International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(17s) 2026 pp. 307-313
https://doi.org/10.70917/ijcisim-2026-4739
Article

# Beyond RAGAS: A Compliance-Aware Evaluation Framework for Retrieval-Augmented Generation in Regulated Sectors

**Ashutosh Rana***

Independent Researcher, USA

**Abstract:** Retrieval-Augmented Generation has become the dominant architecture for enterprise knowledge systems, with adoption spanning healthcare, education, and financial services. Existing evaluation frameworks — most notably RAGAS — measure faithfulness, answer relevance, context precision, and context recall. These metrics are necessary but insufficient for regulated sectors, where legal obligations imposed by the Family Educational Rights and Privacy Act (FERPA), the Health Insurance Portability and Accountability Act (HIPAA), and the General Data Protection Regulation (GDPR) introduce evaluation requirements that generic quality metrics cannot address. This paper introduces CAREF, the Compliance-Aware RAG Evaluation Framework, comprising five formally defined dimensions: Access-Control Precision (ACP), PHI Leakage Rate (PLR), Regulatory Citation Accuracy (RCA), Role-Aware Retrieval Precision (RARP), and Audit Trail Completeness (ATC). RAGAS faithfulness is shown to be statistically uncorrelated with FERPA Access-Control Precision (Pearson $r = 0.12$, $p = 0.44$), confirming that faithfulness optimisation cannot serve as a proxy for compliance. With CAREF applied, the composite Compliance Score improved from 0.51 to 0.93 on a production RAG system serving over 88,000 enrolled students.
**Keywords:** FERPA, HIPAA, GDPR, RAG evaluation, compliance-aware retrieval, access control, PHI leakage detection, regulatory citation accuracy, enterprise RAG, audit trail

## 1. Introduction

Retrieval-Augmented Generation has become the principal architecture for grounding LLM outputs in authoritative organisational knowledge. By pairing a retrieval mechanism — typically a dense vector store or hybrid sparse-dense index — with an LLM generator, RAG systems reduce hallucination rates relative to parametric-only generation (Lewis et al., 2020) [15]. Enterprise adoption has accelerated across higher education, healthcare, and financial services, with 63.6% of enterprise AI deployments relying on GPT-family models paired with FAISS or Elasticsearch retrieval backends (Ozay et al., 2024) [6]. Agentic RAG architectures, which introduce dynamic multi-step retrieval loops, further amplify compliance risk by potentially crossing authorisation boundaries multiple times per user request (Singh et al., 2025) [7].

Three failure modes illustrate the evaluation gap. First, a RAG system may retrieve documents a user is not authorised to access under FERPA identity-scoped access controls [10] — a RAGAS faithfulness score of 1.0 is entirely compatible with this failure. Second, a system may include HIPAA-protected health information (PHI) in a response delivered to an unauthorised recipient [11] — RAGAS scores are unchanged. Third, a system may cite a superseded regulatory provision — cited faithfully because the retrieved document predates the amendment — producing a materially incorrect regulatory recommendation that registers as high faithfulness. The role-based access control challenges in enterprise vector database deployments, documented by Zhong et al. (2025) [16], confirm that most deployed RAG systems treat all retrieved content as equally accessible, failing to enforce the fine-grained access control required in regulated environments.

Five contributions: (1) CAREF — the first compliance-specific RAG evaluation framework; (2) formal definitions of five dimensions (ACP, PLR, RCA, RARP, ATC) with metric formulations and regulatory grounding; (3) correlation analysis demonstrating RAGAS faithfulness cannot substitute for compliance evaluation ($r = 0.12$); (4) production validation on a FERPA/HIPAA/GDPR system at scale (500 queries, 4 user roles, Cohen's $\kappa = 0.89$); (5) open-source implementation of all five dimensions.

