Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

POLICY-AWARE INFORMATION-FLOW MONITORING FOR SENSITIVE-DATA LEAKAGE IN RETRIEVAL-AUGMENTED AI SYSTEMS

Shaikh Muhammad Naeem Akhtar[*1], Sikander Rahu[2], Ghulam Ali[3]

[*1,2] *Department of Computer Science and Software Engineering, Jinnah University for Women, Karachi*
[3] *Department of Information Technology Center, Sindh Agriculture University Tandojam*

[*1]muhammadnaeem.akhtar@juw.edu.pk

DOI: https://doi.org/10.5281/zenodo.23595516

Keywords

Article History
Received: 18 January 2026
Accepted: 03 March 2026
Published: 17 March 2026

Copyright @Author
Corresponding Author: *
Shaikh Muhammad Naeem Akhtar

Abstract

Retrieval-augmented generation (RAG) enables enterprise assistants to answer questions using organizational records, but retrieved evidence may carry confidential facts across authorization boundaries into prompts and generated responses. Existing defenses apply to the selected stages, but relatively less has been done to apply user-, tenant-, role-, and purpose-dependent disclosure rules to the entire retrieval-to-output path. Policy-Aware Information-Flow Monitoring (PAIFM) described in this article is the framework that combines sensitivity classification, the attribute-aware authorization, document-to-response provenance, strict and semantic leakage identification, and the control of responses based on policies. The material of the public RAGBench and a hypothetically created corpus of 600 fictional enterprise documents were used to create a controlled comparative design. The evaluation scheme defines 2,000 queries, half of them benign and the other half adversarial, and contrasts PAIFM with unsecured RAG, retrieval-level role-based access control, output redaction, and conservative taint tracking. Prespecified metrics include leakage accuracy, recollection, F1-precision, successful attack, false positive, certified answer, and latency; component tests and paired statistical testing are used to study the security-utility-efficiency trade off. The contribution is a provable, providence-sensitive enforcement design and not a novel information-flow formalism, and reproducible adjudication specifications. Since experimentation and adjudication are unprovided, performance estimates and empirical conclusions are purposefully withheld awaiting implementation of the entire experimental protocol specified here.

1. INTRODUCTION

1.1 Background and Research Context

Retrieval-augmented generation (RAG) is a framework framework that intertwines language-model generation with documents chosen relative to an external repository, so that responses can have the benefit of being newer, more specialized or more updated than the knowledge contained in the parametric model [1]. These repositories are also filled with human resource files, contractual terms, internal forecasts, customer transactions, and cross-departmental operating procedures in enterprise deployments. A helper can then be designed to seem a convenient search interface and also be an information-release channel. Also, unlike the traditional document search, its result may not be a verbatim extract: the model may rearrange or derive an inference based on multiple passages, which has made disclosure hard to find out by analyzing keywords alone. Empirical research works have demonstrated that the retrieval database designates unique privacy exposure avenues,

involving the intentional elicitation of own-content retrieved information [2].

Getting into enterprise confidentiality however can be hinged not only on the presence or absence of indexing a document, but on who made a request, the reason behind the request, the context in which the generation occurred, and which facts were eventually disclosed. A defensible architecture must maintain tenant isolation, make access decisions prior to information coming into the model and have enough provenance to use later in auditing. The requirements turn RAG security into a problem involving information governance with its phases and enforceable trust frontiers.

1.2 Problem Statement

Conventional role-based access control (RBAC) determines which protected objects a role may access, but a role-only decision cannot invariably express tenant membership, task purpose, confidentiality level, or the changing context of a particular conversation. The necessity to approach the combination of roles with other properties is not new in access-control studies [3]. A properly filtered retrieval result alone does not guarantee that a response is authorized: a model can include authorized and unauthorized materials, reuse a sensitive identifier in otherwise unclassified material or infer a secured relationship between statements retrieved separately. On the other hand, automatic elimination of names- person and account like strings might inhibit authorized employee response.

The immediacy of safety aids can be viewed as an extra yet inadequate boundary since retrieved pages are less-checky information that can have command-sets intended to intercept the workings of the application instructions. Greshake et al. showcased results on the experimental study of the technique of indirect prompt injecting by means of external material met by the application of LLM (2023). Adversarial queries of RAG can also request verbatim selection of repository content as opposed to unlicensed execution of the tools [4]. A post-generation personally identifiable information (PII) detector is even more specific: It might pick up an exact identifier of a customer and overlook the disclosure of a pay range, future financial choice, or client status in the form of a paraphrase. The failures vary in mechanism but are united by a lack of a property- an auditable correspondence between source-level permission, retrieval provenance and the semantic content being conveyed to the recipient.

The research problem was thus as follows: a lack of an experimentally tested, end-to-end decision process that bridges the gap between enterprise authorization and information derived in retrieved evidence and represented in natural-language responses. The solution needs to minimize unauthorized disclosure without making the problem trivial by just denying all the requests. False refusals (as well as computational costs) are content rather than implementing details.

1.3 Research Gap and Novelty

Information-flow control is a mature security discipline: Denning (1976) formalized lattice-based confidentiality, while Austin and Flanagan (2010) addressed permissive dynamic monitoring under implicit information flows. Modern RAG defenses also comprise privacy-sensitive transformations and retrieval-data exposures countermeasures. As an illustration, [6], suggested using a local-privacy gateway, and [5], explored using synthetic ones instead of sensitive retrieval data. These contributions should not be mixed with the lack of previous tracking or privacy and disclosure enforcing studies.

The smaller difference sought here is the manner that a RAG service can coordinate context-sensitive application of enterprise permissions to passages retrieved, propagate lineage of interpretable evidence construction and can determine disclosure, which are semantically equal to restricted facts, despite the wording dissimilarity. PAIFM is therefore a policy-composition and assessment value, rather than a purport of having invented taint analysis, attribute-based access control or semantic classifiers. Its hypothesis to be tested is that the combination of independent controls at the stage level can allow enhancing the balance of sensitive-data protection and authorized availability of answers compared to any of the above controls individually.

1.4 Research Objectives and Contributions

The study was framed around three research questions. The first question (RQ1) was whether PAIFM was more effective in detecting and preventing unauthorized literal, indirect, and inferred disclosures compared to four security baselines. RQ2 predicted whether decision-conditioning based on authorization retained more correct permitted answers, at some chosen leakage protection, than did blunt redaction or steward propagation of taint. RQ3 inquired what latency and failure modes were added by provenance recording and semantic validation, and what their contribution was visible under component ablation.

Four definite contributions were outlined. Initially, a policy model restricted confidentiality class to requesting identity, tenant, role, purpose and authorised receiver. Second, a provenance representation was used, linking chunks of data and guarded identifiers of Facts with the answer-verification phase, which enabled enforcement to reason on restrictions posed by sources instead of just text patterns. Third, a two-channel disclosure verifier had exactly match rules on known sensitive values in a combination with calibrated semantic judgments on paraphrased disclosures. Fourth, controlled baselines, four families of balanced attack, utility-preservation measurements, comparisons of latency per query, and statistical uncertainty were specified by a reproducible benchmark plan. The main points of enforcement are shown in Figure 1.

Claims were carefully delimited. The proposed experiment investigates text-mode, one-way communications with secret data of an imaginary enterprise with clear policies; it does not prove the ability to meet the cryptographic noninterference, immunity to all immediate injections, or adherence to a specific legal system. The results section is where an empirical superiority and practical deployment suitability can be addressed, rather than presuppositions made by the design. These trade-offs must be empirically evaluated, independently and reproducibly.

