[sfeir]

WENVISION

# L'IA générative pour les CIO et les CTO

## Release 8

**TECHWAVES - #GENAI FOR CIO & CTO**

RELEASE 8 - Avril 2024 - Fiches mises à jour signalées par :

MAJ - R8

# “L'IA GÉNÉRATIVE N'EST PAS UNE MENACE POUR LES DEV, AU CONTRAIRE”

## L’ÉDITO

**Olivier Rafal**
**Consulting Director - Strategy**

olivier@wenvision.com

Vous avez tous lu comme moi que l’IA générative allait augmenter la productivité des développeurs jusqu’à, in fine, supprimer des emplois d’ici peu. Voire le métier même de développeur. Vous y croyez, vous ? Pas moi.

Oui, **l’IA générative augmente considérablement la productivité des développeurs. C’est un formidable assistant, qui va réaliser les tâches les plus rébarbatives** : écrire le code pour requêter une API, ajouter des champs dans une interface Web ou traduire un code d’un langage vers un autre. Des tâches normées, fastidieuses, sans valeur ajoutée.

L’assistant ira éventuellement plus loin, générant une grande partie du code applicatif souhaité, à partir de prompts, voire de schémas à main levée - une implémentation moderne, dopée à l’IA, de l’esprit du Model Driven Development, né au début des années 2000, où on envisageait de concevoir l’ossature d’un logiciel à partir de sa modélisation en UML.

Le MDD n’a jamais décollé. Trop ambitieux, trop rigide, pas suffisamment performant, etc. De par leur souplesse, leur performance et leur simplicité d’utilisation, les outils de développement intégrant de l’IA générative ont eux toutes les chances de réussir.

### POURQUOI L'IA NE SUPPRIMERA PAS DE POSTE DE DÉVELOPPEURS

**Cela ne veut pas dire pour autant que les nouveaux outils de dev intégrant de l'IA diminueront le besoin en développeurs** dans les années qui viennent. Trois réflexions à ce sujet :

- **Il faut piloter l’IA**. L’IA génère du code demandé par le développeur, c’est lui qui comprend les souhaits du PO, qui a l’architecture du logiciel et son contexte d’utilisation en tête, et doit affiner les prompts en fonction de ça.
- **Il faut vérifier le code produit par l’IA**. IA générative ou pas, un développeur vérifie et teste le code. Toujours.
- **Il faut créer toujours plus de code**. Penser que l’augmentation de productivité va supprimer des emplois à court ou moyen terme, c’est croire que nous devons produire une quantité finie de code. Alors qu’aujourd’hui tech et business sont indissociables, que le logiciel est partout et surtout qu’il doit évoluer constamment.

Non seulement les développeurs ont encore de beaux jours devant eux, mais en plus, l’IA générative leur simplifiera grandement la tâche. Pour la plus grande satisfaction des DSI qui la mettront en œuvre.

# LA PAROLE AUX EXPERTS

## Consulting

**Marie Fontaine**
**Head of GenAI**

**marie@wenvision.com**

**La rapidité d'évolution**, à la fois technologique et d'usage, de l'IA Générative est sans précédent. Il est devenu difficile de suivre la production de connaissances dans ce secteur et de comprendre les effets possibles sur nos entreprises. L'histoire commence en 2020, lorsqu'OpenAI ouvre les accès à GPT-3 en Bêta, mais c'est avec ChatGPT que le grand public prend conscience de la puissance de cet outil.

**Ce sentiment de dépassement se confronte à celui de l'urgence. Les gains de productivité attendus sont tels qu'aucune entreprise ne peut ignorer cette technologie.** En particulier lorsqu'on parle de développement logiciel. Une étude menée par GitHub montre que les développeurs assistés par Copilot mettent 56% moins de temps à accomplir une tâche. McKinsey projette des gains de productivité d'au moins 30% pour le software engineering. Cela fait naître de nouveaux défis pour les entreprises : choisir les bons outils, concilier l'adoption de cette technologie et accompagner les collaborateurs.

**Chez SFEIR et WENVISION, nous suivons cette tendance depuis de nombreux mois et nous souhaitons partager notre retour d'expérience au travers des TechWaves - #GenAI 4 Dev.** Toute approche stratégique de l'IA Générative nécessite une acculturation, une feuille de route basée sur des cas d'utilisation réels et la mise en place d'une organisation adaptée.

