

Exercice Scientifique et Technique

# RAG : Retrieval-Augmented Generation

Yoann HOERTNER



Une école de



# Sommaire

1

## Le problème des LLMs

Données figées · privées · hallucinations

2

## Qu'est-ce que le RAG ?

Pipeline en 4 étapes · RAG vs fine-tuning

3

## Retrieval : Chunking & Embeddings

Découpage · représentation vectorielle · cosinus

4

## Bases de données vectorielles

HNSW · Pinecone · pgvector · comparatif

5

## Augmentation & Génération

Prompt enrichi · fenêtre de contexte · grounding

6

## Limites & Techniques avancées

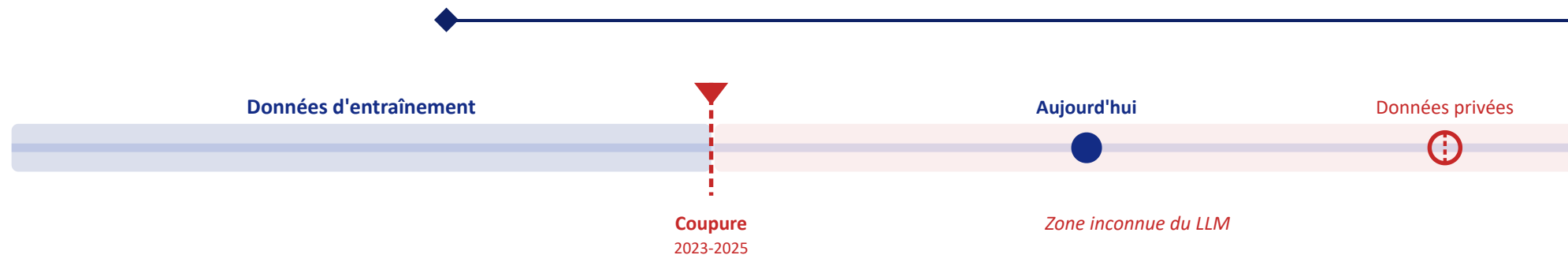
Reranking · HyDE · Agentic RAG · Self-RAG

7

## Conclusion & Références

Bilan · liens cours · bibliographie

# Le problème des LLMs



## Connaissance figée

Date de coupure d'entraînement variable selon les modèles (GPT-4o : avril 2024, Claude 3.7 : début 2025). Tout ce qui est postérieur est totalement inconnu du modèle.

## Données privées

Documents internes, emails, réunions, bases de connaissances d'entreprise : le LLM n'a jamais vu ces données lors de l'entraînement.

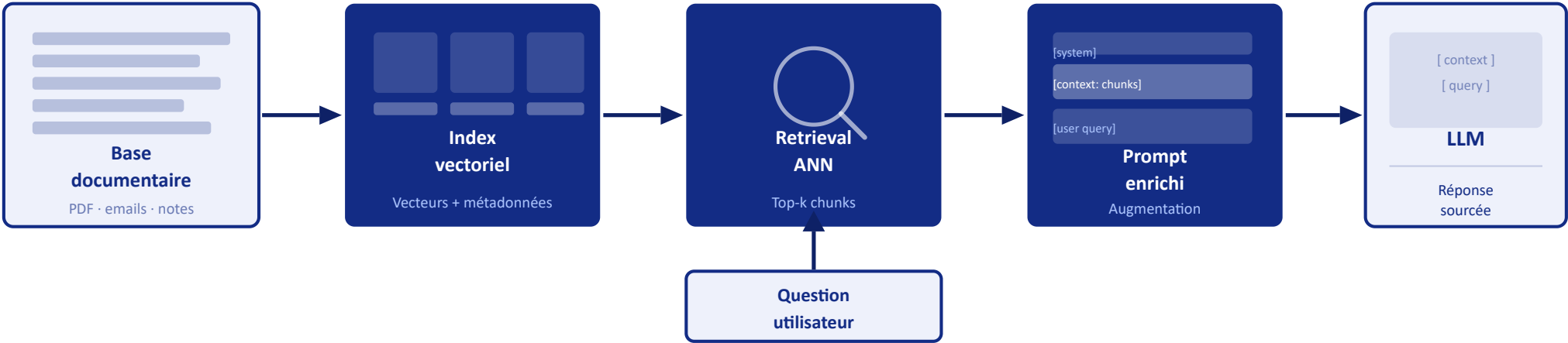
## Hallucinations

Sur des sujets non couverts, le modèle génère une réponse plausible mais incorrecte : sans avertissement ni indicateur de confiance.

