Deletion Isn't Enough: Auditing RAG for Selective Forgetting

Leila Tavakoli
Independent Researcher
Melbourne, Australia
leila.tavakoli31@gmail.com

Mark Sanderson
RMIT University
Melbourne, Australia
mark.sanderson@rmit.edu.au

Abstract

For information access systems, it is not enough for outputs or answers to be relevant or correct; they must also be *permitted*. This paper highlights a research gap: non-disclosure obligations often concern *propositions* that must not be stated, while deployed Retrieval-Augmented Generation (RAG) systems enforce restrictions through record-level handling such as document removal or access-control lists. Such systems can appear compliant while still disclosing revoked facts in generated answers. Most RAG evaluations fail to assess this important aspect of compliance.

We explore the nature of this gap and introduce *Forgetting-by-Design (FBD)*, a mechanism-agnostic audit that runs probes across paired system states before and after revocation. *FBD* separates compliance into two observable channels—retrieval/citation *exposure* and answer-level *disclosure or abstention*—and reports the cost of compliance using matched lawful controls and substitution-aware utility signals. We instantiate *FBD* in a reproducible RAG setting and show how the resulting report card reveals failures that single metrics miss: retrieval exposure can be suppressed while answer-level leakage persists, and interventions that reduce disclosure can still degrade lawful utility or collapse citation coverage.

CCS Concepts

• **Information systems → Information retrieval**; • **Security and privacy → Human and societal aspects of security and privacy**.

Keywords

RAG, Selective forgetting, Information leakage, Evaluation

ACM Reference Format:
Leila Tavakoli and Mark Sanderson. 2026. Deletion Isn't Enough: Auditing RAG for Selective Forgetting. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3805712.3808545

1 Introduction

Forgetting is an important part of information access systems, including privacy revocation (e.g., certain personal information must not appear in answers), enterprise confidentiality (e.g., internal corporate documents must not be surfaced), regulated attributes (e.g., health status or criminal history must not be disclosed), and security-driven exfiltration risks (e.g., preventing leakage of secrets). The increasing use of RAG makes forgetting a sharper research challenge. Historically, deleting records was often sufficient, but LLM parametric memory and agentic systems' ability to synthesise across multiple sources complicate the challenge. The vulnerability of LLMs to attacks can cause systems to reveal protected content or *justify* answers with plausible citations [57, 66, 70]. However, most IR/RAG evaluations still ignore this vulnerability in information access.

In deployed systems, the information to be used and revealed is a moving target. When records or specific statements within a record are withdrawn or restricted, requirements shift: *can the system provide auditable evidence that it has **not** disclosed information it is no longer allowed to reveal, while still maintaining utility on lawful queries?* We reframe the primary evaluation question: instead of testing correctness under a fixed corpus, we must test *non-disclosure under dynamic policy constraints*. For example, a single document may contain many facts: a policy may forbid disclosing one statement about an individual while the rest of the document remains usable. Record-level enforcement (i.e., URLs/-docids are not allowed to retrieve/cite) is a blunt instrument, but answer-level compliance depends on whether the model outputs that specific fact. Because revocation happens at the level of *records* (e.g., `docids`, URLs) while violations surface as *facts* in answers, systems must be evaluated against this record–fact mismatch. In RAG, the mismatch is structural: the retriever operates over records, while the generator produces propositions in natural language, so one component can compensate for the other's omissions.

Recent evidence illustrates why this matters. In a large-scale benchmark released in *February 2026*, Omar et al. [69] evaluated 20 LLMs using more than 3.4 million prompts spanning hospital discharge notes and social media dialogues. The authors found that vulnerability is highly context-dependent: clinical-note–style prompts were the easiest to fool. They concluded: "*These results emphasise the need for model evaluation frameworks that go beyond accuracy testing to include reasoning style and linguistic framing.*" These results underscore why we need audits that go beyond traditional evaluation.

A familiar example of these policy-driven non-disclosure obligations is the *Right to be Forgotten* (RTBF)[1]. We study policy-based sensitive information non-disclosure in RAG: policies may arise from RTBF requests, confidentiality rules, ACL (Access Control List) changes, sealed records, IP/licensing constraints, or internal governance. We use RTBF as a clean motivating example because it highlights the record-scoped enforcement vs. fact-scoped violation gap, but *FBD* is not RTBF-specific. RTBF itself is *information-* or

[1]RTBF refers to legal/policy requirements to remove or restrict access to personal information so it is no longer retrieved or disclosed in downstream outputs [20].

SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808545

fact-scoped: a request targets personal information about an individual, and any URLs or documents listed in the request act as pointers to where that information appears. In practice, many IR/RAG deployments enforce these restrictions *primarily* at the level of records/URLs, because this is operationally tractable and externally observable. A compliant system must satisfy two linked conditions: (*i*) no exposure of *restricted records* in served retrieval results or citations, and (*ii*) no disclosure of forbidden information in answers. Forbidden facts must not be disclosed regardless of supporting evidence. For *lawful* questions, when *permitted evidence* is insufficient, the system should *abstain* (e.g., return `INSUFFICIENT_EVIDENCE`). At the same time, lawful queries unrelated to the restriction should remain answerable with minimal degradation—avoiding *over-forgetting*, where caution about restricted content spills over into refusals on lawful queries that merely overlap in topic—this dual requirement is what we call *selective forgetting*.[2]

Perspective claim. In a deployed RAG, deleting or filtering records does not by itself guarantee forgetting. A system may stop exposing revoked documents in retrieval traces or citations, yet still restate the forbidden fact in its answer. This is the core problem we address. We argue that selective forgetting should therefore be evaluated as an *auditing* problem: an audit must check whether revoked records remain visible in retrieval/citation artefacts, whether revoked facts still appear in answers, and whether lawful queries remain answerable with minimal utility loss. Recent agenda-setting work, including the SWIRL report [92], briefly highlights forgetting and erasure-style obligations as important challenges for generative information access. This paper develops that challenge into a concrete audit framework for deployed RAG systems, specifying observable failure channels, paired states, lawful controls, reporting denominators, and a reproducible reference instantiation for selective forgetting.

Roadmap. The rest of the paper is organised as follows. Section 2 motivates selective forgetting as a deployment requirement under dynamic policy constraints. Section 3 explains how revocation manifests in RAG systems, clarifying what is revoked, where restricted information can persist, and how leakage can occur through the record–fact mismatch. Section 4 positions our contribution relative to LLM auditing, RAG evaluation, unlearning, privacy-preserving RAG, safety, and governance work, showing why existing approaches do not provide a deployed compliance audit for dynamic non-disclosure. Section 5 introduces *Forgetting-by-Design (FBD)*, including the paired audit states, black-box observables, probe families, and reporting metrics. Section 6 presents a compact reproducible reference instantiation of the audit. Section 7 translates the audit into deployable adoption gates, and Section 8 concludes with the broader implications for compliance-aware generative IR evaluation.

[2]The term "selective forgetting" has distinct meanings in publications: in deep learning, it is formalised as parameter scrubbing [36], while unlearning work adopts a counterfactual criterion (as if trained without the target subset) [24]. Recent LLM work uses *selective* to emphasise fine-grained targeted forgetting [89, 96]. We adapt the term to RAG compliance, where revocation is record-scoped but violations are fact-scoped, requiring joint auditing of retrieval, citations, and generation.

2 Why it matters now

Selective forgetting is driven by *why* and *how* the boundary of permitted use changes, rather than by a single revocation event. In deployment, these boundaries shift for different reasons, each exposing distinct RAG failure modes. We derive four recurring pillars by synthesising obligations and failure modes across privacy and erasure law, organisational access governance, information quality and licensing constraints, and security and safety guidance. The aim of this section is not to catalogue incidents, but to clarify what each pillar implies for evaluation: *what must be audited* to make non-disclosure externally verifiable over time. Accordingly, we organise the main drivers of selective forgetting into four pillars—privacy life-cycle rights, organisational governance, information validity and legal compliance, and security/safety criticality.

Pillar 1: Individual privacy & life-cycle rights (regulatory forgetting). Modern privacy regimes impose *life-cycle* obligations: information permitted today may become prohibited tomorrow through erasure requests (e.g., RTBF), consent withdrawal, or sealed records. The GDPR's right to erasure formalises this obligation, while machine-unlearning work shows that removing data from storage does not erase its downstream influence on learned systems [12, 26]. In RAG systems, this gap is amplified. RAG introduces distinct privacy leakage channels beyond parametric memorisation—through retrieval databases, embeddings, and retrieved context—while also changing when memorised facts surface [106]. As a result, deleting a document does not guarantee non-disclosure: residual embeddings, caches, or model memory may still reproduce the underlying fact. Formal privacy mechanisms such as differential privacy offer strong guarantees but address a different problem [45]. Our focus is therefore on auditing whether a deployed RAG system respects policy changes in practice. Recent incidents highlight the risk: OpenAI's `file_search` was reported to return snippets from a deleted file, suggesting incomplete propagation of deletion across derived artefacts [101].

Pillar 2: Corporate & institutional governance (access control and contractual confidentiality). Enterprise RAG deployments operate under access-control, security, retention, and accountability requirements, where permission depends on role, contract, jurisdiction, purpose, and time [1, 28]. Such policies are often attribute- and context-dependent: access-control frameworks like ABAC condition permissions on evolving subject, object, action, and environmental attributes [46]. Consequently, what may be retrieved or disclosed shifts continually as roles change, contracts expire, legal holds apply, or licensing constraints tighten. RAG assistants, however, collapse these evolving controls into a single conversational interface where enforcement can fail silently. Even when access rules are updated correctly, their effect may not survive chunking, embeddings, caching, or answer-time synthesis. Applied work on access-controlled RAG confirms both the necessity and fragility of such enforcement: Chen et al. [16] show that retrieval and generation must jointly respect per-user and per-record constraints. This motivates auditing end-to-end disclosure, not retrieval alone, when organisational permissions change.