## 2. Background and Related Work

### *2.1 RAG Evaluation: State of the Art*

RAGAS (Es et al., 2024) [1] defines five metrics — faithfulness, answer relevance, context precision, context recall, and answer correctness — processing over 5 million monthly evaluations across major cloud deployments. ARES (Saad-Falcon et al., 2024) [2] extends automated RAG evaluation by training LLM judges on synthetic data, achieving high precision in reference-free evaluation across eight knowledge-intensive tasks at NAACL 2024. The BEIR benchmark (Thakur et al., 2021) [3] provides heterogeneous retrieval evaluation across 18 datasets — assessing retrieval quality without access-control or compliance constraints. MetaRAG (2025) [8] applies metamorphic testing for hallucination detection in a black-box, reference-free manner (62+ CEUR-WS citations). MEGA-RAG (Xu et al., 2025) [9] demonstrates a 40% hallucination reduction through multi-evidence guided answer refinement (91+ PMC citations). RAGVue (2025) [17] introduces a diagnostic evaluation paradigm for explainable RAG assessment in high-stakes settings. None of these frameworks addresses access-control correctness, PHI leakage, citation currency, or audit trail completeness.

### *2.2 Privacy and Compliance Challenges in Enterprise RAG*

Privacy challenges in RAG-enhanced LLMs for healthcare contexts are systematically reviewed by Guan et al. (2025) [18], who identify PHI leakage at retrieval, transmission, and generation stages as distinct failure modes requiring targeted mitigation at each pipeline boundary. The review covers 23 healthcare RAG applications and confirms that standard evaluation frameworks provide no detection coverage for any of the three PHI leakage stages. Complementing this, the privacy protection framework proposed in the ScienceDirect study [19] introduces a novel PHI detection methodology specifically designed for RAG outputs, validating it against HIPAA's 18-category specification. Zhong et al. [16] address the access control dimension directly: HONEYBEE, accepted at SIGMOD 2026, demonstrates that RBAC enforcement in vector databases can achieve up to 13.5x lower query latency than row-level security while maintaining recall — making compliance-preserving retrieval practically deployable without the performance penalty that had previously deterred enterprise adoption.

### *2.3 Regulatory Requirements for RAG Systems*

FERPA (34 C.F.R. Part 99) [10] requires identity-scoped education record access — retrieval must be filtered by the querying user's role-permission mapping. HIPAA (45 C.F.R. § 164.514(b)) [11] prohibits disclosure of 18 PHI categories to unauthorised recipients, with civil monetary penalties up to $1.5 million per category per year. GDPR (Regulation (EU) 2016/679) [12] imposes data minimisation under Article 5(1)(c), erasure rights under Article 17, and records-of-processing requirements under Article 30. The EU AI Act (Regulation (EU) 2024/1689) [13], Article 12, mandates automatic logging for high-risk AI systems, with educational AI explicitly high-risk under Annex III. Hacker et al. (2026) [14] provide the compliance architecture framework for AI agent providers under EU law, establishing that Article 12 logging obligations extend to every retrieval and generation step in agentic RAG pipelines.

### *2.4 Agentic RAG and Multiplied Compliance Risk*

Agentic RAG architectures surveyed by Singh et al. (2025) [7] introduce dynamic, goal-oriented retrieval loops executing multiple retrieval steps per request, each potentially crossing authorisation boundaries. This multiplicative effect on compliance risk makes the absence of access-control-aware evaluation not merely inconvenient but architecturally dangerous: a single-step RAG system with ACP = 0.90 becomes a multi-step agentic system whose

probability of at least one access-control violation per session approaches near-certainty as the number of retrieval steps increases.

## 3. The Compliance-Aware RAG Evaluation Framework (CAREF)

### *3.1 Framework Overview*

CAREF comprises five independently evaluable dimensions, collectively sufficient to certify compliance readiness for FERPA, HIPAA, and GDPR deployment contexts. The framework extends RAGAS rather than replacing it — a fully compliant system should score highly on both quality (RAGAS) and compliance (CAREF) dimensions. Improving any single CAREF dimension does not imply improvement in the others; they address distinct regulatory obligations with distinct failure mechanisms.

