

La RAG

la couche d'accès à vos connaissances

PANORAMA 2026 : RAG AGENTIQUE, OFFRES MANAGÉES, SOUVERAINETÉ

En 2026, la RAG n'est plus un simple pipeline « chercher puis répondre » : c'est la façon dont les assistants et les agents vont chercher vos connaissances. Le vrai choix porte moins sur l'outil que sur l'endroit où vivent vos données et sur qui les contrôle. Ce panorama recense les offres clés en main, les frameworks, les bases vectorielles et les briques (embeddings, reranking, parsing, graphes), avec leurs prix relevés en septembre 2026, et ce qui distingue une RAG solide d'une démo.

- RAG AGENTIQUE
- RECHERCHE HYBRIDE
- OFFRES MANAGÉES
- SOUVERAINETÉ
- ÉVALUATION

» REPÈRES CHIFFRÉS

51 % DES DÉPLOIEMENTS D'IA GÉNÉRATIVE UTILISENT LA RAG Contre 31 % un an plus tôt ; le fine-tuning ne concerne que 9 % des modèles en production (enquête Menlo Ventures auprès d'entreprises, novembre 2024).	-67 % D'ÉCHECS DE RÉCUPÉRATION Embeddings contextuels + BM25 + reranking, sur les 20 premiers passages ; 35 % avec les seuls embeddings contextuels (Anthropic, septembre 2024).
1 M DE TOKENS DE CONTEXTE Par défaut sur les derniers modèles Claude et Gemini. Mais les performances se dégradent avec la longueur : 18 modèles testés (context rot, Chroma, juillet 2025).	0,15 \$ PAR MILLION DE TOKENS INDEXÉS Tarif d'indexation de Gemini File Search ; côté OpenAI, 2,50 \$ pour 1 000 recherches et 0,10 \$/Go/jour au-delà du premier gigaoctet gratuit (septembre 2026).

» SOMMAIRE

- 01

La RAG en quatre étapes
- 02

De la RAG naïve à la RAG agentique
- 03

Les outils qui comptent en 2026
- 04

RAG managée ou RAG maison ?
- 05

À quoi ça sert, concrètement
- 06

Ce qui distingue une RAG solide d'une démo
- 07

Évaluer et surveiller
- 08

La RAG comme outil de gestion du contexte
- 09

Ce qui a été dépassé

— LE PRINCIPE / 01

La RAG en quatre étapes

Un modèle de langage ne connaît que ce qu'il a vu pendant son entraînement. La RAG lui donne accès à vos documents au moment de la question, et permet de vérifier d'où vient la réponse.

01

Un modèle de langage ne connaît que ce qu'il a vu pendant **son entraînement** : des données figées, souvent générales, parfois anciennes. Interrogé au-delà, il peut inventer une réponse plausible. La RAG (retrieval-augmented generation, « **génération augmentée par récupération** ») consiste à **aller chercher les bons passages dans vos documents au moment de la question**, puis à les donner au modèle comme contexte. La réponse est alors ancrée dans des sources réelles, et l'on peut remonter à chacune.

01

Les quatre étapes

- **Indexer** : les documents sont découpés en passages, puis convertis en vecteurs (embeddings) rangés dans un index.
- **Chercher** : la question est convertie de la même façon, et l'on récupère les passages les plus proches — idéalement en combinant vecteurs et mots-clés.
- **Augmenter** : les meilleurs passages sont insérés dans le prompt, avec leur référence.
- **Générer** : le modèle rédige la réponse à partir de ces passages, et cite ses sources.

02

Pourquoi c'est intéressant

- **Pas de réentraînement** : ajouter ou retirer un document met la connaissance à jour immédiatement.
- **Traçabilité** : chaque affirmation peut être rattachée à un passage, donc vérifiée.
- **Contrôle d'accès** : on peut filtrer les documents selon les droits de la personne qui interroge.
- **Réduction des inventions**, mais pas suppression : un système mal réglé cite mal, ou ignore les passages fournis.

» ENCADRÉ : D'OÙ VIENT LE MOT « RAG »

01

HISTORIQUE

2020 · L'article fondateur

Le terme vient d'un article de **Facebook AI présenté à NeurIPS 2020** (Lewis et al.), qui couplait un modèle de génération à un index vectoriel de Wikipédia. L'implémentation « Hugging Face RAG » (DPR + BART) qui en découle est ce modèle de recherche d'origine : elle a une valeur historique, mais ce n'est plus un outil de prototypage en 2026.

02

HISTORIQUE

2021 · La première mesure de l'effet

L'année suivante, toujours chez Facebook AI, **Shuster et al.** montrent que l'augmentation par récupération réduit les hallucinations en conversation (arXiv 2104.07567). C'est l'argument qui a porté la RAG dans l'industrie, et il reste valable : réduire, pas supprimer.

— CE QUI A CHANGÉ / 02

De la RAG naïve à la RAG agentique

Depuis 2024, quatre déplacements : l'agent planifie ses recherches, le deep research industrialise l'enquête, les fenêtres de contexte atteignent 1 million de tokens sans régler le problème, et MCP permet d'interroger les systèmes existants plutôt que de tout recopier.

02

La RAG est devenue la **couche d'accès aux connaissances des agents**. Dans la version classique, on cherche une fois et on répond ; désormais un modèle peut planifier ses recherches, reformuler, interroger plusieurs sources et relancer s'il juge les résultats insuffisants. Attention à ne pas surestimer la bascule : selon l'enquête Menlo Ventures

de décembre 2025, **seuls 16 % des déploiements en entreprise** sont de « vrais » agents, et la plupart des systèmes en production restent simples.

