Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

Securing Enterprise Language Model Applications Against Data Leakage and Response Manipulation

Kaik Tavares Barroso
Faculdade Presidente Antônio Carlos, Brazil
Corresponding author: kaikbarroso75@gmail.com
Orcid ID: https://orcid.org/0009-0007-7410-2543

Abstract

Enterprise language-model applications combine internal documents, retrieval, conversation history, and executable tools. This critical narrative and thematic review examines data leakage and response manipulation through targeted discovery completed on September 29, 2026, with a bounded supplementary discovery and source-status audit on October 1, 2026. Primary studies are distinguished from institutional guidance and contextual reviews. The evidence identifies indirect prompt injection, corpus poisoning, retrieval-stage exploitation, log contamination, embedding inversion, and cumulative knowledge-base extraction under different conditions (Chang et al., 2026; Greshake et al., 2023; Morris et al., 2023; Yao et al., 2026). Defensive studies address provenance signaling, structured training, preference optimization, capability enforcement, and action attribution, with configuration-dependent findings (Chen, Piet, et al., 2025; Chen, Zharmagambetov, et al., 2025; Debenedetti et al., 2026; He et al., 2026). The synthesis proposes six assurance boundaries: identity and scope, source authorization, content provenance, model interaction, executable actions, and output release. Provider and inference-data handling, monitoring, and incident response span all six. Authoritative confidentiality and authorization decisions require application or infrastructure controls independent of unconstrained model interpretation. The framework is an author-developed, unvalidated decision aid, not a certification or a NIST/OWASP-endorsed standard. A hypothetical assistant and prospective protocol illustrate permission-preserving

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

processing, cumulative-disclosure budgets, revocation propagation, legitimate-task utility, and verification effort. No original experiment, complete database census, or universal estimate of compromise is claimed.

Keywords: large language models; prompt injection; data leakage; retrieval-augmented generation; access control

1. Introduction

Applications using large language models (LLMs) can answer questions about internal policies, summarize correspondence, and coordinate tools over organizational information. Retrieval-augmented generation (RAG) combines generated answers with external material, while agentic applications extend this arrangement to actions. These architectures create a security question beyond whether the model produces plausible text: which information may enter its context, which instructions may govern its behavior, and which outputs may leave the application? Foundational RAG research explains the combination of parametric and retrieved knowledge, while security benchmarks show that external material can redirect tool-using agents (Debenedetti et al., 2024; Lewis et al., 2020).

Terminology follows the component being discussed. A language model or LLM is the generative component; an assistant is the user-facing application; an agent is a component that proposes or coordinates tool operations; and an agentic application is the surrounding system that supplies state, tools, permissions, and execution services. These operational definitions prevent an observed model response from being confused with an authorized application action. Security operations center (SOC) denotes the organizational log-analysis setting discussed below (Debenedetti et al., 2024; Karanjai et al., 2026).

The problem is especially important for internal-document assistants. A useful assistant may need information about contracts, personnel, financial plans, or technical operations, but legitimate access is not equivalent to unrestricted disclosure or export.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

OWASP identifies sensitive-information disclosure and excessive agency as distinct application risks. Their separation matters: an answer-only assistant can disclose confidential content, whereas an agent can additionally misuse the permissions of connected services. A security design must address both exposure and execution rather than assuming that successful authentication resolves every downstream authorization decision (OWASP Foundation, 2025a, 2025c).

Prompt injection challenges the boundary between instructions and content. In an indirect attack, malicious directions arrive inside material that an application retrieves or processes for a legitimate task. The attacker need not control the user's request or system prompt. Greshake et al. (2023) demonstrated this mechanism in integrated applications, establishing that lower-trust content can influence behavior beyond its evidentiary role. The appropriate interpretation is a demonstrated class of vulnerabilities, not proof that every application has the same exploitability or that every unusual instruction succeeds (Greshake et al., 2023).

Evidence from 2025 and 2026 broadens the problem beyond conspicuous commands. Research has examined targeted answer corruption, suppression of legitimate answers, malicious content optimized for retrieval, and extraction of a knowledge base through repeated questions. These outcomes concern integrity, availability, and confidentiality, respectively. Consequently, a defense that detects a familiar instruction pattern cannot be evaluated as though it also guarantees trustworthy retrieval, prevents cumulative disclosure, or preserves access restrictions across derived data (Chang et al., 2026; Shafran et al., 2025; Yao et al., 2026; Zou et al., 2025).

The supplied reviews provide a relevant conceptual starting point: useful AI performance depends on verification, correction, integration, and governance rather than initial output speed alone. Their productivity findings are not reused as estimates of security effectiveness. Instead, their workflow perspective motivates evaluating the cost and quality of secure completion across the full transaction. This article therefore

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

treats security enforcement and verification effort as components of application performance rather than external overhead that can be omitted from an apparent efficiency gain (Guedes, 2026; Menezes, 2026).

The primary research question is: How can organizations protect internal documents and constrain malicious instructions in applications using language models? Subsidiary questions ask which leakage pathways require different controls, how response manipulation differs from unauthorized action, what existing defenses demonstrate, and how deployment assurance can be tested without confusing benchmark performance with production security. The contribution is a bounded synthesis and a proposed evaluation architecture, not a new attack experiment, a meta-analysis, or a validated compliance standard (Liu et al., 2024; Vassilev et al., 2025).

2. Literature Review

2.1 Separate confidentiality from answer quality

Confidentiality concerns disclosure to an unauthorized recipient or destination. Answer integrity concerns whether the response accurately represents authorized evidence and the intended task. Availability concerns whether legitimate work remains possible. These properties can fail independently: a false answer need not contain a secret, an accurate quotation can still be unauthorized, and a blocked answer can undermine availability without exporting data. This taxonomy is consistent with distinct findings on answer corruption, RAG jamming, and knowledge-base extraction, and prevents a single “safe response” label from obscuring the actual outcome (Shafran et al., 2025; Yao et al., 2026; Zou et al., 2025).

A confidentiality analysis also needs a specified recipient. Data sent to an approved processor under the intended service arrangement is different from data exposed to another tenant or an attacker-controlled destination, although both require governance. Organizational classification, contractual restrictions, and personal-data

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

obligations must therefore be represented separately. The Brazilian legal discussion below concerns personal data, whereas protection of trade secrets and internal operational material may additionally depend on organizational policy and other legal instruments (Brasil, 2018; OWASP Foundation, 2025c).

2.2 Direct injection, indirect injection, and poisoning

Direct prompt injection arrives through a user-facing input; indirect injection arrives through documents, retrieved pages, messages, or tool outputs. The practical distinction is the authority of the content and the attacker-controlled entry point. A legitimate user can unknowingly trigger an indirect attack while asking an ordinary question. Poisoning is broader: altering the retrieval corpus can change an answer without a recognizable instruction to disregard system rules. These mechanisms overlap but should not be collapsed into a single keyword-detection problem (Greshake et al., 2023; Liu et al., 2024; Zou et al., 2025).

The proposal in this article is to represent content origin, user intent, and executable authority as separate application attributes. A document may be approved as evidence about policy while remaining incapable of granting permissions. A retrieved paragraph can legitimately contain imperative language, for example in a procedural manual, without becoming an instruction to the assistant. Conversely, a forged statement claiming administrative authority should not acquire privilege merely because it appears inside a high-ranking search result. This design interpretation follows the separation problem addressed by structured queries and capability-based defenses (Chen, Piet, et al., 2025; Debenedetti et al., 2026).

2.3 The model is not the authorization service

Authentication establishes identity; authorization determines whether that identity may perform an action on a resource. OWASP recommends least privilege, denial by default, and permission validation on every request. Zero trust additionally

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

rejects implicit trust based only on network location. Applied to a document assistant, these principles mean that being inside the corporate network, having access to the chat interface, or presenting an administrative-looking prompt cannot by itself authorize access to every indexed document (OWASP Foundation, n.d.-a; Rose et al., 2020).

The proposed invariant is that no protected document chunk may reach a model-visible component unless a trusted policy decision authorizes the current principal for that resource and operation. This includes model-based reranking, query expansion, summarization, and citation generation whenever they receive protected content, rather than only the final answer model. Retrieval filters restrict candidates, but an independent entitlement check must precede semantic processing that exposes their contents. Identity, tenant, and operation derive from authenticated application state; absent entitlement metadata or an unavailable policy service results in denial. Authoritative confidentiality and authorization decisions must be enforced by application or infrastructure controls independent of unconstrained model interpretation. Generated text may propose an action, but the execution service separately authorizes it (Debenedetti et al., 2026; OWASP Foundation, n.d.-a, n.d.-d).