Table 1. CAREF: Five Compliance Evaluation Dimensions

| Dimension | Abbr. | Regulatory Basis | Formal Metric | Target |
|---|---|---|---|---|
| Access-Control Precision | ACP | FERPA 34 C.F.R. §99.31 | | 1.00 |
| PHI Leakage Rate | PLR | HIPAA 45 C.F.R. §164.514(b) | | 0.00 |
| Regulatory Citation Accuracy | RCA | FERPA/GDPR/HIPAA | \|{c: valid_at_query(c)}\| / \|Cite\| | |
| Role-Aware Retrieval Precision | RARP | FERPA §99.31 + GDPR Art.5 | \|{d: role_auth(role,d)}\| / \|R\| | |
| Audit Trail Completeness | ATC | GDPR Art.30 + EU AI Act Art.12 | \|{e: logged(e)}\| / \|Events\| | 1.00 |

### *3.2 Dimension 1: Access-Control Precision (ACP)*

ACP = |{d ∈ R : authorized(user, d)}| / |R|. The access-control enforcement challenge in vector database RAG systems — specifically the trade-off between query latency and access correctness — is addressed by Zhong et al. [16], whose HONEYBEE framework provides the retrieval-layer enforcement mechanism against which ACP is measured. A RAGAS faithfulness score of 1.0 is compatible with ACP = 0: perfect faithfulness to unauthorised content is still a regulatory violation. Target: ACP = 1.0.

### *3.3 Dimension 2: PHI Leakage Rate (PLR)*

PLR = |{r ∈ Resp : contains_phi(r)}| / |Resp|. Detection uses a two-stage pipeline: rule-based scanning for all 18 HIPAA categories followed by semantic classification for indirect identifiers. Guan et al. [18] identify three leakage stages (storage, transmission, generation); PLR targets the generation stage — the final point at which HIPAA-compliant filtering can be applied. The privacy protection methodology of [19] informs the semantic classification pass. Target: PLR = 0.0.

### *3.4 Dimension 3: Regulatory Citation Accuracy (RCA)*

RCA = |{c ∈ Cite : valid_at_query_date(c)}| / |Cite|. A high-faithfulness RAG system will faithfully reproduce citations from its knowledge base — including citations to statutes that have since been amended. Currency verification requires a maintained regulatory amendment log and citation extraction from generated

responses. This failure mode is entirely invisible to RAGAS, ARES [2], and BEIR [3]. Target: RCA ≥ 0.95.

### *3.5 Dimensions 4 and 5: RARP and ATC*

RARP = |{d ∈ R : role_authorized(role, d)}| / |R|. Enforces functional scope separation at the role level: a financial aid counsellor and academic advisor are both authorised users but should not cross-retrieve from each other's document scopes. The RBAC structure of HONEYBEE [16] provides the enforcement mechanism; RARP measures its effectiveness. Target: RARP ≥ 0.90.

ATC = |{e ∈ Events : logged(e)}| / |Events|. Satisfies GDPR Article 30 records-of-processing requirements and EU AI Act Article 12 automatic logging obligations [13] as interpreted by Hacker et al. [14] for agentic RAG pipelines. Target: ATC = 1.0.

## 4. Regulatory Grounding and Implementation

FERPA compliance in the production system is implemented via a HONEYBEE-aligned metadata filter [16] applied to Vertex AI Matching Engine queries server-side, before any candidate document reaches the application layer. This pre-ranking filter prevents the failure mode described by Zhong et al. [16] in which approximate nearest-neighbour search surfaces access-restricted documents as false positives when filtering is applied post-ranking. Moving ACP from 0.61 to 1.00 required no algorithm change — only the architectural shift from post-ranking to pre-ranking filter application.

PHI leakage prevention applies the 18-category scanner from [19] to every generated response before delivery, with HIPAA-positive detections triggering redaction and an ATC audit event. GDPR data minimisation is enforced by RARP's role-scope limiting. The GDPR Article 17 erasure pipeline removes records from the vector index within 72 hours of a data subject request and records the deletion event in the ATC audit log. Hacker et al. [14] establish that Article 12 logging obligations in agentic RAG extend to every retrieval step — not only the final generation — motivating the ATC dimension's requirement for complete retrieval-event logging across all pipeline stages.

## 5. Validation

### *5.1 Deployment Context*

Production RAG system: US higher education technology provider, 88,860+ enrolled students. Three workflows: financial aid advising, academic advising, and enrollment management. Knowledge base: approximately 14,000 documents with FERPA access-control metadata. Platform: GCP Vertex AI, FAISS hybrid retrieval, Gemini 1.5 Pro generator, CAREF compliance middleware implementing all five dimensions.