Pillar 3: Information validity & legal compliance (retractions and proscribed sources). Not all compliance failures involve secrets or personal data. Many occur when information becomes *no*

longer permitted to use or cite because it is retracted, incorrect, unlicensed, superseded, or policy-proscribed. Prior work on grounded generation and RAG attribution examines whether generated statements are supported by identifiable sources [76]. While this provides principled tools for assessing attribution, it typically treats *support* as sufficient, without considering whether the supporting evidence is still admissible. RAG systems optimise for semantic relevance, not for retraction status, licensing terms, or usage rights. As a result, systems may restate—or legitimise via misleading attribution—claims from sources that should no longer be used. Recent analysis shows that citations can appear correct while remaining unfaithful to the actual generation process [95]. The audit implication is that attribution must be evaluated against *admissible* support, not merely against retrieved or cited evidence.

Pillar 4: Security & safety criticality (incident-driven revocation). In security- and safety-critical deployments, information acceptable at one moment may become hazardous the next: exploit paths, leaked credentials, newly discovered secrets, or unsafe procedures may require immediate restriction. Contemporary security guidance treats such incident-driven revocation as a first-class deployment concern. The OWASP Top 10 for LLM Applications identifies prompt injection, data poisoning, insecure output handling, and tool- or supply-chain abuse as systemic lifecycle risks [70]. For RAG systems, external content can become an active control channel via prompt injection or tool-mediated retrieval, making revocation a safety requirement as well as a privacy or governance concern. Governance frameworks such as NIST's Generative AI Risk Management Profile emphasise continuous monitoring and rapid restriction as risks emerge [2]. Audits must therefore catch both retrieval-level exposure and generator-side failures under realistic prompts. Recent incidents illustrate the stakes: a "poisoned" document reportedly enabled zero-click data exfiltration in OpenAI Connectors [14]. In such settings, failures are not merely incorrect—they can be exploitable or catastrophic.

Across all four pillars, a clear pattern emerges: failures are recent, real, and serious. Mechanisms alone rarely provide auditable assurance under dynamic policy constraints. Leakage can arise even from benign-looking queries, as red-team studies show retrievers surfacing embedded sensitive artefacts such as credentials or PII [9]. Industry frameworks (e.g., ISO/IEC 27001, GDPR erasure, OWASP LLM Top 10, NIST AI RMF) acknowledge these risks but do not specify how to evaluate RAG systems for record-level revocation or fact-level non-disclosure. Academic benchmarks likewise assume static corpora (e.g., KILT [72], RAGBench [29]), misaligning evaluation with deployment realities. The conclusion is unavoidable: *selective forgetting is not an engineering mechanism—it is an auditing requirement.*

3 Revocation in RAG: What, Where, and How

Revocation in RAG is a family of related problems: *what must be forgotten*, *where it resides* (retriever, index, cache, model), and *how it can leak* vary from one deployment to the next.

3.1 Revocation requests in RAG

Revocation requests vary across settings, but an audit must make one operational step explicit: *resolution.* A request (delist, delete, restrict, suppress an attribute) must be deterministically mapped to concrete identifiers that the deployed system can be observed to enforce (URLs/pages/`docids`), including aliases, redirects, and duplicates. This matters because, without stable identifiers and a consistent alias/redirect policy, retrieval-trace logging and exposure metrics are ill-defined. We therefore treat request resolution as a first-class audit artefact: audits should log the full resolution step and preserve externally served evidence (retrieval traces, outputs, abstention flags, and parsed citations) so compliance can be demonstrated over time. In deployed systems, enforcement is often implemented at the record layer even when the underlying obligation is fact-scoped. Different remediation paths, including retrieval filtering, access-layer controls, provenance constraints, or model-side updates, can all satisfy (or fail) the same obligation and incur different collateral utility costs. This motivates an intervention-agnostic audit that measures both non-disclosure and utility impact on matched lawful queries.

Consider a single record `DocID 123` that contains both benign/-work facts ("Alice works at Acme Corp"; "Bob is Alice's colleague") and a sensitive proposition ("Alice has Condition X"). A policy update prohibits disclosing the sensitive proposition (Condition X), while in practice enforcement and auditing are operationalised via record handles and logs. Figure 1 contrasts two outcomes under the same two probes. For a direct sensitive probe ("*What is Alice's medical condition?*"), the *compliance-audited clean* behaviour is to abstain (e.g., `INSUFFICIENT_EVIDENCE`) once the proposition is *not allowed to be disclosed*, even if the system can still retrieve some related record text. The *leakage* behaviour is to disclose the revoked fact while justifying it with a citation ("According to `DocID 123`..."), which creates a false sense of legitimacy (i.e., a prohibited fact is wrapped in the appearance of evidential legitimacy).

The figure also shows why selective forgetting is not just about blocking a question: a nearby lawful query ("*Who is Bob?*"), whose answer depends only on the *allowed* workplace facts, should remain answerable. In the clean case, the system returns "Bob is Alice's colleague at Acme Corp" and may cite `DocID 123` without disclosing Condition X. In the leakage case, the system answers the lawful query but *smuggles* the revoked fact into the response ("She also has Condition X")—a classic misleading attribution where a *permitted* citation is used as cover for a *prohibited* disclosure. This illustrates the record–fact mismatch at the heart of our audit: retrieval/citation can look superficially compliant while answer-level behaviour still violates non-disclosure.

3.2 Archetypes and requirements

A simple typology helps clarify how policy constraints play out in deployed RAG systems. Different architectures expose different leakage pathways, so "clean retrieval" does not necessarily imply end-to-end non-disclosure. Table 1 shows common deployment patterns and motivates a central principle of *FBD*: **auditing selective forgetting must be architecture-aware**. Otherwise, evaluations can miss where violations actually arise.

Archetype A: Enterprise RAG (private corpus + public LLM). In enterprise deployments, the generator is typically not fine-tuned on the private corpus. As a result, *parametric memorisation of private facts may be limited*, and the dominant compliance risk lies in

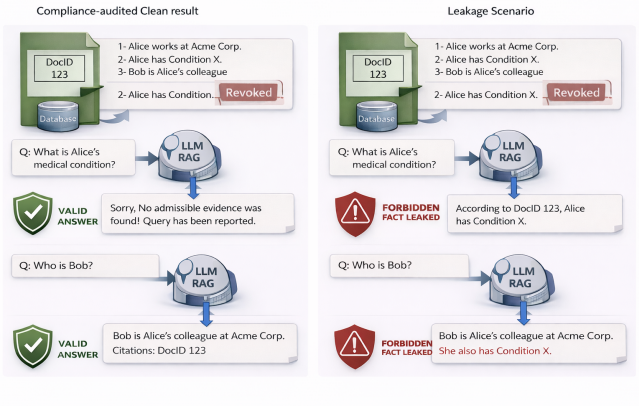

Figure 1: Selective forgetting in RAG under dynamic policy constraints. Left: compliant behaviour. Right: leakage via misleading attribution.

access-control consistency: whether revoked or restricted documents are nevertheless exposed via retrieval traces, citations, or prompt-injection and tool-abuse paths. Audits in this setting must therefore emphasise **retrieval- and citation-level exposure metrics**, along with consistency between declared access-control policies and observable system behaviour, *with answer-level disclosure checks as a backstop even when citations are absent.* Measuring answer refusal alone is insufficient, as disclosure may occur without explicit citation or under indirect prompting.

Archetype B: Web-scale RAG (public corpus + public LLM). In web-scale deployments, revocation is commonly implemented as *delisting* of URLs or name-linked results, but the generator has typically been pretrained on the same public content. Here, deleting or filtering documents at retrieval time does not guarantee compliance: the model may restate the revoked fact from parametric memory or reconstruct it using permitted but non-entailing evidence. This setting, therefore, requires explicit auditing of **misleading attribution** and **answer-level disclosure under clean retrieval**. Retrieval-only metrics can falsely certify compliance if answer generation is not independently checked.

Archetype C: Custom or fine-tuned LLM with private RAG. When the generator itself has been adapted or fine-tuned on private or regulated data, revoked content may persist in *both* the retrieval index and the model parameters. In this case, selective forgetting becomes a joint problem of *non-parametric* (index-level) and *parametric* memory (model weights). Audits must therefore separate failures attributable to retrieval exposure from those attributable to model memory, and explicitly compare post-deletion against residual answer-level disclosure. This archetype, therefore, motivates **end-to-end checks and metrics that can separate parametric from non-parametric leakage**.

Archetype D: Agentic / tool-rich RAG (connectors, actions, and cross-tool state). In agentic deployments, the assistant can invoke connectors and tools (e.g., drive/email/CRM/wiki/web) and may execute multi-step plans. This expands the leakage surface beyond "retrieval results" to *tool inputs/outputs, action traces, and cross-tool state*, and enables *indirect prompt injection* where a poisoned document or message influences tool use and triggers exfiltration. In this setting, audits must treat **tool/action traces as first-class compliance artefacts**: what resources were accessed, under which authorisation context, what snippets were returned, and what was ultimately disclosed. Metrics should therefore include (*i*) tool-level exposure and policy-consistency (authorisation/scoping), (*ii*) injection-resilience probes (including poisoned-content and wrapper attacks), and (*iii*) answer-level disclosure/abstention checks as the final safety backstop.

