

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



DISSERTATION

To obtain the diploma of **Engineering**

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 thesis measures that leakage rather than asserting it. It formalizes a LINDDUN-aligned metric pair (a binary leakage function L and a severity score S) evaluated through a profile-dependent engine across four CVSS-calibrated attacker profiles and a systematic adversarial campaign of eight attack families, over a deterministically generated corpus of 155 synthetic documents (2,190 entity-level annotations of personally identifiable information, PII; inter-annotator $\kappa = 0.875$), and quantifies a layered defense by ablation. The central finding is a genuine but *partial* reduction: through a validated language-model judge, the full stack lowers measured leakage from 28.8% to 17.3%, about forty percent, **not** an elimination. Redaction defeats direct PII extraction, but *re-identification by linkage*, a person reassembled from individually innocuous attributes scattered across documents, is the dominant unreduced residual, and is shown *intractable at the model layer* across three independent mechanisms: redaction, query-time k -anonymity, and ingestion-time k -anonymity all fail or are foreclosed by their information cost. Linkage is therefore a *data-governance* problem, not a model-layer one: access control *solves* the external, unauthorized attacker, whereas the authorized user who brings outside knowledge leaves a residual that can be *bounded*, minimized and audited, but not eliminated. A methodological contribution underwrites these results: the deterministic detection engine is blind to linkage ($\kappa = -0.10$ against human judgment), while the language-model judge, reliable on undefended output ($\kappa = 0.827$), degrades on the defended distribution ($\kappa = 0.306$), so leakage evaluation must be re-validated on the distribution it scores, not only on undefended text. The absolute percentages demonstrate a measurement methodology on a controlled corpus; they are 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.