01

FORTE CROISSANCE

RAG agentique : chercher, vérifier, recommencer

- Le modèle **planifie ses recherches**, reformule, interroge documents, web et bases de données, juge si les résultats suffisent et relance si besoin.
- Les grandes plateformes l'ont intégrée sous le nom d'agentic retrieval : Foundry IQ chez Azure, Bedrock Managed Knowledge Base chez AWS.
- Plus lente et plus coûteuse : à réserver aux questions complexes ou multi-étapes.

02

NOUVEAU STANDARD

Deep research : la RAG à l'échelle d'un rapport

- Les modes « recherche approfondie » (Gemini, 11 décembre 2024 ; ChatGPT, 3 février 2025 ; Claude Research, 15 avril 2025 ; Le Chat de Mistral, 17 juillet 2025) sont des agents RAG qui enchaînent des dizaines de recherches pendant 5 à 30 minutes et rendent un rapport sourcé.
- Le coût est réel : chez Anthropic, un système multi-agents consomme environ **15 fois plus de tokens** qu'une conversation, pour un gain de 90 % sur leur évaluation interne.
- Les sources citées restent à vérifier : ces outils peuvent reprendre des rumeurs.

03

À NUANCER

Long contexte ou RAG ? Les deux

- Avec 1 million de tokens (de l'ordre de 2 000 pages), on peut parfois tout mettre dans le prompt. Anthropic le conseille sous ~200 000 tokens (environ 500 pages), avec la mise en cache du prompt.
- Mais les modèles n'exploitent pas tout le contexte de la même façon : l'information au milieu est moins bien retrouvée (« Lost in the Middle », 2023), et les performances baissent avec la longueur, y compris sur des tâches simples (« context rot », Chroma, juillet 2025, 18 modèles testés).
- Au-delà de quelques centaines de pages, ou quand les documents changent souvent, **la récupération reste nécessaire**.

04

STANDARD OUVERT

MCP : la prise universelle vers les données

- Plutôt que de tout copier dans une base vectorielle, on peut laisser l'assistant **interroger les systèmes existants** (GED, CRM, SharePoint, bases SQL) via le Model Context Protocol.
- Lancé par Anthropic fin 2024, MCP a été confié le 9 décembre 2025 à l'Agentic AI Foundation (Linux Foundation), dont AWS, Google, Microsoft et OpenAI sont membres. La spécification courante est datée du 28 juillet 2026.
- Azure expose déjà ses bases de connaissances via MCP. Chaque serveur MCP est aussi une entrée à contrôler.

POUR ALLER PLUS LOIN **Agents IA : de l'expérimentation à l'industrialisation** ›

Les outils qui comptent en 2026

Cinq familles, de l'offre « RAG en une API » à la brique d'évaluation. Prix et versions relevés le 18 septembre 2026 : ils changent presque chaque mois.

» CLÉ EN MAIN (RAG EN UNE API) — 07 OUTILS

OpenAI File search

OpenAI (US) · fermé, managé

Le plus simple si l'on est déjà chez OpenAI : on dépose des fichiers, la plateforme gère découpage, embeddings, index et citations, en combinant recherche sémantique et mots-clés. Attention : l'Assistants API a fermé le 26 août 2026, les vector stores passent désormais par la Responses API.

0,10 \$/Go/jour (1 Go gratuit) + 2,50 \$ / 1 000 appels

Gemini API File Search

Google (US) · fermé, managé

Annoncé le 6 novembre 2025 : stockage et requêtes sans surcoût, seule l'indexation est facturée, et les citations sont renvoyées automatiquement. Le tarif d'indexation en fait l'offre la moins chère à l'entrée.

Stockage gratuit ; 0,15 \$/M de tokens à l'indexation

Amazon Bedrock Managed Knowledge Base

AWS (US) · fermé, managé

RAG entièrement managée, GA depuis le 17 juin 2026 : six connecteurs natifs (S3, SharePoint, Confluence, Google Drive, OneDrive, Web Crawler), recherche hybride, reranking, agentic retrieval, multimodal, régions UE (Dublin, Francfort, Londres). Depuis septembre 2026, connecteur ServiceNow et synchronisation planifiée.

À l'usage (non détaillé dans l'annonce)

Azure AI Search / Foundry IQ

Microsoft (US) · fermé, managé

« Cognitive Search » est l'ancien nom. Foundry IQ apporte des bases de connaissances pour agents avec agentic retrieval (planification de requêtes), exposées en API ou via MCP. GA dans l'API REST 2026-04-01 ; certaines fonctions, dont la synthèse de réponse, restent en préversion.

À l'usage (non révisé)

RAG Engine / Agent Search

Google Cloud (US) · fermé, managé

L'ex-Vertex AI Search, intégré à la plateforme renommée « Gemini Enterprise Agent Platform » le 22 avril 2026. Pertinent pour une organisation déjà installée sur Google Cloud.

À l'usage (non révisé)

Le Chat Enterprise / Document Library

Mistral AI (France) · fermé, auto-hébergeable

Bibliothèques de documents (bêta) et connecteurs Drive / SharePoint, avec un déploiement possible sur site ou en cloud privé. L'option européenne la plus directe pour une organisation qui veut une offre packagée sans quitter l'UE.