## Engineering

**Florent Legras**
**Engineering Director**

**legras.f@sfeir.com**

Les LLM offrent une solution qui permet **d'améliorer largement les applications existantes**. Par exemple, offrir aux utilisateurs métier la possibilité de bénéficier réellement d'un self-service BI, en dialoguant. Pour les développeurs, ces outils permettent **d'augmenter leur productivité** et de résoudre certaines problématiques plus efficacement.

Le déploiement d'un LLM en entreprise nécessite de passer d'un simple POC à une application prête pour la production. Cette transition revêt une importance cruciale. Bien qu'un **POC puisse démontrer les fonctionnalités de base d'une application LLM**, il peut ne pas tenir compte de facteurs essentiels tels que la scalabilité, les performances et la sécurité. **Or, le passage d'un POC à une application déployable est un exercice complexe en raison de la nature versatile des LLM.**

**Ces applications doivent être conçues et pensées dans un contexte global, intégrées dans une plateforme data au sein d'un système d'information.** Une architecture bien pensée permet la scalabilité, la maintenabilité et l'extensibilité de l'application. Une application prête pour la production comprend des mécanismes pour surveiller et gérer les modèles. Cela implique de suivre les performances, d'identifier les problèmes, ainsi que de garantir la sécurité et la conformité de l'application. Sans ces fonctionnalités, une application LLM risque de générer plus de désagréments que de satisfaction.

# “CODER AVEC UNE IA GÉNÉRATIVE ? J’AI RELEVÉ LE DÉFI”

## Engineering

**Didier GIRARD**
**Co-CEO SFEIR & WENVISION**

**didier@wenvision.com**

Au cours de l'année écoulée, plusieurs activités ont limité mon temps de programmation, réduisant mes sessions de codage à quelques heures par mois.

Je considère qu'il est crucial de coder régulièrement et, à force de ne pas pratiquer, j'ai commencé à perdre certains réflexes, un peu comme un musicien qui arrête de jouer de son instrument et perd peu à peu sa dextérité.

Cela m'a poussé à relever un défi : **serait-il possible de développer une application basée sur une IA générative, malgré mes compétences un peu rouillées ?**

Pour ce faire, j'ai utilisé Replit et son IA générative, un outil pratique qui ne nécessite aucune installation. Mon objectif était de recréer une application que j'avais développée en Flutter il y a plusieurs années pour compenser ma dyslexie : lorsque je suis fatigué, mon handicap se manifeste et je fais des "fautes de frappe".

L'application Flutter originale traduit mon texte du français à l'anglais, puis de l'anglais au français. Elle est mon assistant personnel. Pour la traduction, elle s'appuie sur l'API de Deepl.

J'aime le langage Go, j'ai donc choisi de travailler avec Replit en Go. L'interface utilisateur a également été conçue par Replit, d'abord en utilisant un formulaire basique, puis en s'appuyant sur Bootstrap.

***“J'ai relevé le défi de coder une application avec Replit et son IA générative. Résultat ? Un développement rapide et efficace ! L'IA ne remplacera pas les développeurs, mais nous permettra de coder plus vite, tout comme le rabot électrique a aidé les menuisiers.”***

Le développement était très simple et a pris environ 2 à 3 heures.

**Enthousiasmé par cette rapidité de développement, j'ai décidé d'ajouter une nouvelle fonctionnalité** à mon assistant : la génération de texte. Pour ceci, je me suis appuyé sur GPT-4 d'OpenAI. L'idée étant qu'avec cette évolution, mon assistant personnel m'aide maintenant dans la formulation de mes textes et ne soit plus un simple relecteur.  Avec Replit, intégrer l'API d'OpenAI  ne m'a pris que quelques dizaines de minutes. Pour l'instant, l'intégration est basique, mais au final, mon assistant correspond à mes besoins.

Cette expérimentation avec Replit a confirmé ce que je soupçonnais. L'arrivée de l'IA générative pour aider les développeurs est une belle avancée. D'ici quelques trimestres, les développeurs les plus efficaces utiliseront ce genre d'outils.