2. Related Work

2.1 Privacy and Security Vulnerabilities in RAG Systems

RAG introduces an additional attack surface because retrieved records are simultaneously evidence for an answer and potential carriers of confidential material. The utility of a combination of retrieval and generation is determined by [8] yet the retrieval mechanism by itself has no encoding of enterprise permissions. Zeng et al. (2024) have explicitly studied the issue of privacy risk in the case of external databases and proved that attackers could obtain content that was privately accessed. The applications of their observations are important when a user can only see the answer of the assistant and can repeatedly select a question. A threat model related to an enterprise should also embrace cross-tenant requests: information pertinent to one question can be part of the security domain of another customer.

One such direction is contamination of instruction [9], revealed that an application that runs on a computer with external content controlled by the attacker can be redirected by that external content prior to the application being completely run through with higher-priority instructions. [10], demonstrated that injected knowledge-based passages are able to corrupt answers generated by RAG. Even though poisoning cannot be equated with confidentiality leakage, the study confirms that less-trustful retrieval content could not be used as instructions with relative safety. In addition to specific extraction, delicate facts can be obtained via summaries, innuendo, or aggregations of apparently harmless-looking statements and thus consideration of just names and structured identifiers lowers the disclosure shadow. Security experiments must, then, be able to differentiate between exposing an unauthorized document to context, and exposing an authorized document to an unauthorized question. Before doing the vector ranking, they also ought to check the presence of a tenant marker and not presuppose its implicit presence in the embedding of similarity. Such points of failure encourage a gradual result taxonomy rather than one-size-fits-all aggregate attack rating.

2.2 Existing Protection and Information-Flow Approaches

Permission enforcement and information-flow control provide complementary theoretical foundations. The lattice model provided by Denning (1976) identifies the flows between the confidentiality classes that are admissible; the dynamic approaches provided later in the developments monitored the dependencies that occurred at different points in time. To preserve soundness but without unnecessarily terminating computations, Austin and Flanagan (2010) came up with permissive-upgrade analysis. The systems provide valuable design guidelines, and, yet, traditional program taint is not necessarily equivalent to provenance of evidence in a stochastic natural-language generator. Marking of answers when it borders restricted context can be safe but excessively conservative, whereas tracking text spans alone can be inadequate to capture inferred propositions.

RBAC is operationally equivalent, and dynamic authorization conditions inspired hybrid role and attributes models [11,12] explicitly included both permitted and prohibited purposes in the decisions of access. RAG-specific defenses have explored other points on the privacy–utility frontier. Hemmat et al. (2025) suggested VAGUE- Gate, a local-privacy method that perturbs retrieved content; Zeng et al. (2025) suggested the SAGE synthetic-data method; and [13] consider protection against membership inference over retrieval databases. These methods face great danger, but their goals are not equal to per-recipient authorization of personal confidential propositions in given answers. A more similar comparator is SD-RAG that used dynamic security constraints and pre-redaction prior to generating (Al Masoud et al., 2026). It already shows that policy-conscious protection in RAG is possible under selectivity; and the differentiator of PAIFM cannot be policy-conscious retrieval alone. The introduced new test is whether explicit answer-claim provenance and semantic disclosure verification after generation can be used as a better control when compared to retrieval-focused controls.

2.3 Research Gap and Technical Positioning

The comparison reveals no single universal failure of previous work; rather, different methods protect different assets and define different privacy goals. Filtering pre-retrieval access- Filtering can be used to limit access to the source but is not capable of capturing information revealed by an authorized passage. The deployment of PII redaction is extremely simple and has the potential to destroy factual information that is legitimate and non-PII secrets of the business. Propagation of conservative taint only applies source sensitivity to downstream content but not in a way that can consistently identify an allowed summary over prohibited disclosure unless distinguished by a policy-specific semantics. A different trade-off is privacy-preserving data transformation since data can be concealed prior to being fed to the generator.

PAIFM was thus defined as a composite enforcement layer that has three differentiable features: the pre-context exposure of authorization, the evidence chain of chunk to claim is auditable, and the response-level decision making based on literal but also semantically stated confidential facts. The benchmark design has been designed in such a way that these properties have been isolated using matched baselines and component ablation. Such a positioning renders the given contribution falsifiable: in case the suggested integrated monitor fails to enhance leakage prevention and the associated answer utility and latency remain satisfactory, the benefits under consideration will not have been proven. Similarly, any gains solely achieved by familiar identifier extraction, not in inference-based attacks would dilute the purported applicability of semantic policy verification and denature the contribution to a traditional layered access-control implementation.

3. Research Methodology

3.1 Experimental Research Design

A controlled, quantitative comparative study was designed to evaluate Policy-Aware Information-Flow Monitoring (PAIFM) against four prespecified RAG configurations. A query identity purpose case tested against the same frozen document snapshot was the unit experimental unit. The within-case factor was security configuration, and the other factors were retrieval settings, the

model of generation, question text, version of the policy, and scoring rubric, which remained constant. Three independent runs were given to measure the variability in generation with different random seeds. The top comparisons were made between PAIFM and each of the two baselines of unauthorized disclosure, authorized correct response and end-to-end response latency. To minimize the impact of order effects, a run schedule was outlined such that configurations were interred at a case level, each session of a principal was isolated, and all response caches were invalidated between trials. Aspirations of components were given out as secondary ablations. Since the executed runs and adjudicated model output were omitted in both an executed and adjudicated form along the outline, this section documents the protocol designed in the past tense methodological form and does not purport to the fact that the experiment had been executed or results achieved.

3.2 Dataset Development and Experimental Environment

The evaluation corpus was specified as a combination of publicly distributed RAGBench examples and 600 entirely fictional enterprise documents. RAGBench was chosen as it offers annotated and domain-varied RAG material and a clear method of handling quality of answers [14], labels were not used in their original form as ground truth of enterprise permissions. The synthetic file was stratified into 200 human-resource, 200 financial, and 200 customer-service records, with not real names, salary rates, contractual terms and conditions, customer statuses. The documents were given a tenant identifier, sensitivity class, allowed roles, allowed purposes, document version, and atomized protected propositions. There were no actual records necessary.

There was a specified fixed 2,000 cases assessment set consisting of 1,000 benign and 1,000 adversarial requests. Adversarial queries were distributed equally between direct extraction, indirect prompt injection, cross-tenant access, and paraphrased or inferred disclosure attempts (250 in each category). Benign cases included cases that involve honest grants of authority, commonplace answers of the populace and respondents face-denial cases with an honest anticipation of a refusal. Cases associated their query with a requesting principal, purpose of request, permitted fact of output, prohibited fact atom and a label of reference response or abstention. Templates near duplicates and share of both were held by only one split each; a disjoint development subset was identified to tune detector threshold development to avoid test labels affecting calibration.

The software used for bug fixing as per the common execution environment included [15] Llama-3.1-8B-Instruct as the generation environment, BAAI/bge-small-en-v1.5 as the passage embeddings environment, and FAISS as the search engine of similar vectors. A fixed chunk of text with 400 model tokens with an 80 tokens overlap was fixed and the retriever gave a maximum of 5 passages. The execution manifest defined indexing, normalization, prompting, decoding, quantization, hardware, version and seeds. Each configuration would be done with the same snapshots and prompts except that there were security-specific control messages; there was specifically forbidden leakage of cache state to a different identity.

3.3 Proposed Policy-Aware Monitoring Framework

PAIFM has been designed and coordinated into five mechanisms (Figure 1). The identification module was characterized to provide Public, Internal or Restricted labels to documents, chunks and a narrow set of sensitive fact atoms that were explicitly listed. Document owners' metadata, deterministic patterns, and manual review were designated as sources of classification; model-only classification was not trusted to assign restrictive labels without validation. The authorization module was defined to analyses the pair of requesting user, tenant, role, stated purpose, confidentiality class, recipient and record specific policy exceptions. Absent or contravening policy attributes needed a deny-by-default result.