Sur devis (non révisé)

Albert API (OpenGateLLM)

DINUM, État français · open source

Pile souveraine du secteur public : RAG, OCR et vectorisation, hébergement en France, code ouvert. Utilisée par plus de 70 projets publics. Réutilisable en dehors de l'État en tant que logiciel libre.

Réservé aux administrations

» FRAMEWORKS — 04 OUTILS

LangChain / LangGraph

LangChain (US) · open source (MIT)

L'écosystème le plus large, et LangGraph pour orchestrer une RAG agentique. Version 1.4.1 le 16 septembre 2026 : les anciennes chaînes (RetrievalQA) ont migré vers le paquet de compatibilité langchain-classic, les loaders et vectorstores vers langchain_community et les intégrations dédiées. Les tutoriels de la documentation officielle sont à jour, les copier-coller de 2023 ne le sont plus.

Gratuit ; LangSmith payant

LlamaIndex (+ LlamaParse)

LlamaIndex (US) · open source (MIT) + managé

Le spécialiste des « agents sur vos documents », avec LlamaParse pour les PDF complexes. Version 0.14.24 le 19 août 2026 : ServiceContext a disparu depuis la v0.10 au profit de Settings, et les imports passent par llama_index.core. La documentation a déménagé sur developers.llamaindex.ai.

Framework gratuit ; LlamaCloud payant

Haystack 3

deepset (Allemagne) · open source (Apache 2.0)

Acteur européen orienté pipelines de production : composants explicites, exécution asynchrone et agents depuis la v3.0 (20 juillet 2026, v3.1.1 en septembre). Son tutoriel « first RAG pipeline » reste une bonne première implémentation.

Gratuit ; offre entreprise

RAGFlow

InfiniFlow · open source (Apache 2.0)

Application RAG complète avec interface web, parsing de documents et gestion des bases de connaissances : plutôt qu'une bibliothèque à assembler, un produit à déployer (~ 91 000 étoiles GitHub).

Gratuit (auto-hébergé)

» BASES VECTORIELLES — 06 OUTILS

pgvector

Communauté PostgreSQL · open source

Des vecteurs dans le PostgreSQL que vous avez déjà : index HNSW, types halfvec et sparsevec (v0.8.6). Pas de nouvelle base à exploiter, pas de synchronisation à maintenir, et les filtres SQL habituels pour les métadonnées. Le choix par défaut raisonnable pour la plupart des projets.

Gratuit

Qdrant

Qdrant (Allemagne) · open source (Apache 2.0) + cloud

Performant, européen, avec des options hybrid cloud et private cloud pour garder les données chez soi tout en confiant l'exploitation à l'éditeur. Recherche hybride et filtres riches (v1.19.1).

Gratuit (1 Go) puis à l'usage

Weaviate

Weaviate (Pays-Bas) · open source + cloud

Recherche hybride native, modules d'embeddings et briques agentiques intégrés. Européen également, avec une offre managée à l'entrée abordable (v1.39.5).

Gratuit ; offre Flex dès 45 \$/mois

Milvus / Zilliz

Zilliz · open source (Apache 2.0) + cloud

Conçu pour les très grands volumes : la v3.0 (29 juillet 2026) apporte notamment un BM25 compressé pour la recherche hybride à grande échelle.

Gratuit ; cloud à l'usage

Pinecone

Pinecone (US) · fermé, managé

Serverless et zéro exploitation : historiquement la référence du « je ne veux pas gérer d'index ». Option BYOC (déploiement dans votre compte cloud) sur l'offre Enterprise.

Gratuit ; Builder 20 \$, Standard 50 \$ minimum, Enterprise 500 \$ minimum par mois

Elasticsearch / OpenSearch

Elastic · source disponible / cloud

Recherche plein texte et vectorielle dans un outil souvent déjà en place : BM25 et vecteurs au même endroit, donc recherche hybride sans nouvelle brique. Elastic a intégré les modèles Jina AI (Jina a rejoint Elastic en octobre 2025). ChromaDB et FAISS restent par ailleurs très pratiques pour prototyper en local.

Selon offre (non revérifié)

» BRIQUES : PARSING, EMBEDDINGS, RERANKING, GRAPHS — 06 OUTILS

Docling

IBM / LF AI · open source (MIT)

Conversion de PDF, DOCX et présentations en texte structuré, avec les tableaux et la mise en page : la brique de parsing la plus suivie (~ 66 600 étoiles, v2.128.0 le 16 septembre 2026). À comparer à PyMuPDF ou PDFPlumber pour les documents simples, et à LlamaParse pour les PDF les plus retors.

Gratuit

ColPali / ColQwen

Illuin Technology, communauté ViDoRe · open source (MIT)

Plutôt que d'extraire le texte, on indexe directement l'image de chaque page avec un modèle vision-langage et l'« interaction tardive » (plusieurs vecteurs par page). Décisif pour les PDF riches en tableaux, schémas et scans. ColPali (ICLR 2025) a été conçu en France ; le benchmark ViDoRe V3 couvre 26 000 pages en 6 langues, dont le français.

Gratuit

Qwen3-Embedding / Qwen3-Reranker

Alibaba · poids ouverts (Apache 2.0)

Famille d'embeddings et de rerankers multilingues de 0,6 à 8 milliards de paramètres, parmi les modèles de tête du classement MTEB. La version 0,6B suffit très souvent, et tourne sur un GPU modeste : la meilleure porte d'entrée vers une pile entièrement auto-hébergée. Alternatives ouvertes : BGE-M3 (MIT), EmbeddingGemma-300m, voyage-4-nano (Apache 2.0).