**Question :** Comment connecter un LLM à des données récentes et privées sans le ré-entraîner entièrement ?

# Qu'est-ce que le RAG ?

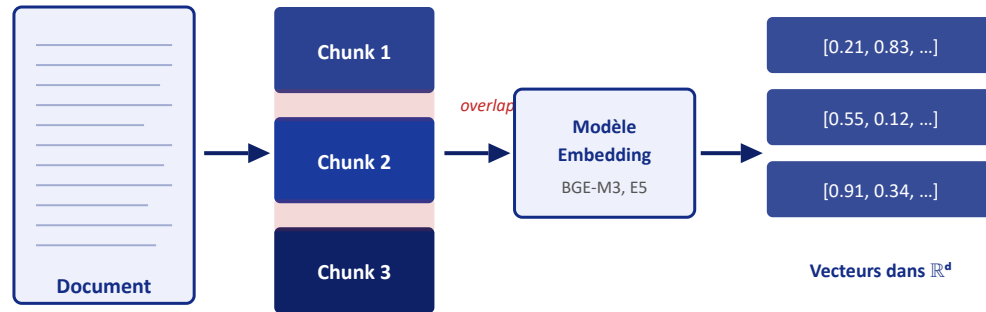
Lewis et al., Facebook AI Research : NeurIPS 2020



| Critère     | RAG                              | Fine-tuning                       |
|-------------|----------------------------------|-----------------------------------|
| Mise à jour | Temps réel, sans ré-entraînement | Ré-entraînement complet requis    |
| Coût        | Faible : indexation seulement    | Très élevé : GPU, temps de calcul |
| Traçabilité | Sources directement citables     | Opaque, connaissance mémorisée    |

# Retrieval : Chunking & Embeddings

## ① Chunking : découpe du document

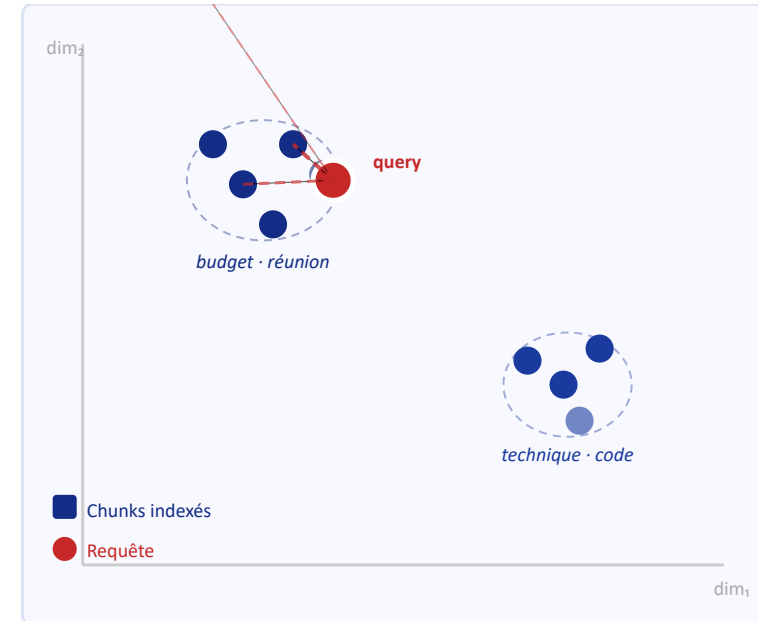


## ③ Similarité cosinus

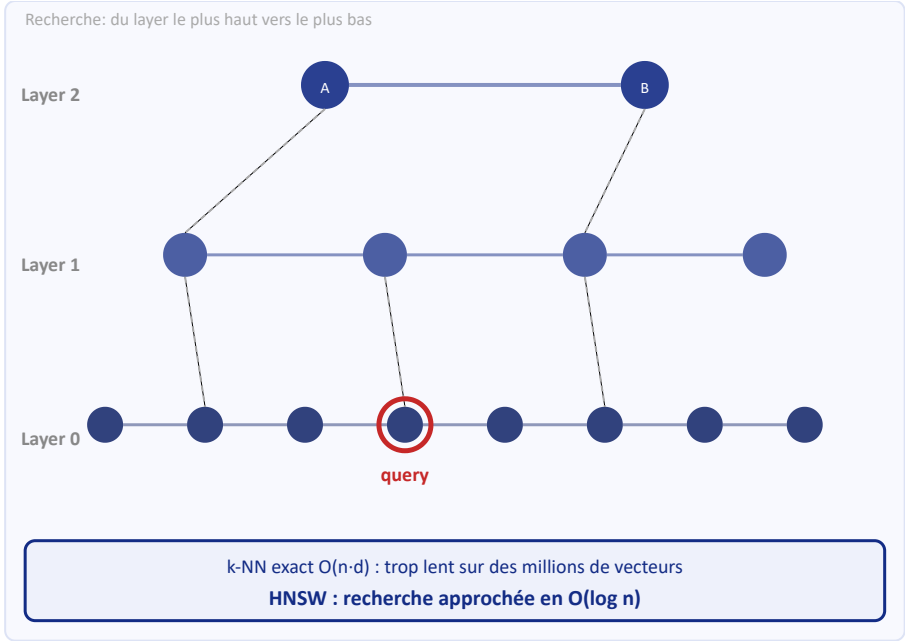
$$\cos(q, c) = (\mathbf{q} \cdot \mathbf{c}) / (\|\mathbf{q}\| \cdot \|\mathbf{c}\|)$$

Retourne les  $k$  chunks les plus proches sémantiquement : invariant à la magnitude.