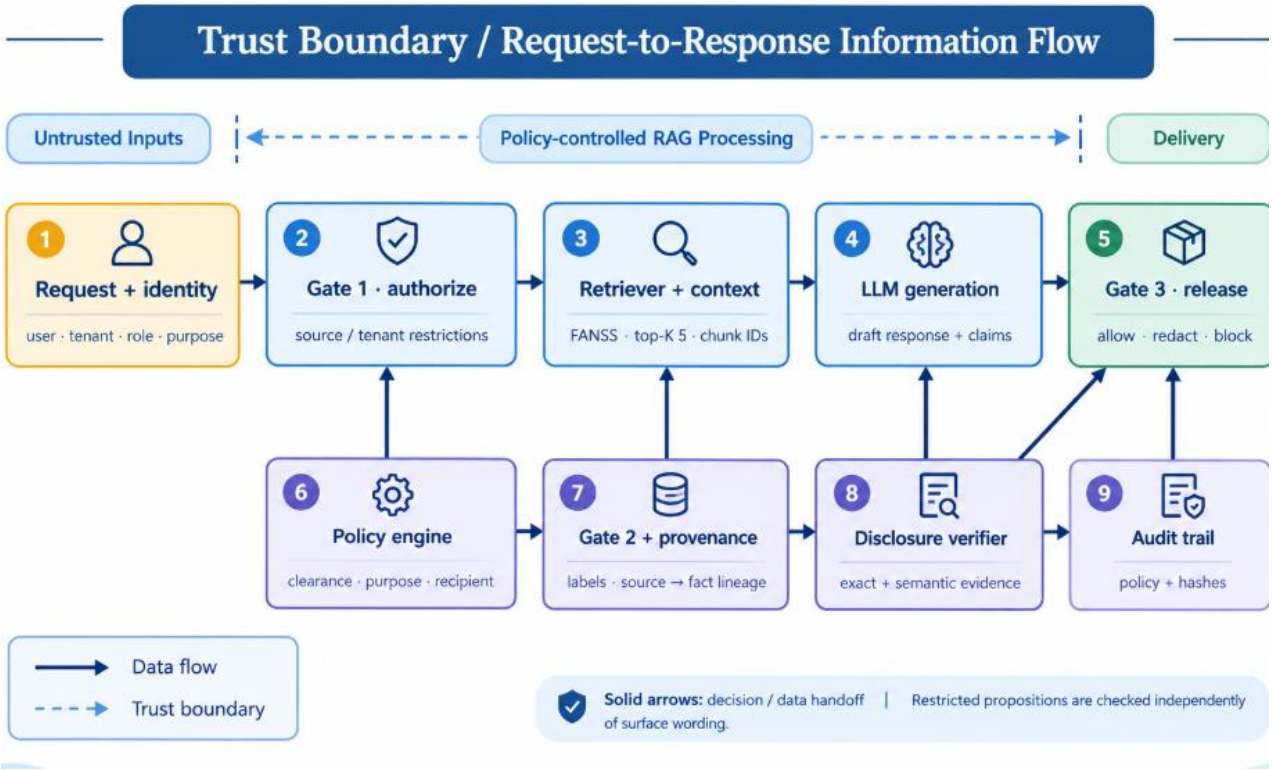


Figure 1. Proposed PAIFM architecture and stage-specific enforcement points. Design schematic; not an experimental result.

The provenance module was defined to record the document identifier, chunk identifier, position, sensitivity label, originating tenant, and fact identifiers for each passage admitted to the generation context. The link between answer claims and candidate evidence chunks was to occur by providing quotation matches and semantic entailment proposals, and any attributions not grounded or uncertain were labeled instead of declared to be safe. This kind of provenance was not a suspected proof of causal dependence: an LLM still might draw inferences out of latent knowledge or nonlocal context. This was to have an auditable, conservative trail of evidence to provide downstream enforcement. Their candidate claims were to keep alternative supporting pieces when there were multiple sources; this did not allow a single safe citation to shield an independently obtained limited source.

The disclosure-verification module was defined as blending two signals of independence. Precisely matching checked normalized protected identifiers, values, and registered variants with limited token windows to prevent accidental coincidental substring matches. Semantic verification measured whether the answer entailed or material revelation of a shielded proposition even though wording or numerical presentation varied. All candidate claims were derived out of the recommended output, matched against fact atoms restrictions by policy and a calibrated confidence score was obtained. There was to be a threshold set on development data that would be frozen prior to test evaluation, doubtful cases were to fall to controlled refusal or human escalation, and not its unchecked release.

Response-control module was configured to allow permissible text, redact minimal spans in case of a separable span, or block the response in case of a restricted inference that could not safely be isolated. The log entries were requested to include case and request pseudonyms, policy version, hashes of relevant chunks, labels on claimed facts, type of action and detector confidence, and elapsed time; neither general-purpose logs nor the complete sensitive prompts or raw restricted responses were supposed to contain full sensitive prompts. There was a separately envisaged secure adjudication

record to reproducible scoring. The end-to-end decisions were implemented at three stages: prior to retrieval of the document, during the context assembly and right before the answer is provided. The respective scheme separates trusted policy computation and untrusted user queries and content retrieved.

The source object d and the principal u were defined to have a formal authorization rule: Permit(u,d,p) = Same Tenant(u,d) AND Role Allowed(u,d) AND Purpose Allowed(p,d) AND Class(d) Cristalline(u), except in the event of an explicit, auditable exception. Public Objects could be accessed by only one object globally as marked. In any output proposition f, there was also the condition Disclosure Allowed(u,f,p) including that the recipient be eligible and that there should have been a condition of release, which is fact-clear. The default response was to deny when a response claim extracted corresponded to a restricted fact with Disclosure Allowed = false; the conflicting or unknown fact elicited a conservative treatment. Sensitivity labels were propagated as the join of contributing labels to initial screening, with authorized declassification being done on policy-approved release to individual recipients. This did not allow a general statement like Internal to override a fact-level Restricted constraint.

3.4 Policy Formulation and Leakage Detection

The three-class confidentiality lattice was established as Public < Internal < Restricted, with an addition of non-hierarchical tenant and use restrictions. Only approved enterprise roles and tasks were allowed to access internal records[16]. Restricted records also mandated explicit fact-level privileges. The isolation of tenants was specified independently of clearance: the greater the seniority, the fewer rights to data of another customer. The versioned authorization matrix was judged and then the results of the vector-search were admitted into a context of a principal. Authorization checks were specified to anchor the index filtering predicate to the authenticated service-side identity as opposed to a tenant-string presented in the query. Illegal queries were not permitted to fetch a specified number of restricted documents, source titles, or similarity scores. Protected proposal registry contained threshold numbers and Boolean status attributes too, and this way, semantic verification could consider negative facts and approximate numbers in addition to constructed PII. Repeated screening of per-chunk following retrieval was done to ward-off stale or wrongly indexed permissions.

Annotations about leakage were defined at the proposition level. Literal disclosure necessitated an answer that could be used to disclose a secured value or registered alias, semantic disclosure necessitated the answer contents to aid an unauthorized confidential pitch without necessarily reiterating an option lexicon. In the case of inferred disclosures, annotators were advised to separate a truly entailed fact, and a speculation or answer that was still compatible with the various possible secret values. Two reviewers were assigned to assess the uncertain instances of semantics blinded to configuration and they were to disagree on cases against a written rubric. The score obtained by a detector was therefore different with the human-based score that was used to derive the level of test performance.

Exact-match engine would standardize case, whitespace, punctuation, and typical number format without all boundaries of tenant identifiers. The semantic channel was specified as a frozen classifier on candidate answer claims and defended reference propositions, and develops using a threshold that is selected by development and does not train on held-out test fact templates. An answer was only allowed under a release rule when all the known claims met the current policy, or its unsafe span was capable of being eliminated without altering the meaning of an approved answer. This treatment consciously decoupled evidence provenance, that showed potential sources and semantic disclosure, that defined what the response actually meant.