Gratuit (auto-hébergé)

Gemini Embedding 2 / Voyage 4

Google, Voyage AI (MongoDB) · fermé

Les nouveautés 2026 côté modèles fermés : Gemini Embedding 2 (annoncé le 10 mars 2026) est multimodal (texte, image, vidéo, audio, PDF) ; Voyage 4 (15 janvier 2026) partage un même espace d'embedding entre ses tailles, ce qui permet d'indexer avec le gros modèle et d'interroger avec le petit. Cohere Embed v4 et OpenAI text-embedding-3 restent des valeurs sûres (3-small à 0,02 \$/M de tokens, 3-large à 0,13 \$/M).

À l'usage ; Voyage : 200 M de tokens gratuits

Cohere Rerank 4.0

Cohere (Canada) · fermé

Le reranker multilingue de référence, en deux variantes (rerank-v4.0-pro et -fast). Un reranker relit les 20 à 100 passages candidats et les reclasse finement : c'est le meilleur rapport gain / effort de tout le pipeline. Alternatives : Qwen3-Reranker (poids ouverts), rerankers Voyage et Jina.

À l'usage (non revérifié)

GraphRAG / LightRAG

Microsoft Research, Université de Hong Kong · open source (MIT)

Le découpage en passages perd les liens entre entités (qui dirige quoi, quel contrat cite quelle norme). GraphRAG (v3.1.2 en août 2026) extrait un graphe de connaissances et des résumés par « communautés » pour répondre aux questions globales (« quels sont les grands thèmes de ce corpus ? »). Son indexation coûte cher : la variante LazyGraphRAG la ramène à 0,1 % de ce coût, au niveau d'une RAG vectorielle. LightRAG (~ 40 000 étoiles) propose une version plus légère et incrémentale.

Gratuit (coût LLM à l'indexation)

» ÉVALUATION — 04 OUTILS

Ragas[Vibrant Labs](#) · [open source](#)

La bibliothèque de référence pour mesurer une RAG : fidélité de la réponse aux passages, pertinence des passages, rappel du contexte. Le dépôt a migré vers [vibrantlabsai/ragas](#) ; dernière version 0.4.3 (13 janvier 2026), à un rythme désormais ralenti.

Gratuit**DeepEval**[Confident AI](#) · [open source](#)

Cadre d'évaluation façon tests unitaires (v4.2.3) : on écrit des cas de test sur ses propres questions et l'on fait échouer la CI quand une métrique se dégrade. Le complément naturel de Ragas quand l'évaluation doit entrer dans le pipeline d'intégration.

Gratuit**Évaluateurs des plateformes**[LangSmith](#), [Phoenix](#), [MLflow](#) · [mixte](#)

Les plateformes d'observabilité intègrent leurs propres évaluateurs et la trace complète des requêtes (passages récupérés, prompt final, réponse). Utile en production, où le problème n'est plus de mesurer une fois mais de détecter une dérive. Fonctionnalités non revérifiées individuellement en septembre 2026.

Selon offre**Un LLM juge, sur vos questions**[Méthode](#) · [gratuit](#)

La brique la plus importante n'est pas un outil : c'est un jeu de 50 à 200 questions issues de vos vrais utilisateurs, avec les réponses attendues. Un modèle récent tient le rôle de juge pour noter fidélité et pertinence. Sans cela, aucune des métriques ci-dessus n'est interprétable.

Coût des appels au modèle

04

— LE VIS-À-VIS / 04

RAG managée ou RAG maison ?

Le vrai arbitrage n'est pas technique : il porte sur le délai, le contrôle et l'endroit où vivent vos documents.



La RAG managée : déposer des fichiers et avancer

En production en quelques jours

Découpage, embeddings, index, reranking et citations sont fournis. Aucune base vectorielle à exploiter, aucun modèle d'embedding à choisir.

Un coût d'entrée très bas

0,15 \$/M de tokens à l'indexation chez Gemini File Search, 1 Go de stockage gratuit chez OpenAI : une preuve de concept coûte quelques euros.

Des connecteurs déjà écrits

SharePoint, Confluence, Google Drive, OneDrive, ServiceNow, web crawler : côté Bedrock, la synchronisation planifiée est fournie.

Les fonctions avancées incluses

Recherche hybride, reranking et agentic retrieval arrivent par mise à jour de la plateforme, sans travail d'intégration.

Une exploitation qui n'est pas la vôtre

Disponibilité, montée en charge et mises à jour de modèles sont assurées par le fournisseur.



La RAG maison : garder la main sur ses données

Vos documents restent chez vous

Une offre managée place vos documents chez le fournisseur. Une région UE ne suffit pas toujours : le Cloud Act s'applique aux entreprises américaines, même pour des données hébergées en Europe.

Une pile entièrement ouverte existe

Modèle ouvert (Mistral, Qwen, Llama, Gemma), embeddings ouverts (Qwen3-Embedding, BGE-M3), base auto-hébergée (pgvector, Qdrant, Milvus, Weaviate) et framework open source : rien d'indispensable ne manque.

Le contrôle fin du pipeline

Découpage adapté à vos documents, filtres de métadonnées liés à vos droits d'accès, reranker choisi, graphe si vos questions sont relationnelles : autant de réglages qu'une offre packagée ne propose pas.

