





► o (+ o) ► o (+ o) ► o (+ o) ► o (+ o) ► o (+ o) ► o (+ o) ► o (+ o) ► o (+ o) ► o (+ o) ► o (+ o) ► o (+ o) ► o (+ o) ► o (+ o) ► o (+ o) ► o (+ o)

—

Maitriser le RAG

- Connecter les modèles d'IA génératives aux données de l'entreprise

Comptoir Octo

16/01/24



Intervenants OCTO



Nicolas Cavallo
Head of **Generative AI**

There is a better way

OCTO

T e c h n o l o g y

Part of Accenture

1000 OCTOs

Dont plus de 900
consultant.e.s expert.e.s

Tou.te.s cultivant des expertises tech et méthodo pour toujours trouver de meilleures façons de résoudre vos problématiques complexes.

Best Workplaces™

De 250 à 1 000 salariés

Great Place
To Work.

FRANCE
2021

Certified



Corporation™

Planet Tech'Care
- SIGNATAIRE -



Sommaire

1. **Le RAG ?**
2. Le “**bon**” RAG ?
3. **Construire** son RAG
4. **Retours** d’expérience
5. **Demain** le RAG



01

Retrieval Augmented Generation

RAG



En une phrase

Retrieval-Augmented Generation

.... On **augmente** les **modèles de langages** (LLMs)

...avec de la **données** qui peut être privée, à jour etc..

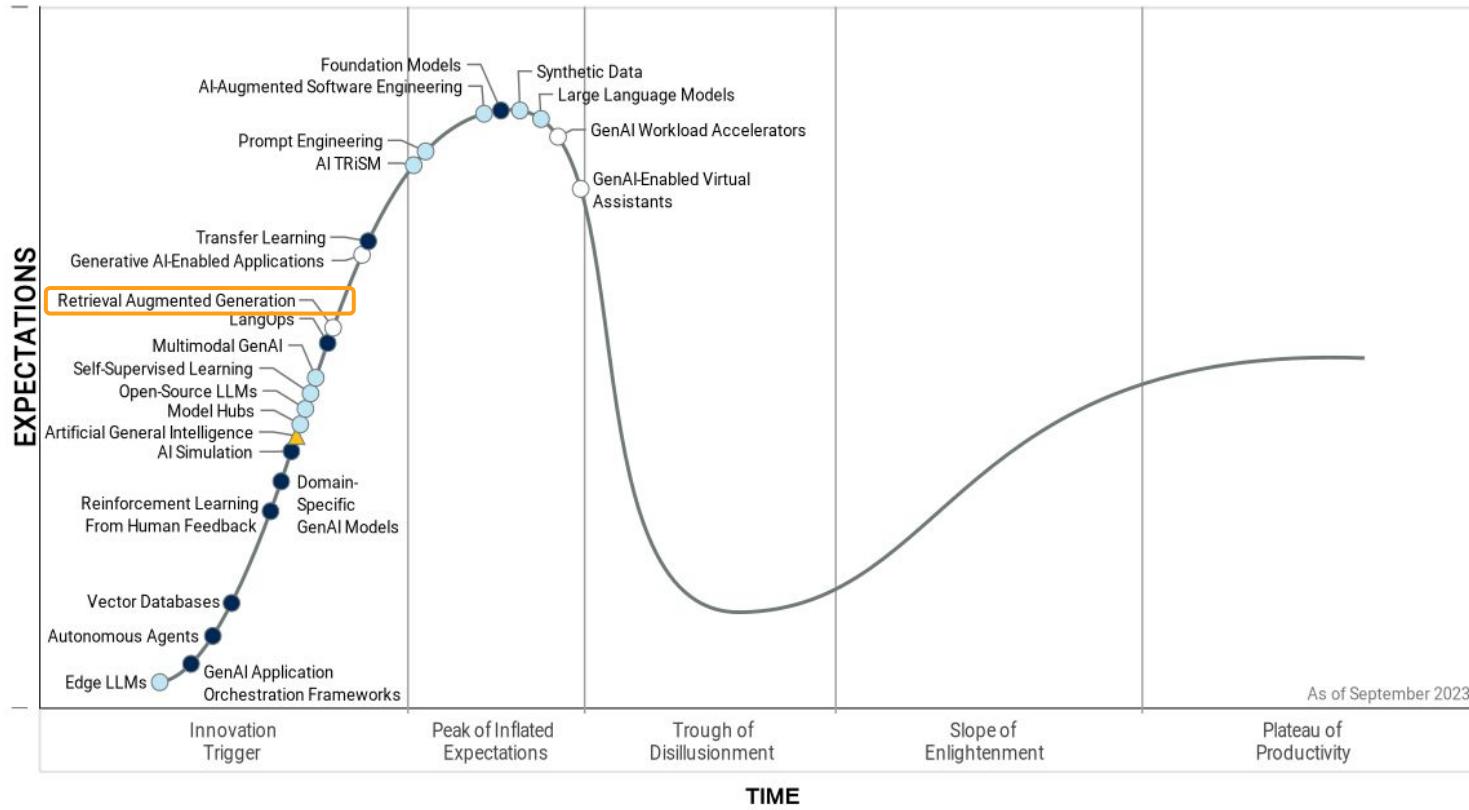
*“Je peux discuter avec un agent conversationnel qui est **à jour**, comprend **ma donnée** **mon vocabulaire d’entreprise**, qui peut **interagir avec des données sensibles, stratégiques**”*



On en met partout



Hype Cycle for Generative AI, 2023



Plateau will be reached: ○ <2 yrs. ● 2–5 yrs. ● 5–10 yrs. ▲ >10 yrs. ✘ Obsolete before plateau

Gartner



OpenAI en met dans ses GPTs

There is a better way

GPTs

Discover and create custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills.

Q Search public GPTs

Top Picks DALL-E Writing Productivity Research & Analysis Programming Education Lifestyle

Featured

Curated top picks from this week

**AllTrails**
Find trails that fit your nature for your next hike, ride, or run.
By alltrails.com

**Consensus**
Your AI Research Assistant. Search 200M academic papers from Consensus, get science-based...
By Consensus

My GPTs

 **Create a GPT** Beta
Customize a version of ChatGPT for a specific purpose

Name
Name your GPT

Description
Add a short description about what this GPT does

Instructions
What does this GPT do? How does it behave? What should it avoid doing?

Conversation starters

Knowledge
If you upload files under Knowledge, conversations with your GPT may include file contents. Files can be downloaded when Code Interpreter is enabled

Upload files

Capabilities
 Web Browsing
 DALL-E Image Generation
 Code Interpreter ?

Actions
Create new action



Google aussi...

Quand **Google Bard** ajoute la **fonctionnalité** de répondre en se **basant sur le contenu de vidéos youtube...**
C'est du RAG !

There is a better way

YouTube ▾

Regenerate draft ⚙️ 🔍

Here is the answer to your question according to the video:

Dan uses rum instead of vodka because rum has toffee and caramel notes that complement coffee, while vodka is very clean and doesn't add much flavor. The aged rum that Dan uses picks up these flavors through the aging process.

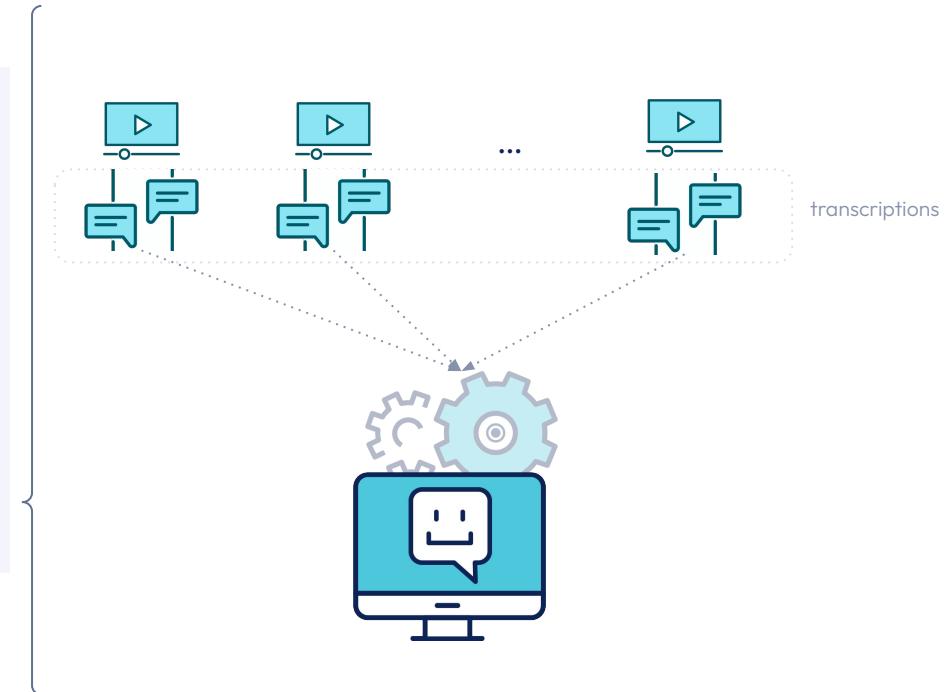
YouTube Why does Dan use rum instead of vodka?