Toutefois, certaines compétences sont nécessaires pour utiliser correctement l'IA générative dans le développement. D'après mon expérience, bien que limitée à quelques heures, je pense que la capacité à déboguer est essentielle. En effet, **l'IA ne génère pas tout parfaitement du premier coup, et être capable d'identifier rapidement les problèmes pour demander à l'IA de générer une nouvelle solution est crucial**.

Je vous encourage à essayer ces outils. Le travail du développeur va changer radicalement avec l'IA, mais elle ne remplacera pas les développeurs. Tout comme le rabot électrique n'a pas remplacé le charpentier, l'IA nous permettra simplement de coder plus rapidement et plus efficacement.

A suivre...

Pour les curieux, voici le repo GitHub : https://github.com/dgirard/wizwizwiz17/tree/main

-> Ce n'est probablement pas le plus beau code, mais disons que c'est une IA qui code depuis moins d'un an qui l'a produit. Dans les années à venir, elle fera mieux !

# Plus vite, plus haut, plus fort

MAJ - R8

## Consulting

**Aurélien Pelletier**
**Platform director**
**aurelien@wenvision.com**

A quelques mois des JO de Paris, la nouvelle génération de LLMs a adopté la devise olympique.

**Plus rapides**

La majorité des nouveaux LLM se déclinent désormais en plusieurs tailles, chacune offrant un ratio coût/vitesse/pertinence adapté à différents cas d'usage. C'est le cas par exemple de Google Gemini 1.0 (nano, Pro, Ultra), d'Anthropic Claude 3 (Haiku, Sonnet, Opus) ou encore de Mistral (Small, Large).

Groq présente de nouveau processeur pour générer du texte dix fois plus rapidement.

**Plus grands volumes de données**

Google Gemini 1.5 annonce une fenêtre de contexte de 1 000 000 de tokens, soit près de 8 fois plus que GPT-4 (128K tokens). Cela permettrait d'ingérer d'un coup l'intégralité du Seigneur des Anneaux, une heure de vidéo ou 30 000 lignes de code.

**Multimodalité : au-delà du texte**

La conversation ne se limite plus seulement à du texte mais peut désormais contenir des images, de l'audio, voir de la vidéo. Cette polyvalence ouvre la voie à de nombreuses applications innovantes, comme l'analyse d'images médicales, la reconnaissance vocale ou encore la génération de contenus multimédias.

**Montée en puissance de l'open source**

Les modèles Open Source gagnent aussi en puissance. Mistral a publié Mixtral, pour Mixture Of Expert, tandis que Databricks a présenté DBRX comme le plus performant des modèles ouverts.

**Plus fort**

Selon la chatbot Arena où Hugging face organise "une bataille des LLMs" Claude 3 Opus a détrôné GPT-4. Et notre usage quotidien confirme cette prouesse.

Désormais tout le monde attend la réponse d'OpenAI et un probable GPT-5...

# L'actu Gen AI de ces dernières semaines

MAJ - R8

## Consulting

**Baptiste Pugnaire**
**Consultant IA / Data**
**baptiste@wenvision.com**

### Bataille au sommet

La course à l'innovation dans le domaine de l'IA générative s'intensifie. Les leaders multiplient les annonces :

- **Anthropic** dévoile Claude 3 (04/03), promettant des performances supérieures à GPT-4.
- **Mistral** annonce son modèle "Large" (26/04), une première avec des poids non open-source, accessible uniquement via API.
- **Meta** avait commencé l'entraînement de LLama 3 en janvier et la sortie des premiers petit modèles serait imminente.
- **Google** annonce Gemini 1.5 (15/02) avec une fenêtre de contexte record, ainsi que les modèles open-source Gemma (21/02).
- Whisper, le modèle TTS **d'OpenAI** est désormais disponible sur Microsoft Azure (13/03). D'autre part, un article du blog officiel annonçant la sortie de GPT-4.5 pour juin aurait leaké (13/03).

Dans cette course effrénée, les innovations se succèdent et chaque acteur cherche à prendre l'avantage. Cependant, pendant que les géants de la tech multiplient les effets d'annonce, les entreprises sont attendues et il est maintenant impératif de démontrer des résultats tangibles pour éviter le "over promise, under deliver".