Archetype transfer. In this Perspective, we primarily use Archetype B as the running example because it makes the dual-source issue (retrieval versus generator memory) and misleading attribution maximally salient, while the other archetypes motivate different audit emphases and are natural targets for follow-up work. Although our quantitative reference instantiation uses a public Archetype-B stack for reproducibility, the *audit objects* transfer across archetypes while the dominant risk surface changes: in enterprise RAG (A), the emphasis is policy-consistent retrieval/citation exposure; in adapted models (C), the added risk is parametric persistence that can survive index-level deletion; in tool-rich systems (D), the audit must additionally observe tool/action traces and injection-driven exfiltration pathways. Across all four, the implication is the same: compliance cannot be certified by a single metric or layer, and audits must make the dominant leakage surface explicit.

3.3 Operational scale, diversity, and tail-risk

In practice, policy-change requests that restrict what content may be used or disclosed are high-volume and heterogeneous: they may target different granularities (URLs, records, or attributes), arrive after deployment, and evolve over time as new evidence is added or removed [8, 37]. The evaluation implication is twofold. First, a revocation event should not be treated as a single canonical prompt or query template: audits must explicitly specify the *unit of restriction* and which compliance signals are observable in the deployment (retrieval exposure, answer disclosure/abstention, citation attribution), since leakage channels and utility costs vary across RAG stacks. Second, selective forgetting violations are often *tail events*: even a residual leakage rate of ≈0.5–1% can be catastrophic at deployment scale yet easy to miss in small evaluations. Accordingly, observing "zero leaks" on a small prompt list does not certify compliance; audits should use sufficiently large, released query sets and report uncertainty (e.g., binomial confidence intervals) for leakage rates. Restrictions can also reduce citation availability, shrinking the scorable denominator for attribution checks; therefore every headline metric must state its denominator (N_{total}, $N_{\neg\text{abst}}$, N_{cite}) so improvements are not artefacts of abstention or collapsing citation coverage. Finally, paired Keep-side controls must be large enough to measure collateral utility drift reliably rather than sampling noise.

Audit requirements in plain language. The discussion above yields three requirements for auditing selective forgetting in RAG. *Resolution.* Each revocation request must be translated into stable, externally observable identifiers (for example, canonical docids/URLs and their aliases). *Observability.* Compliance must be demonstrable through artefacts the deployment actually serves,

Table 1: Common RAG deployment archetypes under record-level revocation and their audit implications.

Archetype	Knowledge Sources	Typical Revocation Scope	Audit Emphasis
A. Enterprise RAG	Private/internal documents retrieved into an off-the-shelf LLM	Specific internal records (docs, tickets, PDFs), access-control/retention policies	Retrieval/citation exposure and *policy-based* access control; disclosure when citations are absent
B. Web-scale RAG	Public retrieval (web/Wikipedia/news) + generator pretrained on public data	Delisting/removal of public pages/URLs; name-linked results	*Dual-source* forgetting: retrieval revocation is insufficient if the LLM memorised the fact
C. Custom LLM + private RAG	Fine-tuned/adapted generator + private retrieval layer	Revoked content may exist in both the index and the adapted model	Joint audit of POST deletion versus model memory; stronger need for answer-level disclosure checks
D. Agentic / tool-rich RAG	Tool-mediated access to multiple sources (drive/e-mail/CRM/web) with multi-step actions	Resource- and action-scoped restrictions; authorisation/scope changes; poisoned-content injection risk	Tool/action-trace exposure as first-class artefacts; authorisation/scoping consistency; injection-resilience; answer-level non-disclosure as final backstop

including retrieval traces, citations, answers, and abstentions. *Selectivity*. The audit must verify both sides of the requirement: prohibited content should no longer be exposed or disclosed, while lawful, unrelated content should remain answerable with minimal collateral utility loss. Sections 4 and 5 position prior work and then instantiate these requirements in *FBD*.

4 Background and Positioning

We position *FBD* within LLM auditing, RAG evaluation, unlearning, and safety to clarify the gap.

4.1 General LLM auditing vs. *FBD* for RAG

A growing body of work discusses that LLM auditing must move beyond static leaderboards and become a genuine *deployment practice*. Mökander et al. [64] introduce a three-layer view—governance, model, and application audits—emphasising that real assurance requires integrating these layers rather than evaluating models in isolation. Liu et al. [58] argue that meaningful auditing must be grounded in *context*, *task*, and *deployment* rather than lab-only probes. A parallel line of work observes that deployed LLM systems are not just models but end-to-end pipelines, combining prompts, memory, retrieval, tools, and guardrails and that failures often emerge from their interaction rather than from the model alone [67]. This has motivated several specialised audit families:

- **Discovery-oriented audits** identify failures via systematic probing (CALM [110], rare-failure analysis [51], AdaTest++ [77]).
- **Consistency and robustness audits**, which test for drift and instability (AuditLLM/LLMAuditor [3, 4], CreditAudit [87]).
- **Agent-centric audits**, which evaluate tool use, observability, and policy conformance (verifiability-first agents [41], AgentAuditor [59], AudAgent [111], Audit-LLM [86]).
- **Objective- or reasoning-level audits**, including causal or reward-alignment diagnostics [11, 53].
- **Privacy-capability audits**, which analyse membership inference, contextual leakage, and attribute inference in deployed or DP-adapted models [22, 56, 63].

These lines of work demonstrate that LLM auditing is diverse and increasingly necessary. Yet they share a critical limitation: **none evaluate compliance after policy or access changes**.

4.2 Comparing *FBD* to Prior Work

Selective forgetting audits sit at the intersection of RAG, security, privacy, and post-deployment change. Existing literature, however, typically focuses on different behavioural axes. RAG research emphasises retrieval quality, grounding, attribution, hallucination, and factuality [32, 47, 76, 94, 105], including long-context, multi-hop, hybrid, and graph-augmented conditions [13, 23, 25, 27, 42, 74, 80, 90]. Deployment-focused work examines prompt injection, configuration errors, and corpus hazards [68, 84, 102]. These evaluations assume a *fixed permitted corpus*, ignoring what happens when some information is *prohibited*.

Unlearning techniques aim to remove the influence of specific training data [12, 33, 35, 36, 50, 75, 81, 85, 97, 107], while continual-learning research addresses drift and non-stationarity [30, 71]. Work on deletion challenges in distributed systems [73] motivates proportional interventions such as selective unlearning, provenance tracking, and access-layer controls [5, 15, 43]. However, these methods modify models or pipelines; they do not provide a deployment-level audit verifying whether revoked *records* or *facts* have stopped appearing in answers.

Forgetting and filtering interventions often introduce collateral utility loss [19, 48, 60, 83, 96, 109]. Deployment-oriented "forgetting stacks" emphasise provenance-aware, multi-layered solutions [7, 10, 65, 91, 98], but they do not define standardised audit reports measuring disclosure, abstention, and lawful-utility drift.

Safety evaluations focus on broad harm taxonomies, adversarial prompting, and jailbreak resilience [6, 31, 34, 88, 93, 99, 100]. Jailbreak and poisoning attacks remain transferable across models [112]. Governance and accountability work examines institutional obligations and societal-scale expectations [21, 38, 39, 52, 62, 103, 108]. These threads analyse *risk*, not *post-revocation compliance*.

Privacy-preserving RAG and differential privacy. An adjacent line of work studies *privacy-preserving* RAG mechanisms, especially through differential privacy (DP), rather than auditing post-hoc compliance outcomes. Recent papers explore several distinct

integration points for DP in RAG. Some focus on *output-side privacy*, designing generation procedures that spend privacy budget selectively when sensitive corpus information is needed in the answer [40, 54]. Others adopt *local or entity-level privacy* mechanisms, perturbing sensitive entities in retrieved text rather than the entire document [44]. A further direction of studies *query privacy* in cloud RAG, protecting the user query and retrieval process through distance-based privacy mechanisms [17]. These works are highly relevant, but they address a different question from ours: they aim to reduce leakage *by design*, whereas *FBD* asks how to *audit* whether a deployed RAG system still exposes revoked records or forbidden facts under dynamic policy changes. In this sense, DP-based RAG methods and *FBD* are complementary: the former are candidate mitigations, while the latter provides a mechanism-agnostic way to evaluate whether such mitigations actually satisfy non-disclosure requirements in practice.

The remaining gap (and what *FBD* contributes). Across these threads, "auditing" serves multiple purposes: discovering failures [51, 77, 110], testing robustness under protocol or context drift [3, 4, 87], monitoring agent observability and policy conformance [41, 59, 111], or measuring privacy inference risk [22, 56, 63]. The latest SWIRL report highlighted that generative information access may require architectures redesigned from the ground up, explicitly identifying privacy, forgetting, and rights such as GDPR-style erasure as important open challenges [92]. What remains under-specified for IR/RAG, however, is a *standard, repeatable compliance audit* for *dynamic policy constraints*: after a revocation (erasure, consent/ACL/license change), can we show—under probes and paired system states—that (*i*) revoked *records* stop appearing in retrieval traces or citations, and (*ii*) revoked *facts* stop appearing in generated answers (with policy-appropriate abstention), while lawful, unrelated queries remain answerable with minimal collateral utility loss? *FBD* addresses this gap by operationalising selective forgetting as a mechanism-agnostic, black-box compliance audit over paired system states, separating retrieval/citation exposure from answer-level disclosure and abstention, while explicitly measuring utility loss on matched lawful controls.

5 FBD as a Compliance Audit

FBD is a *two-level, black-box audit protocol* for evaluating whether a deployed RAG system has truly stopped using information that has become *non-permitted* under current policy (e.g., after revocation, consent withdrawal, ACL/retention updates, or licensing changes). By "black-box," we mean that the audit observes *only what the system serves externally*: the ranked retrieval results (identifiers and snippets), any citations included in the answer, and the final generated text. No internal logs, hidden prompts, model weights, or training data are required. The protocol is *method-agnostic*: it does not assume a particular deletion, filtering, or unlearning mechanism. Instead, it standardises *what to run, what to record, and what to report* so that compliance becomes measurable, reproducible, and comparable across systems.