 How to Make the Best Espresso Martini | What's Eating Dan?
America's Test Kitchen
Dan demonstrates how to make a complex, flavorful espresso martini with aged rum.

Enter a prompt here

Bard may display inaccurate info, including about people, so double-check its responses. [Your privacy & Bard](#)

src img :
<https://www.theverge.com/2023/11/22/23972636/bard-youtube-extension-update-search-video-content>





Techniquement



Une réponse contextualisée

ChatGPT3.5 sans RAG

**You**

C'est quoi le RAG ?

**ChatGPT**

Le terme "RAG" peut avoir différentes significations en fonction du contexte. Voici quelques-unes des significations possibles :

1. Rapport Annuel de Gestion : En comptabilité et gestion, le "RAG" peut faire référence au Rapport Annuel de Gestion, qui est un document présentant les résultats financiers et la performance d'une entreprise ou d'une organisation au cours de l'année fiscale.
2. Risque Aggravé de Garantie : Dans le domaine des assurances, le "RAG" peut signifier Risque Aggravé de Garantie. Il s'agit d'une situation dans laquelle une personne présente un risque plus élevé pour l'assureur en raison de facteurs tels que la santé, l'âge ou d'autres circonstances, ce qui entraîne des primes d'assurance plus élevées.
3. Ragdoll : Le "RAG" peut également faire référence à une race de chat appelée "Ragdoll". Les Ragdolls sont connus pour leur taille impressionnante, leur pelage soyeux et leur nature docile.



Si vous avez un contexte spécifique en tête ou si vous cherchez des informations sur un autre sujet lié au "RAG", n'hésitez pas à fournir plus de détails pour obtenir une réponse plus précise.



ChatGPT3.5 avec RAG

**You**

C'est quoi le RAG ?

**OctoGPT**

Le RAG, ou Retrieval Augmented Generation, est une approche qui permet d'utiliser un agent conversationnel pour répondre aux questions en utilisant des données spécifiques. Il consiste en plusieurs étapes, notamment la division du corpus de textes en sous-parties, la transformation de ces sous-parties en vecteurs numériques, le stockage de ces vecteurs dans une base de données, et la recherche des informations sémantiquement proches de la question posée. L'objectif du RAG est de personnaliser la réponse en utilisant les données de l'utilisateur ou de l'entreprise.



source:

<https://blog.octo.com/le-chatbot-docto-langchain-rag-et-code-associe>

Fonctionnement

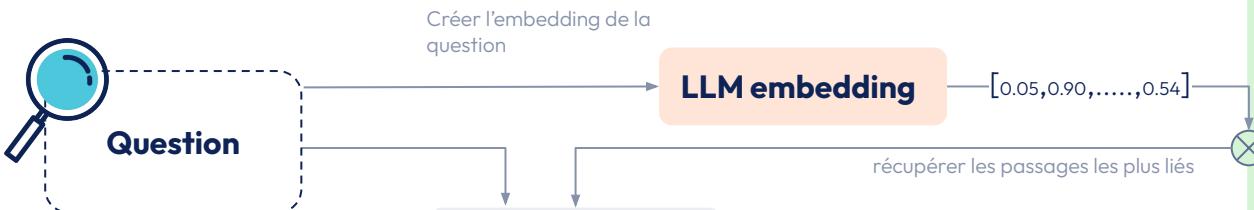
1

Ingestion
de l'existant



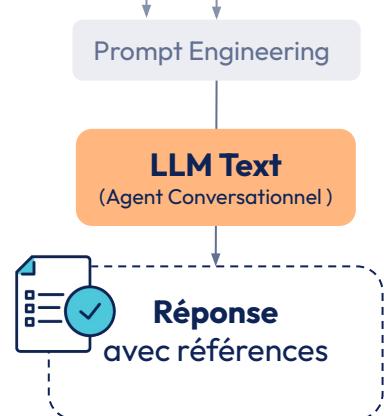
2

Recherche
des informations
utiles



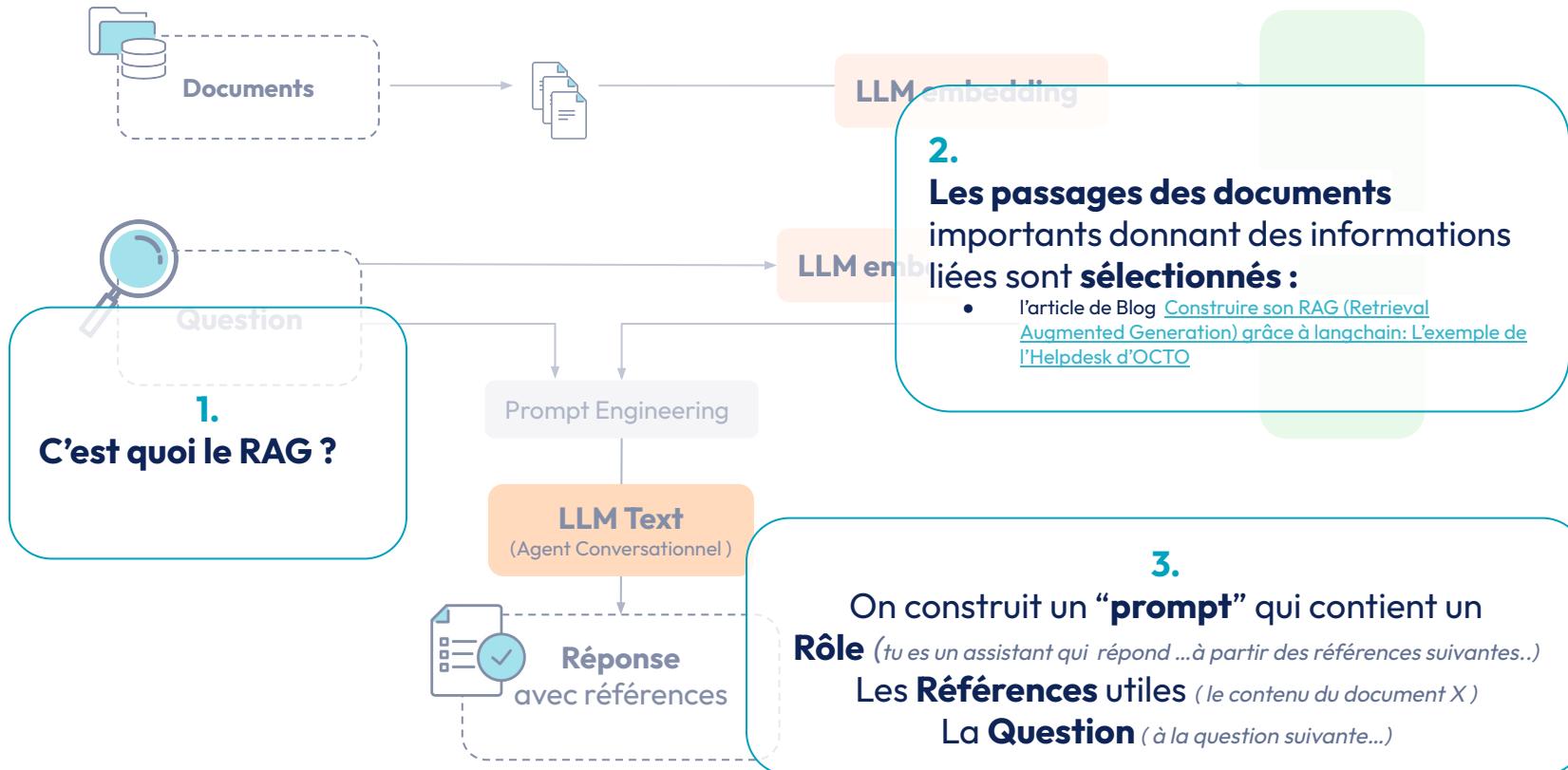
3

Formulation
d'une réponse
avec ses sources



L'exemple pour l'assistant Octo

There is a better way





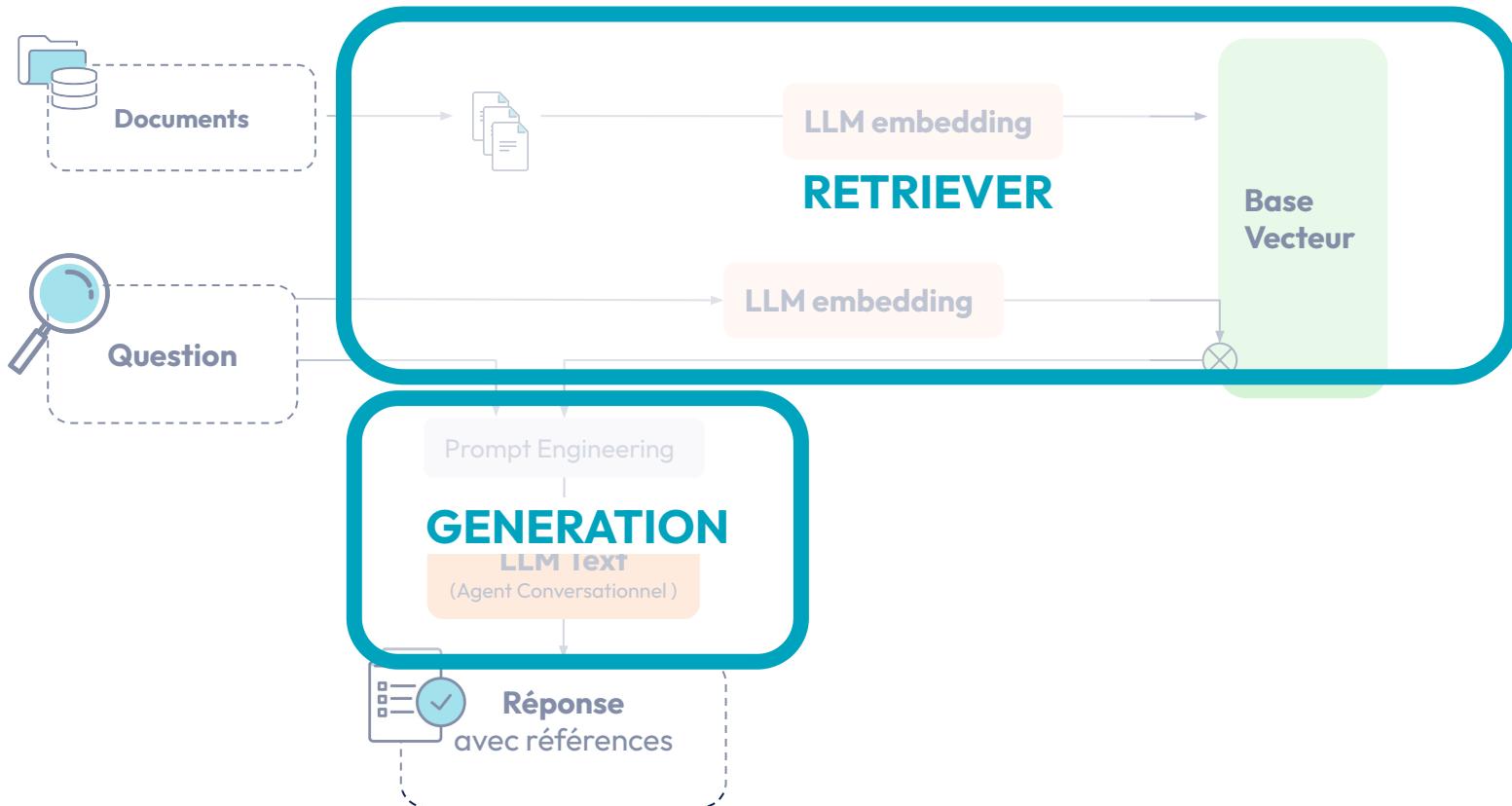
02

un **bon RAG ?**



Evaluation RAG

QUALITÉ : Deux composants à étudier dans un RAG



Complexité d'évaluer la qualité des réponses

Question

C'est quoi un RAG ?

Un article qui parle de RAG

Construire son RAG (Retrieval Augmented Generation) grâce à langchain: L'exemple de l'Helpdesk d'OCTO

Posté le 17/10/2023 par Florian Bastin, Nicolas Cavallo

Un article qui évoque le RAG

Sous le capot des bases de données vectorielles

Posté le 14/12/2023 par Aurélien Massiot, Philippe Stepniewski

Réponse 1

Le **RAG**, ou Retrieval Augmented Generation, est une approche qui permet **d'utiliser un agent conversationnel** pour répondre aux **questions en utilisant des données spécifiques**. Il consiste en plusieurs étapes, notamment la division du corpus de textes en sous-parties, la transformation de ces sous-parties en vecteurs numériques, le stockage de ces vecteurs dans une base de données, et la recherche des informations sémantiquement proches de la question posée. L'objectif du RAG est de personnaliser la réponse en utilisant les données de l'utilisateur ou de l'entreprise.

Réponse 2

Le **RAG** (Retrieval Augmented Generation) est une approche qui permet **d'exploiter la puissance d'un agent conversationnel en utilisant nos propres données**, en lien avec les bases de données vectorielles. Cependant, le contexte ne fournit pas de définition précise du RAG

RETRIEVER

deux **documents différents** peuvent contenir une **information suffisante**

GENERATION

Deux **réponses** dont la **formulation** et les **mots** sont **totalelement différents** peuvent contenir une **bonne réponse**.

Définir les bonnes métriques

Un **domaine** en **pleine évolution**

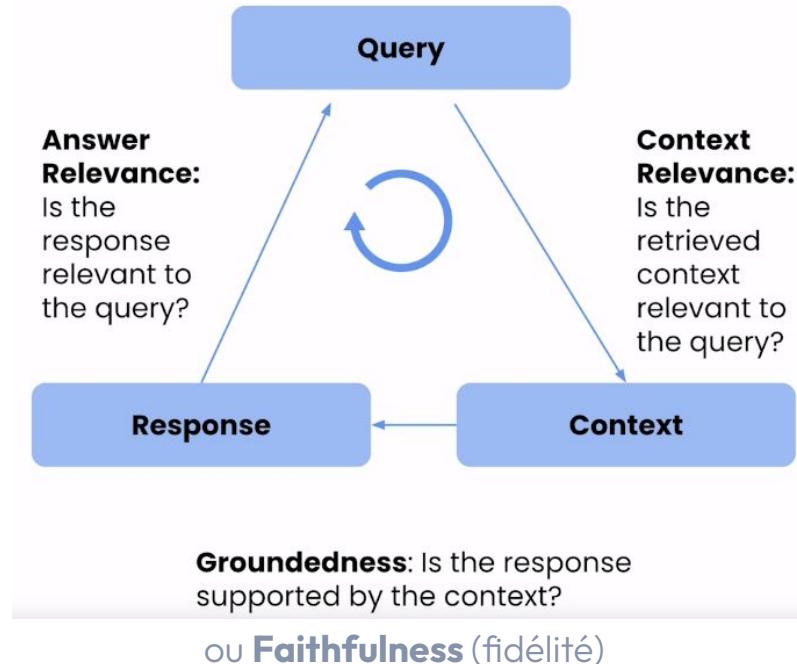
- Des méthodologies d'évaluation qui utilisent des LLMs 
- Des “frameworks” qui emergent



Framework : Trulens

Trulens The RAG - triad

The RAG Triad



<https://github.com/truera/trulens>



Framework : Ragas

Un score qui prend en compte les **différents axes d'évaluation**

GENERATION

Faithfulness

(fidélité)

Answer relevancy

(Pertinence de la réponse)

RETRIEVER

Context precision

(Pertinence du contexte)

Context Recall

(Rappel du contexte)

ragas : <https://github.com/explodinggradients/ragas>



Framework : Ragas

Un score qui prend en compte les **différents axes d'évaluation**

GENERATION

- **Faithfulness** (fidélité) : mesure la précision de la réponse générée par rapport au contexte fourni.
 - > Etape 1 : Pour le couple “**question/ réponse générée**”, nous utilisons un LLM pour **déterminer** les **affirmations** que la réponse générée fait.
 - > Etape 2 : Nous utilisons un LLM pour qualifier la liste des affirmations (Etape 1). **Pour chacune des affirmations** nous vérifions qu'elle est **bien contenue dans le contexte** fourni . Le score est la division du nombre d'affirmations validées par le nombre total d'affirmations de la réponse générée.
- **Answer relevancy** (Pertinence de la réponse) : Mesure la pertinence de la réponse par rapport à la question.
 - > Nous utilisons un LLM pour déterminer une liste de questions probables auxquelles la réponse générée serait une réponse. Nous calculons la similarité de ces questions probable avec la réelle question.