### OpenAI révolutionne la création vidéo avec Sora

C'était la grosse annonce qui secoue le monde de la création vidéo, Open AI a dévoilé Sora le 16 février dernier. Sora ("ciel" en japonais) est un modèle de génération de vidéo, il permet, sur la base d'un prompt simple, de créer des extraits vidéo photo réalistes et cohérents de plusieurs dizaines de secondes.
Des studios hollywoodiens tels que Warner, Paramount et Universal ont d'or et déjà été conviés à des démonstrations techniques de ce nouveau modèle.

### Go big AND go open source

Jusqu'à récemment, modèle open-source était souvent synonyme de "petite taille" (max 70B paramètres) mais récemment xAI, Databricks et Cohere ont sortis des modèles performants et relativement massifs : respectivement Grok-1 (314B), DBRX (132B) et Command R+ (104B). Ce changement est une excellente nouvelle pour la communauté de recherche et d'innovation en IA.

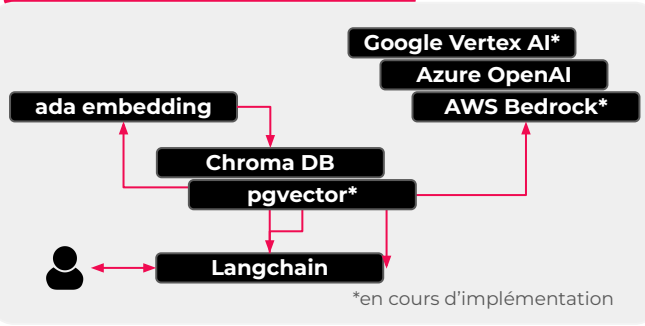
### Microsoft consolide son emprise sur l'écosystème IA

Récemment, Microsoft a recruté Mustafa Suleyman et Karén Simonyan (ex-Inflection AI et Deep Mind) pour dirigé sa branche Microsoft AI. Historiquement partenaire privilégié d'Open AI, l'entreprise **multiplie les accords** : citons notamment Mistral AI dont le dernier modèle n'est accessible que sur le cloud Microsoft.

# RETOUR D'EXPÉRIENCE : VEOLIA SECURE GPT

MAJ - R5

## L'ARCHITECTURE MISE EN PLACE




*en cours d'implémentation

*"C'est une révolution dans l'IT.*

*Il y aura forcément des déceptions, des retours en arrière sur tel ou tel PoC, mais globalement, la GenAI va drastiquement changer notre façon de travailler et de développer des logiciels.*

*Notre devoir est de maîtriser ces nouvelles technos pour accompagner la transition de Veolia."*

Au printemps 2023, la direction IT groupe de Veolia a décidé la création de Secure GPT, avec l'objectif de maîtriser cette technologie et d'éviter l'usage d'outils grand public. L'ambition était de l'ouvrir d'abord aux 5 000 collaborateurs IT du groupe. Au vu des résultats (taux d'adoption, scalabilité, maîtrise des coûts…), l'application est désormais accessible aux quelque 200 000 collaborateurs de Veolia dans le monde, avec plusieurs mois d'avance sur le calendrier.

Secure GPT offre aujourd'hui 3 services :

- un chat conversationnel, dont le niveau de créativité peut être ajusté ;
- la traduction en 95 langues ;
- la possibilité d'uploader des PDF et fichiers textes pour les interroger en chat (le temps d'une session).

Les premiers développements ont utilisé Google Cloud pour le front et Azure OpenAI pour le LLM et l'embedding, orchestrés via Langchain sur Google Cloud Run. Veolia travaille aujourd'hui à étendre sa plateforme, avec l'implémentation d'AWS Bedrock (Llama2, Claude2, Stability) et de Google Vertex AI (Gemini Pro 1.0 [màj Février 2024]), ainsi que la mise en œuvre de pgvector dans PostgreSQL, ce qui offrira de la persistance pour interroger les bases de connaissance Veolia.

A terme, l'enjeu pour Veolia sera d'investir sur ce qui apportera de la valeur aux collaborateurs et ne risquera pas de se retrouver dans la roadmap des éditeurs SaaS. En d'autres termes, Veolia n'ira pas optimiser un CRM, mais se focalisera par exemple sur des solutions spécifiques au bon fonctionnement de ses usines.