### *5.2 Experimental Design*

500 queries across 4 user roles (125 per role) and 3 workflow categories. Two annotators established ground-truth access-control labels (Cohen's $\kappa = 0.89$ — indicating high inter-annotator agreement). PHI ground truth applied the full HIPAA 18-category specification. Citation currency ground truth verified against the institution's compliance office amendment log. RAGAS v0.2 for quality metrics; CAREF open-source pipeline for compliance metrics.

### *5.3 Results*

Table 2. CAREF Validation Results — Production RAG System (n=500 queries, 4 roles)

| Metric | Pre-CAREF | Post-CAREF | Target | Regulatory Basis |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Access-Control Precision (ACP) | 0.61 | **1.00** | 1.00 | FERPA §99.31 |
| PHI Leakage Rate (PLR) | 0.08 | **0.00** | 0.00 | HIPAA §164.514(b) |
| Reg. Citation Accuracy (RCA) | 0.72 | **0.97** | | FERPA/GDPR/HIPAA |
| Role-Aware Retrieval Prec. (RARP) | 0.74 | **0.93** | | FERPA §99.31 + GDPR Art.5 |
| Audit Trail Completeness (ATC) | 0.51 | **1.00** | 1.00 | GDPR Art.30 + EU AI Act Art.12 |
| **Composite Compliance Score (CCS)** | **0.51** | **0.93** | | **All above** |
| RAGAS Faithfulness | 0.87 | 0.89 | | Quality only |
| Pearson r (Faithfulness vs. ACP) | 0.12 (p=0.44) | — | N/A | Correlation analysis |

The pre-CAREF composite score of 0.51 falls below the production-readiness threshold despite a RAGAS faithfulness score of 0.87 — validating the central claim. The ACP improvement from 0.61 to 1.00 was achieved by shifting the FERPA metadata filter from post-ranking to pre-ranking retrieval enforcement, consistent with the HONEYBEE architectural recommendation [16]. The PLR reduction from 0.08 to 0.00 required deploying the 18-category PHI scanner middleware at the generation boundary. ATC improvement from 0.51 to 1.00 required extending the logging layer to capture retrieval events previously omitted from the audit trail — precisely the gap identified by Hacker et al. [14] in agentic RAG logging architectures.

## 6. Discussion

### *6.1 The RAGAS-Compliance Gap*

The Pearson r = 0.12 between RAGAS faithfulness and ACP confirms that faithfulness optimisation is orthogonal to access-control correctness. Faithfulness asks whether the generated answer is grounded in retrieved content; it cannot ask whether retrieval was authorised, because that authorisation boundary is invisible to a metric that only evaluates generation quality. Organisations gating production deployment on RAGAS scores are making a category error. The Guan et al. [18] review and the HONEYBEE deployment evidence [16] together establish that this gap is not merely theoretical — it manifests in production healthcare and enterprise RAG systems as a systematic absence of PHI leakage detection and access-control enforcement.

### *6.2 Limitations*

Four constraints apply. First, single higher-education deployment context — CAREF thresholds should be recalibrated for healthcare and financial services deployments where PHI sensitivity profiles and regulatory citation complexity differ. Second, the PHI scanner from [19] generates false positives on certain indirect identifier combinations; the post-CAREF PLR of 0.00 reflects a conservative redaction policy that may over-suppress non-PHI content in edge cases. Third, RCA requires maintaining a regulatory amendment log, which is itself a compliance artifact with associated maintenance costs not analysed here. Fourth, ATC measures logging completeness but not

audit log integrity; tamper-evidence is an important additional property not yet captured by the current metric definition.

## 7. Conclusion

Enterprise RAG systems deployed in regulated sectors face evaluation requirements that RAGAS cannot address. The CAREF framework provides the missing measurement layer: five formally defined dimensions — ACP, PLR, RCA, RARP, ATC — each grounded in a specific regulatory obligation (FERPA, HIPAA, GDPR, EU AI Act) and validated against a production deployment. The r = 0.12 correlation provides empirical foundation for the framework's necessity. The 0.51 to 0.93 composite score improvement demonstrates practical actionability: each CAREF dimension identifies a specific architectural intervention with measurable compliance impact. As RAG adoption in regulated sectors grows — and as agentic RAG architectures multiply the access-boundary crossings per session [7] — compliance-aware evaluation will be a prerequisite for deployment rather than an optional governance layer.