3. Methodology

3.1 Review design and actual discovery route

This article is a critical narrative and thematic review. The original targeted discovery and bibliographic checks were completed on September 29, 2026, using domain-focused web queries and direct consultation of conference proceedings, author publication records, repositories, and official security documentation. A bounded supplementary public-record discovery and metadata audit was performed on October 1, 2026, and is reported separately in Supplementary Material S1. Proceedings establish publication metadata; repositories identify accessible research versions; institutional pages supply guidance. The supplementary route used reproducible query strings targeting public ACM Digital Library and IEEE Xplore

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

records and a Scopus-domain verification query. It was a web-based check, not a native database search or exported result set. No direct structured query of Scopus, Web of Science, IEEE Xplore, or ACM Digital Library is claimed, and the October audit does not establish exhaustive coverage.

Actual query families included “prompt injection RAG data exfiltration security 2025 2026 site:usenix.org,” “site.usenix.org PoisonedRAG 2025 retrieval augmented generation,” “site.aclanthology.org 2024 InjecAgent,” “site.proceedings.neurips.cc AgentDojo 2024,” “site.nist.gov AI 600-1 generative artificial intelligence profile 2024 adversarial machine learning 2025,” and “site.cisa.gov 2025 AI data security best practices.” Additional title-specific queries checked embedding inversion, training-data extraction, StruQ, CaMeL, and spotlighting. These are representative executed queries, not a reproducible export from a subscription database. English research and guidance were prioritized; Brazilian legislation was read in Portuguese for the jurisdictional discussion.

3.2 Eligibility and source roles

Sources were retained when they directly addressed a leakage mechanism, adversarial manipulation, relevant defense, authorization practice, incident response, or regulatory context. Priority was given to primary conference research from 2023–2026 and official technical guidance. Earlier sources were retained for foundational RAG, training-data extraction, zero trust, and Brazilian data-protection law. General news, commercial promotional claims, Wikipedia, and noninstitutional blogs were excluded from the evidentiary synthesis. The two supplied articles were retained only as contextual reviews supporting the full-workflow perspective.

The revised bibliography contains 36 sources: 17 primary security research studies, comprising 16 conference publications and one retained preprint; one foundational RAG paper; two supplied contextual reviews; and 16 institutional, technical-guidance, or legal sources. The supplementary check added SecAlign as

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

primary conference research. CaMeL is classified as IEEE SaTML 2026 conference research, with its accessible arXiv version distinguished from its publication status; Spotlighting remains the only retained preprint. Source categories describe principal evidentiary roles. Twenty-eight of the 36 sources are dated 2021–2026 (77.8% of all sources); five living OWASP documents are undated and carry October 1, 2026 retrieval dates. Among the 31 dated sources, 28 (90.3%) fall within 2021–2026. Older foundational sources were retained for their relevance rather than removed to inflate recency.

3.3 Extraction and interpretive rules

For primary research, extraction records the threat asset, attacker capability, corpus control, model access, tool and egress privileges, benchmark scale, attack budget, principal endpoint, adaptive testing, source status, and transfer limitation. Table 1 supplies the focal synthesis; Supplementary Material S1 expands the evidence map and identifies fields not extracted rather than assuming an absent capability. Institutional documents support recommended safeguards and are not counted as experimental demonstrations. Task counts, attack success, corpus coverage, and authorized completion have different denominators and are not pooled. Findings are interpreted only within the available evidence and reported conditions.

This preparation used a single evidence-selection and synthesis workflow without independent duplicate screening or coding. No inter-rater reliability statistic was calculated. Candidate discovery was iterative, and no complete initial-identification, deduplication, or exclusion ledger was retained; consequently, no PRISMA flow count or claim of exhaustive coverage is presented. The final corpus and its source roles are stated explicitly, while the absence of a complete discovery ledger remains a limitation. A future systematic extension should preserve database exports and decisions prospectively rather than inventing retrospective counts.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

The revision includes a targeted source-status audit of CaMeL and Spotlighting, the four cited USENIX 2026 records, the additional SecAlign publication, and official NIST and OWASP records. This audit belongs to the same synthesis workflow and does not constitute independent duplicate appraisal. No second-person verification, inter-rater statistic, or retrospective selection ledger is claimed. Supplementary Material S1 distinguishes completed public-record checks from database access and classification checks not performed, preserving the limitation instead of attributing procedures to an unavailable independent reviewer.

3.4 Narrative appraisal and synthesis boundaries

Evidence is appraised qualitatively according to clarity of the threat model, direct observation of the relevant endpoint, realism of retrieval and tool permissions, availability of comparators, adaptation of attacks to the defense, and reproducibility of source metadata or artifacts. These criteria are an author-developed narrative appraisal, not GRADE or a validated security-maturity score. Demonstration of a mechanism is separated from estimation of its prevalence. Likewise, institutional endorsement of a practice does not establish its effect size in a particular deployment (Debenedetti et al., 2024; Liu et al., 2024).

Recommendations introduced below are explicitly identified as proposed deployment rules or illustrative design choices. Their supporting citations identify research or guidance informing the proposal; they do not imply that the complete architecture has been experimentally evaluated. This distinction is essential because standard software-security principles, model-specific defenses, and evolving agent architectures operate at different layers and supply different kinds of assurance (Chen, Piet, et al., 2025; Rose et al., 2020; Souppaya et al., 2022).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

4. Results

4.1 Evidence on Leakage and Manipulation

4.1.1 From external content to adversarial behavior

InjecAgent supplies 1,054 cases spanning 17 user tools and 62 attacker tools, with objectives including private-data exfiltration and harm to users. Its evaluation demonstrates that indirect content can redirect tool-integrated agents under specified prompting configurations. AgentDojo complements this approach with 97 legitimate tasks and 629 security test cases in an extensible environment. Importantly, it evaluates the fact that agents can fail legitimate tasks even without attacks. An application evaluation should therefore measure ordinary task success alongside adversarial success, rather than celebrating a defense that merely prevents the assistant from doing anything useful (Debenedetti et al., 2024; Zhan et al., 2024).

These benchmarks differ from evidence of an incident in a deployed organization. Their value is controlled comparison and explicit tasks, not population-level estimates of breach frequency. Liu et al. (2024) further evaluated five attacks and ten defenses across ten models and seven tasks, providing a framework for comparisons while illustrating the diversity of evaluation conditions. The synthesis consequently reports mechanisms and applicable assumptions rather than combining benchmark outcomes into a universal “prompt injection rate” (Liu et al., 2024).

4.1.2 Retrieval is part of the attack chain

PoisonedRAG shows that a small number of malicious texts can induce attacker-selected answers in large knowledge databases under its evaluated conditions. Its reported 90% success with five injected texts per target question is a result for that attack setup, not a forecast for all enterprise RAG systems. The study challenges the assumption that retrieving a source automatically makes an answer

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

trustworthy; the source-selection and corpus-integrity processes themselves require scrutiny (Zou et al., 2025).

Chang et al. (2026) address a separate gap: malicious content must actually be retrieved under a natural query before it can influence the model. Their experiments span eleven retrieval benchmarks and eight embedding models, and include end-to-end exploitation in a tool-using workflow. The important generalizable mechanism is that adversarial content can be engineered to cross the retrieval barrier. The numerical result from a particular exfiltration workflow should not be transferred to read-only assistants with different permissions, corpora, or egress controls (Chang et al., 2026).

4.1.3 Manipulation includes suppression and cumulative extraction

Shafran et al. (2025) demonstrate RAG jamming through blocker documents that cause a system to withhold legitimate answers. Their method does not depend on ordinary instruction injection. This distinction is operationally important: excessive refusal can be an adversarial outcome rather than evidence that a system is behaving securely. Availability testing must therefore include benign questions associated with adversarially modified corpora, and examine whether the system can recover useful authorized information (Shafran et al., 2025).

Yao et al. (2026) examine adaptive knowledge-base stealing rather than a single visible secret in one answer. Their RAGCrawler evaluation measures corpus coverage under a query budget across specified corpora, generators, and retrieval settings. This supports an additional confidentiality concern: repeated apparently legitimate answers can cumulatively expose a substantial collection. The application-level recommendation is to evaluate disclosure over sessions and accounts, as well as per response, while recognizing that acceptable quotation and bulk extraction must be defined by the organization's policy (Yao et al., 2026).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

4.1.4 Logs and derived representations are sensitive surfaces

Karanjai et al. (2026) demonstrate passive prompt injection in network-security logs, including payloads distributed across entries. Their study evaluates activity concealment, false positives, exfiltration, and output hijacking. Layered mitigations reduced attacks substantially in the reported evaluation but left residual vulnerability. Thus, operational logs should not be assumed to contain trustworthy instructions merely because a security team collected them. The relevant trust level is the origin of individual fields, not the institutional importance of the surrounding workflow (Karanjai et al., 2026).