## Les enseignements de Fouad Maach, Head of Architecture and Industrialization for Veolia Group IS&T

**Quelles sont les compétences requises ?**

Une personne bien choisie peut faire le boulot mais il faut très vite paralléliser : on peut démarrer avec une très petite équipe pour initier le développement, mais la montée à l'échelle demande la création d'une vraie organisation, où on va faire monter les gens en compétences.

On n'a pas forcément besoin d'être data scientist, mais il faut comprendre les concepts de base, avoir des compétences de développement back et de fortes compétences Cloud - et y ajouter des compétences front et DevOps pour le passage à l'échelle. Les développeurs doivent aussi s'adapter à un changement de paradigme : on passe d'une programmation algorithmique classique à des agents Langchain qui gèrent une partie des décisions : c'est un changement à faire accepter.

**Quels autres conseils pouvez-vous donner ?**

Soigner la partie juridique : l'équipe "legal" nous a grandement aidé. Les choses auraient été problématiques sans leur validation. Il faut aussi accompagner les utilisateurs : on croit que tout le monde sait utiliser de la GenAI, mais nous nous sommes rendu compte qu'il y avait un besoin de formation ; nous allons déployer un module de e-learning.

**Langchain est au cœur de l'architecture ; le framework a donc donné satisfaction ?**

Oui, il est simple à utiliser et très puissant. Il répond bien aux attentes d'agnosticité : on peut changer de LLM sans souci. Il y a une petite courbe d'apprentissage, mais après, ce n'est que du bonheur !

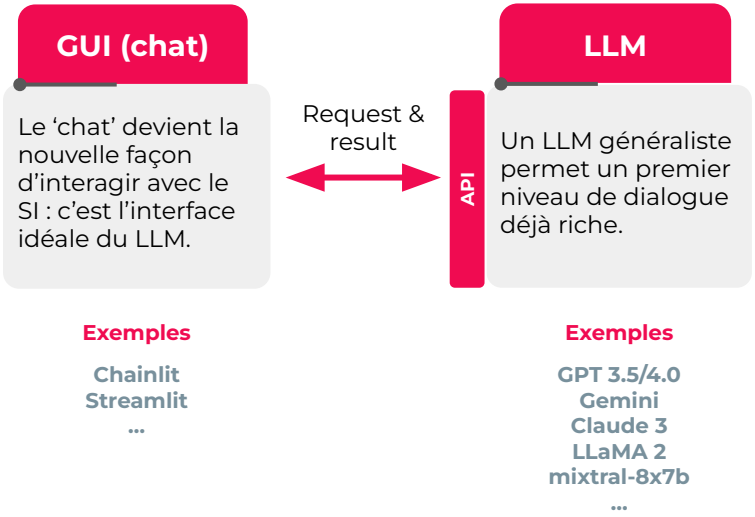
# TECHWAVES #GENAI FOR DEV, MODE D'EMPLOI

[sfeir]