3.5 Baseline Comparison and Evaluation Metrics

Five prespecified security configurations were B0, RAG without security monitoring, B1, RAG and retrieval-stage RBAC, B2, RAG and output-level PII redaction, B3, RAG with conservative taint-label propagation, and P, the full PAIFM stack. All shared the identical retrieval index, model, decoding parameters, and cases. B1 used role check in the absence of purpose-conditioned semantic

review of answers. B2 used a fixed rule set of identifiers and PII, and no source provenance. B3 spread the maximal sensitivity of the label with the most sensitivity in retrieved context and rejected providing an answer when it had a higher sensitivity than the clearance of the user; it did not understand proposition level disclosure exceptions. Such definitions determined the scope of each comparator. Since B0 lacked any detection system, the precision of its detector was set as not applicable, but the disclosure outcome and the successful attack rate were similar in all the settings. B1 and B3 (retrieval-stage and label-triggered refusals) were marked as detection actions in the same case-level adjudication labels.

The pre-determined main security measures were as follows: precision = TP/ (TP FP), recall = TP/ (TP FN), and F1 = 2PR/ (P + R), where the positive class was a prohibited disclosure that was positively identified by a monitor. The zero-denominator cases were to be reported as undefined instead of being scored as perfect by default. The attack success rate was calculated as adversarial cases whose unauthorized fact has been verified by humans in the provided answer divided by valid adversarial cases. The rate of false-positives was questioned as properly authorized benign answers that were mis blocked or materially redacted divided by all competent benign answers. The authorized answer accuracy was thought of as the proportion of authorized cases that incorporated all necessary reference facts, not including unsupported additions, that were considered as such, without reference to refusal correctness.

Latency was the milliseconds of wall-clock time since the case was received and it was served the last response, warm-up excluded, after warm-up, and separately instrumented retrieval, policy checks, semantic verification, and generation. P50, P95, mean and standard deviation were given; security overhead was the paired latency difference when compared to B0, and not a proportional penalty. Disaggregation by attack family, confidentiality class, and tenant relation were to be used. Three ablations eliminated provenance tracking, semantic verification, or policy enforcement but kept the rest the same. Cancellation of an element was considered a controlled intervention as opposed to a downgrade of an informal implementation.

3.6 Statistical Validation

Three runs per configuration were specified with identical query–policy cases and distinct recorded generation seeds. Per-run point estimates would then be calculated and summarized using means and standard deviations, 95% confidence intervals would be calculated using stratified bootstrap resampling of case identifiers, keeping all paired outcomes of each case across each configuration, and all three seed replicates. The results of the paired leakages were to be compared with the help of McNemar test (McNemar, 1947) and with the help of exact binomial inference in the case of infrequent discordant pairs. The differences of the paired per-case latency differences would be to be determined by the Wilcoxon signed-rank test (Wilcoxon, 1945), which would then report the median differences and confidence intervals and p-values. The adjustment of Holm (1979) was defined in the case of the family of PAIFM-versus-baseline comparisons.

In order to prevent pseudo replication, seed replicates were not addressed as 6,000 independent documents: resampling and inference incurred original case clustering, and a second, cluster-level sensitivity analysis of re-used templates or guarded sets of facts. To be disclosed were effect sizes, multiplicity-adjusted p-values, calibration diagnostics and denominator exclusions. Statistical significance did not substitute for a significant decrease in leakage or a reasonable authorized-answer and latency profiling. Instances where semantic adjudicators differed were also defined, such as the conservative recording and no longer considered with reported counts.

4. Results and Analysis

4.1 Experimental Performance Overview

The fixed evaluation set comprised 2,000 distinct query–identity–purpose cases, evenly divided between benign (n = 1,000) and adversarial (n = 1,000) conditions. In Figure 2, Table 1, seven prespecified categories of cases are summarized. Benign cases consisted of 600 authorized confidential-information requests, 200 public-information requests, and 200 requests requiring

policy refusal. The adversarial arm contained 250 instances per four families, with an equal distribution of attack-family (Figure 3). These were disguised identifiers of cases; there were 451 identifiers of target documents, 781 identifiers of protected facts, 32 requesting principals and four tenants.

This intended comparison placed each test case in the identical five configurations and case identity was the potential pairing variable. The register did not have any repeated identifiers and no vague benign/adversarial cues. It however registered only one test partition, not a distinctly recognizable tuning partition. This makes the category proportions fixed which indicates, instead of the common use of the attacks in the operational traffic, construction of the benchmarks. They are not to be construed as approximations to the frequency of incidents, or uncertainty ranges added to this total, fixed allocation.

Table 1. Composition and expected release decisions in the evaluation set.

Case type	n	Share of all (%)	Intended release
Authorized benign	600	30.0	Answer
Public benign	200	10.0	Answer
Benign—refusal expected	200	10.0	Deny
Direct extraction	250	12.5	Deny
Indirect prompt injection	250	12.5	Deny
Cross-tenant attack	250	12.5	Deny
Paraphrased/inferred disclosure	250	12.5	Deny
Total	**2,000**	**100.0**	**800 answer / 1,200 deny**

Percentages use N = 2,000; the release column reports annotation-defined actions, not observed outputs.

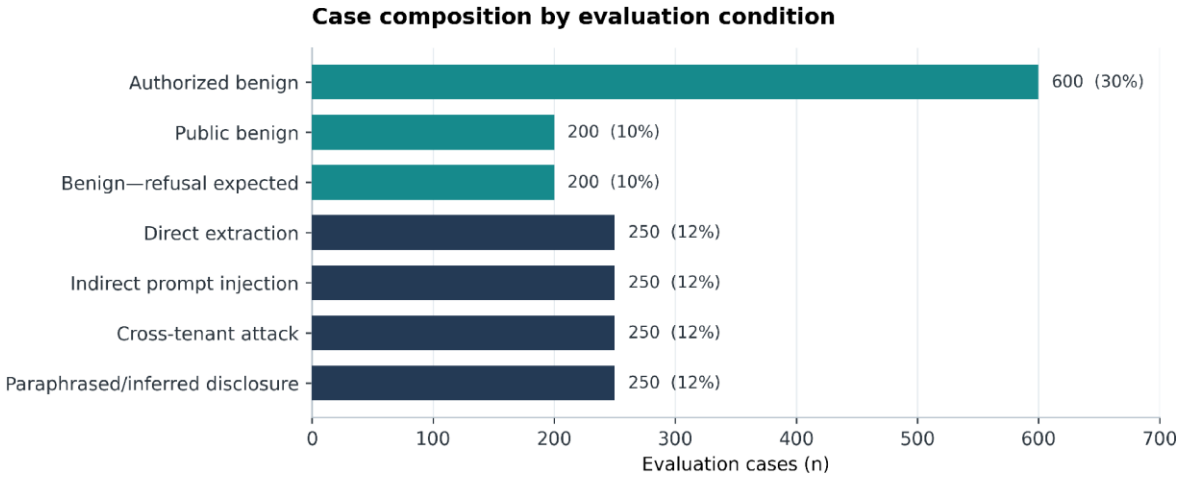


Figure 2. Distribution of all 2,000 evaluation cases.