RETRIEVER

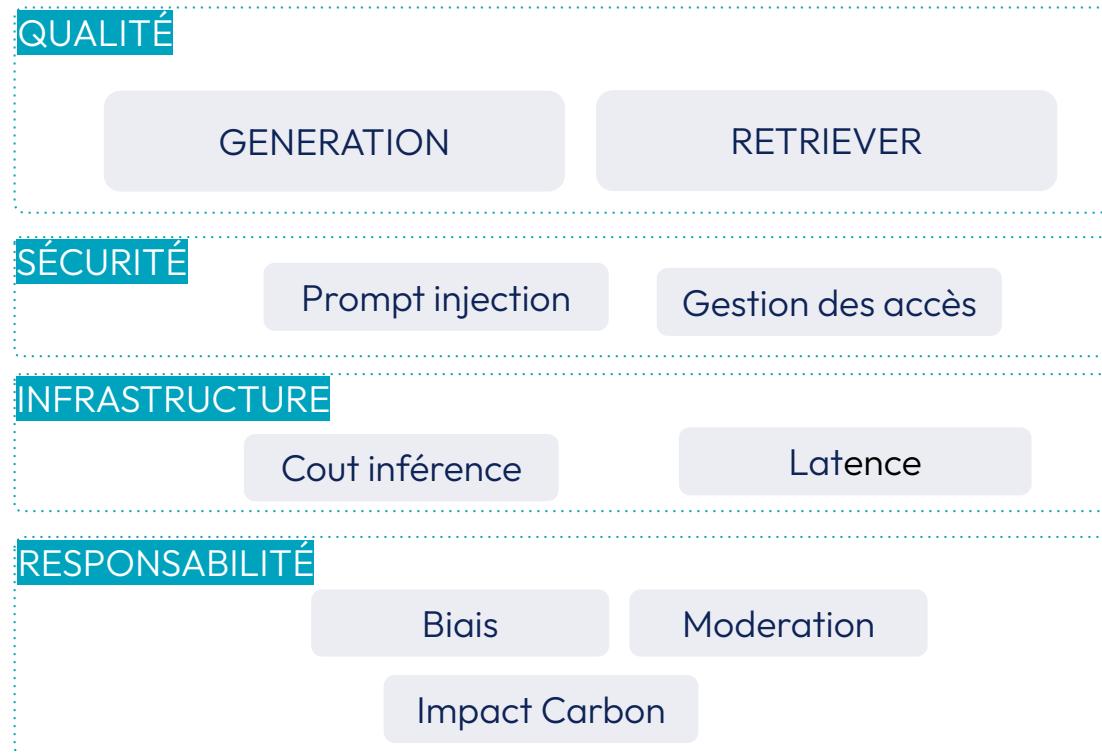
- **Context precision** (Pertinence du contexte) : Mesure le poids des informations utiles comparativement au bruit dans le contexte récupéré.
 - > Nous utilisons un LLM pour déterminer les phrases du contexte récupéré qui sont nécessaires pour répondre à la question. Le score est le ratio entre le nombre de phrases requises et le total des phrases du contexte récupéré.
- **Context Recall** (Rappel du contexte) : Mesure la capacité du “Retriever” à retrouver toutes les informations nécessaires pour répondre à la question.
 - > Nous utilisons un LLM pour lister l'ensemble d'affirmations de la réponse “ground truth” fournie. Le score est la division entre le nombre d'affirmations contenues (d'après un LLM) dans le contexte récupéré et le total d'affirmations de la réponse “ground truth”.



Pas seulement la qualité...

Toutes les évaluations sont fausses, mais certaines sont utiles.

Des **évaluations à plusieurs niveaux**,
dont l'importance dépend du cas
d'usage

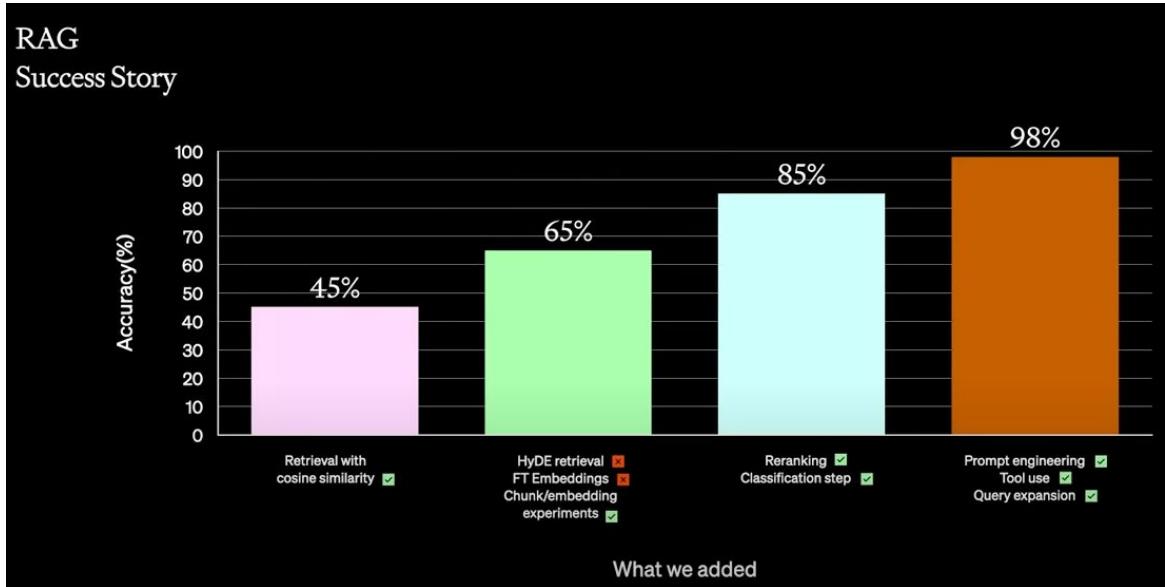


OpenAI diffuse son retour d'expérience pour construire un RAG

“ ... because with prompt engineering and RAG, it sounds like these things can be quite simple, but they're really quite hard. It takes a lot of iterations and a lot of testing and learning to actually make this happen for real.” OpenAI nov-2023 <https://www.youtube.com/watch?v=ahnGLM-RC1Y>

There is a better way

1
ÉVALUATION



98 % d'accuracy
sans fine tuning !