Pas d'enfermement

Changer de modèle ou de base se fait sans réécrire l'application ; les documents et l'index restent des actifs internes.

Mais un vrai coût d'ingénierie

Il faut du temps de développement, de l'exploitation, un jeu d'évaluation et une personne qui maintient tout cela dans le temps. C'est le prix du contrôle.

» À RETENIR

Cinq questions à se poser avant de choisir

- 01 Où vivent vos documents, et le fournisseur envisagé a-t-il le droit d'y accéder ?
- 02 Quelle est la sensibilité du corpus : public, interne, données personnelles, secret industriel ?
- 03 Quelle fraîcheur exige l'usage : indexation mensuelle, quotidienne, ou requête directe dans le système source via MCP ?
- 04 Combien de questions par jour, et pour combien d'utilisateurs ? En dessous de quelques milliers, l'offre managée est presque toujours moins chère qu'une pile maison.
- 05 Qui maintiendra le système dans deux ans, et avec quel jeu d'évaluation pour vérifier qu'il ne s'est pas dégradé ?

À quoi ça sert, concrètement

Cinq familles d'usages où la RAG a fait ses preuves, et un démonstrateur du CATIE sur un corpus réglementaire.

05

01

Support et service client
Un assistant branché sur la documentation produit, les FAQ et l'historique des tickets répond aux questions de premier niveau en citant la page exacte. **Impact › Temps de résolution réduit, et réponses cohérentes avec la documentation officielle plutôt qu'avec l'habitude de chacun**

02

Gestion des connaissances internes
Interroger en langage naturel des années de comptes rendus, de notes techniques et de procédures, sans savoir dans quel dossier chercher. **Impact › Capitalisation réelle des connaissances, y compris celles des personnes parties, avec un contrôle d'accès aligné sur les droits existants**

03

Corpus réglementaire et normatif
Code, normes, conventions collectives, référentiels qualité : des textes stables, structurés, découpés en articles, et dont la citation exacte fait foi. **Impact › Le terrain le plus favorable à la RAG : la réponse doit pointer l'article, et la traçabilité est l'exigence principale**

04

Veille et analyse documentaire
Rapports annuels, publications scientifiques, données de marché : les modes deep research enchaînent les recherches et rendent une synthèse sourcée. **Impact › Des heures de lecture économisées, à condition de vérifier les sources citées avant d'en faire un livrable**

05

Formation et accompagnement
Un tuteur adossé aux supports de cours d'un établissement, ou un assistant d'intégration adossé aux procédures internes d'une entreprise. **Impact › Des explications ancrées dans les supports réels de l'organisation, et non dans la culture générale du modèle**

» ENCART : CE QUE DISENT LES CHIFFRES D'ADOPTION

01

ENQUÊTE

La RAG, technique majoritaire
Fin 2024, la RAG était l'architecture de **51 %** des déploiements d'IA générative en entreprise, contre 31 % un an plus tôt, loin devant le fine-tuning (9 % des modèles en production). Fin 2025, elle reste la deuxième technique de personnalisation après le prompt. Source : enquêtes annuelles de Menlo Ventures, un fonds d'investissement — à lire comme telles.

02

À NUANCER

Les agents, encore minoritaires
Dans la même enquête de décembre 2025, **16 %** seulement des déploiements en entreprise sont de « vrais » agents. La RAG agentique est un pattern en forte croissance, devenu la norme des assistants de recherche et des offres managées, mais la majorité des systèmes en production reste simple. Commencer simple n'est donc pas un retard.

— POUR LES ÉQUIPES TECHNIQUES / 06

Ce qui distingue une RAG solide d'une démo

Le pipeline de 2023 (découper, vectoriser, top-k) fonctionne en démonstration et déçoit en production. Six points font la différence.

Un prototype se monte en une journée ; c'est la qualité de la récupération qui décide du reste. Chez Anthropic, passer d'embeddings simples à **embeddings contextuels + BM25 + reranking** réduit de **67 %** les échecs de récupération sur les 20 premiers passages, contre 35 % avec les seuls embeddings contextuels. Aucun changement de modèle de génération n'a cet effet.

01

AMONT

1 · Parsing : ce qui entre détermine tout

- Un PDF mal extrait (tableaux aplatis, colonnes mélangées, en-têtes répétés) condamne tout le reste.
- Docling, LlamaParse et les parseurs OCR des éditeurs gèrent la structure ; PyMuPDF ou PDFPlumber suffisent pour des textes simples.
- Ou bien **ne pas parser du tout** : indexer l'image des pages avec ColPali / ColQwen pour les documents visuels.

02

AMONT

2 · Découpage : ni trop gros, ni orphelin

- Découper sur la structure du document (titres, sections, articles) plutôt qu'à la longueur fixe.
- Un recouvrement entre passages évite de couper une idée en deux.
- **Embeddings contextuels** : préfixer chaque passage d'une ou deux phrases situant sa place dans le document, ce qui à lui seul réduit d'un tiers les échecs de récupération.

03

INDISPENSABLE

3 · Recherche hybride : le minimum

- Les vecteurs captent le sens, BM25 capte les termes exacts : références, codes produit, noms propres, numéros d'article.
- Combiner les deux n'est plus une option avancée, c'est **le socle de 2026**.
- La plupart des bases (Weaviate, Qdrant, Milvus, Elasticsearch, pgvector) le font nativement.