WENVISION

# COMMENT LES TECHNOS SE POSITIONNENT ENTRE ELLES : NOTRE VISION D'UNE "POC LLM APP STACK"

MAJ - R8



# DÉMARRER AVEC UN POC

MAJ - R6

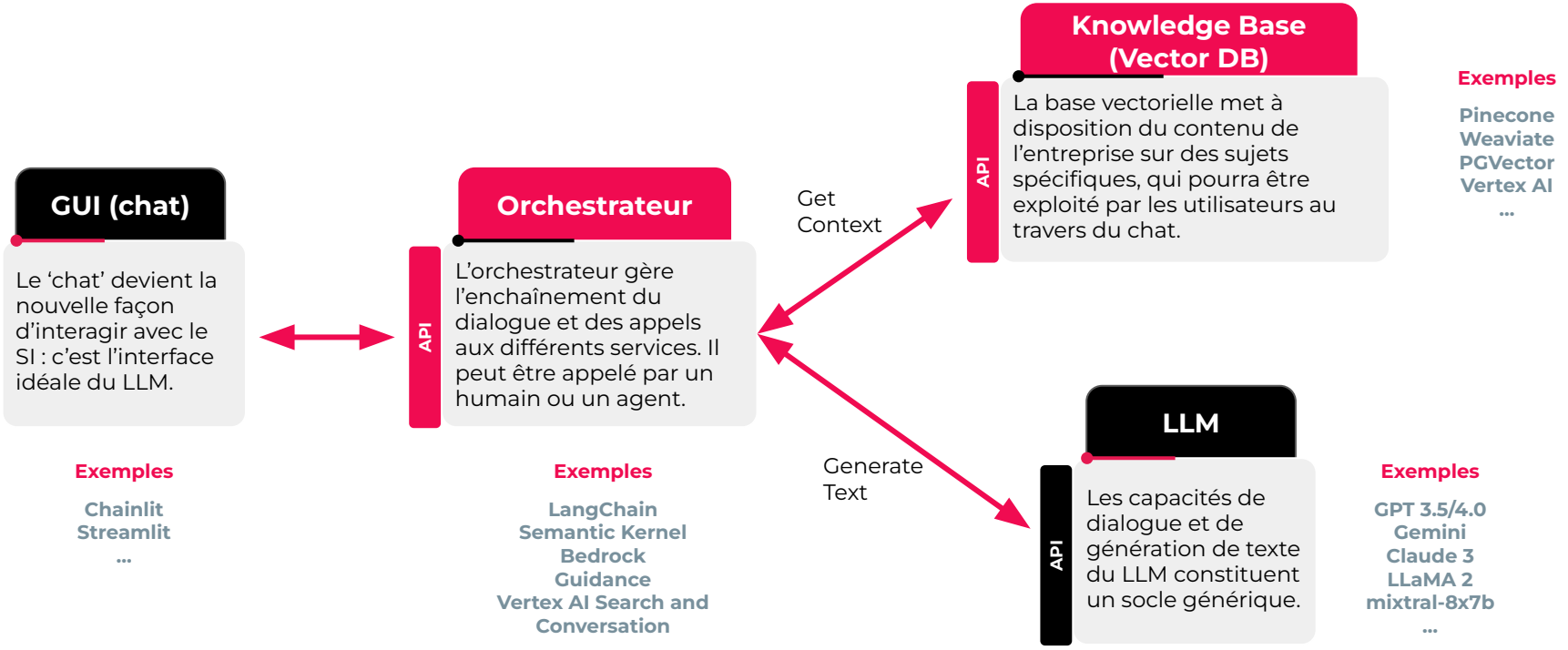
Le premier enjeu pour une entreprise est de pouvoir mettre à disposition de ses salariés une version sécurisée d'un agent conversationnel type ChatGPT. Pour ce faire, elle doit déployer une instance de LLM généraliste dans son environnement et le connecter à une interface de type chat. Cette première étape permet de valider de premières hypothèses telles que le modèle de langage le plus approprié, ou encore l'ajustement de certains paramètres qui jouent sur la créativité du modèle (la température par exemple). Il s'agit bien d'un PoC, avec une population utilisatrice restreinte.

Cette première étape est également un moyen pour les entreprises de collecter les cas d'usage des collaborateurs à travers les questions posées, et ainsi d'améliorer les prochaines versions de ce produit grâce à diverses techniques de fine tuning.

Le risque de fuite des données vers l'extérieur est levé, néanmoins le travail sur la pertinence des informations transmises par le LLM reste à approfondir dans les cycles de développement ultérieurs.