2

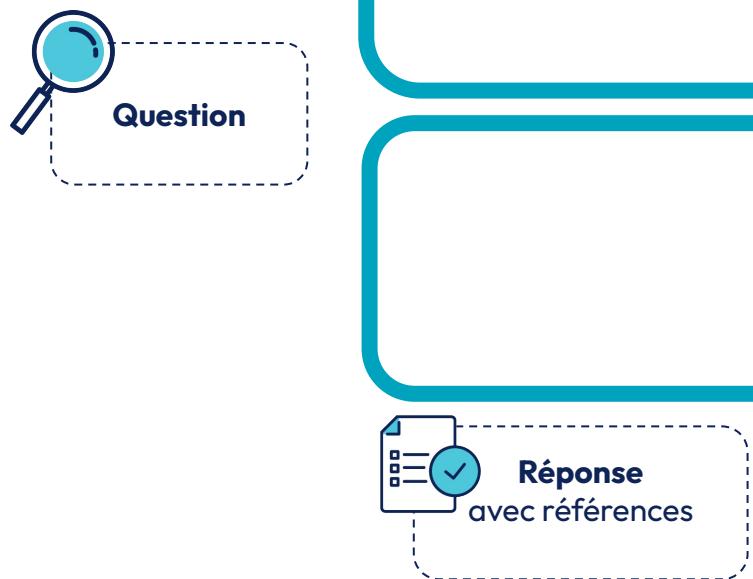
TESTER, TESTER ... TESTER



Un **RAG**
ça peut être **complexe**
du RAG basique ... au RAG avancé



Deux composants dans un RAG





Aller plus loin sur le RETRIEVER

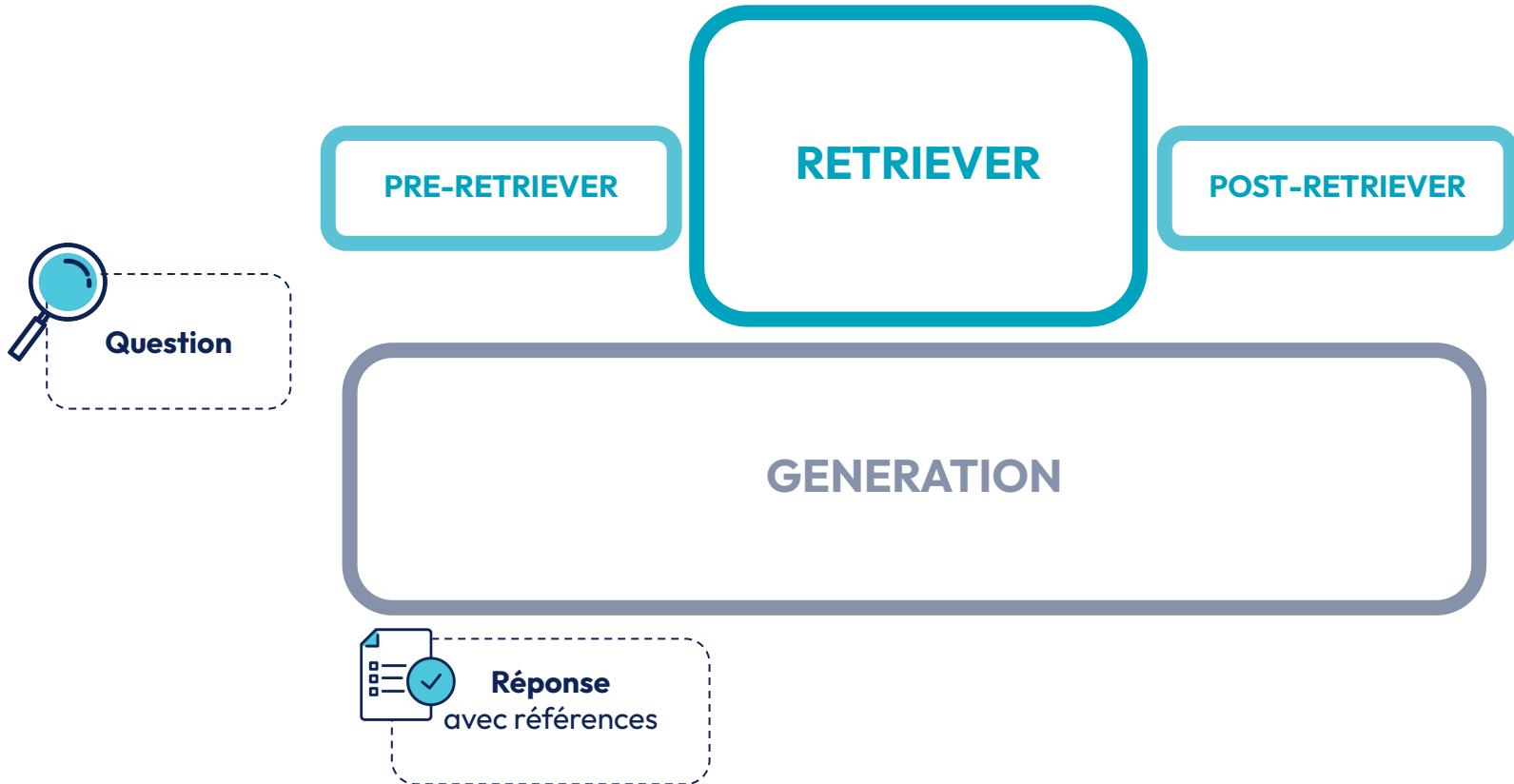


Le Retriever ce n'est pas une base vecteur

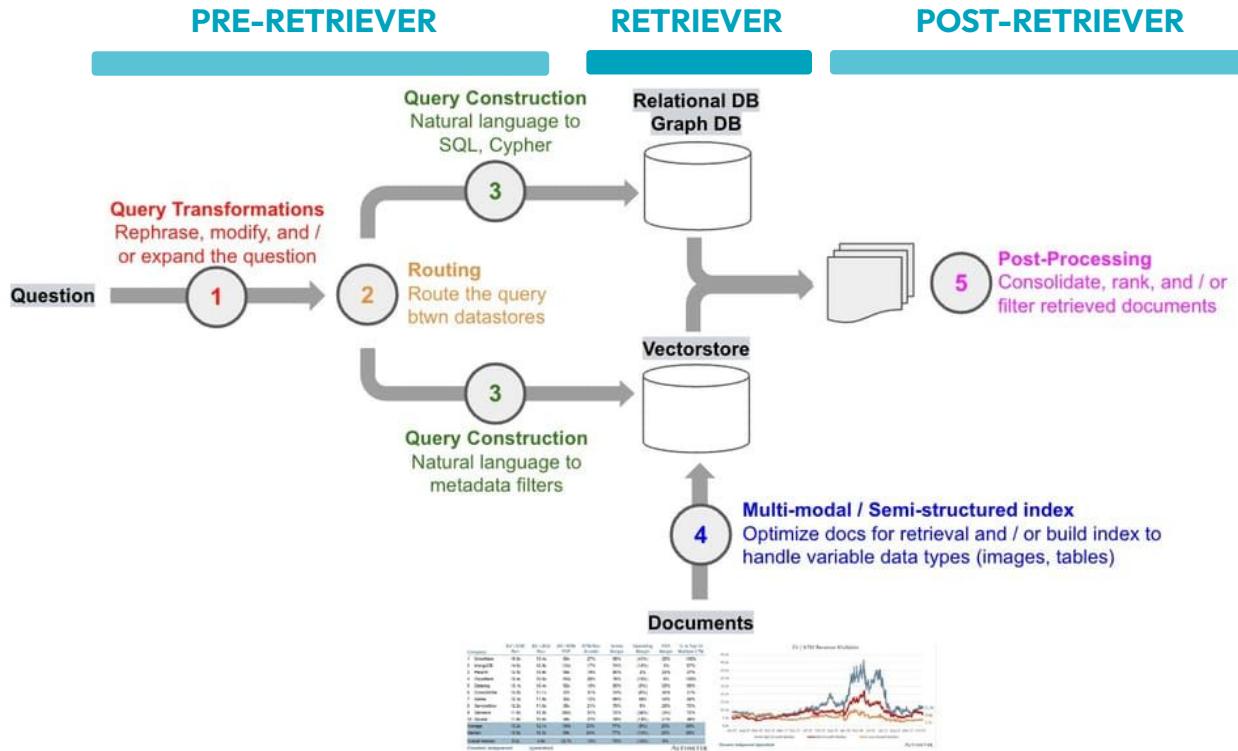
There is a better way

Build a search engine not a vector DB

Un “RETRIEVER” Avancé



Un “RETRIEVER” Avancé

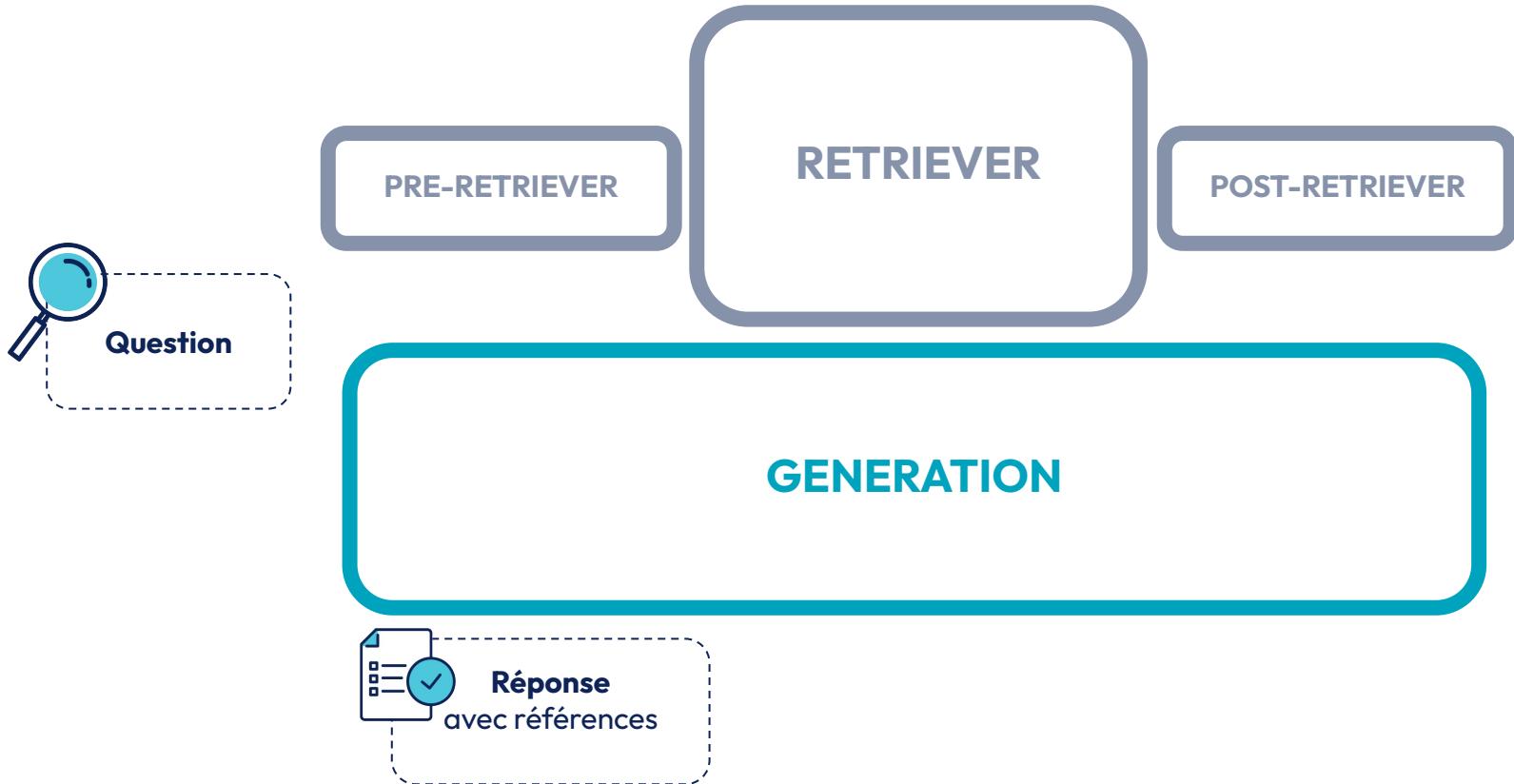


src img : langchain blog



Aller plus loin sur la GENERATION

Une “GENERATION” Avancé





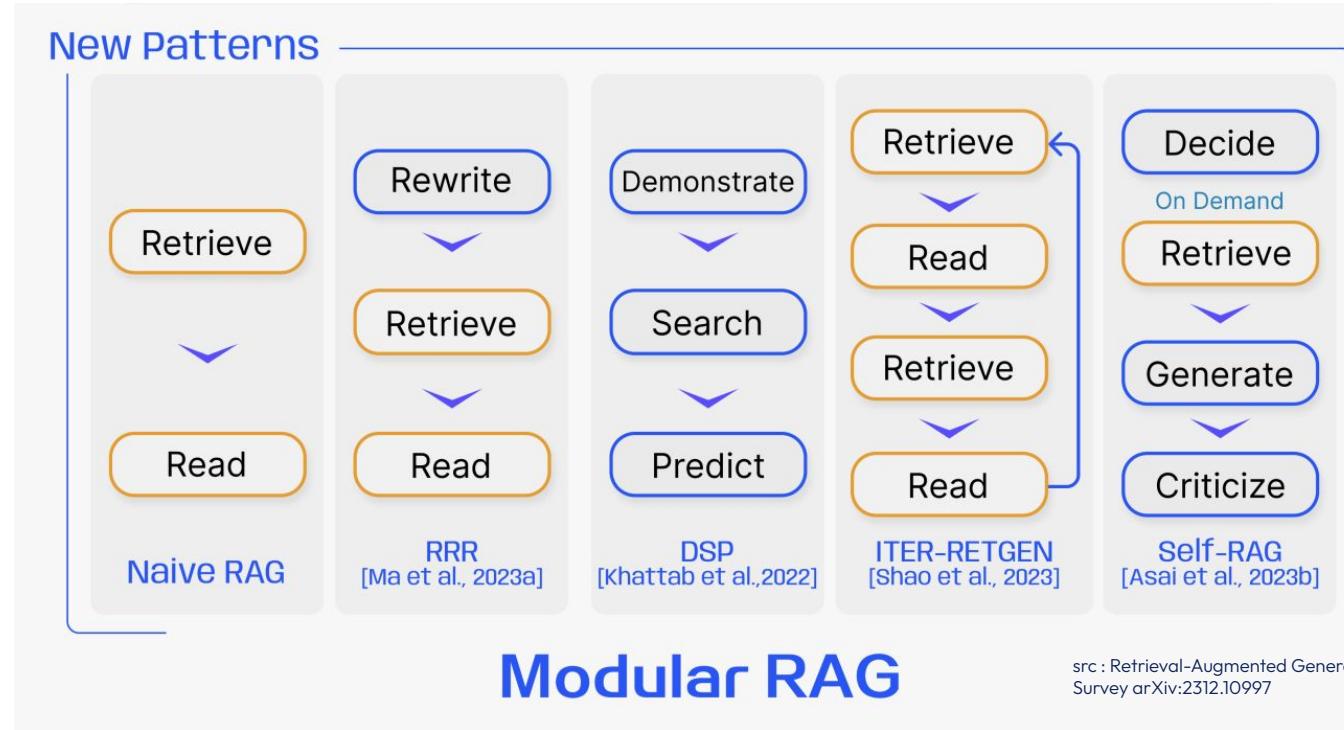
Une GÉNÉRATION avancée

- **Prompt Engineering** : Modifier/ Adapter le Rôle
- **Adapter-modifier le modèle LLM text :**
 - > GPT3.5 <-> GPT4 <-> Mistral , etc..
 - > Utiliser des modèles alignés pour le RAG
- Ajouter des **modèles - garde-fous** :
 - > Modèle de modération
 - > Modèle de relecture
 - > Utiliser plusieurs modèles et analyser la cohérence entre les différentes réponses
 - > etc..

jusqu'au
RAG modulaire

Modular RAG

L'idée est de permettre de choisir les **differents modules à activer** en fonction des questions.



03

Construire son RAG

Commencer techniquement un projet RAG

On a déjà validé la valeur potentiel du projet



DÉFINITION

- Fixer les pré-requis :
 - **Données**
 - **Environnement technique**
- Fixer la GÉNÉRATION