Embeddings also require confidentiality analysis. Morris et al. (2023) recovered substantial original text from particular embedding representations, including personal information in clinical-note experiments. This does not establish that every embedding is fully invertible, but it refutes treating a vector as inherently anonymized. The proposed organizational policy is to classify derived vectors according to their source sensitivity unless a separate, applicable privacy evaluation justifies a different classification. Vector-store permissions, exports, snapshots, and diagnostic access should therefore appear in the data inventory (Morris et al., 2023).

4.1.5 Training data and runtime context require different controls

Carlini et al. (2021) demonstrated extraction of memorized training examples from GPT-2, and Nasr et al. (2025) examined extraction from aligned production models. These studies concern information encoded during training, which is distinct from material supplied through a contemporary RAG context. A document exposed through a retrieved chunk need not have entered training, and a model-generated memorized passage need not have come from the application's index. Incident classification must preserve this distinction to identify the relevant remediation path (Carlini et al., 2021; Nasr et al., 2025).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

This article proposes separate requirements for training-use restrictions, inference retention, source authorization, and output release. A prohibition on training use addresses one pathway without resolving access to retained prompts, service logs, caches, or downstream copies. The provider and inference-data requirement is therefore transversal to the six boundaries: document permitted training use, inference and diagnostic retention, administrative access, processing location, subprocessors, deletion, and contractual obligations. Each deployment must check actual service configuration and data paths, including exceptions and dependencies. Removal from retrieval storage is distinct from erasure from model parameters (Autio et al., 2024; National Security Agency et al., 2025; OWASP Foundation, 2025c).

4.2 Findings on Existing Defenses

4.2.1 Provenance signaling and structured training

Spotlighting transforms lower-trust inputs to give the model a continuing signal of their provenance. Hines et al. (2024) report substantial attack reduction with limited task impact in their GPT-family experiments. The source is retained as a preprint, and its outcomes are not presented as a universal guarantee. The useful design insight is that provenance should remain visible throughout content processing, rather than appearing once in a short instruction followed by a large unmarked block (Hines et al., 2024).

StruQ combines a secure formatting front end with a specially trained model that distinguishes the instruction and data channels. Its defense is therefore more substantial than adding delimiters to an arbitrary existing model. Chen, Piet, et al. (2025) report improved resistance with little or no utility loss in the evaluated settings. Deployment interpretation must preserve the training dependency: an organization cannot assume that a similarly formatted prompt reproduces StruQ's findings without the corresponding model behavior and evaluation (Chen, Piet, et al., 2025).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

SecAlign adds a training-based approach using preference optimization to favor authorized instructions over injected alternatives. The supplementary check identified its ACM CCS 2025 publication and included it as primary research. Its contribution concerns model behavior under the evaluated training and attack conditions; it does not independently establish document entitlements, authorize tool operations, or control release destinations. Consequently, a trained defense remains complementary to application enforcement, and its findings should not be transferred to an untrained model using similar prompt text (Chen, Zharmagambetov, et al., 2025).

4.2.2 Capability enforcement and action attribution

CaMeL separates control and data flows and uses capabilities to constrain information flows during tool execution. Its publication status was updated to IEEE SaTML 2026, confirmed against the official conference program and the authors' publication record; the accessible arXiv manuscript remains a research-access route rather than the source-status label. Its guarantees depend on its architecture, policies, and stated assumptions. It illustrates enforcement in a surrounding system layer and does not imply that an arbitrary agent becomes secure after receiving a natural-language description of capabilities (Debenedetti et al., 2026).

AttriGuard instead evaluates the causal support for proposed tool invocations using counterfactual execution. He et al. (2026) report no successful static attacks in their tested models and benchmarks, alongside adaptive-attack evaluation. This is evidence of a promising defense under explicit conditions. It is not evidence that the probability of compromise is zero in production, nor does it automatically cover separate document-authorization failures or training-data leakage. Organizations should examine utility, overhead, and coverage of their actual tool environment before adoption (He et al., 2026).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

4.2.3 Defense claims need coverage statements

Table 2 organizes controls by the boundary they protect and the residual failure they cannot independently resolve. This is a proposed comparison, not a ranking of universal effectiveness. Prompt-oriented defenses principally address behavioral redirection; permissions constrain available resources; output handling constrains release; provenance protects traceability; incident processes support containment. The article's central synthesis is that these layers are complementary because they answer different security questions, not because the literature supplies a single proven optimal combination (Chen, Piet, et al., 2025; He et al., 2026; Hines et al., 2024).

5. Discussion

5.1 Protecting Internal Documents Across Their Lifecycle

5.1.1 Inventory the complete document path

The 2025 joint AI-data guidance emphasizes provenance, trusted infrastructure, and integrity of data throughout the AI lifecycle. Applied here, the proposed inventory follows each approved source into extracted text, chunks, embeddings, indexes, conversational context, caches, traces, and exports. Every representation should have an owner and a retention purpose. This inventory is a design synthesis: it prevents the narrow assumption that protecting the original file also protects every copy made by the assistant's processing pipeline (National Security Agency et al., 2025).

Data minimization should be task-specific. For an internal-policy question, the proposed retrieval budget is the smallest collection of authorized excerpts sufficient to answer, rather than the largest context window available. Sensitive attributes unrelated to the task should be removed or replaced before submission where feasible. However, pseudonymization should not be described as guaranteed anonymization, and masking must not erase facts needed to interpret a decision correctly. The

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

appropriateness of processing personal information remains a separate governance question (Autio et al., 2024; Brasil, 2018).

5.1.2 Preserve permissions through transformation

OWASP's RAG guidance calls for access-control metadata to accompany chunks and for permission checks at retrieval time. An indexed copy must not preserve an outdated entitlement after the source permission changes. The article proposes that a chunk carries its tenant, source identifier, version, classification, authorization reference, and expiry where applicable. Permission evaluation should occur before sensitive content reaches model-visible context, including any model-based reranker, query-expansion assistant, or citation-generating component (OWASP Foundation, n.d.-d).

A proposed robust implementation uses authenticated application identity to generate a policy-restricted retrieval request and independently validates candidate entitlements before prompt assembly. This can be implemented using filtered indexes, physically separated collections, or equivalent enforcement appropriate to the platform. The requirement is the protected outcome, not a particular vendor feature. If the authorization service is unavailable or metadata is missing, the proposed behavior is to stop the protected operation rather than broaden the search. This recommendation derives from denial-by-default and request-level authorization principles (OWASP Foundation, n.d.-a).

5.1.3 Govern ingestion and integrity without overstating provenance

Source onboarding should document who may submit or amend material, how versions are approved, and which connectors can import documents. The proposed ingestion process records approved versions and integrity checks, isolates parsing from sensitive systems, and distinguishes externally supplied attachments from authoritative organizational policies. Signed or hashed provenance can establish origin

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

or detect changes, but it does not prove that the approved author supplied accurate or harmless content. Integrity verification and semantic scrutiny therefore answer different questions (National Security Agency et al., 2025).

The same reasoning applies to hidden or transformed text. The proposed review compares relevant visible content with extracted content when high-impact documents are ingested, and flags unusual text or instructions for review without assuming that every imperative sentence is malicious. False positives matter for procedural manuals and security reports that legitimately discuss attacks. Input inspection should be an additional safeguard, not the service responsible for granting document access or deciding whether an export is permitted (Chen, Piet, et al., 2025; OWASP Foundation, n.d.-b).

5.1.4 Keep secrets outside language-model context

Credentials should be handled by a secrets-management service and scoped to the operation that needs them. The model generally needs to request an approved action, not read the token used to authenticate that action. The proposed architecture injects credentials only within the execution boundary, limits their lifetime and scope, and supports revocation and rotation. This design follows OWASP's least-privilege and secret-lifecycle guidance and reduces the quantity of directly exposable authentication material in prompts and traces (OWASP Foundation, n.d.-e).

System prompts should consequently contain behavior specifications rather than authentication secrets. The proposal treats disclosure of a prompt as an integrity or configuration concern according to its content, while maintaining separate enforcement of real permissions. A system whose security depends on an undisclosed sentence has not supplied an independent authorization boundary. The relevant risk reduction comes from removing unnecessary sensitive content and controlling executable privileges, rather than expecting all internal instructions to remain permanently unreadable (Debenedetti et al., 2026; OWASP Foundation, 2025b).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

5.1.5 Revocation includes caches and memory

Document protection must include deletion and de-permissioning. OWASP's RAG guidance explicitly addresses derived indexes and response caches. The proposed policy requires an entitlement change to invalidate reusable answers containing affected material, and requires historical conversation context to be checked before reuse. Where stored transcripts remain subject to legitimate retention, access must still be restricted. A person who loses access to a document should not regain its content through a cached answer generated while another user had permission (OWASP Foundation, n.d.-d).