# COMMENT LES TECHNOS SE POSITIONNENT ENTRE ELLES : NOTRE VISION D'UNE "MVP LLM APP STACK"

MAJ - R5



# SE FORGER DE PREMIÈRES CONVICTIONS SUR UN LLM CUSTOMISÉ AVEC LES DONNÉES DE L'ENTREPRISE

MAJ - R6

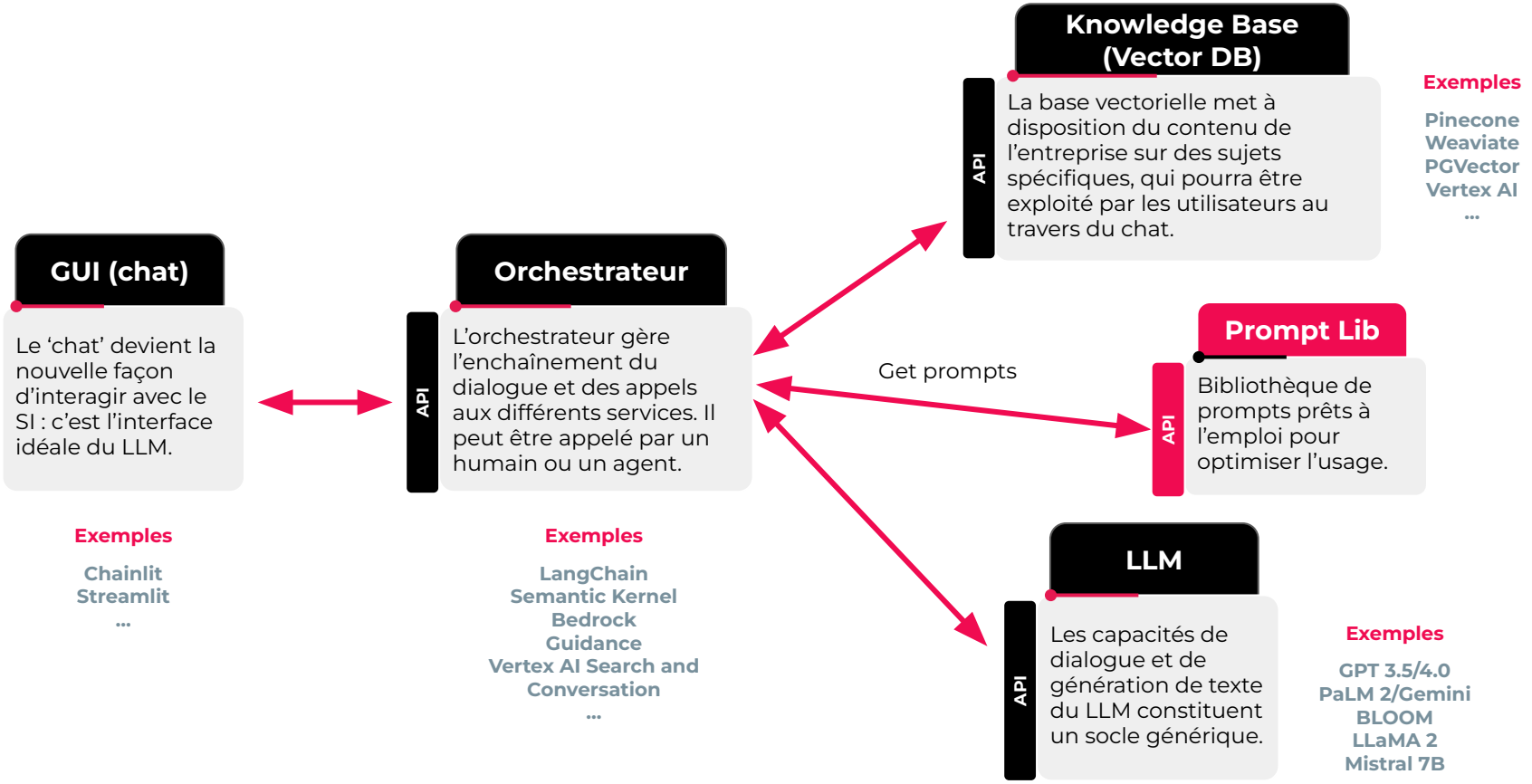
L'étape du MVP a pour objectif principal de personnaliser le contenu généré avec des données de l'entreprise. Pour ce faire, deux nouveaux composants entrent en jeu dans cette architecture : les bases de données vectorielles et un orchestrateur.

La base de connaissance, encapsulée dans les bases de données vectorielle, est externalisée du LLM qui n'a pour fonction que celle de la rédaction, synthèse d'un texte en langage naturel. L'orchestrateur assure l'enchaînement entre les différents services pour assurer la continuité du dialogue. L'orchestrateur peut également conserver les sources d'information et les restituer à l'interface, ce qui permet à l'utilisateur de retrouver le document sur lequel la réponse a été construite.