 - Définir le **LLM text utilisable**
- Fixer un RETRIEVEUR
 - base vecteur, LLM embedding



EXPERIMENTATION SIMPLE

- Commencer avec un **sous périmètre**.
- Définir un moyen d'**évaluer**
- Créer le **RAG simple** et tester quelques questions



AMÉLIORATION

- Prendre la **totalité du périmètre**.
- Passer à un **RAG avancé** (voire modulaire)

Un cadrage technique

Un MVP en quelques semaines

Construire le projet RAG

Pipeline RAG



LangChain



LlamaIndex



python™

Base vecteur

Des **dizaines de bases vecteur possibles** (PostgreSQL, Elasticsearch, chroma, etc..)

- Comparaison : [Vector DB Feature Matrix](https://docs.google.com/spreadsheets/d/170HErOyOkLDjQfy3TJ6a3XXXM1rHvw_779Sit-KT7uc/edit#gid=0) :
https://docs.google.com/spreadsheets/d/170HErOyOkLDjQfy3TJ6a3XXXM1rHvw_779Sit-KT7uc/edit#gid=0
- Construire sa base vecteur [Sous le capot des bases de données vectorielles - OCTO Talks !](https://blog.octo.com/sous-le-capot-des-bases-de-donnees-vectorielles-(vector-databases))
[https://blog.octo.com/sous-le-capot-des-bases-de-donnees-vectorielles-\(vector-databases\)](https://blog.octo.com/sous-le-capot-des-bases-de-donnees-vectorielles-(vector-databases))

LLM embedding

- API externe : OpenAI embedding,
- Open Source : HuggingFace Embedding Models (Bert, etc..)

LLM text

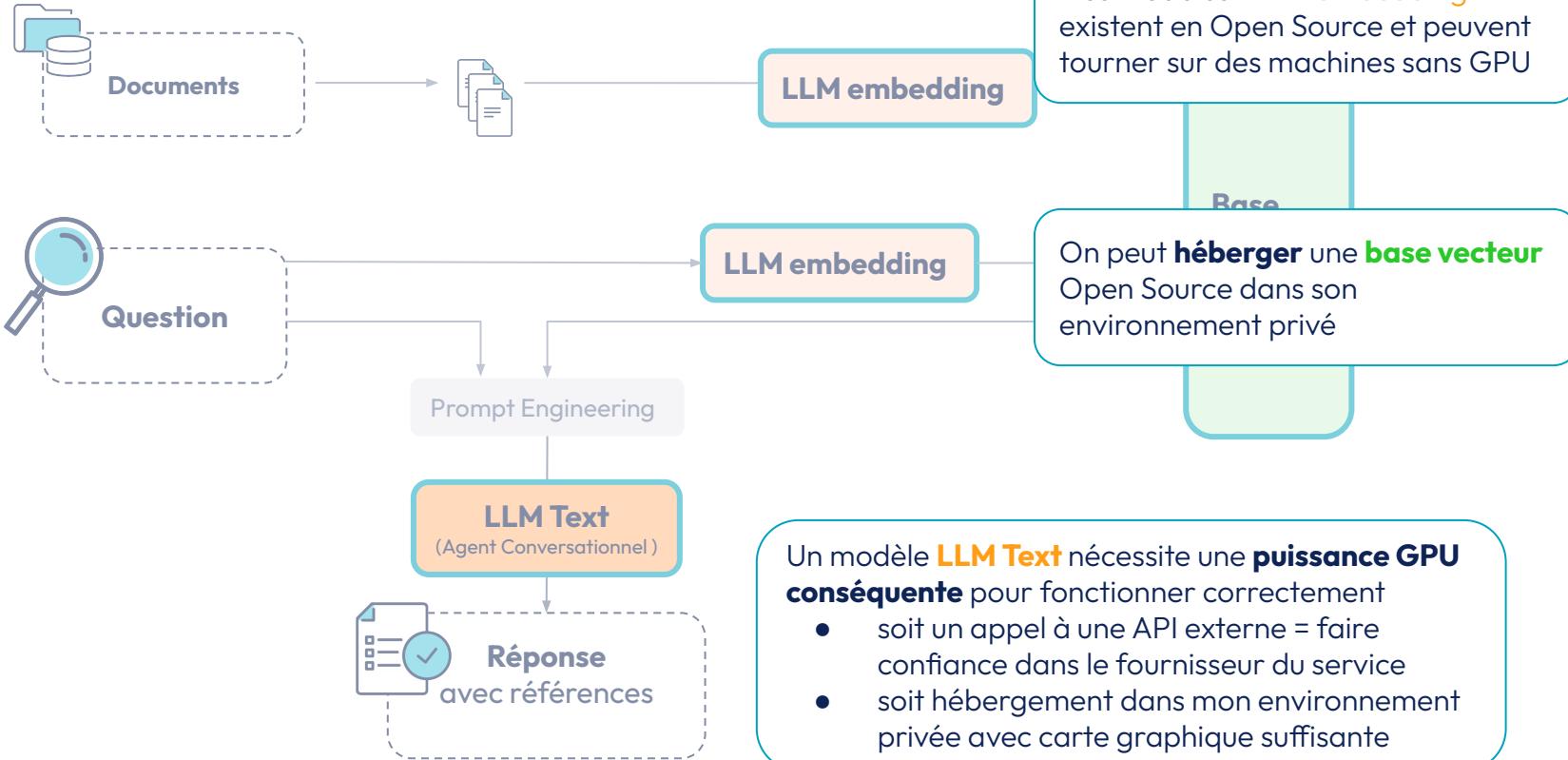


Bard AI



...