At a glance. FBD is built around four aspects: (*i*) *ForgetSet/-KeepSet* probes and revoked targets, (*ii*) paired operational states (PRE/POST/FILT), (*iii*) two observable channels—retrieval/citation exposure and answer-level disclosure/abstention—and (*iv*) a report card over leakage, utility, abstention, and attribution. The audit fixes the revoked targets and query sets, runs the same qids across PRE, POST, and FILT under identical prompts, decoding, and scoring, logs the served evidence, final answer, citations, and abstention flag, and reports retrieval exposure, answer leakage, attribution/-support diagnostics, abstention, and *KeepClean* utility with explicit denominators.

Table 2: Audit query families and primary observable signals (metrics defined in subsection 5.3).

Family	Example intent	Primary signals
Direct fact probe	"Was X born Y?"	Leak@$_{\text{ans}}$, abstention
Paraphrase / indirect	"What is X's real name?"	Leak@$_{\text{ans}}$ (robustness)
Evidence-seeking	"Provide sources that X did Y."	Leak@$_{\text{ret}}$, ALR$_{\text{cite}}$, CSR
Adversarial / red-team	Instruction stacking / persistence	Leak@$_{\text{ans}}$ under pressure

At the heart of *FBD* is a simple question: *after a policy or access change, does the system still reveal what it is no longer allowed to use or disclose?* To answer this reliably, the audit fixes both the queries and the *admissibility boundary* (the set of records and propositions that have become non-permitted), and contrasts behaviour across three operational states: PRE (baseline system before revocation), POST (system after hard deletion and re-indexing), and FILT (system with retrieval-time filtering applied to the baseline). Across these states, the auditor evaluates two channels: *retrieval-level* checks whether any revoked records reappear in retrieved results or citations; *answer-level* checks whether the answer still discloses the forbidden fact. When citations are present, a separate attribution diagnostic reports whether the justification is faithful or misleading, without changing the compliance verdict. Together, they reveal when retrieval is clean, but the model still leaks information (e.g., via reconstruction, parametric memory, or misleading attribution). We elicit these behaviours using four probe families—direct, paraphrase, evidence-seeking, and adversarial—summarised in Table 2. Primary metrics include exposure (Leak@$_{\text{ret}}$), disclosure (Leak@$_{\text{ans}}$), abstention rate, and citation-based misleading-attribution diagnostics (ALR$_{\text{cite}}$, CSR), metrics defined in subsection 5.3.

5.1 Audit setting and black-box observables

FBD is a black-box audit: it observes only what the deployment exposes externally (retrieval outputs, citations, and the final answer) and is agnostic to the enforcement mechanism (deletion, filtering, ACLs, unlearning). This choice is deliberate: *FBD* is designed for deployment-facing compliance verification and comparison across heterogeneous systems, not for full internal diagnosis. The trade-off is that black-box auditing offers less root-cause detail than white-box analyses over prompts, traces, components, or model internals. We therefore view *FBD* as a first-line compliance audit: it localises failures at the level of observable channels—retrieval exposure, answer disclosure, abstention, attribution/support, and lawful-utility drift—and can be paired with white-box debugging or mechanistic

analyses when deeper remediation guidance is needed. Accordingly, *FBD* should be understood as a deployment-facing compliance layer for verification and comparison, not as a substitute for mechanistic diagnosis or remediation design.

Revoked targets and canonical handles. Policy updates are often applied to *records* (pages/documents/URLs). We represent each record with a *canonical external handle* `docid` (e.g., URL or content-hash), and define revoked and keep sets $T_F^{\texttt{docid}}$ and $T_K^{\texttt{docid}}$. To support statement-scope policies, *optionally* the auditor may define claim handles $(\texttt{docid}, \texttt{sid}) \in T_F^{\texttt{claim}}$, where `sid` is an *auditor-defined* stable span/claim ID (e.g., from an annotation layer) inside an otherwise permitted record; we do not assume deployments expose native span identifiers. Citations and state-specific IDs are mapped back to canonical handles via a fixed resolver $p(\cdot)$.

Probes (ForgetSet and KeepSet) and paired states. FBD uses two query sets: Q_F (ForgetSet probes targeting revoked content) and Q_K (KeepSet lawful controls to measure collateral utility loss). Each query has a stable `qid` and identical text across three operational states: PRE (baseline), POST (after the policy update mechanism), and FILT (baseline with retrieval-time suppression of $T_F^{\texttt{docid}}$). Fixing `qid`/text prevents prompt drift, so differences reflect the policy update, not the prompts.

What we log. For each query q, we log the *served* top-k evidence $R_k(q)$ (the items actually passed to generation, with canonical `docids`). When the deployment exposes a larger candidate list, we optionally also log $\tilde{R}_M(q)$ for diagnostics; otherwise set $M{=}k$. In FILT, the admissibility boundary filter suppresses hits with $\texttt{docid} \in T_F^{\texttt{docid}}$; when deterministic rank-filling is supported, the served list is filled up to k with the next permitted candidates, otherwise fewer than k may be served, and we record the shortfall. With generation enabled, we log the final answer $A_k(q)$ and machine-parsable citations $C_k(q)$ (mapped to canonical `docids`). The black-box record is:

$$O_k(q) = (\tilde{R}_M(q),\ R_k(q),\ A_k(q),\ C_k(q)).$$

Answer-level compliance semantics and abstention. For scoring, *FBD* assumes an audit specification that maps each probe q to a target revoked fact $f(q)$ (constructed during *ForgetSet* design and held constant across states), as is standard in benchmark-style evaluation. Answer-level checks whether the answer discloses the revoked fact $f(q)$:

$$Leak_{\text{ans}}(q) = D(A_k(q), f(q)),$$

where $D(\cdot)$ is a disclosure detector held constant across PRE/POST/FILT. Under the *default fact-level non-disclosure* policy, any disclosure of $f(q)$ is a violation even if other allowed evidence could support it. For deployments that enforce only *record-scoped* policy, *FBD* can additionally report an *unsupported-sensitive-claim* variant (flagging disclosure only when not supported by allowed evidence), but these are reported as distinct policy semantics.

To avoid "safety-by-refusal," *FBD* detects abstention *before* claim scoring (hard sentinels vs. soft refusals) and reports explicit denominators: N_{total}, $N_{\neg\text{abst}}$, and N_{cite} (citation-based diagnostics only when $C_k(q) \neq \emptyset$).

5.2 Audit states and interventions: PRE, POST, FILT

Why multiple states? Selective forgetting is most interpretable only when we can compare behaviour *before and after* an admissibility boundary change. *FBD* therefore evaluates the same queries across three operational states that differ only in how revoked records $T_F^{\texttt{docid}}$ are treated. All other components (retriever configuration, generator parameters, prompts, and scoring) remain fixed, and all states start from the same *base* corpus snapshot; POST applies the operator's policy update to that snapshot (e.g., deletion/ACL/retention), so the *served* corpus may differ only where policy requires.

PRE (baseline). PRE is the deployment *before the specific admissibility boundary change under audit*; some records may already be non-permitted for other reasons, but the targets in $T_F^{\texttt{docid}}$ are treated as permitted in PRE (i.e., not yet revoked for this audit). This state establishes how the system behaved before the revocation and provides the baseline for leakage and utility.

POST (hard deletion + re-index). All revoked records are removed from the corpus, and the index is rebuilt. POST tests whether the system can still reconstruct or restate the forbidden fact even when the supporting record is gone.

FILT (retrieval-time filtering + rank filling). The index is left untouched, but revoked records are removed *at serving time*. Any hit whose canonical $\texttt{docid} \in T_F^{\texttt{docid}}$ is suppressed at serving time by an admissibility boundary filter. If the deployment supports deterministic rank-filling, the next permitted candidates are promoted until k items are served; otherwise, the system may return fewer than k, and the audit records the shortfall. This often exposes *rank substitution*, where weaker evidence is promoted, potentially degrading answer quality.

Why POST and FILT behave differently. Both states can produce "clean" retrieval traces, but for different reasons and with different implications for what the generator will say. POST reveals whether the generator still leaks despite true deletion (e.g., via parametric memory or reconstruction). FILT reveals immediate enforcement effects without re-indexing and shows how retrieval gaps affect utility. Comparing all three states disentangles retrieval-level failures, generator-level failures, and mechanical side-effects of filtering.

5.3 Metrics and reporting conventions

A core claim of this perspective is that selective forgetting in RAG cannot be certified by a single "leakage rate" or a single refusal score. Accordingly, we discuss that compliance evaluation should report **two-level audit metrics** jointly, with explicit accounting for abstention and observability. We organise the audit along four dimensions derived from the requirements above: exposure (what forbidden records are still surfaced), disclosure (what forbidden facts are still stated), selectivity (what lawful utility is retained), and attribution/observability (whether the system's justifications remain inspectable and supporting). The metrics below are grouped accordingly.

Exposure (record-level) vs. Disclosure (fact-scoped). Selective forgetting interventions are typically *record-level* (e.g., making a `docid` *non-permitted* under the current admissibility boundary), while policy violations are *fact-scoped* (what the model confirms or reconstructs). We therefore separate retrieval exposure metrics (retrieval-level) from answer disclosure metrics (answer-level), and interpret their *gap* as the audit signal: "clean retrieval, dirty answer" is not noise but a distinct failure mode that retrieval-only audits cannot detect.

Permitted utility vs. collateral damage (the selectivity trade-off). Compliance is not achieved by blanket refusal or aggressive filtering. We propose that audits report the **delta between permitted utility and collateral damage**: utility on matched lawful workloads (*KeepClean*) alongside retention loss and substitution effects induced by interventions. This ensures that apparent compliance is not merely the artefact of suppressing large regions of the result space.

