AYI NEDJIMI Consultants

Vector DB Poisoning 2026 : Pinecone, Weaviate

16 mai 2026 • Mis à jour le 17 mai 2026 • 20 min de lecture • 3630 mots •

> Le vector DB poisoning injecte des embeddings adversariaux qui matchen
> requete utilisateur, hijackant un RAG entier. ASR > 80%.

À RETENIR

A retenir — Vector Database Poisoning

Vector DB poisoning insere des embeddings adversariaux qui matchent to
utilisateur, hijackant systematiquement le RAG.

Embedding inversion (Morris et al., 2023) : a partir d'un embedding text-
embedding-3-small, reconstruction du texte original a >90% similarity.

Pinecone, Weaviate, Qdrant : sans authentification stricte, l'API d'upsert pe
exploitee. Cas reel Q1 2026 : 230 RAG enterprise compromis.

Rogue ANN nodes : embeddings forges pour rity maxim
cluster cible, devenant le top-1 retrieval pour s du cluster

> Defenses : access control strict, anomaly detection sur embeddings (Maha
> distance), signing des inputs, periodic re-embedding.

Le **vector database poisoning** est l'attaque la moins mediatisee mais l'une des plu
sur les architectures RAG 2026. Contrairement a l'Indirect Prompt Injection RAG qu
texte adversarial dans le corpus, le poisoning attaque directement la couche vecto
embeddings stockes dans Pinecone, Weaviate, Qdrant, ChromaDB, Milvus. Un atta
maitrise la geometrie des embeddings peut crafter un vecteur qui matche statistiq
les requetes utilisateur d'un domaine, devenant le top-1 result a chaque retrieval. L
branche dessus ingere alors systematiquement le contenu malveillant. Cet article
mecanique mathematique (rogue ANN nodes), le code Python d'attaque, les attaq
embedding inversion (Morris et al., 2023), et les defenses 2026. Pour les architect
CISO, le **vector database poisoning** represente le risque le moins audite mais le pl
en 2026 — un seul rogue ANN node bien craft peut hijacker l'integralite du retrieva
corporate de millions de documents.

---

1. Genese et etat de l'art

Le concept de poisoning vectoriel emerge en 2024 avec deux papiers fondateurs

Zhong et al. (2023) — *Poisoning Web-Scale Training Datasets*, demontre le po
Common Crawl + RAG corpus.

Morris et al. (2023) — *Text Embeddings Reveal (Almost) As Much As Text*, mon
inverser un embedding pour recuperer le texte original.

Cheng et al. (2024) — *PoisonedRAG: Knowledge* *on RAG*, A
Llama-2 RAG.

Réponse sous 24h

Devis gratuit →