Peut fonctionner dans un environnement privé



Les outils de suivi d'un pipeline LLM

Stable et fonctionnel



Solutions émergentes (mieux connectées aux spécificités LLM-RAG)





Cout financier & Environnemental



Mesurer le coût !

Coût = LLM text : représente rapidement >99% du coût de RUN du projet RAG

FINANCIER :

- Coût principalement au **nombre de tokens**

| Tokens | Characters |
|---------------------|------------|
| 7 | 19 |
| C'est quoi le RAG ? | |

- **GPT-3.5-turbo est 30 fois moins cher que GPT4** (10 fois moins que GPT4 turbo)

<https://docs.google.com/spreadsheets/d/1NX8ZW9Jnfpv88PC2d6Bwla87JRiv3GTeqwXoB4mKUs/edit#gid=0>

- > input : 1\$ pour 1M de tokens
- > output : 2\$ pour 1M de tokens

ENVIRONNEMENTAL :

- Action principale = **mesurer l'impact carbon** de l'usage de ce **LLM text**.
 - > Si **modèle Open Source hébergé** :
 - + CodeCarbon (consommation de l'application) * Empreinte carbone en France (50 g.CO2eq/kWh)
 - > Si modèle externe (ex GPT4) : ...
 - + ~**coûts financiers corrélés** avec le **coût environnemental**



04

Retours d'expérience

Simple l'exemple “Onboarding”

Je viens d'arriver à octo que dois je faire ?

Bienvenue, voici les sujets que tu dois découvrir en priorité :

1. Le CRA : connaître les bases pour bien démarrer
2. Modifier mon CRA
3. Les Note de frais
4. Pointage des absences

CRA ?

Le CRA : **Compte Rendu d'Activité**, c'est la page d'accueil de **Octopod** et c'est là où tu dois indiquer ce que tu as fait : missions client, intercontrat, projets internes, etc.

ROLE : Tu es l'assistant qui ne se base que sur le contexte fourni pour répondre.

CONTEXT :

page_confluence : Octopod - Onboarding : Bienvenue sur cette page ! Que tu sois nouveau chez Octo ou même que tu démarres en tant que consultant, retrouve ici le plein d'infos pour t'y retrouver...

...

...

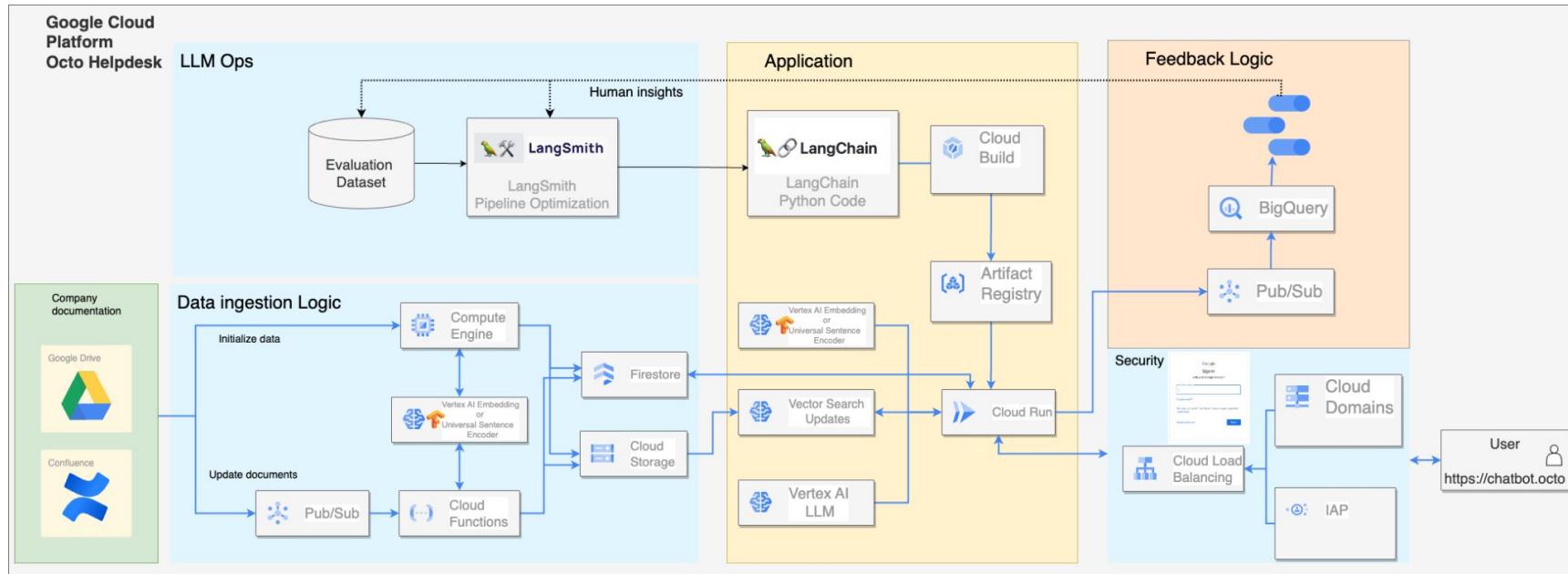
QUESTION : Je viens d'arriver à octo que dois je faire