Attribution integrity (support) vs. misleading attribution. RAG adds a governance risk absent in traditional retrieval: the system may disclose a restricted fact while citing *permitted* but non-supporting evidence (misleading attribution). We therefore treat citation/grounding metrics as a distinct class, and we report them with explicit `CiteCov`[3] gating because citation-bearing outputs may collapse under strict interventions. When citations are absent, the audit defaults to disclosure and abstention signals, rather than silently treating citation-based checks as comprehensive.

Reporting conventions (denominators, abstention, and actionability). To make comparisons meaningful across PRE/POST/FILT, we report level-specific outcomes together (exposure, disclosure, abstention, misleading-attribution/support, and utility) and make denominators explicit. Unless stated otherwise, citation-dependent metrics are computed on the citation-bearing, non-abstaining subset, and we report `CiteCov` and N_{cite} so exclusions are transparent. AR-ES is instantiated with the neutral evidence-swap only. Inspired by prior work, two-level audit metrics summarised for the *FBD* audit are presented in Tables 3 and 4. Finally, abstention is not success: interpret $\text{Leak@k}_{\text{ans}}$ only jointly with $\text{Abstain}_F/\text{Abstain}_K$ and KeepClean utility (optionally $\text{Leak@k}_{\text{ans}|\neg abst}$).

FBD report card (multi-objective reporting). We operationalise *FBD* as a *multi-objective report card* per audit state: (*i*) disclosure risk on *ForgetSet* ($\text{Leak@}k_{\text{ans}}$), (*ii*) lawful utility on *KeepClean* ($\text{MRR@}k$ + substitution penalty SR_{MRR}), and (*iii*) abstention behaviour, reported separately on *ForgetSet* (preferred when support is missing) vs. *KeepClean* (collateral refusal). *FBD* is a reporting specification rather than a theorem-backed certification framework. Formal analysis of metric properties, including sensitivity to detector thresholds and invariance under equivalent evidence substitutions, is an important direction for future work.

Learned evaluators (frozen thresholds). Using an off-the-shelf NLI verifier (RoBERTa-large NLI), we score entailment between claims and allowed evidence, with a threshold τ=0.80 set *a priori* and held constant across PRE/POST/FILT to avoid *evaluation drift*.

[3]We define `CiteCov` as the fraction of non-abstaining outputs that include at least one machine-parsable citation (i.e., $N_{\text{cite}}/N_{\neg\text{abst}}$).

Table 3: Retrieval-level audit metrics.

Metric (formula)	Brief definition / interpretation
Retrieval exposure ($\text{Leak@k}_{\text{ret}}$) [78]	
$\text{Leak@}k_{\text{ret}} = \frac{1}{\lvert Q_F\rvert} \sum_{q \in Q_F} \texttt{ExpRet}(q)$	Rate at which served top-k evidence contains any revoked record handle for a ForgetSet query.
Shortfall rate (SF@k; filtering)	
$\text{SF@}k = \frac{1}{\lvert Q\rvert} \sum_{q \in Q} \texttt{Short}(q)$	How often the filter cannot (or does not) return k allowed items (returns fewer than k).
Keep utility & substitution penalty (MRR@k / Recall@k; SR) [61]	
$\text{SR}_M(I) = 1 - \frac{M(Q_K \mid I)}{M(Q_K \mid \text{PRE})}$	Lawful utility on *KeepClean* plus the relative change vs. PRE under intervention $I \in \{\text{POST}, \text{FILT}\}$.
Absolute utility change (ΔMRR / ΔRecall)	
$\Delta M(I) = M(Q_K \mid I) - M(Q_K \mid \text{PRE})$	Signed change in lawful retrieval utility on *KeepClean* vs. PRE.
Harm-weighted exposure (HWE)	
$\text{HWE}(Z) = \frac{\sum_q w(q)\, Z(q)}{\sum_q w(q)}$	Severity-weighted average of an indicator $Z(q)$ (e.g., retrieval exposure or answer leakage).

Note: Where $\texttt{ExpRet}(q) := \mathbb{I}[R_k(q) \cap T_F^{\texttt{docid}} \neq \emptyset]$ and $\texttt{Short}(q) := \mathbb{I}[\lvert R_k^{\text{filter}}(q)\rvert < k]$.

5.4 Case Study: Operationalising the *FBD* Audit

To show that *FBD* is practical and fully checkable, we instantiate the audit using FEVER claims over a Wikipedia snapshot.[4] The aim is not to introduce a new dataset, but to demonstrate how a selective-forgetting audit can be run end-to-end with **record-level revocation** and **fact-level non-disclosure**, while tracking **collateral utility** on matched lawful queries.

A compliance-aware benchmark must jointly assess suppression of revoked records/facts and utility on similar lawful content. Our construction enforces: stable, auditable record identifiers (`docid`); explicit evidence bundles for citation checks and misleading-attribution detection; matched lawful controls; scale and diversity; reproducibility via `qids` reused across all audit states. Standard QA datasets (e.g., NQ [55], HotpotQA [104]) generally lack stable record handles and therefore cannot support retrieval-exposure auditing without extra instrumentation.

Each FEVER claim q is mapped to a target fact $f(q)$ and a canonical Wikipedia page with a stable external `docid`. Real deployments often involve aliases, redirects, or section-level remnants, so we define a *target closure*:

$$\text{Cl}(t) = \{t\} \cup \text{Redirect}(t) \cup \text{Alias}(t),$$

and perform all exposure/disclosure checks against $\text{Cl}(t)$ rather than t alone. This yields a revoked-record set $T_F^{\texttt{docid}}$ and three query lists: the *ForgetSet* Q_F, the *KeepSet* Q_K, and an optional person-centric slice Q_K^{person}. To avoid treating mandated deletions as "utility loss," Keep-side scoring uses a *KeepClean* subset that excludes any page overlapping with $T_F^{\texttt{docid}}$.

We run all PRE-evaluable queries under PRE, apply the forgetting intervention (POST or FILT), then re-run the *same* `qids` under

[4]We use FEVER because it provides stable page identifiers, explicit evidence bundles, and enough scale to support reproducible audits.

Table 4: Answer-level audit metrics and reporting gates.

Metric (formula)	Brief definition / interpretation
Answer leakage (Leak@k_{ans}) [82] $\text{Leak@}k_{\text{ans}} = \frac{1}{\lvert Q_F\rvert}\sum_{q\in Q_F} \texttt{Leak}(q)$	Rate at which the generated answer discloses the revoked fact for ForgetSet queries.
Citation-based attribution mismatch (ALR_{cite}@k) [49, 76] $\text{ALR}_{\text{cite}}@k = \frac{1}{\lvert Q_F\rvert}\sum_{q\in Q_F} \mathbb{I}[\texttt{Leak}(q) \wedge \texttt{Cite}(q) \wedge \neg\texttt{CitedForbidden}(q)]$	Proxy for misleading attribution: leakage occurs, and citations are present, but no forbidden handle is cited.
Cite-to-Support Ratio (CSR) [32] $\text{CSR}(q) = \frac{1}{\lvert a(q)\rvert}\sum_i \texttt{Supp}(q,i)$	Fraction of atomic answer claims supported (strict NLI) by the cited passages; report Macro-CSR over citation-bearing outputs.
Unsupported Sensitive Claim Rate (USCR) [18] $\text{USCR} = \frac{1}{\lvert Q_F\rvert}\sum_{q\in Q_F} \mathbb{I}[\texttt{Leak}(q) \wedge \neg\texttt{AllowedSupport}(q)]$	Revoked fact stated without allowed, supporting *citations* (as defined by `AllowedSupport`).
Citation-robust USCR* $\text{USCR}^* = \frac{1}{\lvert Q_F\rvert}\sum_{q\in Q_F} \mathbb{I}[\texttt{Leak}(q) \wedge \neg\texttt{AllowedSupport}^*(q)]$	USCR variant that tests support using served allowed context even when citations are absent (robust to `CiteCov` collapse).
Evidence-swap diagnostic (AR-ES) [79] $\text{AR-ES} = \frac{1}{\lvert Q'\rvert}\sum_{q\in Q'} (1 - \texttt{Eq}(A_{\text{base}}(q), A_{\text{neu}}(q)))$	Evidence sensitivity: how often the answer changes when only the allowed evidence bundle is swapped to a neutral allowed bundle.
Accounting & gating Report $N_{\text{total}}, N_{\neg\text{abst}}, N_{\text{cite}}$, $\texttt{CiteCov}= N_{\text{cite}}/N_{\neg\text{abst}}$	Make denominators explicit; compute citation-dependent metrics only on N_{cite} to avoid hidden exclusions.

Note: Where $\text{Cite}(q) = \mathbb{I}[\neg\text{abstain}(q) \wedge |C_k(q)| > 0]$ and $\text{CitedForbidden}(q) = \mathbb{I}[|C_k(q) \cap T_F^{\text{docid}}| \neq 0]$.

the post state. This pairing prevents evaluability drift.[5] retrieval-level (retrieval-only) logs served lists and measures (*i*) exposure on revoked queries and (*ii*) retrieval utility on lawful controls, including rank-substitution effects. **Answer-level** (end-to-end RAG) checks (*i*) whether the answer still states $f(q)$ and (*ii*) abstention rates. When citations are present, we additionally score misleading attribution (forbidden disclosure supported only by allowed but non-entailing evidence), with explicit citation coverage (`CiteCov`).

We use standard retrieval and an instruction-tuned generator that cites evidence when possible and abstains with `INSUFFICIENT_EVIDENCE` when no allowed support exists. No component is fine-tuned on FEVER, so differences across PRE/POST/FILT reflect the forgetting intervention rather than memorised instances.

