

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
المدرسة العليا للإعلام الآلي - 08 ماي 1945 - بسيدي بلعباس
Ecole Supérieure en Informatique
08 Mai 1945 – Sidi Bel Abbès



THESIS

To obtain the Master's degree

Field: **Computer Science**

Specialty: **Artificial Intelligence and Data Science**

Theme

**Generative AI, Privacy and Personal Data:
Mapping and Reducing Data Leakage in
Retrieval-Augmented Generation Systems**

Presented by:

TADJ Mohamed Walid

Defended on: **Thursday, 9 July 2026**

In front of the jury composed of:

Dr. ELHANNANI Souad

President

Dr. BOUDIEB Wassim

Examiner

Dr. CHAIB Souleyman

Supervisor

Pr. AÏMEUR Esma

Co-Supervisor

Academic Year: 2025/2026

ABSTRACT

Enterprise retrieval-augmented generation (RAG) makes a sensitive document corpus conversationally accessible, and in doing so can disclose personal data to users and attackers not authorized to see it. This report establishes the conceptual and methodological foundation for measuring that leakage rather than asserting it. It surveys the field across forty-nine primary sources spanning five areas (RAG-specific vulnerabilities, PII memorization in language models, prompt-injection taxonomies, defense and guardrail architectures, and LLM-as-judge evaluation) and, through a comparative analysis, locates the research gap precisely at their intersection. On that basis it sets out a unified measurement framework: a LINDDUN-aligned metric pair (a binary leakage function L and a severity score S) evaluated through a profile-dependent engine, four CVSS-calibrated attacker profiles, a taxonomy of eight attack families, and a validated language-model judge as the leakage adjudicator. The empirical realization of that framework (a deterministically generated synthetic corpus, the adversarial campaign, the ablation of a layered defense, and the validation of the judge) is carried out in the companion final-year (PFE) report and is referenced here only as context: its headline companion finding is that the defenses achieve a genuine but *partial* reduction (about forty percent, **not** an elimination), with *re-identification by linkage* the dominant unreduced residual. The contribution of *this* report is the survey, the comparative gap analysis, and the design of that framework; any absolute percentages the framework later produces demonstrate a measurement methodology on a controlled corpus, not a safety certification for any particular deployment.

Keywords: Generative AI; Retrieval-Augmented Generation; privacy leakage; re-identification; quasi-identifiers; LLM-as-judge; privacy-utility trade-off; data governance.

RÉSUMÉ

Les systèmes de génération augmentée par récupération (RAG) d'entreprise rendent un corpus documentaire sensible accessible de manière conversationnelle et peuvent, ce faisant, divulguer des données personnelles à des utilisateurs et à des attaquants non autorisés à les consulter. Ce rapport établit le fondement conceptuel et méthodologique permettant de *mesurer* cette fuite plutôt que de l'affirmer. Il dresse un état de l'art de quarante-neuf sources primaires couvrant cinq domaines (les vulnérabilités propres au RAG, la mémorisation des données personnelles par les modèles de langage, les taxonomies d'injection de requête, les architectures de défense et de garde-fous, et l'évaluation par modèle-juge) puis, au moyen d'une étude comparative, situe précisément le verrou scientifique à leur intersection. Sur cette base, il définit un cadre de mesure unifié : une paire de métriques alignée sur LINDDUN (une fonction binaire de fuite L et un score de gravité S) évaluée par un moteur dépendant du profil, quatre profils d'attaquant calibrés selon le CVSS, une taxonomie de huit familles d'attaques, et un modèle-juge validé comme arbitre de la fuite. La réalisation empirique de ce cadre est menée dans le rapport de fin d'études (PFE) qui l'accompagne et n'est rappelée ici qu'à titre de contexte : son résultat principal est une réduction réelle mais *partielle* (environ quarante pour cent, et **non** une élimination), la *ré-identification par recoupement* constituant le résidu dominant non réduit. La contribution propre de *ce* rapport est l'état de l'art, l'analyse comparative des verrous, et la conception de ce cadre ; les pourcentages absolus produits ultérieurement démontrent une méthodologie de mesure sur un corpus contrôlé, et non une certification de sûreté pour un quelconque déploiement.

Mots-clés : IA générative ; génération augmentée par récupération ; fuite de données personnelles ; ré-identification ; quasi-identifiants ; modèle-juge ; compromis sécurité-utilité ; gouvernance des données.