Revocation should be evaluated as a propagation process. The proposed clock starts at the authoritative entitlement-change event and stops when denial is effective across all affected indexes, embeddings, caches, conversation history, and reusable generated artifacts. Tests should record the delay for each surface, the maximum observed propagation time, failed propagation, and any disclosure during the interval. An organization should prespecify its acceptable delay and apply a fail-closed or suspension policy where a stale derivative cannot be safely reauthorized. No universal time threshold is proposed, and revocation cannot undo information already received by a user (OWASP Foundation, n.d.-a, n.d.-d).

The article also proposes scoped conversational memory: retained preferences should not silently become policy instructions, and information from one tenant or task should not be imported into another without an authorized purpose. This is an architectural extension rather than an experimentally validated memory defense. It responds to the broader finding that lower-trust material can affect later agent behavior, and to the need to preserve origin and authority when content is transformed or replayed (Greshake et al., 2023; Zhan et al., 2024).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

5.2 Constraining Responses and Tool Actions

5.2.1 Validate proposed actions outside the model

Excessive agency can arise from unnecessary functionality, excessive permissions, or excessive autonomy. OWASP specifically distinguishes a read requirement from a connector that also permits modification and deletion. The proposed starting point for a document assistant is therefore read-only access with a minimal tool set. Adding export, messaging, write, or execution capability should require a documented business purpose and a corresponding policy boundary rather than a broad grant inherited from an administrator's account (OWASP Foundation, 2025a).

For each proposed action, this article recommends an external check of the authenticated principal, operation, resource, destination, payload classification, and task scope. A syntactically valid tool call is not necessarily authorized. A generated identifier or recipient must be resolved against trusted application state, and a policy decision must precede execution. Parameters should be parsed into constrained data structures rather than interpreted as unrestricted executable text. These implementation recommendations synthesize ordinary secure-development practice with research on capabilities and action attribution (Debenedetti et al., 2026; He et al., 2026; Souppaya et al., 2022).

5.2.2 Apply destination and release controls

A disclosure path may involve the answer channel, an export feature, a connected service, or content rendered by a client. The proposed release service checks what is leaving, who is receiving it, and whether that destination is authorized for the classification. The same service should govern bulk download and external sharing rather than leaving those operations to conversational persuasion. A model's statement that an export is necessary is evidence of a proposal, not proof of user authorization (OWASP Foundation, 2025a; OWASP Foundation, 2025c).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

The article recommends disabling automatic fetching of arbitrary external resources in generated content unless it is required and independently constrained. Rendering and outbound requests are part of the application, so restricting only the backend tool list may leave a separate release path. Generated structured output should also be checked against the application's schema and policy before downstream use. This recommendation addresses output-mediated disclosure and manipulation as deployment concerns, informed by OWASP's prompt-injection and RAG prevention guidance (OWASP Foundation, n.d.-b; OWASP Foundation, n.d.-d).

5.2.3 Protect answer integrity and informed human decisions

Response validation should examine whether material claims are supported by authorized sources and whether the cited passage actually entails the claim. Citation presence alone is insufficient when the cited corpus can be poisoned. The proposed interface distinguishes documented evidence, unresolved uncertainty, and suggested next actions. For contested or conflicting policies, it should preserve source version and ownership information rather than inventing a reconciled rule. This recommendation follows the contrast between retrieval grounding and knowledge-corruption risk (Lewis et al., 2020; Zou et al., 2025).

Human approval requires visibility of the actual resource, operation, destination, and payload, together with information classification and relevant evidence. The proposed approval interface resolves these fields from trusted application state and binds confirmation to the displayed transaction. A subsequent change in destination or payload invalidates the previous approval and requires reevaluation. The model's explanation is supplementary, not the authoritative transaction record. Longitudinal evaluation should examine approval errors and review effort under repeated requests to detect routine approval or habituation; this is a prospective research target, not an experiment performed here (Autio et al., 2024; Debenedetti et al., 2026).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

5.3 A Proposed Deployment Assurance Framework

5.3.1 Status and scope of the framework

This article proposes six deployment-assurance boundaries: identity and scope, source authorization, content provenance, model interaction, executable actions, and output release. Provider and inference-data handling, monitoring, and incident response span all six. Provider handling is a transversal responsibility, not a seventh sequential boundary. It requires evidence on training use, inference retention, administrative access, processing location, subprocessors, deletion, and contracts. The framework is an author-developed, unvalidated decision aid; it is not a validated score, standardized risk scale, regulatory certification, or formally NIST/OWASP-endorsed standard (Autio et al., 2024; National Security Agency et al., 2025; Vassilev et al., 2025).

The proposed boundaries are intended to expose specific failure points. For example, successful injection detection does not repair unauthorized retrieval, and accurate retrieval does not authorize external transmission. Each deployment should attach its own policy owner, enforcement component, test evidence, and residual limitation to every boundary. Table 2 operationalizes this mapping. The novelty claimed here is the organization of evidence into a reviewable assurance argument; the underlying controls and attack mechanisms derive from established guidance and primary research (Chang et al., 2026; Rose et al., 2020; Zhan et al., 2024).

5.3.2 Context determines required assurance

The proposal uses contextual escalation rather than a numerical risk score. Required assurance increases when documents are more sensitive, recipients are less trusted, actions are less reversible, tool privileges are broader, or decisions have greater consequences. A system answering questions about an approved public handbook differs from one summarizing personnel files or sending contract material externally. Local legal and organizational responsibilities determine whether an

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

operation is permitted at all; additional monitoring cannot compensate for an operation that lacks authorization (Brasil, 2018; OWASP Foundation, 2025a).

For low-consequence read-only use, the proposed minimum includes documented data classification, verified entitlements, source-linked answers, restricted retention, and a functioning incident path. For sensitive internal use, the proposal adds stronger tenant separation, revocation tests, cumulative-disclosure monitoring, and explicit destination restrictions. For consequential actions, it adds transaction-bound approval and independently enforced execution policies. These tiers of implementation effort are recommendations for local adaptation rather than validated thresholds or universal deployment categories (National Cyber Security Centre & Cybersecurity and Infrastructure Security Agency, 2023; Nelson et al., 2025).

5.3.3 Hard gates and residual-risk decisions

The proposed release decision uses hard gates for authorization and controlled execution. A test revealing cross-tenant content or an unapproved export blocks deployment for the affected workflow until corrected; faster answers do not offset those violations. Other outcomes, such as answer completeness and latency, should remain separate measures with locally defined tolerances. This prevents the economically attractive but invalid aggregation in which reduced task time cancels a serious confidentiality failure (Menezes, 2026; OWASP Foundation, n.d.-a).

The proposed assurance record identifies the model and version, prompts and configuration, embedding model and version, retriever version and parameters, corpus snapshot, enabled tools and privileges, permission-policy version, release destinations, attack families, attempt budgets, and excluded surfaces. It also records provider processing and retention configuration, the responsible parties, and the evidence date. A risk owner reviews residual uncertainty for a stated use case and review date. Passing finite tests does not establish the absence of unknown vulnerabilities or applicability after architectural changes (National Cyber Security

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

Centre & Cybersecurity and Infrastructure Security Agency, 2023; Vassilev et al., 2025).

5.4 Hypothetical Internal Document Assistant

5.4.1 Legitimate task and protected assets

Consider a hypothetical accounting organization deploying an assistant over approved procedure manuals, client contracts, and restricted personnel records. An employee asks for the current expense-reimbursement procedure. This scenario is illustrative and uses no real organization or participant data. The desired outcome is a source-supported answer from authorized procedures, without exposure of personnel information or client material irrelevant to the task. The example applies the reviewed authorization and document-lifecycle principles rather than reporting an implemented system (National Security Agency et al., 2025; OWASP Foundation, n.d.-a).

Before retrieval, the application authenticates the employee and obtains the organizational and task scope from trusted state. It queries only authorized collections and validates the entitlement of each candidate passage. The retriever returns approved procedure excerpts with source version and ownership metadata. A personnel document is excluded before any language-model processing, including reranking. This illustrates the proposed source-authorization invariant; asking the model to ignore a restricted passage after providing it would not demonstrate equivalent protection (OWASP Foundation, n.d.-d).

5.4.2 Lower-trust content cannot create authority

Assume an imported attachment associated with an otherwise relevant record contains instructions to disclose unrelated internal material or change the conclusion. The illustrative pipeline preserves its origin and treats it as evidence to inspect, not a grant of authority. The model receives a clear task specification and marked source content. Input safeguards may flag the suspicious text, but the principal confidentiality

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

control remains that unrelated restricted material is unavailable. The example separates behavioral resistance from resource confinement in the manner motivated by indirect-injection and structured-query research (Chen, Piet, et al., 2025; Greshake et al., 2023).