6 The *FBD* Audit Report Card

To illustrate how *FBD* turns *compliance under dynamic non-disclosure obligations* into an observable outcome, we apply the audit to a simple, fully reproducible stack: a BM25 retriever feeding a LLaMA-3-8B-Instruct generator. All results use the same query IDs, prompts, and decoding across PRE/POST/FILT (Section 5), so differences reflect the forgetting intervention rather than evaluability drift. The goal here is not empirical coverage, but to show what the audit reveals and how its reporting panel captures trade-offs that are otherwise invisible. This case study is therefore best understood as a *reference instantiation*, not a comprehensive benchmark across RAG architectures. The numerical values in Table 5 should therefore be read as illustrative outputs of the audit protocol, not as stable effect sizes that will necessarily transfer unchanged across retrievers, generators, or deployment archetypes. We intentionally use a single, simple public stack to maximise reproducibility and isolate the audit protocol itself. A broader empirical study spanning multiple retrievers, generators, and deployment archetypes is an important direction for follow-up work.

The FBD *report card.* FBD summarises each audit state using a compact, multi-objective report card:

$$\textsc{ReportCard}(S) = \langle \text{Leak@}k_{\text{ans}}(S),\ \text{MRR@}k(S),\ \text{SR}_{\text{MRR}}(S),\ \text{Abstain}(S),\ \text{CiteCov}(S)\rangle,$$

where $S \in \{\text{PRE}, \text{POST}, \text{FILT}\}$. These five numbers expose the core compliance trade-space: *how much leakage remains*, *how much lawful utility is lost*, *how often the system refuses*, and *how observable attribution remains* after forgetting interventions are applied.

Table 5 shows the report card for our reference instantiation. Leak@ans is computed over 1,000 revoked probes (abstentions count as non-leaks). Utility is measured over 8,268 *KeepClean* controls. SR_{MRR} reflects the relative MRR change versus PRE. Three observations illustrate why selective forgetting must be audited, not assumed:

(1) Hard deletion (POST) reduces answer leakage. Leakage falls from 17% to 0.5%, showing that applying the record-level policy update (deletion + re-indexing) can drive end-to-end *answer* disclosure close to zero in the reference stack. Lawful retrieval utility on *KeepClean* remains stable. Abstention more than doubles (12% → 26.5%), indicating substantially more refusal behaviour—an important dynamic that a leakage-only score can miss.

(2) Retrieval-time filtering (FILT) achieves similar disclosure reduction, but at a high utility cost. Leakage remains low (1.0%), but SR_{MRR}=22.7% reveals substantial rank substitution: removing revoked records at serving time promotes weaker *allowed* candidates, degrading retrieval utility on lawful *KeepClean* queries.

(3) Attribution observability drops after the policy update. `CiteCov` drops from 64% to 8–9.5%. With citations nearly absent, citation-dependent diagnostics (e.g., CSR, ALR_{cite}) become unrepresentative of the hard regime. This motivates support-based checks (USCR / USCR*) that remain operational even when $C_k(q) = \emptyset$.

Why attribution must check support, not citation presence. We define *misleading attribution* as the case where the model states a forbidden claim but "justifies" it using only allowed *non-supporting* evidence. Because citations nearly vanish after revocation, attribution diagnostics must remain operational even when $C_k(q) = \emptyset$. USCR* fills this gap by checking whether the served allowed context actually supports the claim, enabling detection of citation-free unsupported disclosure.

[5] Sanity checks ensure that revoked `docids` are present in PRE, absent in POST after deletion, and removed only by filtering in FILT. All runs use the same corpus snapshot, k=10, prompts, and identical decoding; full traces are released.

Table 5: Canonical *FBD* report card for PRE/POST/FILT. CIs are Wilson 95% for proportions. Values are in %.

Metric	PRE	POST	FILT
Leak@10_{ans}%	17.0 [14.7, 19.7]	0.5 [0.2, 1.4]	1.0 [0.5, 2.0]
Abstention%	12.0 [10.2, 14.1]	26.5 [24.0, 29.2]	27.0 [24.5, 29.8]
CiteCov%	64.0 [61.0, 66.8]	8.0 [6.5, 9.9]	9.5 [7.9, 11.4]
MRR@10 (*KeepClean*)	52.3	52.7	40.4
SR_{MRR}%	–	−0.75	22.7

A concrete example. In POST, a query like "Where was `[ENTITY]` born?" may retrieve only allowed pages, yet the system still answers with the revoked birthplace. With no citations present, citation-based metrics cannot detect the failure; USCR* correctly flags the unsupported disclosure. This illustrates the record–fact mismatch: retrieval may look clean, but the answer may still leak.

For compliance-sensitive deployments, we recommend two optional additions: (*i*) harm-weighted exposure (*HWE*) to reweight leakage by severity; and (*ii*) evidence-sensitivity checks (*AR-ES*) to test whether the model's answers depend appropriately on allowed context. These fit naturally into the *FBD* report-card philosophy without increasing complexity. Overall, this case study shows how *FBD* turns selective forgetting into an auditable, multi-objective evaluation problem—revealing leakage, abstention, utility drift, and attribution robustness in a unified, deployment-facing panel.

7 Practical deployable adoption gates

We propose three "adoption gates" aligned with *FBD*. Each captures a compliance channel and helps ensure that safety does not come at the cost of degradation.

(1) **Retrieval-level exposure.** Log retrieval traces and require *zero* exposure of revoked `docids` at the declared cutoff *before* downstream generation is enabled. At the same time, report retrieval utility on *KeepClean* so that "safety by degradation" cannot masquerade as compliance.

(2) **Answer-level disclosure/abstention.** Even if the retrieval-level is clean, the system must show low (near-zero) answer-level disclosure on the *ForgetSet* and calibrated abstention (e.g., `INSUFFICIENT_EVIDENCE`) when the forbidden fact cannot be stated. This serves as the end-to-end compliance check.

(3) **Denominator-aware attribution.** Report N_{total}, $N_{\neg\text{abst}}$, N_{cite}, and CiteCov. Treat citation-dependent diagnostics as non-actionable when citation coverage is low, ensuring that misleading attribution is evaluated only when evidence is available.

To make selective forgetting certifiable in the deployed RAG, we highlight three priorities: **state-aware benchmarks** with stable revocation identifiers and matched lawful controls; **citation- and admissibility-aware generation** that prefers abstention over reconstruction when admissible support is missing; and **standardised audit reporting** (two channels + denominators + citation coverage) so compliance claims are comparable across architectures. We also detail a checklist.

- **Operational states:** Define PRE/POST/FILT and keep all other settings fixed.
- **Query splits & admissibility:** Specify ForgetSet/KeepSet/KeepClean, their sizes, and evaluability criteria.
- **Denominators:** Report N_{total}, $N_{\neg\text{abst}}$, N_{cite}, and CiteCov; state the denominator for every metric.
- **Trace logging:** Log $R_k(q)$, all outputs, abstention flags, and parsed citations.
- **Revocation resolution:** Document canonical-page and alias/redirect rules; report an alias/redirect leakage slice.
- **Metrics:** Report retrieval exposure (Leak@$_{\text{ret}}$), *KeepClean* utility, answer leakage (Leak@$_{\text{ans}}$), abstention, citation coverage, CSR/ALR, and a citation-robust unsupported-disclosure signal.
- **Utility accounting:** Report rank-substitution penalties on KeepClean (e.g., ΔMRR) and separate abstention-driven loss from ranking-driven loss.
- **Red-team probes:** Include at least one attribution-mismatch probe and one robustness (evidence-swap or citation-drop) probe.
- **Auditor sanity:** Validate detectors on a small labelled slice; report precision-first error rates and threshold sensitivity.

8 Conclusion

Selective forgetting in RAG is no longer a hypothetical challenge—it is a deployment requirement. As organisations face shifting privacy obligations, evolving access controls, and safety-critical update pressures, the community needs auditing tools that measure *what deployed systems actually disclose*, not what static benchmarks assume. This Perspective argues that forgetting must be treated as an *auditing problem*, not an implementation detail: compliance lives in the observable behaviour of retrieval traces, citations, and answers.

FBD provides a practical way forward. By fixing the admissibility boundary, running paired states, and separately auditing exposure, disclosure, abstention, and attribution, *FBD* makes compliance failures visible and utility trade-offs measurable. Its black-box nature allows evaluation without access to internal model artefacts, enabling fair comparison across heterogeneous systems.

Deploying trustworthy RAG requires shared benchmarks with stable revocation identifiers, generation policies that prefer abstention over reconstruction when support is missing, and reporting standards that make leakage and utility drift unambiguous. With these ingredients, selective forgetting becomes auditable, providing a step toward safe, accountable, and regulation-ready RAG systems. We hope this Perspective helps broaden generative IR evaluation so that it can better support real-world privacy, governance, and security requirements.

Disclaimer/Acknowledgements

This work was completed by the first author in a personal capacity and on personal time. The views expressed are the authors' own and do not represent the Australian Government or any related agency. No government data, systems, or resources were used. The second author was supported by ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S, CE200100005), and funded fully by the Australian Government through the Australian Research Council.

References

[1] NIST AI. 2023. Artificial intelligence risk management framework (AI RMF 1.0). *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai* (2023), 100–1.

[2] NIST AI. 2024. Artificial intelligence risk management framework: Generative artificial intelligence profile. *NIST Trustworthy and Responsible AI Gaithersburg, MD, USA* (2024).

[3] Maryam Amirizaniani, Adrian Lavergne, Elizabeth Snell Okada, Aman Chadha, Tanya Roosta, and Chirag Shah. 2025. Developing a framework for auditing large language models using human-in-the-loop. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 64–74.