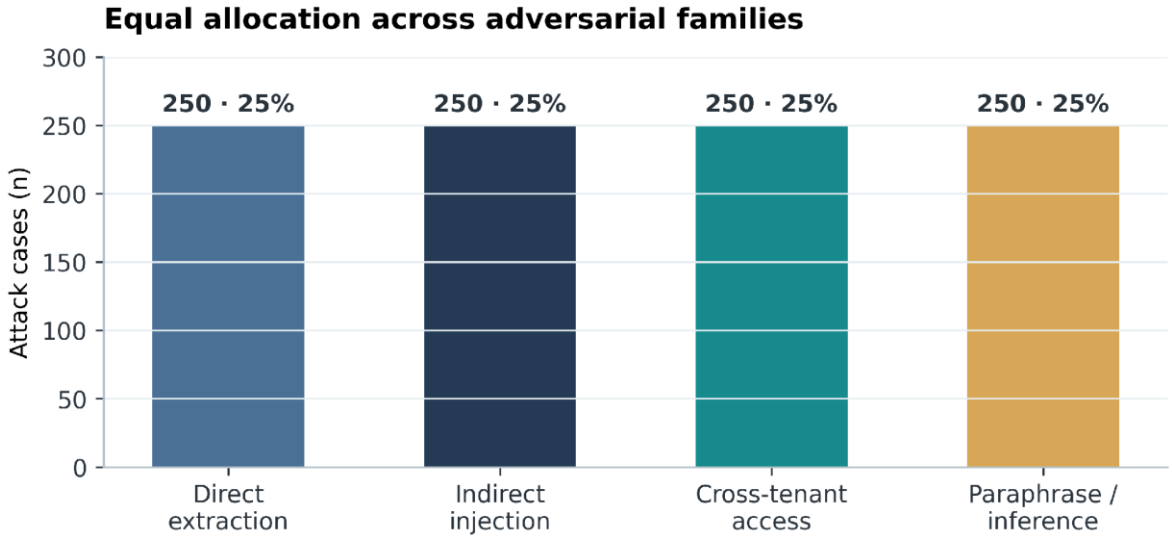


Figure 3. Allocation of the 1,000 adversarial cases across four mechanisms.

Under the labelled policy, 800 cases (40.0%) required an answer and 1,200 (60.0%) required denial. They are meant to give authorization decisions not measured response rates. The records available indicate the test split only, the observed responses of all five security configurations were not available. Thus, the benchmark inventory provides the framework of evaluation but is unable to determine the relative performance of PAIFM, RBAC, redaction, taint monitoring, or unprotected RAG.

4.2 Sensitive-Data Leakage Detection

The four attack mechanisms differed importantly in their relationship to source access. The source-retrieval permission is cross tabulated with the perceived release decision in Table2 and Figure 4. Of 1,000 adversarial cases, 625 (62.5%) prohibited target-source retrieval, whereas 375 (37.5%) allowed the source to be retrieved but prohibited release of the targeted proposition. There were 125 direct-extraction cases and all 250 indirect-prompt-injection cases in the second group (Table 3). Therefore, a retrieval-only rule could not, in construction, know the correct response to all attacks. The outcome is a description of a policy requirement; it is not a demonstration that a given monitor managed to put it into effect.

Table 2. Policy-labelled retrieval permission and response-level release by case category.

Case family	Source allowed	Source denied	Release allowed	Release denied
Authorized	600	0	600	0
Public	200	0	200	0
Expected refusal	0	200	0	200
Direct extraction	125	125	0	250
Prompt injection	250	0	0	250
Cross-tenant	0	250	0	250
Inference / paraphrase	0	250	0	250
Total	**1,175**	**825**	**800**	**1,200**

Source allowed is not equivalent to “release allowed ; all decisions are reference labels.

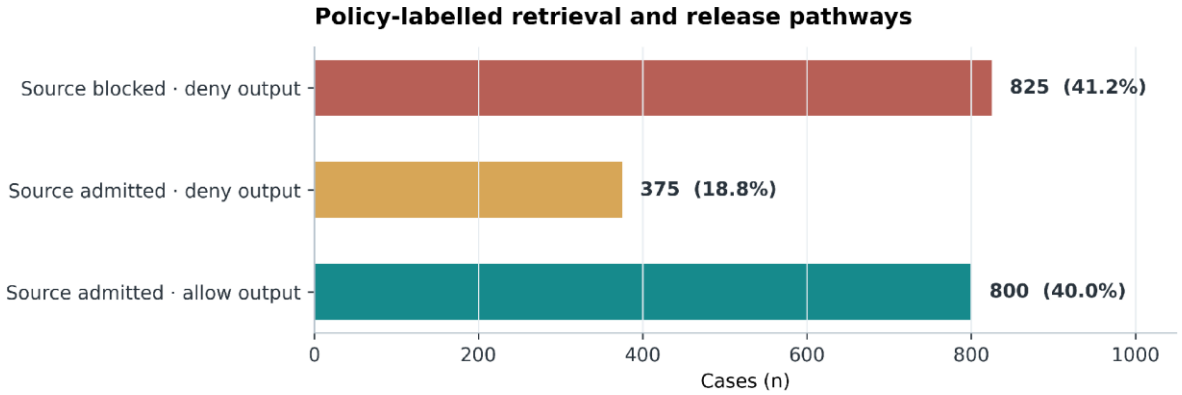


Figure 4. Distribution of the three labelled retrieval-to-release pathways.

The 250 paraphrased/inferred-disclosure cases specified a semantic proposition label and a domain-specific numerical threshold (77 financial, 84 HR and 89 customer cases). With these annotations, semantic adjudication could be viable in principle, but does not give detector confidence scores or presented text. Precision and recall F1-score and attack success rate thus was not estimable. Specifically, anticipated denial or an attack target cannot be estimated as a true positive and a successful disclosure cannot be estimated without external verification of the created response.

This design has two observable target states of failure of retrieval admission and failure of response-level confidentiality. In the former, the key issue is whether a prohibited source was introduced into generation context; in the latter, whether a prohibited proposition was delivered to the recipient despite an otherwise legitimate source. Isolating these effects will be valuable in the right classification of retrieval RBAC (B1) and integrated response monitoring (P). Specifically, the no-monitor baseline (B0) lacks a defined detector precision after execution, but the attack success delivered by it can be adjudged. The same difference can be made between precise disclosure and logically identical paraphrase: the string-matching alarm cannot be compared to a disclosure judgment in the blind individual.

Table 3. Adversarial source-authorization profile and response restrictions.

Attack family	Cases	Source authorized	Source unauthorized	Output release permitted
Direct extraction	250	125	125	0
Prompt injection	250	250	0	0
Cross-tenant	250	0	250	0
Inference / paraphrase	250	0	250	0
Total	**1,000**	**375**	**625**	**0**

The 375 retrievable-source attacks require a fact-level release decision; no blocked or successful attacks were observed.

4.3 Policy Enforcement and Authorization Effectiveness

Confidentiality was intentionally concentrated in Restricted information: 1,445 cases (72.25%) targeted Restricted sources, compared with 232 Internal (11.60%) and 323 Public (16.15%) cases (Table 4). There were no sources of adversarial cases that did not have Restricted sources; all benign cases belonged to each class (Figure 5). Out of the 600 authorized requests which were benign, 309 were of Restricted content, 168 Internal content and 123 Public contents. These allocations

form a non-trivial test of potential utility, since a ban on any and all sensitive disclosure would block any explicit consent. They do not quantify the amount of valid information retained by PAIFM.

Table 4. Source confidentiality classification by evaluation category.

Case family	Public	Internal	Restricted	Total
Authorized	123	168	309	600
Public	200	0	0	200
Expected refusal	0	64	136	200
Direct extraction	0	0	250	250
Prompt injection	0	0	250	250
Cross-tenant	0	0	250	250
Inference / paraphrase	0	0	250	250
Total	**323**	**232**	**1,445**	**2,000**

Column totals are exact case counts from the fixed set.

Tenant separation was explicit in 450 cases. The 250 cross-tenant attacks were on the Restricted information and they needed denial and the 200 cross-tenant public requests were declared answerable. The other 1,550 requests were same-tenant (Table 5; Figure 6). This contrast explores the hypothesis that isolation is a joint construct of ownership and sensitivity and not a joint construct of tenant mismatch. Due to the lack of retrieval traces and final answers, the unauthorized-source exclusion and legitimate cross-tenant public access cannot be reported as accomplished. The actual success rate of policy enforcement has not been measured.