If the assistant proposes a new export or message, the execution layer checks whether that operation is part of the authenticated request and whether the destination is permitted. In the scenario, the employee requested an answer and no export capability was issued, so the proposal cannot execute. An authorized user could initiate a separate export workflow, but the destination and data would require their own policy checks. This example illustrates capability scoping; it does not assume that intent can be safely inferred solely from the model's explanation (Debenedetti et al., 2026; OWASP Foundation, 2025a).

5.4.3 Useful responses and controlled recovery

The output service checks that the answer is grounded in the approved procedure and that quotations remain within the task's release rules. The response includes the policy identifier and version, flags unresolved conflicts, and avoids silently converting a document instruction into a new organizational rule. Where source evidence is unavailable, the illustrative assistant reports that limitation and directs the employee to the responsible policy owner. The intended outcome is useful authorized assistance, not indiscriminate refusal or fabricated certainty (Shafran et al., 2025; Zou et al., 2025).

Suppose the policy is later withdrawn or the employee changes departments. The proposed revocation procedure updates retrieval entitlements and invalidates associated answer caches before they can be reused. The security test repeats the earlier question under the changed identity and verifies that historical context does not restore access. Table 3 includes this scenario as a release test. It is a design example

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

informed by derived-data retention guidance, not evidence that every indexing platform already implements automatic revocation correctly (OWASP Foundation, n.d.-d).

5.5 Prospective Evaluation and Measurement

5.5.1 Define the unit of evaluation

The recommended unit is a complete authorized task executed against a specified corpus, identity, tool set, and policy state. An evaluation should separately record whether attacker material entered retrieval results, entered model context, changed the answer, triggered a tool proposal, executed an unauthorized operation, or caused an unauthorized release. These events are stages of a chain, not interchangeable definitions of success. Retrieval-oriented research and tool-agent benchmarks support measuring intermediate stages rather than scoring only the final textual response (Chang et al., 2026; Zhan et al., 2024).

Stage-specific reporting should use a declared eligible set of attack attempts. Retrieval success is the number of attempts returning attacker-controlled material divided by eligible retrieval attempts; context inclusion is counted among successful retrievals; behavioral manipulation is counted among contexts containing that material. Unauthorized proposals, executions, and releases are reported as separate counts, with denominators stated for both the relevant preceding stage and end-to-end eligible attempts when meaningful. Paths that bypass retrieval should be reported separately. Conditional rates describe the tested workflow, not independent probabilities to multiply into a universal compromise rate. Each result requires the success rule, numerator, denominator, exclusions, and attempt budget (Chang et al., 2026; Liu et al., 2024; Zhan et al., 2024).

The proposed protocol combines legitimate tasks, adversarial modifications of those tasks, benign imperative documents, cross-tenant access tests, permission changes, and multi-turn disclosure attempts. Each test has an external expected result

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

defined before generation. Synthetic records and canary values should replace real secrets in evaluation. Human adjudication should inspect ambiguous outcomes, while deterministic checks establish whether a protected resource or action crossed a policy boundary. Table 3 presents the recommended matrix, which remains a prospective protocol requiring implementation and validation (Debenedetti et al., 2024; Liu et al., 2024).

5.5.2 Separate security, utility, and effort

Attack success should be reported with its exact numerator, denominator, and definition. Unauthorized-data release should specify the recipient, protected content, and channel. Answer manipulation should be checked against independent source-grounded criteria, while legitimate-task completion and excessive refusal should be measured separately. A defense can lower attack success by preventing all work, so security claims must disclose utility loss. This concern is directly visible in the agent-benchmark distinction between benign performance and robustness under attack (Debenedetti et al., 2024; Shafran et al., 2025).

False denial should be operationalized as an authorized, answerable task incorrectly blocked under the prespecified entitlement and evidence state, divided by such eligible tasks. Correct refusal of an unauthorized request must be counted separately, as must inability to answer because authorized evidence is unavailable. Excessive refusal, incorrect completion, verification time, and correction effort remain distinct outcomes. This prevents a deny-all configuration from acquiring an artificial security advantage and keeps availability failures visible (Debenedetti et al., 2024; Shafran et al., 2025).

The proposed utility analysis also records retrieval relevance, answer completeness, verification time, correction effort, approval effort, latency, and resource consumption. These measures should not be compressed into one unvalidated “secure productivity” score. Comparisons require matched tasks and the same

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

authorization requirements in defended and baseline systems. The two supplied reviews motivate inclusion of downstream verification costs, while the protocol extends that principle to security controls without claiming that productivity estimates from those articles measure injection resistance (Guedes, 2026; Menezes, 2026).

5.5.3 Test adaptation and repeated attempts

The recommended evaluation distinguishes fixed attacks from attacks adapted to the defense. A known payload set can estimate regression on that set, but it cannot alone establish resistance to adversaries who change their approach. Repeated generations and multi-turn sessions should also be tested under declared attempt budgets. This recommendation is informed by benchmark comparison, adaptive capability analysis, and action-attribution research; it does not require publishing operational attack instructions or exposing actual internal documents (He et al., 2026; Liu et al., 2024; Yao et al., 2026).

Cumulative-disclosure tests should prespecify a session and account query budget, an observation window, and a synthetic protected corpus. The proposed record includes protected-source coverage, unique sensitive facts or tokens released, disclosure per session and account, and aggregation across sessions. It should distinguish authorized quotation from prohibited reconstruction using an externally defined policy. Limits are organization-specific; a query limit alone is not proof of confidentiality. Tests should retain both within-budget and attempted beyond-budget outcomes and assess whether paraphrasing or distributed sessions evade the stated release rules (Yao et al., 2026; OWASP Foundation, 2025c).

Finite tests require qualified conclusions. Observing no unauthorized export in a test set means no export was observed under those conditions; it does not establish zero production risk. Evaluators should report uncertainty appropriate to the sampling and dependence structure, rather than treating repeated variants as independent organizational incidents. Likewise, a nonsignificant difference in utility does not prove

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

formal noninferiority without a prespecified margin and adequate precision. These are proposed interpretive rules for the protocol, motivated by the heterogeneous units and conditions in the reviewed evaluations (He et al., 2026; Liu et al., 2024).

5.5.4 Evaluate changes and defense dependencies

Model changes, embedding updates, connector changes, prompt revisions, and permission-policy revisions should trigger evaluation of the affected boundary. The proposed release record identifies which components changed and which existing assurance evidence remains applicable. A defense trained into a particular model should not be assumed to transfer unchanged to a new model, and an authorization guarantee should not be attributed to model behavior. Secure-development guidance explicitly treats changes across the AI lifecycle as security-relevant, while StruQ illustrates a defense whose behavior depends on its training design (Chen, Piet, et al., 2025; National Cyber Security Centre & Cybersecurity and Infrastructure Security Agency, 2023).

5.6 Governance and Incident Response

5.6.1 Assign responsibility across providers and operators

Organizational responsibility should distinguish the application owner, data owner, identity administrator, connector operator, model provider, and incident-response lead. The proposed governance record assigns each party an enforceable obligation and identifies evidence of configuration, rather than relying on a general assurance that a product is enterprise-ready. The NIST generative-AI profile supplies voluntary risk-management support; it is not a certification that a particular application protects its documents or resists injection (Autio et al., 2024).

Hosted and local deployments should be assessed through actual data paths and assigned responsibilities. The provider and inference-data register should state permitted training use; inference, cache, diagnostic-log and backup retention;

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

administrative access and auditability; processing locations; relevant subprocessors; deletion scope, timing and exceptions; and applicable contractual controls. The application owner documents configured options, the data owner approves permitted processing, and provider/operator obligations are recorded separately. Changes to these terms or settings trigger review of the affected boundaries. Local hosting can remove an external transfer without correcting unsafe document or tool permissions; a no-training clause does not establish authorization (Autio et al., 2024; National Security Agency et al., 2025; Rose et al., 2020).

5.6.2 Protect logs while retaining useful evidence

OWASP recommends purpose-based application logging and cautions against excessive sensitive content in logs. The proposed audit record includes principal, policy decision, source identifiers, tool requests, execution outcomes, and release destinations, with sensitive payloads minimized or separately protected. Logging raw prompts and full documents by default can create a second confidential corpus. Access to traces should therefore be limited, and retention should reflect a defined operational and legal purpose rather than debugging convenience alone (OWASP Foundation, n.d.-c).

Security analysts must also treat event fields as lower-trust data when submitting logs to a language model. The operational recommendation is to preserve raw evidence in controlled storage, keep AI interpretation separate from authoritative event facts, and require independent confirmation before high-impact containment actions. This responds to the demonstrated manipulation of log-analysis assistants and avoids making a potentially compromised summary the sole basis for an operational decision (Karanjai et al., 2026).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

5.6.3 Contain the affected workflow and restore assurance

NIST's 2025 incident-response guidance integrates preparation, detection, response, and recovery with cybersecurity risk management. Applied to this article's architecture, the proposed playbook can suspend an affected connector, disable exports, quarantine a corpus version, revoke exposed credentials, or stop a tool capability while preserving evidence. These are illustrative containment actions to be selected according to the confirmed exposure and business context; they are not a claim that every manipulation requires shutting down the entire application (Nelson et al., 2025).