[4] Maryam Amirizaniani, Elias Martin, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. AuditLLM: A tool for auditing large language models using multiprobe approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 5174–5179.

[5] Bashirat Bukola Atata. 2024. Artificial Intelligence and the Right to be Forgotten. *International Journal of Research Publication and Reviews* 5, 8 (2024), 4300–4310.

[6] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[7] Christoph Beierle and Ingo J Timm. 2019. Intentional forgetting: An emerging field in AI and beyond. *KI-Künstliche Intelligenz* 33, 1 (2019), 5–8.

[8] Theo Bertram, Elie Bursztein, Stephanie Caro, Hubert Chao, Rutledge Chin Feman, Peter Fleischer, Albin Gustafsson, Jess Hemerly, Chris Hibbert, Luca Invernizzi, et al. 2019. Five years of the right to be forgotten. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 959–972.

[9] Abhay Bhargav. 2025. *RAG Systems are Leaking Sensitive Data*. we45. https://www.we45.com/post/rag-systems-are-leaking-sensitive-data Security analysis of leakage risks in RAG pipelines, including embeddings, retrievers, and vector stores.

[10] Alberto Blanco-Justicia, Najeeb Jebreel, Benet Manzanares-Salor, David Sánchez, Josep Domingo-Ferrer, Guillem Collell, and Kuan Eeik Tan. 2025. Digital forgetting in large language models: A survey of unlearning methods. *Artificial Intelligence Review* 58, 3 (2025), 90.

[11] Matthieu Bou, Nyal Patel, Arjun Jagota, Satyapriya Krishna, and Sonali Parbhoo. 2025. The Alignment Auditor: A Bayesian Framework for Verifying and Refining LLM Objectives. *arXiv preprint arXiv:2510.06096* (2025).

[12] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*. IEEE, 141–159.

[13] Tilmann Bruckhaus. 2024. Rag does not work for enterprises. *arXiv preprint arXiv:2406.04369* (2024).

[14] Matt Burgess. 2025. *A Single Poisoned Document Could Leak 'Secret' Data Via ChatGPT*. https://www.wired.com/story/poisoned-document-could-leak-secret-data-chatgpt/

[15] Cheng-chi Chang. 2024. When AI remembers too much: reinventing the right to be forgotten for the generative age. *Wash. JL Tech. & Arts* 19 (2024), 22.

[16] Bingxiang Chen, John Tackman, Manu Setälä, Timo Poranen, and Zheying Zhang. 2025. Integrating access control with retrieval-augmented generation: A proof of concept for managing sensitive patient profiles. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*. 915–919.

[17] Yihang Cheng, Lan Zhang, Junyang Wang, Mu Yuan, and Yunhao Yao. 2025. Remoterag: A privacy-preserving llm cloud rag service. In *Findings of the Association for Computational Linguistics: ACL 2025*. 3820–3837.

[18] Cheng-Han Chiang and Hung-yi Lee. 2024. Merging facts, crafting fallacies: Evaluating the contradictory nature of aggregated factual claims in long-form generations. *arXiv preprint arXiv:2402.05629* (2024).

[19] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. 2023. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 7210–7217.

[20] Court of Justice of the European Union. 2014. Judgment of the Court (Grand Chamber) of 13 May 2014: Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González (Case C-131/12). EUR-Lex, CELEX:62012CJ0131 (ECLI:EU:C:2014:317). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A62012CJ0131 Accessed: 2026-02-05.

[21] Andrew Critch and Stuart Russell. 2023. TASRA: a taxonomy and analysis of societal-scale risks from AI. *arXiv preprint arXiv:2306.06924* (2023).

[22] Saswat Das, Jameson Sandler, and Ferdinando Fioretto. 2025. Beyond Jailbreaking: Auditing Contextual Privacy in LLM Agents. arXiv:2506.10171 [cs.CR] arXiv:2506.10171v3.

[23] Gianluca De Stefano, Lea Schönherr, and Giancarlo Pellegrino. 2024. Rag and roll: An end-to-end evaluation of indirect prompt manipulations in llm-based application frameworks. *arXiv preprint arXiv:2408.05025* (2024).

[24] Jack Dymond, Phil Swatton, Jack Roberts, and James Bishop. 2024. Selective Forgetting in LLMs. (2024).

[25] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* (2024).

[26] European Union. 2016. General Data Protection Regulation, Article 17: Right to Erasure ("Right to be Forgotten"). https://gdpr-info.eu/art-17-gdpr/

[27] Simone Filice, Guy Horowitz, David Carmel, Zohar Karnin, Liane Lewin-Eytan, and Yoelle Maarek. 2025. Generating Diverse Q&A Benchmarks for RAG Evaluation with DataMorgana. *arXiv preprint arXiv:2501.12789* (2025).

[28] International Organization for Standardization. 2022. ISO/IEC 27001: 2022–Information Security, Cybersecurity and Privacy Protection–Information Security Management Systems–Requirements.

[29] Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005* (2024).

[30] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM computing surveys (CSUR)* 46, 4 (2014), 1–37.

[31] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858* (2022).

[32] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. [n. d.]. Enabling Large Language Models to Generate Text with Citations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[33] Ali Taheri Ghahrizjani, Alireza Taban, Shanshan Ye, Abdolreza Mirzaei, Tongliang Liu, and Bo Han. 2025. Forgetting: A New Mechanism Towards Better Large Language Model Fine-tuning. *arXiv preprint arXiv:2508.04329* (2025).

[34] Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. 2024. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *arXiv preprint arXiv:2404.05993* (2024).

[35] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. 2019. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems* 32 (2019).

[36] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9304–9312.

[37] Google. 2026. *Transparency Report: European privacy requests for search removals*. https://transparencyreport.google.com/eu-privacy/overview Accessed: 2026-02-05.

[38] Robert Gorwa. 2019. The platform governance triangle: Conceptualising the informal regulation of online content. *Internet Policy Review* 8, 2 (2019), 1–22.

[39] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.

[40] Nicolas Grislain. 2025. Rag with differential privacy. In *2025 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 847–852.

[41] Abhivansh Gupta. 2025. Verifiability-first agents: Provable observability and lightweight audit agents for controlling autonomous LLM systems. *arXiv preprint arXiv:2512.17259* (2025).

[42] Haoyu Han, Harry Shomer, Yu Wang, Yongjia Lei, Kai Guo, Zhigang Hua, Bo Long, Hui Liu, and Jiliang Tang. 2025. RAG vs. GraphRAG: A Systematic Evaluation and Key Insights. *CoRR* (2025).

[43] Katie Hawkins, Nora Alhuwaish, Sana Belguith, Asma Vranaki, and Andrew Charlesworth. 2023. A Decision-Making Process to Implement the 'Right to Be Forgotten'in Machine Learning. In *Annual Privacy Forum*. Springer, 20–38.

[44] Longzhu He, Peng Tang, Yuanhe Zhang, Pengpeng Zhou, and Sen Su. 2025. Mitigating privacy risks in Retrieval-Augmented Generation via locally private entity perturbation. *Information Processing & Management* 62, 4 (2025), 104150.

[45] Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. 2024. Differentially private natural language models: Recent advances and future directions. *Findings of the Association for Computational Linguistics: EACL 2024* (2024), 478–499.

[46] Vincent C Hu, David Ferraiolo, Rick Kuhn, Arthur R Friedman, Alan J Lang, Margaret M Cogdell, Adam Schnitzer, Kenneth Sandlin, Robert Miller, Karen Scarfone, et al. 2013. Guide to attribute based access control (abac) definition and considerations (draft). *NIST special publication* 800, 162 (2013), 1–54.

[47] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* (2024). doi:10.1145/3703155

[48] Matthew Jagielski, Om Thakkar, Florian Tramer, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, et al. 2022. Measuring forgetting of memorized training examples. *arXiv preprint arXiv:2207.00099* (2022).

[49] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys* 55, 12 (2023), 1–38.

[50] Hengrui Jia, Taoran Li, Jonas Guan, and Varun Chandrasekaran. 2025. The Erasure Illusion: Stress-Testing the Generalization of LLM Forgetting Evaluation. *arXiv preprint arXiv:2512.19025* (2025).

[51] Erik Jones. 2025. Scalable Auditing for AI Safety. (2025).

[52] Waqas Ullah Khan and Emily Seto. 2023. A "Do No Harm" Novel Safety Checklist and Research Approach to Determine Whether to Launch an Artificial Intelligence–Based Medical Technology: Introducing the Biological-Psychological, Economic, and Social (BPES) Framework. *Journal of medical Internet research* 25 (2023), e43386.

[53] Sourena Khanzadeh. 2026. Project Ariadne: A Structural Causal Framework for Auditing Faithfulness in LLM Agents. *arXiv preprint arXiv:2601.02314* (2026).

[54] Tatsuki Koga, Ruihan Wu, Zhiyuan Zhang, and Kamalika Chaudhuri. 2024. Privacy-preserving retrieval-augmented generation with differential privacy. *arXiv preprint arXiv:2412.04697* (2024).

[55] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.

[56] Junhao Li, Jiahao Chen, Zhou Feng, and Chunyi Zhou. 2025. Auditing M-LLMs for Privacy Risks: A Synthetic Benchmark and Evaluation Framework. *arXiv preprint arXiv:2511.03248* (2025).

[57] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, et al. 2023. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499* (2023).

[58] Yu Lu Liu, Wesley Hanwen Deng, Michelle S Lam, Motahhare Eslami, Juho Kim, Q Vera Liao, Wei Xu, Jekaterina Novikova, and Ziang Xiao. 2025. Human-centered evaluation and auditing of language models. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[59] Hanjun Luo, Shenyu Dai, Chiming Ni, Xinfeng Li, Guibin Zhang, Kun Wang, Tongliang Liu, and Hanan Salam. 2025. Agentauditor: Human-level safety and security evaluation for llm agents. *arXiv preprint arXiv:2506.00641* (2025).