## ② Espace vectoriel : similarité sémantique



## HNSW : Hierarchical Navigable Small World



Ref. Malkov & Yashunin, *IEEE TPAMI* 2020 · Algorithmique BDD

## Comparatif des solutions

| Solution | Type            | Avantage principal       |
|----------|-----------------|--------------------------|
| Pinecone | SaaS managé     | Déploiement immédiat     |
| Weaviate | Open-source     | Hybride vectoriel + BM25 |
| Qdrant   | Open-source     | Très performant (Rust)   |
| Chroma   | Open-source     | Prototypage rapide       |
| pgvector | Ext. PostgreSQL | Si déjà sous PostgreSQL  |

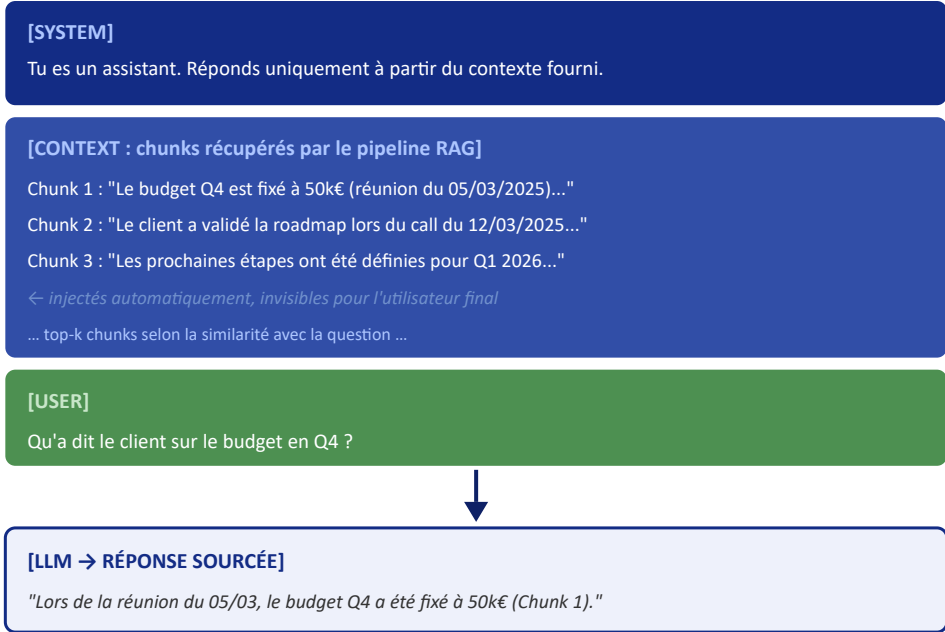
### Critères de choix

Volume de données · latence requise · infrastructure existante · besoin de recherche hybride (mots-clés + vecteurs)

### Cas d'usage typique

Base documentaire d'entreprise, FAQ, emails, notes : interrogeables en langage naturel.

## Anatomie du prompt enrichi



## Fenêtre de contexte



## Grounding du LLM

Le modèle est contraint à répondre depuis le contexte fourni : réduit les hallucinations et permet la traçabilité des sources.

# Limites & Techniques avancées

## Limites du RAG standard

### Retrieval défaillant

Mauvais chunking ou embeddings insuffisants : des chunks non pertinents remontent et polluent la réponse.

### Hallucination persistante

Le LLM peut ignorer ou contredire le contexte fourni en s'appuyant sur sa connaissance interne.

### Coût & latence en production

Embedding + recherche ANN + appel LLM : 3 étapes supplémentaires par requête, critiques en production.

## Solutions avancées

### Reranking

Un Cross-Encoder reclasse les chunks par pertinence fine, après la recherche approximative initiale.

### HyDE (Gao et al., 2022)

Génère un document hypothétique en réponse à la question, puis l'embedde pour améliorer le rappel.

### Agentic & Self-RAG (Asai et al., 2023)

L'agent itère (cherche → raisonne → re-cherche) et s'auto-évalue sur la qualité de sa réponse.



## Conclusion & Références

### Flexible

Modèle inchangé : la base documentaire se met à jour sans ré-entraînement

### Économique

Coût d'indexation seulement : pas de GPU, pas de fine-tuning

### Traçable

Chaque réponse cite ses sources : vérifiable et auditable

### Privé

Données internes jamais exposées lors de l'entraînement du modèle

### Liens avec les cours

NLP

Algèbre linéaire

Deep Learning

Algorithmique

BDD

Systèmes distribués

*"RAG + agents + mémoire long-terme = prochaine frontière de l'IA en production"*

### Références essentielles

Lewis et al. (2020). *RAG for Knowledge-Intensive NLP Tasks*. NeurIPS.

Malkov & Yashunin (2020). *Efficient ANN Search Using HNSW*. IEEE TPAMI.

Gao et al. (2022). *Precise Zero-Shot Dense Retrieval (HyDE)*. ACL.

Asai et al. (2023). *Self-RAG: Learning to Retrieve, Generate, and Critique*. NeurIPS.