04

INDISPENSABLE

4 · Reranking : le meilleur rapport gain/effort

- Récupérer large (20 à 100 passages), puis faire relire et reclasser par un reranker, et ne garder que les meilleurs.
- Rerankers actuels : **Cohere Rerank 4.0** (pro / fast, multilingue), **Qwen3-Reranker** (poids ouverts, Apache 2.0), rerankers Voyage et Jina.
- Quelques lignes de code, un gain souvent spectaculaire sur la précision.

05

PRODUCTION

5 · Métadonnées et droits d'accès

- Filtrer par date, service, type de document, version : indispensable dès que le corpus contient plusieurs millésimes d'une même procédure.
- Le filtre de droits d'accès doit s'appliquer **à la recherche**, pas à l'affichage : un passage récupéré est un passage divulgué.

06

SELON LE CAS

6 · Graphes, si vos questions sont relationnelles

- « Quels sont les grands thèmes de ce corpus ? », « quel contrat cite quelle norme ? » : ces questions globales échappent au top-k.
- GraphRAG extrait un graphe et des résumés par communautés ; LazyGraphRAG ramène le coût d'indexation à 0,1 % de la version complète ; LightRAG est plus léger et incrémental.
- À n'engager que si le besoin est démontré : l'indexation coûte cher en appels de modèle.

Choisir un modèle d'embedding. Le classement MTEB (plus de 100 tâches, multilingue) est utile pour dégrossir, et ses modèles de tête sont de gros modèles ouverts ou propriétaires. Il ne remplace pas un test sur votre propre corpus : un modèle de 0,6 milliard de paramètres bien choisi (Qwen3-Embedding-0.6B, BGE-M3) suffit très souvent, et se garde en interne. Deux critères l'emportent sur le rang : le multilingue réel sur vos langues, et la taille de vecteur que votre base peut absorber. Les nouveautés 2026 sont multimodales (Gemini Embedding 2) ou modulaires (Voyage 4, dont les tailles partagent un même espace d'embedding : indexer avec le gros modèle, interroger avec le petit).

Choisir un modèle de génération. Pas de nom de modèle ici : ils changent tous les trimestres. Le critère utile est stable — un modèle récent à long contexte, capable d'appeler des outils, et dont la juridiction d'hébergement convient à votre corpus. Les familles à poids ouverts (Mistral, Qwen, Llama, Gemma) permettent un déploiement entièrement interne. Voir les pages « Chatbots IA » et « Évolutions des LLMs » pour l'état du marché.

Ne pas écrire le pipeline à la main. Les tutoriels officiels sont la seule référence qui reste à jour : « RAG » dans la documentation LangChain, « starter » chez LlamaIndex, « first RAG pipeline » chez Haystack. Les exemples de code qui circulent depuis 2023 ne s'exécutent plus : LangChain 1.x a déplacé les anciennes chaînes dans le paquet langchain-classic, LlamaIndex a supprimé ServiceContext depuis la v0.10, et les modèles OpenAI que ces exemples appellent (gpt-3.5-turbo, gpt-4-turbo) ferment le 23 octobre 2026.

Ordre de grandeur des coûts. L'indexation est le poste le moins cher : chez OpenAI, text-embedding-3-small coûte 0,02 \$ par million de tokens (3-large : 0,13 \$) — un corpus de quelques millions de mots s'indexe pour moins d'un euro. L'ancien ada-002, encore tarifé 0,10 \$/M, n'est pas déprécié mais vieillissant : il n'y a plus de raison de le choisir. Le stockage vectoriel va de gratuit (pgvector sur une base existante, offres d'entrée de Qdrant ou Weaviate) à quelques dizaines de dollars par mois. La génération domine la facture en régime permanent, et la RAG agentique la multiplie : un système multi-agents consomme environ quinze fois plus de tokens qu'une conversation simple.

Exploitation. Mettre en cache les passages et les réponses fréquentes ; prévoir une réindexation incrémentale plutôt qu'un rechargement complet ; tracer chaque requête (passages récupérés, prompt final, réponse) pour pouvoir expliquer une mauvaise réponse ; surveiller la latence de la recherche, qui se dégrade avec le volume et avec le reranking. Côté sécurité, les passages récupérés sont du contenu non fiable qui entre dans un prompt : l'injection de prompt par document piégé est un risque réel, comme le contrôle d'accès appliqué trop tard dans la chaîne.

— MESURER / 07

Évaluer et surveiller

Sans jeu de questions de référence, on ne sait pas si un changement améliore ou dégrade le système. Les métriques se séparent en deux familles : la récupération et la génération.

Une RAG se mesure en deux temps : a-t-on retrouvé les bons passages, et le modèle s'en est-il servi correctement ? Confondre les deux conduit à corriger le prompt quand le problème vient de la recherche.

01

ÉTAPE 1

Mesurer la récupération

- **Précision@k** : parmi les k passages remontés, combien sont pertinents.
- **Rappel du contexte** : les passages nécessaires à la réponse ont-ils tous été retrouvés.
- **MRR / NDCG** : le bon passage arrive-t-il en tête de liste.
- C'est ici que se gagnent les points : recherche hybride, embeddings contextuels, reranking.

02

ÉTAPE 2

Mesurer la génération

- **Fidélité** : chaque affirmation de la réponse est-elle bien étayée par un passage fourni.
- **Pertinence** : la réponse répond-elle à la question posée.
- **Qualité des citations** : la source indiquée est-elle la bonne.
- Ces trois mesures se font le plus souvent avec **un LLM juge**, dont il faut vérifier l'accord avec un échantillon annoté à la main.