Recovery should examine source changes, entitlement propagation, model-visible context, tool execution, output destinations, and downstream caches. The proposed root-cause record distinguishes a failed detector from a missing authorization check or an uncontrolled release channel. Reinstatement should require correction of the affected boundary and targeted retesting, with broader evaluation when the architectural assumptions changed. Secret exposure additionally requires an appropriate revocation and rotation process; removing the offending prompt alone cannot invalidate a credential already obtained (Nelson et al., 2025; OWASP Foundation, n.d.-e).

5.6.4 Brazilian personal-data context

For organizations within the scope of Brazil's General Personal Data Protection Law, purpose, necessity, security, prevention, and accountability affect the treatment of personal data. Article 46 requires appropriate technical and administrative protection against unauthorized access and other unlawful or inadequate processing. The compiled law also addresses international transfers and security-incident communication. This discussion is limited to the Brazilian jurisdiction and is neither legal advice nor a complete compliance assessment; the governance recommendation

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

is to evaluate the actual processing operation, lawful basis, roles, and applicable rules with qualified organizational review (Brasil, 2018).

The proposed technical controls can provide evidence for that review: data classification shows what is processed, authorization records show who may access it, retention records show how long derivatives persist, and release logs show where information went. Technical success is nevertheless distinct from lawful processing; a perfectly isolated system may still process unnecessary information or use it for an incompatible purpose. Regulatory and contractual conclusions should therefore remain separate from benchmark claims about resistance to prompt injection (Autio et al., 2024; Brasil, 2018).

5.7 Discussion and Research Priorities

The synthesis identifies a practical distinction between making the model less susceptible to manipulation and preventing manipulation from producing an unauthorized consequence. Provenance signaling and structured training address behavioral confusion; capability systems and execution controls address available actions and flows. Action attribution examines whether proposed operations reflect the user's intent. None of these alone supplies all document-lifecycle protections. The recommended design therefore states the boundary each control protects and tests that boundary directly (Chen, Piet, et al., 2025; Debenedetti et al., 2026; He et al., 2026; Hines et al., 2024).

An unresolved research need is evaluation under realistic organizational entitlements and evolving corpora. Many useful benchmarks emphasize attacker-controlled content or tool misuse, whereas an enterprise must also handle revoked access, conflicting document versions, sanctioned sharing, and cumulative extraction. Future studies should report these policy states explicitly and compare defenses under identical resource permissions. This agenda is motivated by distinct evidence on retrieval optimization, knowledge-base stealing, and access-control inheritance, not by

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

an assumption that a single benchmark already represents every enterprise workflow (Chang et al., 2026; OWASP Foundation, n.d.-d; Yao et al., 2026).

Longitudinal studies should examine whether operators preserve controls during updates and whether high request frequency produces approval habituation. Prospective measures include the proportion of unauthorized transactions approved, the proportion of authorized transactions wrongly rejected, review time, inspection of the actual payload, and changes across repeated sessions. Monitoring should also be evaluated for expansion of confidential-data exposure. Verification, correction, approval, and incident-handling effort should remain distinct outcomes, alongside legitimate-task utility. These research priorities require independent prospective validation and were not tested in this review (Autio et al., 2024; Guedes, 2026; Menezes, 2026).

5.8 Limitations

This review is selective rather than a comprehensive census of language-model security research. Original discovery relied on targeted web retrieval; the October 1, 2026 supplement added bounded public-record queries and metadata verification without native major-database exports. It does not repair the absence of a complete candidate, deduplication, and exclusion ledger. Scopus/Web of Science coverage remains unverified, and public search snippets cannot establish full database coverage. One synthesis workflow without independent duplicate screening or appraisal leaves selection and interpretation bias. Expanded evidence fields and source-status checks improve transparency but do not eliminate these limitations.

Primary studies differ in models, tools, corpora, budgets, and endpoints. Controlled demonstrations establish mechanisms under their assumptions rather than global incident prevalence or universal effectiveness. Several recommendations extend institutional principles to a proposed architecture rather than reproduce one evaluated system. CaMeL is now treated as IEEE SaTML 2026 conference research,

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

while Spotlighting remains the single retained preprint in the audited corpus. The October 1, 2026 metadata check does not guarantee that later versions or deployment changes will preserve these interpretations (Debenedetti et al., 2026; Hines et al., 2024; Liu et al., 2024).

The focus is primarily text-based enterprise document assistance, RAG, and tool integration. Multimodal perception, training-pipeline security, cryptographic privacy mechanisms, and specialized regulated systems receive limited coverage. Geographic and language coverage is also uneven; English research and international institutional guidance dominate, with Brazilian law included for a bounded legal context. The proposed framework requires adaptation to local entitlements, infrastructure, professional responsibilities, and jurisdiction rather than direct adoption as a universal compliance checklist (Brasil, 2018; Vassilev et al., 2025).

No new experiment validates the architecture, release gates, or worked example. The article supplies a defensible design and evaluation proposal, not a guarantee that malicious instructions cannot affect a model. Practical assurance remains conditional on correct implementation, declared policy, adequate test coverage, and continued operation of controls. Future work must compare the complete proposed workflow against realistic baselines and examine both residual compromise and legitimate-task performance (Debenedetti et al., 2024; He et al., 2026).

6. Conclusion

Protecting internal documents in language-model applications requires separate treatment of document entitlements, lower-trust instructions, tool authority, and output destinations. Primary research demonstrates several distinct failure mechanisms, including indirect injection, targeted corpus corruption, answer suppression, derived-representation exposure, and cumulative extraction. These findings support a layered security argument while preventing the misleading inference

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

that one attack percentage or one defense result characterizes every application (Morris et al., 2023; Shafran et al., 2025; Yao et al., 2026; Zou et al., 2025).

Authoritative confidentiality and action-authorization decisions should be enforced by application or infrastructure controls independent of unconstrained model interpretation. Source authorization precedes every model-visible use of protected content; execution depends on separately verified capabilities; release depends on recipient and destination policy. Provider and inference-data handling is a transversal requirement alongside monitoring and incident response. Model-oriented defenses can reduce susceptibility within their evaluated conditions. The author-developed six-boundary framework remains unvalidated, is not a validated score or certification, and carries no formal NIST or OWASP endorsement (Autio et al., 2024; Chen, Piet, et al., 2025; Debenedetti et al., 2026; OWASP Foundation, n.d.-a).

Deployment assurance requires evidence of permitted access, useful authorized completion, stage-specific adversarial outcomes, cumulative-disclosure limits, revocation propagation, and operational recovery. These measures should be interpreted separately, with exact definitions and recorded configurations. A violation of a hard confidentiality or execution gate cannot be offset by productivity. The hypothetical assistant and prospective protocol illustrate how to test these responsibilities; neither is an implemented or experimentally validated system. Practical assurance remains conditional on implementation, policy, coverage, and ongoing reevaluation (Debenedetti et al., 2024; Menezes, 2026; Nelson et al., 2025).

Data Availability

No original participant dataset or attack experiment was generated. Publications and official documents are listed in the references. Tables 1–3 provide the primary synthesis and proposed evaluation matrix. Supplementary Material S1 contains the bounded supplementary discovery record, source-status audit, and expanded

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

evidence map. The hypothetical scenario contains no real confidential organizational records.

Declaration on Generative AI Use

Generative AI was used to assist with targeted literature discovery, bibliographic checking, organization, revision, and academic English editing. The author is responsible for reviewing the manuscript, verifying its claims and references against the cited sources, and approving the final submission, and takes full responsibility for its content and interpretation.