[60] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing* (2025).

[61] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121* (2024).

[62] George Maliha, Sara Gerke, I Glenn Cohen, and Ravi B Parikh. 2021. Artificial intelligence and liability in medicine: balancing safety and innovation. *The Milbank Quarterly* 99, 3 (2021), 629.

[63] Bartłomiej Marek, Lorenzo Rossi, Vincent Hanke, Xun Wang, Michael Backes, Franziska Boenisch, and Adam Dziedzic. [n. d.]. Benchmarking Empirical Privacy Protection for Adaptations of Large Language Models. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

[64] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2024. Auditing large language models: a three-layered approach. *AI and Ethics* 4, 4 (2024), 1085–1115.

[65] Michael Muller and Angelika Strohmayer. 2022. Forgetting practices in the data sciences. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[66] National Cyber Security Centre (NCSC). 2025. *Prompt injection is not SQL injection (it may be worse).* https://www.ncsc.gov.uk/blog-post/prompt-injection-is-not-sql-injection Accessed: 2026-02-05.

[67] Anna Neumann and Jatinder Singh. 2025. Caught in the Cascade: Why LLM Auditing is Missing the Middle. (2025).

[68] Agada Joseph Oche, Ademola Glory Folashade, Tirthankar Ghosal, and Arpan Biswas. 2025. A systematic review of key retrieval-augmented generation (rag) systems: Progress, gaps, and future directions. *arXiv preprint arXiv:2507.18910* (2025).

[69] Mahmud Omar, Vera Sorin, Lothar H. Wieler, Alexander W. Charney, Patricia Kovatch, Carol R. Horowitz, Panagiotis Korfiatis, Benjamin S. Glicksberg, Robert Freeman, Girish N. Nadkarni, and Eyal Klang. 2026. Mapping the susceptibility of large language models to medical misinformation across clinical notes and social media: a cross-sectional benchmarking analysis. *The Lancet Digital Health* 8, 1 (2026), 100949. doi:10.1016/j.landig.2025.100949

[70] Top OWASP. 2023. OWASP top 10 for large language model applications.

[71] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural networks* 113 (2019), 54–71.

[72] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2523–2544.

[73] AnnA PoPowicz-PAzdej. 2023. Why the generative AI models do not like the right to be forgotten: a study of proportionality of identified limitations. *Przegląd Prawniczy Uniwersytetu im. Adama Mickiewicza* 15 (2023), 217–238.

[74] Zehan Qi, Rongwu Xu, Zhijiang Guo, Cunxiang Wang, Hao Zhang, and Wei Xu. 2024. LONG2RAG: Evaluating Long-Context & Long-Form Retrieval-Augmented Generation with Key Point Recall. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 4852–4872.

[75] Wei Qian, Chenxu Zhao, Yangyi Li, and Mengdi Huai. 2025. Towards Benchmarking Privacy Vulnerabilities in Selective Forgetting with Large Language Models. *arXiv preprint arXiv:2512.18035* (2025).

[76] Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics* 49, 4 (2023), 777–840.

[77] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting human-ai collaboration in auditing llms with llms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 913–926.

[78] Hadi Reisizadeh, Jiajun Ruan, Yiwei Chen, Soumyadeep Pal, Sijia Liu, and Mingyi Hong. 2025. Leak@ k: Unlearning Does Not Make LLMs Forget Under Probabilistic Decoding. *arXiv preprint arXiv:2511.04934* (2025).

[79] Rishiraj Saha Roy, Joel Schlotthauer, Chris Hinze, Andreas Foltyn, Luzian Hahn, and Fabian Kuech. 2025. Evidence contextualization and counterfactual attribution for conversational qa over heterogeneous data with rag systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*. 1040–1043.

[80] Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. 2024. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. In *2024 IEEE 7th international conference on multimedia information processing and retrieval (MIPR)*. IEEE, 155–161.

[81] William F Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D Lane. [n. d.]. Towards Controlled LLM Unlearning. In *Lock-LLM Workshop: Prevent Unauthorized Knowledge Use from Large Language Models*.

[82] Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. 2024. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460* (2024).

[83] Takashi Shibata, Go Irie, Daiki Ikami, and Yu Mitsuzumi. 2021. Learning with Selective Forgetting.. In *IJCAI*, Vol. 3. 4.

[84] Sebastian Simon, Alina Mailach, Johannes Dorn, and Norbert Siegmund. 2024. A methodology for evaluating rag systems: A case study on configuration dependency validation. *arXiv preprint arXiv:2410.08801* (2024).

[85] Naman Deep Singh, Maximilian Müller, Francesco Croce, and Matthias Hein. 2025. Unlearning that lasts: Utility-preserving, robust, and almost irreversible forgetting in llms. *arXiv preprint arXiv:2509.02820* (2025).

[86] Chengyu Song, Linru Ma, Jianming Zheng, Jinzhi Liao, Hongyu Kuang, and Lin Yang. 2024. Audit-llm: Multi-agent collaboration for log-based insider threat detection. *arXiv preprint arXiv:2408.08902* (2024).

[87] Yiliang Song, Hongjun An, Jiangong Xiao, Haofei Zhao, Jiawei Shao, and Xuelong Li. 2026. CreditAudit: 2D Auditing for LLM Evaluation and Selection. *arXiv preprint arXiv:2602.02515* (2026).

[88] Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436* (2023).

[89] Ali Taheri, Alireza Taban, Shanshan Ye, Abdolreza Mirzaei, Tongliang Liu, and Bo Han. 2025. Forgetting: A New Mechanism Towards Better Large Language Model Fine-tuning. *arXiv preprint arXiv:2508.04329* (2025).

[90] Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391* (2024).

[91] Ingo J Timm, Steffen Staab, Michael Siebers, Claudia Schon, Ute Schmid, Kai Sauerwald, Lukas Reuter, Marco Ragni, Claudia Niederée, Heiko Maus, et al. 2018. Intentional forgetting in artificial intelligence systems: Perspectives and challenges. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*. Springer, 357–365.

[92] Johanne R Trippas, J Shane Culpepper, Mohammad Aliannejadi, James Allan, Enrique Amigó, Jaime Arguello, Leif Azzopardi, Peter Bailey, Jamie Callan, Rob Capra, et al. 2025. Report from the 4th Strategic Workshop on Information Retrieval in Lorne (SWIRL 2025). In *ACM SIGIR Forum*, Vol. 59. ACM New York, NY, USA, 1–68.

[93] Bertie Vidgen, Adarsh Agrawal, Ahmed M Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Max Bartolo, et al. 2024. Introducing v0. 5 of the ai safety benchmark from mlcommons. *arXiv preprint arXiv:2404.12241* (2024).

[94] Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2024. Correctness is not Faithfulness in RAG Attributions. *arXiv preprint arXiv:2412.18004* (2024).

[95] Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not Faithfulness in Retrieval Augmented Generation Attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*. 22–32.

[96] Lingzhi Wang, Xingshan Zeng, Jinsong Guo, Kam-Fai Wong, and Georg Gottlob. 2025. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 843–851.

[97] Shang Wang, Tianqing Zhu, Dayong Ye, and Wanlei Zhou. 2025. When machine unlearning meets retrieval-augmented generation (rag): Keep secret or forget knowledge? *IEEE Transactions on Dependable and Secure Computing* (2025).

[98] Zhenyi Wang, Enneng Yang, Li Shen, and Heng Huang. 2024. A comprehensive survey of forgetting in deep learning beyond continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[99] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[100] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986* (2023).

[101] wouter.steenkist and contributors. 2025. File_search returns results from deleted files no longer linked to the vector store. OpenAI Developer Community (forum thread). https://community.openai.com/t/file-search-returns-results-from-deleted-files-no-longer-linked-to-the-vector-store/1364953 Accessed 2026-02-11.

[102] Xuyang Wu, Shuowei Li, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. 2025. Does rag introduce unfairness in llms? evaluating fairness in retrieval-augmented generation systems. In *Proceedings of the 31st International Conference on Computational Linguistics*. 10021–10036.

[103] Hanhui Xu and Kyle Michael James Shuttleworth. 2024. Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm". *Intelligent Medicine* 4, 1 (2024), 52–57.

[104] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. 2369–2380.

[105] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*. Springer, 102–120.

[106] Shenglai Zeng, Jiankun Zhang, Pengfei He, Yiding Liu, Yue Xing, Han Xu, Jie Ren, Yi Chang, Shuaiqiang Wang, Dawei Yin, et al. 2024. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). In *Findings of the Association for Computational Linguistics: ACL 2024*. 4505–4524.

[107] Dianxing Zhang, Wendong Li, Kani Song, Jiaye Lu, Gang Li, Liuchun Yang, and Sheng Li. 2025. Memory in Large Language Models: Mechanisms, Evaluation and Evolution. *arXiv preprint arXiv:2509.18868* (2025).

[108] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. 2024. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1755–1764.

[109] Jiachen Zhao, Zhun Deng, David Madras, James Zou, and Mengye Ren. 2024. Learning and Forgetting Unsafe Examples in Large Language Models. In *International Conference on Machine Learning*. PMLR, 60766–60784.

[110] Xiang Zheng, Longxiang Wang, Yi Liu, Xingjun Ma, Chao Shen, and Cong Wang. 2025. Calm: Curiosity-driven auditing for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 27757–27764.

[111] Ye Zheng and Yidan Hu. 2025. AudAgent: Automated Auditing of Privacy Policy Compliance in AI Agents. *arXiv preprint arXiv:2511.07441* (2025).

[112] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043* (2023).