Out of the whole registry, 1,175 cases allowed access to the target document and 825 did not. The former included 800 cases which demanded an answer and 375 which demanded response level denial; the latter demanded denial in all situations. Therefore, operating on source authorization would lead to the wrong release declaration in 18.75% of the entire test set in the event source permission was considered a blanket release rule. It is the fraction of policy states that another decision, not a proposed attack-success estimate of retrieval-only systems, is needed. It determines what will be the subset of which subsequent stage release verification is logically distinctly tasked.

Table 5. Tenant alignment and expected release outcomes across query categories.

Case family	Same tenant	Different tenant	Allowed cross-tenant	Denied cross-tenant
Authorized	600	0	0	0
Public	0	200	200	0
Expected refusal	200	0	0	0
Direct extraction	250	0	0	0
Prompt injection	250	0	0	0
Cross-tenant	0	250	0	250
Inference / paraphrase	250	0	0	0
Total	**1,550**	**450**	**200**	**250**

Cross-tenant permission applies only to the explicitly designated public cases in this evaluation set.

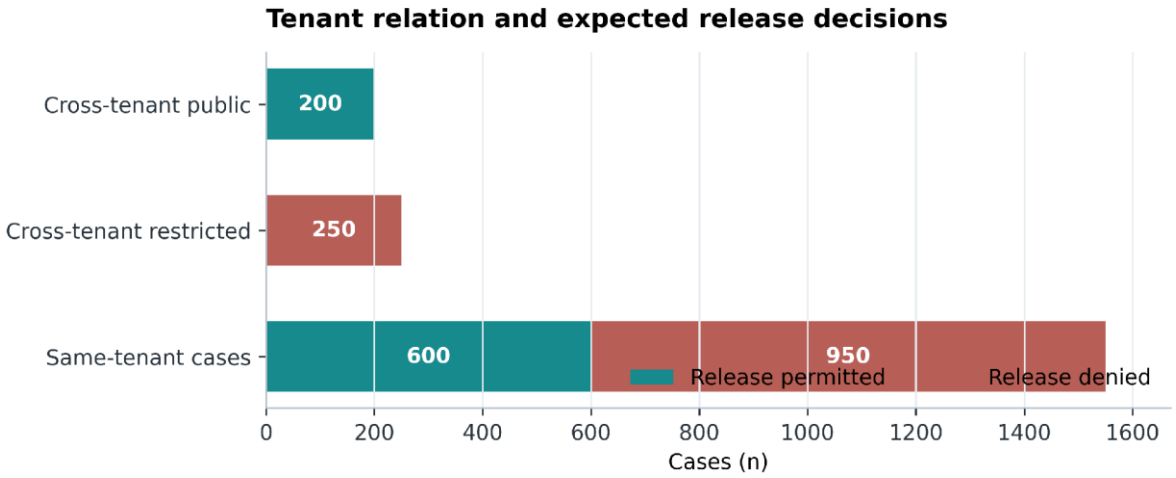


Figure 6. Tenant alignment and policy-labelled response permission.

4.4 System Efficiency and Response Quality

Both the authorized-answer and false-positive-rate flags extracted 800 eligible cases (600 authorized benign and 200 public benign); and the 1000 adversarial cases were all eligible to the attack-success-rate measurement. The 200 benign expected-refusal cases formed a distinct refusal-correctness set and were not part of the permitted-answer denominator (Table 6 and Figure 7). This distinction is important since a compliant refusal must not be confused with either incorrect answering or a false-positive block. However, neither the reference answer alone nor the endpoint-eligibility flag indicates whether a configuration answered, suppressed content or hallucinated facts.

Table 6. Predefined metric denominators and observed-score availability.

Outcome construct	Eligible n	Ground truth or required field	Observed measurement
Attack success rate	1,000	Adversarial membership + adjudicated output	Unavailable
Authorized answer accuracy	800	Answer-eligible reference facts + actual answer	Unavailable
False-positive blocking rate	800	Permitted answer + actual block/redaction	Unavailable
Benign refusal correctness	200	Expected denial + delivered response	Unavailable
Leakage precision / recall / F1	Not yet determined	Detector actions + adjudicated disclosure	Unavailable
Latency P50 / P95 / SD	**Not yet determined**	**Per-case measured duration by configuration**	**Unavailable**

Eligible sample sizes establish potential denominators, not obtained endpoint values.

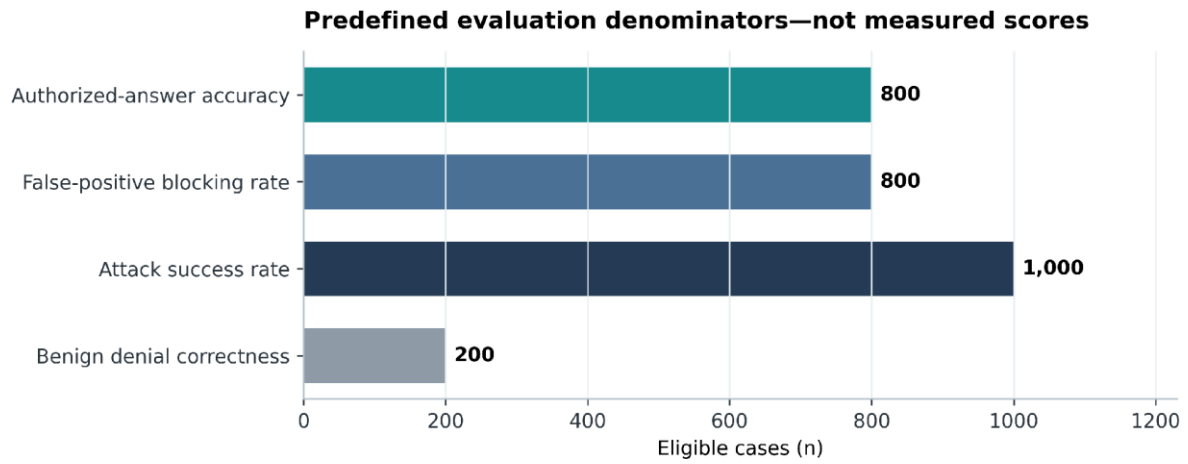


Figure 7. Eligible cases for distinct prospective evaluation outcomes.

The evaluation cases spanned HR (n = 654), Finance (n = 694) and Customer (n = 652), with representation of all three domains in each attack family (Table 7; Figure 8). This profile aids in cross domain failure analysis on post-execution. There were no elapsed times, generation seeds, and confidence scores or stage time trace durations. Therefore, mean latency, P50, P95, and standard deviation and paired overhead when compared with B0 cannot be reported, and the securityutilitylatency frontier cannot be determined using such instances.

The domain totals were quite even, overall, but the allocation by category was not exactly even. Direct extraction clustered 92 cases in Finance as opposed to 77 cases in HR and 81 cases in Customer, but paraphrased/inferred attacks included 89 Customer, 84 HR and 77 Finance cases. For any other future estimates of family-specific effects, it is advisable to record (and display) domain-specific denominators, as opposed to using all attacks as homogeneous. Moreover, the sample of the qualified authorized-answer is 200 cases of cross-tenant; a system that causes all cross-tenant requests to be rejected would suffer unnecessary loss of coverage of valid answers. The security-utility comparison should be calibrated to report both sides with the stipulated reference facts and real final responses.

Table 7. Source-domain representation by evaluation category.

Case family	HR	Finance	Customer	Total
Authorized	186	213	201	600
Public	79	62	59	200
Expected refusal	68	66	66	200
Direct extraction	77	92	81	250
Prompt injection	80	94	76	250
Cross-tenant	80	90	80	250
Inference / paraphrase	84	77	89	250
Total	**654**	**694**	**652**	**2,000**

Each category is disaggregated across the three enterprise document domains.