03

OUTILLAGE

Les outils

- **Ragas** (v0.4.3, janvier 2026) : les métriques de référence, prêtes à l'emploi.
- **DeepEval** (v4.2.3) : l'évaluation comme des tests unitaires, intégrable en CI.
- Les évaluateurs intégrés à LangSmith, Phoenix ou MLflow ajoutent la trace des requêtes en production.

04

PRÉREQUIS

Le jeu de questions, d'abord

- 50 à 200 questions issues de vos vrais utilisateurs, avec la réponse attendue et le passage attendu.
- Sans ce jeu, aucune métrique n'est interprétable et chaque modification devient une intuition.
- Le constituer est le meilleur investissement d'un projet RAG, avant tout choix d'outil.

» À RETENIR

Les pièges les plus fréquents

- 01 Croire qu'un plus grand contexte règle le problème : au-delà de quelques centaines de pages, les performances se dégradent (context rot), et l'information placée au milieu est la moins bien retrouvée.
- 02 Juger le système sur trois questions de démonstration, choisies parce qu'elles fonctionnent.
- 03 Confondre une réponse bien écrite et une réponse étayée : sans vérification des citations, le style masque l'invention.
- 04 Oublier que les passages récupérés entrent dans le prompt : un document piégé peut y injecter des instructions.
- 05 Appliquer les droits d'accès à l'affichage plutôt qu'à la recherche, et divulguer par la citation ce que l'on croyait filtrer.
- 06 Indexer une fois et ne jamais réindexer : un corpus qui vit produit vite des réponses périmées, citées avec assurance.

— CONTEXT ENGINEERING / 08

La RAG comme outil de gestion du contexte

Le mot d'ordre de 2025-2026 : l'essentiel n'est plus de bien formuler le prompt, mais de choisir ce qui entre dans le contexte.

Le mot d'ordre de 2025-2026 : l'essentiel est de **choisir ce qui entre dans le contexte** — « le plus petit ensemble possible de tokens à fort signal », écrit Anthropic en septembre 2025. Au lieu de tout récupérer à l'avance, l'agent charge l'information « juste à temps » via des outils, résume ce qu'il a lu (compaction) et prend des notes externes. La RAG n'est plus l'architecture du système : elle devient **un outil parmi d'autres** de cette gestion du contexte, au même titre que la recherche web, l'accès à une base SQL ou un serveur MCP.

01

Ce que cela change dans la conception

- On n'optimise plus un pipeline unique, mais le **choix de la source** pour chaque question.
- La récupération devient un outil que l'agent décide d'appeler, une ou plusieurs fois.
- La compaction et les notes externes remplacent l'idée d'un contexte qui accumulerait tout.
- Le coût devient variable et doit se piloter : une question complexe peut consommer quinze fois plus qu'une simple.

02

Ce qui ne change pas

- La qualité de la récupération reste le facteur déterminant, quel que soit l'orchestrateur.
- Un jeu d'évaluation sur vos propres questions reste indispensable.
- La traçabilité des sources reste ce qui distingue une réponse vérifiable d'une réponse plausible.
- La question de l'endroit où vivent les données reste entière, et se pose désormais aussi pour chaque serveur MCP branché.

POUR ALLER PLUS LOIN **Agents IA : orchestration, mémoire et outils** ›

— ARCHIVE / 09

Ce qui a été dépassé

Des repères de 2020-2025 encore cités sur le sujet, mais dépassés, fermés ou trompeurs.

« **Hugging Face RAG** » comme outil de prototypage. L'implémentation DPR + BART de la bibliothèque Transformers est le modèle de recherche d'origine (Lewis et al., NeurIPS 2020) qui a donné son nom à la RAG. Elle a une valeur historique, mais ne constitue pas un point de départ en 2026 : les frameworks actuels ou les offres managées font le même travail en quelques lignes.

Les fenêtres de contexte « 4k, 8k, 32k, 128k ». Ce repère décrivait l'état de 2023-2024. Les modèles de pointe offrent désormais jusqu'à **1 million de tokens** par défaut (derniers Claude et Gemini), et les modèles courants 128 à 200 k. Mais plus de contexte ne veut pas dire meilleure réponse : la dégradation avec la longueur (*context rot*) a été mesurée sur 18 modèles en juillet 2025.

Les exemples de code de 2023. Les extraits LangChain (``from langchain.document_loaders ...`, `RetrievalQA``) et LlamaIndex (``from llama_index import ...`, `ServiceContext``) que l'on trouve encore partout **ne s'exécutent plus** : LangChain 1.x a déplacé les anciennes chaînes dans ``langchain-classic``, LlamaIndex a supprimé ``ServiceContext`` depuis la v0.10 au profit de ``Settings``, et les imports passent par ``llama_index.core``. Les modèles OpenAI qu'ils appellent (``gpt-3.5-turbo``, ``gpt-4-turbo``) ferment le 23 octobre 2026. Se référer aux tutoriels officiels, qui sont maintenus.

text-embedding-ada-002. Souvent présenté comme déprécié, il ne l'est pas chez OpenAI au 18 septembre 2026 : il reste tarifié (0,10 \$/M de tokens). Il est simplement **ancien** (décembre 2022) et moins bon, et cinq fois plus cher que `text-embedding-3-small` (0,02 \$/M). Chez Azure, un calendrier de retrait propre existe. Aucune raison de le choisir pour un nouveau projet.

jina-reranker-v1. Périmé comme référence de reranking. Les rerankers actuels sont Cohere Rerank 4.0 (pro / fast), Qwen3-Reranker (poids ouverts), et les modèles Voyage ou Jina — Jina AI ayant rejoint Elastic en octobre 2025.