Author Responsibility Statement

The submitting author is responsible for independently verifying the claims and references, confirming authorship contributions, and ensuring that all authors approve the final manuscript and agree to its submission.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

Table 1

Primary security research and boundaries of interpretation

Source and status	Asset and evaluated setting	Observed mechanism or contribution	Principal transfer limitation
Carlini et al. (2021) Conference	Training examples; GPT-2 queries	Extracted memorized sequences from model parameters.	Historical model and training-data setting; not an estimate of enterprise RAG leakage.
Greshake et al. (2023) Conference	Integrated applications; external content	Demonstrated indirect instructions affecting behavior and connected application functions.	Case demonstrations establish mechanisms, not incident prevalence.
Morris et al. (2023) Conference	Text embeddings; evaluated models and clinical-note material	Reconstructed text and sensitive attributes from particular vector representations.	Results depend on embedding access and evaluated representation; not universal inversion.
Liu et al. (2024) Conference	LLM-integrated tasks; multiple models, attacks and defenses	Formalized injection and compared five attacks and ten defenses over seven tasks.	Different task and attack conditions prevent a universal effectiveness ranking.
Zhan et al. (2024) Conference	Tool-integrated agents; InjecAgent	1,054 cases examine user harm and private-data exfiltration.	Controlled agent cases; configuration-specific rather than deployment prevalence.
Debenedetti et al. (2024) Conference	Tool-using agents; AgentDojo	Extensible benchmark with 97 legitimate tasks and 629 security cases.	Benign task failures and attack outcomes must be distinguished.
Hines et al. (2024) Preprint	Untrusted text in evaluated GPT-family workflows	Spotlighting marks provenance through content transformations.	Model and test-specific findings; no general confidentiality guarantee.
Chen, Piet, et al. (2025) Conference	Instruction and data channels; StruQ	Secure formatting plus special training improves instruction separation.	Formatting alone does not reproduce the evaluated training-dependent defense.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

Source and status	Asset and evaluated setting	Observed mechanism or contribution	Principal transfer limitation
Chen, Zharmagambetov, et al. (2025) ACM CCS conference	Trained instruction-following models; SecAlign	Preference optimization favors authorized instructions over injected alternatives.	Behavioral defense depends on training and tested attacks; does not authorize resources or execution.
Debenedetti et al. (2026) IEEE SaTML conference	Tool agents; CaMeL	Separates control and data flows; enforces capability policies.	Assurance is conditional on architecture, policies and stated assumptions.
Nasr et al. (2025) Conference	Aligned production models; training-data extraction	Demonstrated limits of alignment in preventing extraction.	Parameter memorization differs from runtime retrieval exposure.
Shafran et al. (2025) Conference	RAG answers; blocker documents	Demonstrated denial of legitimate answers without ordinary instruction injection.	Availability endpoint; refusal is not synonymous with confidentiality protection.
Zou et al. (2025) Conference	RAG knowledge bases; target questions	Poisoned corpus material induces attacker-selected answers.	Corpus insertion and target-query assumptions govern applicability.
Chang et al. (2026) Conference	Retrieval pipelines and agent workflows	Evaluated crossing the retrieval barrier and subsequent compromise.	Privileges and egress determine consequences; cannot transfer unchanged to read-only systems.
He et al. (2026) Conference	Proposed agent tool calls; AttriGuard	Counterfactual attribution evaluates whether actions follow trusted intent.	Reported zero static-attack success concerns tested cases, not zero future risk.
Karanjai et al. (2026) Conference	Security-log analysis assistants	Passive injection and cross-entry contamination; layered defense leaves residual risk.	SOC log configurations and tested objectives define the transfer boundary.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

Source and status	Asset and evaluated setting	Observed mechanism or contribution	Principal transfer limitation
Yao et al. (2026) Conference	RAG corpora; repeated black-box queries	Adaptive questioning accumulates knowledge-base coverage.	Coverage under query budgets differs from one-response leakage or incident probability.

Note. Status refers to the publication records audited on October 1, 2026. CaMeL is IEEE SaTML 2026 conference research; Spotlighting remains a preprint. The added SecAlign row is ACM CCS 2025 conference research. Benchmark sizes are not samples of organizational incidents, and no attack probability is pooled. Supplementary Material S1 expands the threat and evaluation fields and identifies unextracted values.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

Table 2

Proposed assurance boundaries and complementary controls

Boundary	Proposed external enforcement	Evidence required before release	Residual issue and supporting sources
Identity and scope	Bind principal, tenant and operation to authenticated application state.	Negative tests for impersonation, altered identity and out-of-scope operations.	Authenticated users can still lack resource permissions (Rose et al., 2020; OWASP Foundation, n.d.-a).
Source authorization	Check current entitlements before protected chunks reach any model-visible stage, including reranking and query expansion.	Cross-user/tenant isolation; missing-metadata denial; preprocessing and revocation propagation tests.	Revocation must cover derived indexes and reusable answers (OWASP Foundation, n.d.-d).
Content provenance	Record source, owner and version; inspect approved ingestion transformations.	Traceable source versions; tests of tampering and suspicious extracted content.	Authentic origin does not prove truth or harmlessness (National Security Agency et al., 2025; Zou et al., 2025).
Model interaction	Mark lower-trust data; evaluate provenance signaling, structured training and preference optimization.	Benign-task, fixed-attack and adapted-attack evaluation with declared configuration.	Behavior depends on model and training (Hines et al., 2024; Chen, Piet, et al., 2025; Chen, Zharmagambetov, et al., 2025).
Executable actions	Validate schemas; externally authorize the actual resource, operation, destination and payload; bind human approval to that transaction.	Tests of unauthorized tool proposals and actual execution, recorded separately.	A correct-looking proposal is not authorization (Debenedetti et al., 2026; He et al., 2026).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

Boundary	Proposed external enforcement	Evidence required before release	Residual issue and supporting sources
Output release	Enforce recipient/classification policy for answers, exports, rendering and external fetching; apply local cumulative-disclosure limits.	Destination tests; protected-source coverage; sensitive-fact/token release; session/account budgets and cross-session aggregation.	Per-response checks may miss aggregate extraction (OWASP Foundation, 2025c; Yao et al., 2026).
Transversal provider and inference handling	Specify training use, inference retention, administrative access, processing location, subprocessors, deletion and contracts.	Dated configuration/contract record; responsibility map; retention/deletion checks and review after changes.	Application authorization does not settle provider processing risk (Autio et al., 2024; National Security Agency et al., 2025).
Transversal operations	Protect telemetry; propagate revocation to indexes, caches, memory and generated artifacts; test containment and recovery.	Measured revocation delay; minimized logs; connector suspension, credential rotation and recovery drill.	Logs can themselves carry injection or sensitive material (Karanjai et al., 2026; Nelson et al., 2025).

Note. This author-developed framework has not been experimentally validated as a complete architecture. It is not a validated score, certification, or formally NIST/OWASP-endorsed standard. The first six rows are sequential assurance boundaries; provider and inference-data handling and operations are transversal requirements. Supporting sources do not validate the complete matrix.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

Table 3

Prospective evaluation matrix for internal document assistants

Test family	Controlled comparison or policy state	Primary outcome	Interpretation and evidence base
Legitimate tasks	Matched questions and entitlements, with and without each defense.	Correct authorized completion; false-denial count/rate; excessive refusal and unavailable evidence separately.	Utility must remain visible (Debenedetti et al., 2024; Shafran et al., 2025).
Retrieval-stage attacks	Clean versus adversarial corpus under natural queries.	Retrieved attempts; context inclusions; behavioral manipulation; proposals; unauthorized execution and release, each with explicit denominators.	Retrieval and behavioral success are distinct stages (Chang et al., 2026).
Response manipulation	External expected answer and authorized source version.	Unsupported policy changes, fabricated claims or source misrepresentation.	Citation presence does not guarantee trustworthy evidence (Zou et al., 2025).
Indirect tool manipulation	Same user task and tools; modified lower-trust observations.	Unauthorized proposals and unauthorized execution separately.	Evaluate permissions and intent boundaries (Zhan et al., 2024; He et al., 2026).
Access isolation	Principals with deliberately different tenant and resource permissions.	Unauthorized chunks reaching model context or output.	A deployment gate; no trade-off against speed (OWASP Foundation, n.d.-a).
Permission revocation	Repeat questions after a timestamped authoritative revocation across all reusable representations.	Per-surface and maximum denial-propagation delay; failed propagation; disclosure during the interval.	Prespecify a local delay objective; test caches, indexes, history and reusable artifacts (OWASP Foundation, n.d.-d).
Outbound disclosure	Synthetic protected records and controlled destinations.	Release to an unauthorized answer recipient, export target or renderer channel.	Define destination and policy explicitly (OWASP Foundation, 2025a, 2025c).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