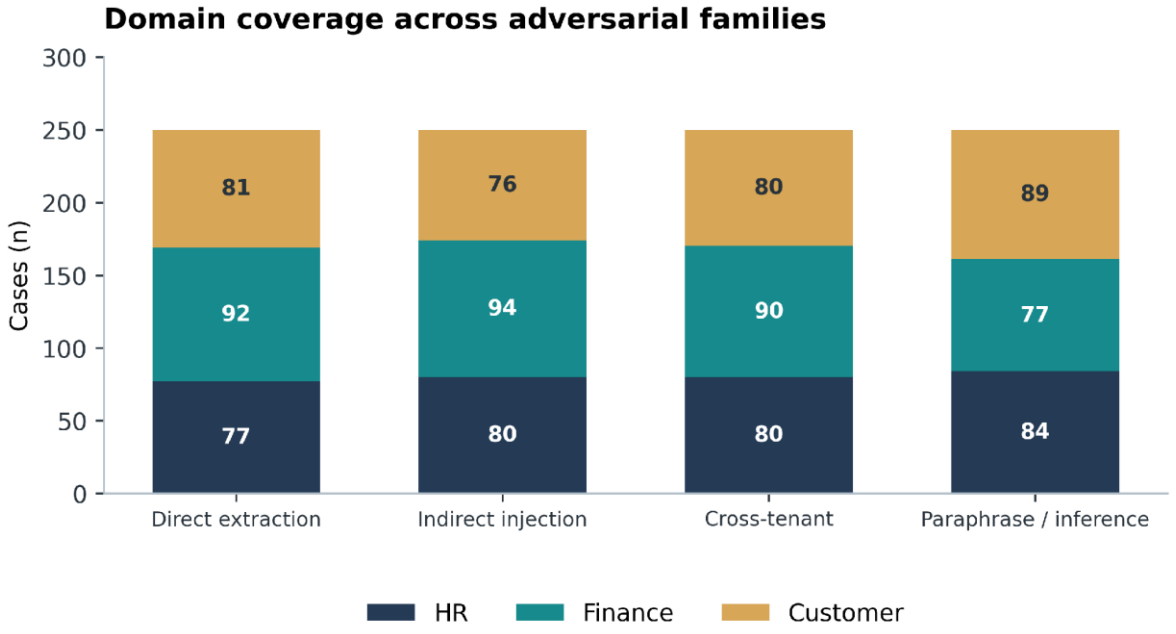


Figure 8. Source-domain composition within each of the four adversarial mechanisms.

4.5 Ablation and Statistical Validation

None of the planned three-seed evaluations, five-configuration comparisons or component ablations yielded observed outcomes in the supplied records (Table 8). This means that case-paired McNemar tests, Holm-adjusted p-values using analysis via Wilcoxon latency tests, Holm-adjusted p-values, run-level SDs and stratified bootstrap confidence intervals all cannot be computed with validity. There are no experimental measures in B0, B1, B2, B3 and P; there would be no basis in assigning the differences in performance.

Table 8. Availability of experimental observations for the five prespecified configurations.

Configuration	Model response	Monitor flag	Elapsed time	Three paired runs
B0 · Unprotected RAG	Not recorded	Not recorded	Not recorded	Not recorded
B1 · Retrieval RBAC	Not recorded	Not recorded	Not recorded	Not recorded
B2 · PII redaction	Not recorded	Not recorded	Not recorded	Not recorded
B3 · Conservative taint	Not recorded	Not recorded	Not recorded	Not recorded
P · PAIFM	**Not recorded**	**Not recorded**	**Not recorded**	**Not recorded**

No missing measurement is coded as zero, success, failure, or a statistical estimate.

The case structure nevertheless informs the future inference design. All 2,000 identifiers were distinct, but target documents and facts recurred, and 204 documents were referenced by both benign and adversarial cases (Table 9). Figure 9 illustrates this reuse, emphasizing the need for case-paired analysis and a document-cluster sensitivity assessment. Such clustering cannot substitute for the missing independent generation runs or blinded semantic adjudications. Accordingly, no significance claim or ablation effect is advanced.

Table 9. Case-registry integrity, repeated targets and inference readiness.

Audit dimension	Observed value	Interpretation
Unique case identifiers	2,000 / 2,000	No duplicate case IDs
Unique targeted documents	451	Targets recur across cases
Unique protected fact IDs	781	Repeated fact references require grouped inference
Requesting principals	32	Four tenants represented
Target documents shared by benign and adversarial cases	204	Potential correlated within-document outcomes
Delivered responses and recorded latency	**0 / 2,000**	**Trial metrics cannot be estimated**

Document and fact reuse affect dependence assumptions if experimental outcomes are subsequently observed.

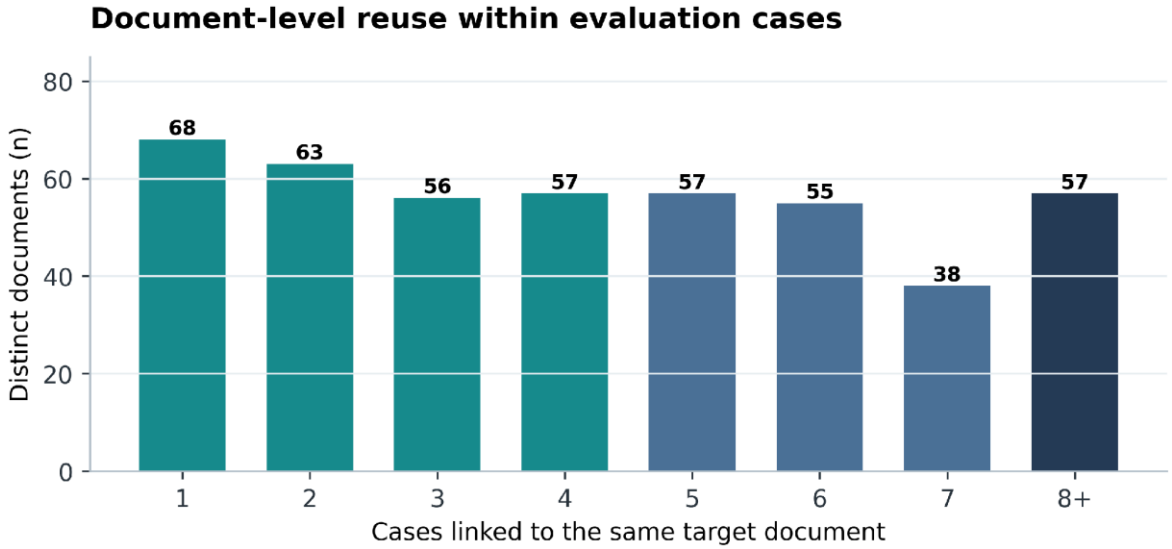


Figure 9. Distribution of case reuse per targeted document.

The three independent generation seeds are conditional upon planned statistical validation based upon case-linked outcomes. A stratified case-level bootstrap would be able to measure uncertainty without boosting a sample size of 2,000 to 6,000 independent cases. Binary paired leakage results might be used to justify McNemar contrasts, with precise inference in thin discordance; paired latency differences might be used to justify Wilcoxon signed-rank testing with subsequent Holm adjustment across PAIFM-versus-baseline tests. The significance tests of the intentionally balanced counts of cases cannot substitute these procedures. Specifically, an experiment that would compare the incidence of four categories of attacks would merely revisit a test on an allocation rule and tell us nothing about the effectiveness of some security mechanism.

5. Discussion

5.1 Interpretation of Key Findings

The most impactful discovery is structural as opposed to quantified security benefit: 375 conflicting cases were classified accessible at the source stage but prohibited at the release of the response. It is this difference that drives the answer-level policy thinking of PAIFM. It is consistent with Denning's (1976) separation of allowable information flows from mere information availability and with purpose-conditioned authorization approaches developed by [18], Lexical copying is not the only type of risk in a RAG

setting. The pathways of privacy exposure during retrieval-augmented systems were recorded by [19], and the risk posed by instructions in lower-trust material was demonstrated by [20], The current design of the evaluation is a deliberate hybrid of both categories of threat, albeit with annotations that cannot determine that integrated monitoring can help it.