Agent Conversationnel

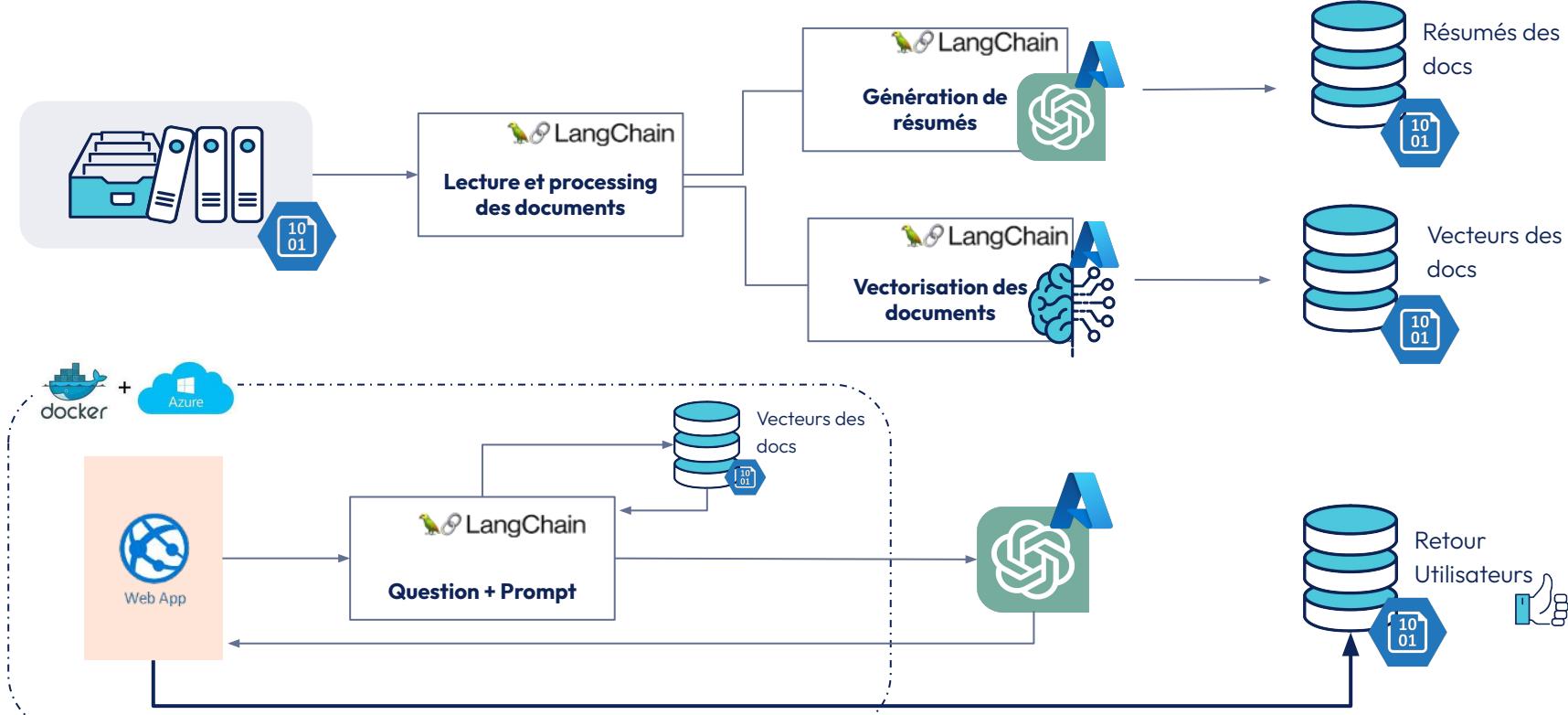
Helpdesk Octo sur GCP

<https://github.com/BastinFlorian/RAG-on-GCP-with-VertexAI>



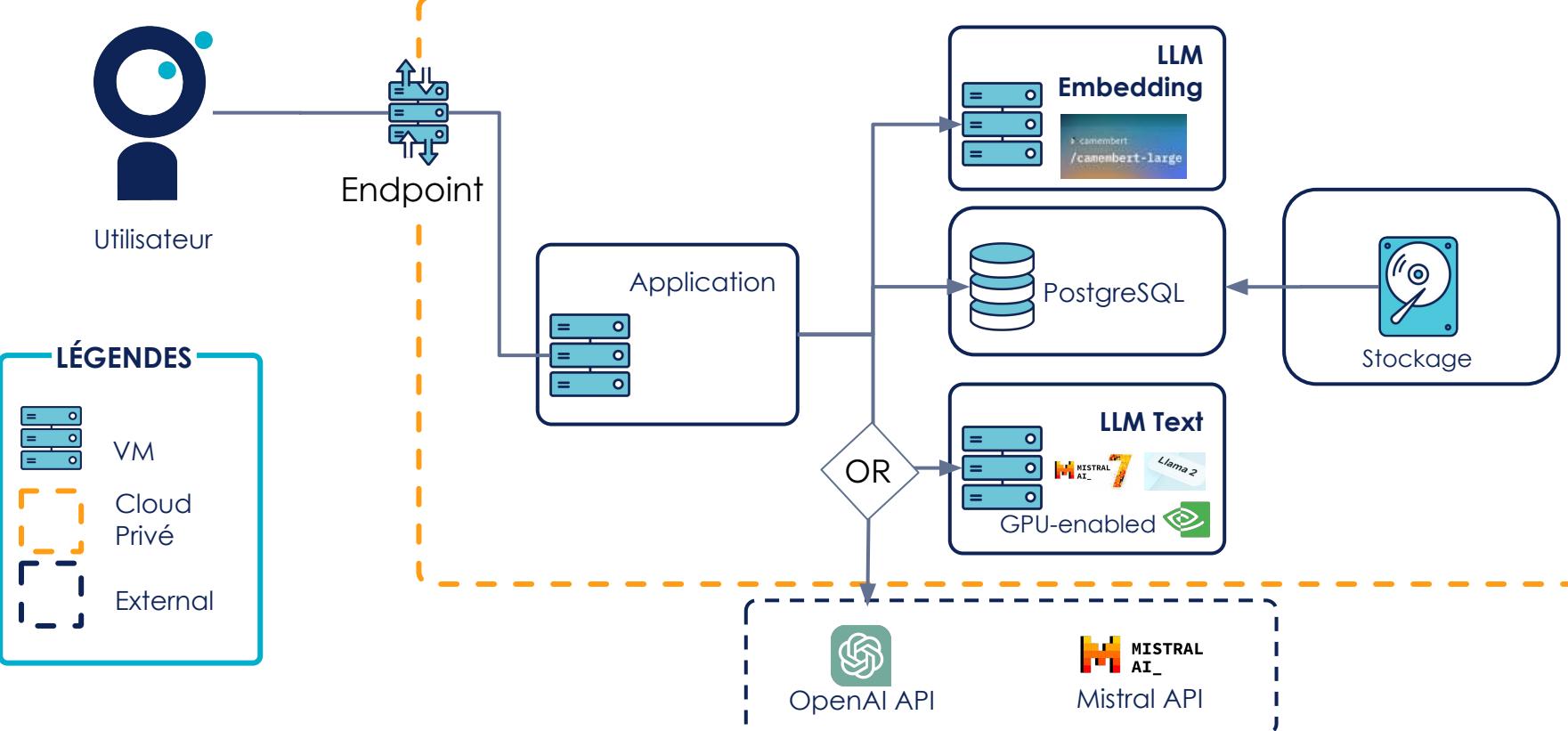
RAG sur des document : Environnement Azur

Protection de vos données !



Assistant augmenté

Architecture environnement Privée



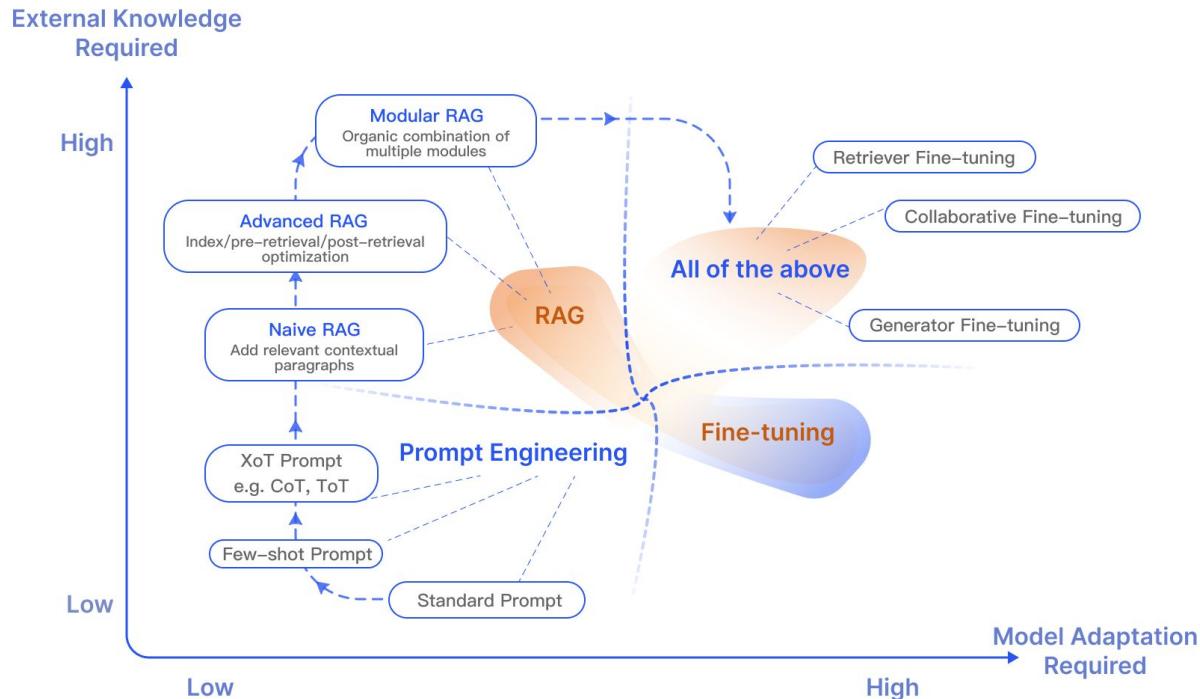
01

Le futur du RAG ?

RAG ou Fine-Tuning

C'est RAG et fine tuning !

There is a better way



src : <https://github.com/Tongji-KGLLM/RAG-Survey>



Mais le RAG demain ?

Une **méthodologie modifiée** :

- Les “**LLM text**” s’améliorent : (le Retriever peut être moins bon...)
- Leur **taille du contexte augmente** (~128k GPT4 , 32k Mixtral) : 10-30 fois plus en un an.
- Les **coûts** de ces **LLM text** se **réduisent** fortement : ~ divisé par 10-20 en un an .

Une **architecture modifiée** : Demain le **LLM text** peux tourner **en local**

Mais l'idée principale devrait rester : **Exposer ses données et des services** à un **agent IA** devient le futur de nos systèmes d'information !



*There
is
a Better
Way*



La Grosse Conf :

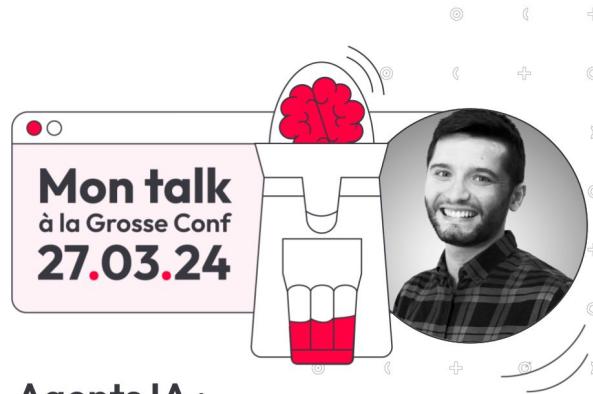
la conférence DATA & IA by OCTO



**RAG : au delà de
la démonstration
simpliste**

Philippe Prados
Pionnier de l'informatique

Infos et Billetterie sur lagrosseconf.com



**Agents IA :
Les Nouveaux Maîtres
de votre système
d'information.**

par Nicolas Cavallo
Head of Generative AI
OCTO Technology



Les Actus OCTO



Comptoir.

06.02.2024
Comptoir OCTO

Comment OCTO transforme la CSRD en un outil stratégique pour un impact positif

Lien d'inscription dans le chat de la conférence ou sur octo.com, catégorie "Événements"



Publication.

Culture Data & Industrie

De l'atelier au Big Data et à l'IA

Lien dans le chat ou sur octo.com catégorie "Ressources"



Formation.

Pour aller plus loin,
découvrez notre formation

Sensibilisation à l'Intelligence Artificielle générative et à ses enjeux

Identifier les opportunités de l'IA générative dans son organisation

Rdv sur octo.academy



Recrutement.

Vous vous sentez l'âme d'un(e) Octo ?
Rejoignez-nous !

Rdv sur rejoins.octo.com ou envoyez un mail à candidature@octo.com