Test family	Controlled comparison or policy state	Primary outcome	Interpretation and evidence base
Cumulative extraction	Synthetic protected corpus; session/account query budgets; declared window and cross-session aggregation.	Protected-source coverage; unique sensitive facts/tokens; session/account disclosure and budget enforcement.	Local quotation/reconstruction policy; no universal acceptable threshold (Yao et al., 2026).
Derived-data access	Controlled permissions for vectors, traces and snapshots.	Unauthorized access to representations and reconstruction feasibility where applicable.	Embedding privacy is representation-specific (Morris et al., 2023).
Log interpretation	Clean, malicious and benign imperative event fields.	False negatives, false positives, exfiltration and output hijacking.	Do not trust event instructions as analyst authority (Karanjai et al., 2026).
Workflow effort	Matched tasks including verification, approval and correction.	Net completion and verification time; correction and approval effort; latency; legitimate utility.	Separate these measures from security violations (Guedes, 2026; Menezes, 2026).
Change and recovery	Model, embedding, retriever, connector or policy update; corpus snapshot and controlled incident drill.	Regression, containment time, revocation and successful authorized recovery.	Reevaluate affected assurance claims (National Cyber Security Centre & Cybersecurity and Infrastructure Security Agency, 2023; Nelson et al., 2025).
Provider data handling	Declared hosting, training-use, retention, access, location, subprocessor and deletion settings.	Documented responsibility and contract/configuration consistency; retention/deletion exceptions.	Transversal assurance; no-training does not prove authorization (Autio et al., 2024; National Security Agency et al., 2025).
Approval habituation	Repeated matched authorized and unauthorized transactions; visible resource, destination, operation and payload.	Unauthorized approvals, false rejections, payload inspection and review time over sessions.	Prospective longitudinal target; transaction changes require new approval (Autio et al., 2024; Debenedetti et al., 2026).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

Note. All tests are proposed, not performed. Use synthetic secrets and approved environments. Prespecify success definitions, numerators, denominators, session/account query budgets, observation windows, revocation-delay objectives, policy states, and adjudication rules. Record model, embedding, retriever, corpus, tool, permission-policy, and provider configuration. Report security, false denial, utility, verification effort, and correction effort separately.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

References

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

Brasil. (2018). *Lei nº 13.709, de 14 de agosto de 2018: Lei Geral de Proteção de Dados Pessoais*. Presidência da República. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709compilado.htm

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX Security Symposium* (pp. 2633–2650). USENIX Association. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

Chang, H., Bao, E., Luo, X., & Yu, T. (2026). Overcoming the retrieval barrier: Indirect prompt injection in the wild for LLM systems. In *35th USENIX Security Symposium* (pp. 1567–1586). USENIX Association. https://www.usenix.org/conference/usenixsecurity26/presentation/chang-hongyan

Chen, S., Piet, J., Sitawarin, C., & Wagner, D. (2025). StruQ: Defending against prompt injection with structured queries. In *34th USENIX Security Symposium* (pp. 2383–2400). USENIX Association. https://www.usenix.org/conference/usenixsecurity25/presentation/chen-sizhe

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

Chen, S., Zharmagambetov, A., Mahloujifar, S., Chaudhuri, K., Wagner, D., & Guo, C. (2025). SecAlign: Defending against prompt injection with preference optimization. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (pp. 2833–2847). Association for Computing Machinery. https://doi.org/10.1145/3719027.3744836

Debenedetti, E., Shumailov, I., Fan, T., Hayes, J., Carlini, N., Fabian, D., Kern, C., Shi, C., Terzis, A., & Tramèr, F. (2026). Defeating prompt injections by design. In IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). IEEE. https://arxiv.org/abs/2503.18813

Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. *Advances in Neural Information Processing Systems, 37*. https://doi.org/10.52202/079017-2636

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. Association for Computing Machinery. https://doi.org/10.1145/3605764.3623985

Guedes, M. A. P. (2026). Artificial intelligence tools for productivity enhancement: A narrative and thematic literature review. *Nexus Science Review, 1*(4), 1–43. https://doi.org/10.66104/4hndmk68

He, Y., Zhu, H., Li, Y., Shao, S., Yao, H., Liu, Z., & Qin, Z. (2026). AttriGuard: Defeating indirect prompt injection in LLM agents via causal attribution of tool invocations. In *35th USENIX Security Symposium* (pp. 1547–1566). USENIX Association. https://www.usenix.org/conference/usenixsecurity26/presentation/he-yu

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

Hines, K., Lopez, G., Hall, M., Zarfati, F., Zunger, Y., & Kiciman, E. (2024). *Defending against indirect prompt injection attacks with spotlighting* (Preprint). arXiv. https://doi.org/10.48550/arXiv.2403.14720

Karanjai, R., Lu, Y., Madhavarao, H. H., Xu, L., & Shi, W. (2026). Context contamination in LLM analysis of network security logs: Poison with passive prompt injection and mitigation evaluation. In *35th USENIX Security Symposium* (pp. 1507–1526). USENIX Association. https://www.usenix.org/conference/usenixsecurity26/presentation/karanjai

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474. https://proceedings.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Liu, Y., Jia, Y., Geng, R., Jia, J., & Gong, N. Z. (2024). Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium* (pp. 1831–1847). USENIX Association. https://www.usenix.org/conference/usenixsecurity24/presentation/liu-yupei

Menezes, R. R. (2026). Generative AI assistants in software development: Productivity gains, code quality, and security trade-offs. *Nexus Science Review, 1*(5), 1–21. https://doi.org/10.66104/z67gjz21

Morris, J. X., Kuleshov, V., Shmatikov, V., & Rush, A. M. (2023). Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12448–12460). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.765

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

Nasr, M., Rando, J., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C., Tramèr, F., & Lee, K. (2025). Scalable extraction of training data from aligned, production language models. In *International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2025/hash/cce0e917b050208170151f77b497fc71-Abstract-Conference.html

National Cyber Security Centre, & Cybersecurity and Infrastructure Security Agency. (2023). *Guidelines for secure AI system development*. https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/guidelines-for-secure-ai-system-development

National Security Agency, Cybersecurity and Infrastructure Security Agency, Federal Bureau of Investigation, Australian Signals Directorate's Australian Cyber Security Centre, New Zealand National Cyber Security Centre, & United Kingdom National Cyber Security Centre. (2025). *AI data security: Best practices for securing data used to train and operate AI systems*. Cybersecurity information sheet. https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/4192332/nsas-aisc-releases-joint-guidance-on-the-risks-and-best-practices-in-ai-data-se/

Nelson, A., Rekhi, S., Souppaya, M., & Scarfone, K. (2025). *Incident response recommendations and considerations for cybersecurity risk management: A CSF 2.0 community profile* (NIST SP 800-61r3). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-61r3

OWASP Foundation. (n.d.-a). Authorization cheat sheet. Retrieved October 1, 2026, from https://cheatsheetseries.owasp.org/cheatsheets/Authorization_Cheat_Sheet.html

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

OWASP Foundation. (n.d.-b). LLM prompt injection prevention cheat sheet. Retrieved October 1, 2026, from https://cheatsheetseries.owasp.org/cheatsheets/LLM_Prompt_Injection_Prevention_Cheat_Sheet.html

OWASP Foundation. (n.d.-c). Logging cheat sheet. Retrieved October 1, 2026, from https://cheatsheetseries.owasp.org/cheatsheets/Logging_Cheat_Sheet.html

OWASP Foundation. (n.d.-d). Retrieval-augmented generation (RAG) security cheat sheet. Retrieved October 1, 2026, from https://cheatsheetseries.owasp.org/cheatsheets/RAG_Security_Cheat_Sheet.html

OWASP Foundation. (n.d.-e). Secrets management cheat sheet. Retrieved October 1, 2026, from https://cheatsheetseries.owasp.org/cheatsheets/Secrets_Management_Cheat_Sheet.html

OWASP Foundation. (2025a). *Excessive agency (LLM06:2025)*. https://genai.owasp.org/llmrisk/llm062025-excessive-agency/

OWASP Foundation. (2025b). *Prompt injection (LLM01:2025)*. https://genai.owasp.org/llmrisk/llm012025-prompt-injection/

OWASP Foundation. (2025c). *Sensitive information disclosure (LLM02:2025)*. https://genai.owasp.org/llmrisk/llm022025-sensitive-information-disclosure/

Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero trust architecture* (NIST SP 800-207). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-207

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 01/10/2026
Vol: 01.05 - October
DOI: 10.66104/3m1m5x77
Pages: 1-32

Shafran, A., Schuster, R., & Shmatikov, V. (2025). Machine against the RAG: Jamming retrieval-augmented generation with blocker documents. In *34th USENIX Security Symposium* (pp. 3787–3806). USENIX Association. https://www.usenix.org/conference/usenixsecurity25/presentation/shafran

Souppaya, M., Scarfone, K., & Dodson, D. (2022). *Secure software development framework (SSDF) version 1.1: Recommendations for mitigating the risk of software vulnerabilities* (NIST SP 800-218). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-218

Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., & Hamin, M. (2025). *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations* (NIST AI 100-2e2025). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-2e2025

Yao, M., Zhang, Z., Luo, N., Li, S., Cai, Y., Chen, X., Guo, Y., & Li, D. (2026). Connect the dots: Knowledge graph–guided crawler attack on retrieval-augmented generation systems. In *35th USENIX Security Symposium* (pp. 6465–6484). USENIX Association. https://www.usenix.org/conference/usenixsecurity26/presentation/yao-dots

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

Zou, W., Geng, R., Wang, B., & Jia, J. (2025). PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium* (pp. 3827–3844). USENIX Association. https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag

CC BY