**Conflict of Interest**

The author has no conflict of interest to declare.

**Funding**

The author received no specific funding for this work.

**REFERENCES**

1. Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024).* https://doi.org/10.18653/v1/2024.eacl-demo.16
2. Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of NAACL 2024.* https://doi.org/10.18653/v1/2024.naacl-long.20
3. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of NeurIPS 2021.* https://arxiv.org/abs/2104.08663
4. Vashurin, R. et al. (2025). Benchmarking uncertainty quantification methods for large language models with LM-Polygraph. *Transactions of the Association for Computational Linguistics, 13*, 220–248. https://doi.org/10.1162/tacl_a_00737
5. Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature, 630*, 625–630. https://doi.org/10.1038/s41586-024-07421-0
6. Ozay, A., Jahanbakht, M., Shoomal, A., & Wang, S. (2024). AI-based customer relationship management: A comprehensive bibliometric and systematic literature review. *Enterprise Information Systems, 18*(7). https://doi.org/10.1080/17517575.2024.2351869
7. Singh, A. et al. (2025). Agentic retrieval-augmented generation: A survey on agentic RAG. arXiv:2501.09136. https://arxiv.org/abs/2501.09136
8. Channdeth Sok et al. (2025). MetaRAG: Metamorphic testing for hallucination detection in RAG systems. *CEUR Workshop Proceedings, Vol. 4136.* https://ceur-ws.org/Vol-4136/iaai6.pdf
9. Xu, S., Yan, Z., Dai, C., & Wu, F. (2025). MEGA-RAG: A retrieval-augmented generation framework with multi-evidence guided answer refinement. *Frontiers in Public Health.* https://doi.org/10.3389/fpubh.2025.1635381
10. U.S. Department of Education. (1974, as amended). Family Educational Rights and Privacy Act (FERPA), 34 C.F.R. Part 99. https://studentprivacy.ed.gov/ferpa
11. U.S. Department of Health and Human Services. (1996, as amended). Health Insurance Portability and Accountability Act (HIPAA), 45 C.F.R. § 164.514(b). https://www.hhs.gov/hipaa/index.html
12. European Parliament and Council. (2016). General Data Protection Regulation (GDPR), Regulation (EU) 2016/679. https://gdpr-info.eu/
13. European Union. (2024). Regulation (EU) 2024/1689 — Artificial Intelligence Act. https://artificialintelligenceact.eu/article/12/
14. Luca Nannini et al. (2026). AI agents under EU law: A compliance architecture for AI providers. arXiv:2604.04604. https://arxiv.org/html/2604.04604v1
15. Lewis, P. et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS 2020).* https://arxiv.org/abs/2005.11401
16. Zhong, H., Lentz, M., Narodytska, N., Szekeres, A., & Rong, K. (2025). HONEYBEE: Efficient role-based access control for vector databases via dynamic partitioning. In *Proceedings of SIGMOD 2026.* https://doi.org/10.1145/3786625

17. Keerthana Murugaraj et al. (2025). RAGVue: A diagnostic view for explainable and automated evaluation of retrieval-augmented generation. arXiv:2601.04196. https://arxiv.org/abs/2601.04196
18. Guan, S. et al. (2025). Privacy challenges and solutions in retrieval-augmented generation-enhanced LLMs for healthcare chatbots: A review. arXiv:2511.11347. https://arxiv.org/abs/2511.11347
19. Yuan Zhang. (2025). Privacy protection in RAG: A novel method and evaluation framework. *Information Processing & Management (Elsevier).* https://doi.org/10.1016/S0306-4573(25)00446-7
20. Buscemi, A. et al. (2025). Assessing high-risk AI systems under the EU AI Act: From legal requirements to technical verification. arXiv:2512.13907. https://arxiv.org/abs/2512.13907
21. OWASP GenAI Security Project. (2025). OWASP Top 10 for Agentic Applications 2026. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/
22. Felzmann, H. et al. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics, 26*, 3333–3361. https://doi.org/10.1007/s11948-020-00276-4