(1), 475–515. https://doi.org/10.1080/19416520802211651
53. RSA & Knostic. (2024). Shadow AI in the enterprise: Security professionals survey. RSA Conference 2024. n = 250+ security professionals. https://www.knostic.ai/research/shadow-ai
54. Schneider, J., Kuss, P., Abraham, R., & Meske, C. (2024). Governance of generative artificial intelligence for companies. arXiv preprint arXiv:2403.08802 [cs.AI]. https://doi.org/10.48550/arXiv.2403.08802
55. Sengar, S. S., Hasan, A. B., Kumar, S., & Carroll, F. (2025). Generative artificial intelligence: A systematic review and applications. Multimedia Tools and Applications, 84(21), 23661–23700. https://doi.org/10.1007/s11042-024-20016-1
56. Shanmugarasa, Y., Ding, M., Chamikara, M. A. P., & Rakotoarivelo, T. (2025). SoK: The privacy paradox of large language models: Advancements, privacy risks, and mitigation. Proceedings of the 20th ACM Asia Conference on Computer and Communications Security (ASIA CCS '25), 425–441. https://doi.org/10.1145/3708821.3733888
57. Shao, S., Ren, Q., Qian, C., Wei, B., Guo, D., JingYi, Y., Song, X., Zhang, L., Zhang, W., Liu, D., & Shao, J. (2026). Your agent may misevolve: Emergent risks in self-evolving LLM agents. In Proceedings of ICLR 2026.
58. Sivalingam, E. (2026). GENXAI: A responsible AI governance framework for generative models in regulated industries. IJCSBS, 4(1), 1–34. https://doi.org/10.34218/IJCSBS_04_01_001
59. Spânu, A., & Shadwell, T. (2025). Preventing URL-based data exfiltration in language-model agents. OpenAI Technical Report, 20 January 2025. https://cdn.openai.com/pdf/dd8e7875-e606-42b4-80a1-f824e4e11cf4/prevent-url-data-exfil.pdf
60. Sydorenko, T. (2025, January 2). AI and cognitive offloading: Sharing the thinking process with machines. UX Design. https://uxdesign.cc/ai-and-cognitive-offloading-sharing-the-thinking-process-with-machines-2d27e66e0f31
61. Taeihagh, A. (2025). Governance of generative AI. Policy and Society, 44(1), 1–22. https://doi.org/10.1093/polsoc/puaf001
62. Templeton, A., et al. (2024). Tracing thoughts: Language model scaling and the localization of knowledge. Anthropic Research. https://www.anthropic.com/research/tracing-thoughts-language-model
63. Tonic.ai. (2025, October 28). Preventing training data leakage in AI systems. Tonic.ai Blog. https://www.tonic.ai/blog/prevent-training-data-leakage-ai
64. Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. British Journal of Management, 14(3), 207–222. https://doi.org/10.1111/1467-8551.00375
65. Tredence. (2025, February 5). Building responsible AI: A guide to effective LLM governance. Tredence Blog. https://www.tredence.com/blog/llm-governance
66. Tricco, A. C., Lillie, E., Zarin, W., et al. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Annals of Internal Medicine, 169(7), 467–473. https://doi.org/10.7326/M18-0850
67. Uddin, M., Ul Arfeen, S., Alanazi, F., Hussain, S., Mazhar, T., & Rahman, Md. A. (2025). A critical analysis of generative AI: Challenges, opportunities, and future research directions. Archives of Computational Methods in Engineering. https://doi.org/10.1007/s11831-025-10355-z
68. van der Heijden, J. (2021). Risk as an approach to regulatory governance: An evidence synthesis and research agenda. SAGE Open, 11(3). https://doi.org/10.1177/21582440211032202
69. Ve3 Global. (2024, July 29). The growing threat of data leakage in generative AI apps. VE3 Blog.
70. Versa Networks. (2025, March 17). Shadow AI & data leakage: How to secure generative AI at work. Versa Networks Blog.
71. Vincent, J., Taiwo, P., & Alamu, R. (2025, October). Securing large language models: Addressing data privacy and prompt injection attacks. Obafemi Awolowo University, Nigeria.
72. WIPO. (2024). Generative AI: The main concepts. In Patent Landscape Report on Generative Artificial Intelligence (Chapter 1). https://www.wipo.int/web-publications/patent-landscape-report-generative-artificial-intelligence-genai/en/1-generative-ai-the-main-concepts.html
73. Yang, H., Sutan Ahmad Nawi, H., & Zhang, Y. (2025). Artificial intelligence and large language models in government document management. JLISS, 12(4), 129–145. https://doi.org/10.33168/JLISS.2025.0408
74. Yang, L., Ye, A., Liu, Y., Lu, W., & Huang, C. (2026). LLM-APTDS: A high-precision advanced persistent threat detection system. Future Generation Computer Systems, 178, Article 108315.

75. Zhan, Q., Liang, Z., Ying, Z., & Kang, D. (2024). InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 10471–10506). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.624
76. Zhang, Z., Yu, X., Yang, S., Bai, Y., Wang, Y., & Yang, S. (2025). Research on methods of large language models in the field of sensitive data governance. In Proceedings of the 2024 2nd International Conference on Artificial Intelligence, Systems and Network Security (AISNS '24) (pp. 271–276). Association for Computing Machinery. https://doi.org/10.1145/3714334.3714380