L'Assistants API d'OpenAI. Fermée le **26 août 2026**. Le `*file search*` et les vector stores passent désormais par la Responses API. De même, « Azure Cognitive Search » est l'ancien nom d'Azure AI Search, et « Vertex AI Search

» celui de RAG Engine / Agent Search dans la plateforme renommée Gemini Enterprise Agent Platform le 22 avril 2026.

« **La RAG est morte, vive le long contexte.** » Argument récurrent depuis 2024, démenti par la pratique : le coût, la fraîcheur des documents, la traçabilité des sources et la dégradation mesurée sur les longs contextes maintiennent la récupération indispensable. La bonne question n'est pas « RAG ou long contexte » mais « quelle part du corpus doit entrer dans le prompt, et quelle part doit être cherchée ».

— RÉFÉRENCES / 10

Sources et références

10

- [01] Lewis et al. – Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (NeurIPS 2020)
- [02] Shuster et al. – Retrieval Augmentation Reduces Hallucination in Conversation (avr. 2021)
- [03] Anthropic – Introducing Contextual Retrieval (19 sept. 2024)
- [04] Anthropic – Context windows (1 M de tokens, context rot, compaction)
- [05] Anthropic – Effective context engineering for AI agents (29 sept. 2025)
- [06] Anthropic – How we built our multi-agent research system (13 juin 2025)
- [07] Chroma – Context Rot (14 juil. 2025, 18 modèles testés)
- [08] Liu et al. – Lost in the Middle: How Language Models Use Long Contexts (TACL)
- [09] Google – Gemini API Long context
- [10] Singh et al. – Agentic Retrieval-Augmented Generation: A Survey (avr. 2026)
- [11] Menlo Ventures – 2024: The State of Generative AI in the Enterprise (20 nov. 2024)
- [12] Menlo Ventures – 2025: The State of Generative AI in the Enterprise (9 déc. 2025)
- [13] OpenAI – Deprecations (Assistants API, gpt-3.5-turbo, gpt-4-turbo)
- [14] OpenAI – API Pricing (embeddings, file search)
- [15] OpenAI – File search guide (Responses API)
- [16] Google – Introducing the File Search Tool in Gemini API (6 nov. 2025)
- [17] Google – Gemini API File Search (documentation)
- [18] AWS – Amazon Bedrock Managed Knowledge Base is now generally available (17 juin 2026)
- [19] AWS – Bedrock Managed Knowledge Base : connecteur ServiceNow (sept. 2026)
- [20] Microsoft Learn – What is Foundry IQ?
- [21] Microsoft Learn – Foundry IQ FAQ (agentic retrieval, API 2026-04-01)
- [22] Google Cloud – Gemini Enterprise Agent Platform (ex-Vertex AI)
- [23] Mistral AI – Document Library (bêta)
- [24] Mistral AI – Le Chat Enterprise
- [25] DINUM – Albert API (OpenGateLLM)
- [26] LangChain – RAG (documentation officielle)
- [27] LangChain – RetrievalQA dans langchain-classic
- [28] LlamaIndex – Documentation du framework
- [29] LlamaIndex – Migration ServiceContext > Settings
- [30] deepset – Haystack
- [31] deepset – Haystack, tutoriel « first RAG pipeline »
- [32] InfiniFlow – RAGFlow (dépôt GitHub)
- [33] pgvector (dépôt GitHub)
- [34] Qdrant – Tarifs
- [35] Weaviate – Tarifs
- [36] Milvus – Release v3.0.0 (29 juil. 2026)
- [37] Pinecone – Tarifs
- [38] Elastic – Vector database
- [39] Elastic – Elastic and Jina AI (9 oct. 2025)
- [40] Microsoft – GraphRAG (documentation)
- [41] Microsoft Research – LazyGraphRAG (25 nov. 2024)
- [42] Guo et al. – LightRAG
- [43] Faysse et al. – ColPali: Efficient Document Retrieval with Vision Language Models (ICLR 2025)
- [44] Hugging Face – Organisation ViDoRe (ColQwen2.5, ViDoRe V3)
- [45] Docling (dépôt GitHub)
- [46] Hugging Face – Qwen3-Embedding-8B
- [47] Hugging Face – Qwen3-Reranker-8B
- [48] Hugging Face – BAAI/bge-m3
- [49] Google – Gemini Embedding 2 (10 mars 2026)
- [50] Voyage AI – The Voyage 4 model family (15 janv. 2026)
- [51] Cohere – Rerank overview (Rerank 4.0)
- [52] MTEB – Leaderboard officiel
- [53] Ragas – Documentation
- [54] DeepEval – PyPI
- [55] Linux Foundation – Formation of the Agentic AI Foundation (9 déc. 2025)
- [56] Model Context Protocol – Versioning (spécification 2026-07-28)
- [57] Google – Try Deep Research (11 déc. 2024)
- [58] Anthropic – Claude takes research to new places (15 avr. 2025)
- [59] Mistral AI – Le Chat dives deep (17 juil. 2025)
- [60] AWS – Qu'est-ce que la génération augmentée par récupération ?
- [61] Red Hat – Qu'est-ce que la génération augmentée de récupération ?
- [62] CATIE – Accompagnement et expertise IA