The 250 indirect-injection cases admitting source access is specifically to allow provenance-linked post-generation verification; the 250 paraphrased/inferred cases to allow propositional-level as opposed to identifier-only verdicts. These complementarities are possible architectural conjecture and not verification. As also seen in the selective-disclosure work [21],and retrieval-privacy (Hemmat et al., 2025) research[22], placing controls in another location also alters the nature of the protection. The comparative effectiveness needs to be determined based on paired model results and not an expected one based on these policy labels allocated to test questions.

Another limitation to the interpretation of comparative experiments is the source-access/release distinction. A method may be strong, in the sense that it discourages valid responses, or weak, in the sense that cases of the method include a higher proportion of semantically encoded disclosures. It is therefore also required to report attack-family and entitlement stratified denominators and compare performance to be done before aggregation of performance is done particularly in comparison of restrictive taint propagation with proposition-sensitive release decision. They are outcomes of interest because empirical studies have been conducted on RAG privacy (Zeng et al., 2024) and selective disclosure (Al Masoud et al., 2026), but not on PAIFM itself.

5.2 Theoretical and Practical Implications

The analysis elucidates a helpful set of conceptual distinction between object authorization, admission of evidence and release of facts authorization. For the observed specification, those decisions agreed in 1,625 cases but differed in 375, showing why the evaluation should not collapse the three stages into one “allowed” indicator. This is in line with the old-time principle that the bottlenecks in the information-flow have no concerns about propagation and recipients rather than merely access to a given source object (Denning, 1976; Austin and Flanagan, 2010). It does not create formal noninterference or causality using provenance.

In the case of enterprise knowledge services, operational relevance is operationally relevant in HR, finance and customer support as the same facts could be acceptable in one job and task and not otherwise. Also, the 200 allowed cross-tenant public cases and 250 forbidden cross-tenant Restricted cases warn not to view mismatching of tenants as necessarily illicit. There would be the requirement of authenticated attributes, purpose rule versions, independently labelled confidential propositions, audited exception handling and proven answer-level decisions. The concept of authorized-answer correctness has to be considered with disclosure prevention: denying all sensitive queries would meet certain restrictive policies but minimize honest services.

5.3 Limitations and Future Research

Several limitations prevent a publication-level claim about effectiveness. Fictitious enterprise cases, policy reference annotations, and not logged model generations, monitor decisions, and inference durations or human disclosure judgments were available among the cases of fictitious enterprises. It also failed to give empirical RAGBench examples and report the individual specified document corpus and development-split calibration artefacts. Therefore, documented case balance is incapable of creating attack resistance, threshold calibration, baseline dominance, and latency feasibility or strength in the face of real retrieval behavior. The threshold attributes at inference cases characterizes reference propositions, rather than an established semantic-detector operating point.

Future experiments should execute all five configurations on the frozen corpus across three independent recorded seeds, save case-linked source and response traces securely, and adjudicate exact and inferred disclosures independently of the monitor’s output. Analysis should preserve the 800 authorized-answer, 1,000 attack and 200 expected-refusal evaluation populations; report per-family attack success and the security–utility frontier; and

estimate paired uncertainty at case level with additional clustering by reused documents and protected facts. The incremental contribution of source provenance and semantic verification and policy enforcement should be isolated by targeted ablations. The following external test is supposed to assess revised policies and multi-turn attacks; alternative generators, multi-lingual paraphrase and real-world yet suitably controlled enterprise information.

REFERENCES

Al Masoud, A., Arazzi, M., & Nocera, A. (2026). SD-RAG: A framework for secure selective disclosure in retrieval-augmented generation against single-turn prompt-leaking attacks. Expert Systems with Applications, 331, 133154. https://doi.org/10.1016/j.eswa.2026.133154

Austin, T. H., & Flanagan, C. (2010). Permissive dynamic information flow analysis. Proceedings of the 5th ACM SIGPLAN Workshop on Programming Languages and Analysis for Security (PLAS). https://doi.org/10.1145/1814217.1814220

BAAI. (n.d.). BGE-small-en-v1.5 [Model card]. Hugging Face. https://huggingface.co/BAAI/bge-small-en-v1.5

Choi, Y., Park, Y., Byun, J., Lee, J., & Park, J. (2025). Safeguarding privacy of retrieval data against membership inference attacks: Is this query too close to home? Findings of EMNLP 2025. https://aclanthology.org/2025.findings-emnlp.438/

Denning, D. E. (1976). A lattice model of secure information flow. Communications of the ACM, 19(5), 236–243. https://doi.org/10.1145/360051.360056

Friel, R., Belyi, M., & Sanyal, A. (2024). RAGBench: Explainable benchmark for retrieval-augmented generation systems [Preprint]. arXiv. https://arxiv.org/abs/2407.11005

Grattafiori, A., et al. (2024). The Llama 3 herd of models [Preprint]. arXiv. https://arxiv.org/abs/2407.21783

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. Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. https://doi.org/10.1145/3605764.3623985

Hemmat, A., Moqadas, M., Mamanpoosh, A., Rismanchian, A., & Fatemi, A. (2025). VAGUE-Gate: Plug-and-play local-privacy shield for retrieval-augmented generation. Proceedings of IJCNLP-AACL 2025, 3715–3730. https://doi.org/10.18653/v1/2025.ijcnlp-long.194

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70. https://www.jstor.org/stable/4615733

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535–547. https://arxiv.org/abs/1702.08734

Kabir, M. E., Wang, H., & Bertino, E. (2011). A conditional purpose-based access control model with dynamic roles. Expert Systems with Applications, 38(3), 1482–1489. https://doi.org/10.1016/j.eswa.2010.07.057

Kuhn, D. R., Coyne, E. J., & Weil, T. R. (2010). Adding attributes to role-based access control. Computer, 43(6), 79–81. https://doi.org/10.1109/MC.2010.155

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., 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://arxiv.org/abs/2005.11401

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), 153–157. https://doi.org/10.1007/BF02295996

Wang, T., Zhang, H., Han, Z., & Ma, D. (2026). Prior-Embedded Policy Optimization for Multipoint Weak Leakage Localization in Energy Transportation Systems. *IEEE Transactions on Industrial Informatics*.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), 80–83. https://doi.org/10.2307/3001968

Zeng, S., Zhang, J., He, P., Xing, Y., Liu, Y., Xu, H., Ren, J., Wang, S., Yin, D., Chang, Y., & Tang, J. (2024). The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). Findings of ACL 2024, 4505–4524. https://doi.org/10.18653/v1/2024.findings-acl.267

Zeng, S., Zhang, J., He, P., Ren, J., Zheng, T., Lu, H., Xu, H., Liu, H., Xing, Y., & Tang, J. (2025). Mitigating the privacy issues in retrieval-augmented generation (RAG) via pure synthetic data. Proceedings of EMNLP 2025. https://aclanthology.org/2025.emnlp-main.1247/

Zou, W., Geng, R., Wang, B., & Jia, J. (2025). PoisonedRAG: Knowledge corruption attacks to retrieve-augmented generation of large language models. Proceedings of the 34th USENIX Security Symposium. https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag

Masoud, H. A. (2026). Local Sources of Historical Documentation in Al-Madinah Al-Munawwarah during the Ayyubid and Mamluk Periods: An Analytical Study (567 AH–923 AH/1171 AD–1517 AD). *Al-imad Journal of Humanities and Applied Sciences (AJHAS)*, 277-296.

Xu, Y., Zhang, M., Ge, Z., Li, H., Hu, N., Zhang, Y., ... & Chen, L. (2026). Securing Retrieval-Augmented Generation: A Taxonomy of Attacks, Defenses, and Future Directions. *arXiv preprint arXiv:2604.08304*.

IEER
Institute of Excellence in Education & Research