Cette étape permet d'expérimenter l'articulation entre un corpus de données sélectionné par l'entreprise et un LLM. La qualité des résultats reposera ici grandement sur les prompts formulés en entrée, dans l'interface, et donc sur les compétences de prompt engineering des utilisateurs.

# COMMENT LES TECHNOS SE POSITIONNENT ENTRE ELLES : NOTRE VISION D'UNE "CLASSIC LLM APP STACK"

MAJ - R5



# TIRER LES BÉNÉFICES DU PROMPT ENGINEERING AVEC UNE CLASSIC LLM STACK

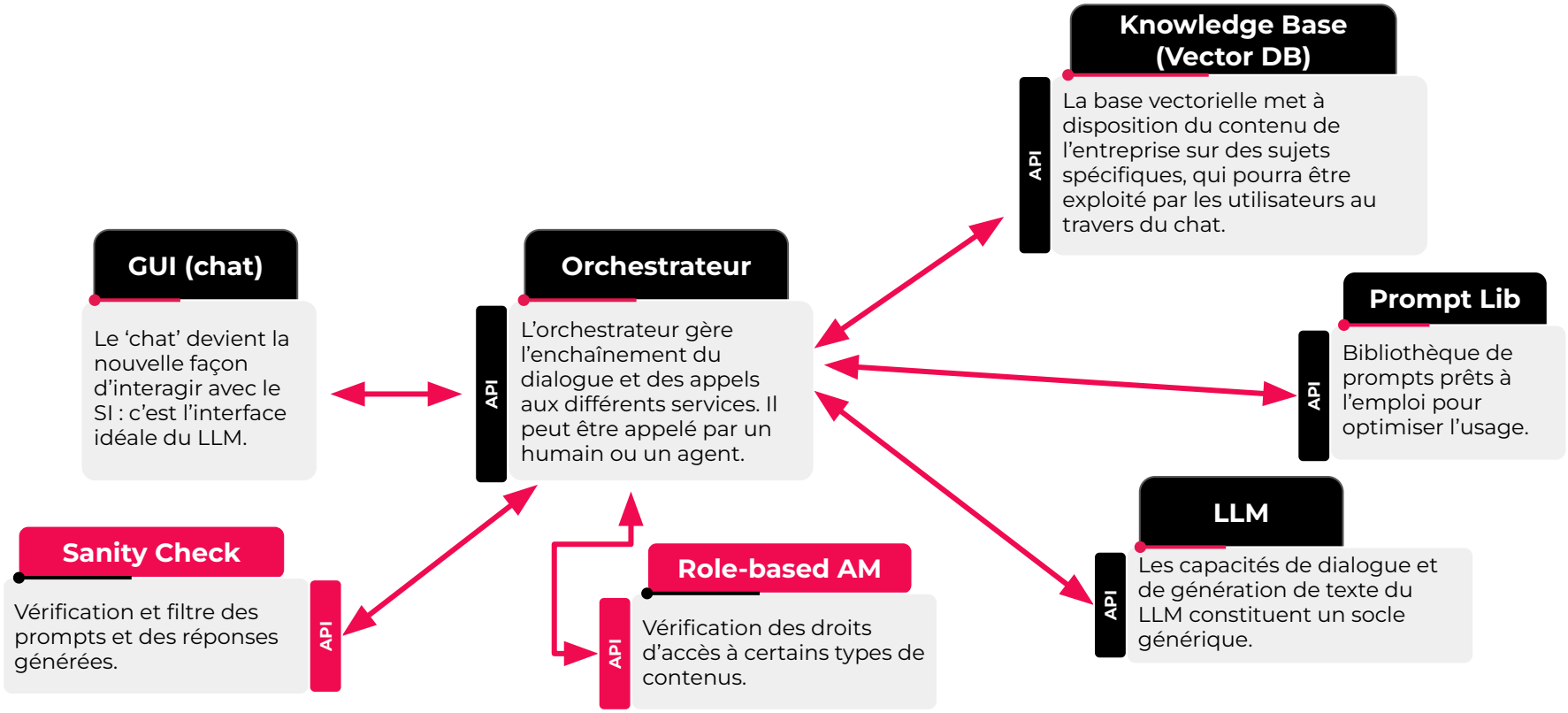
MAJ - R6

L'intelligence est dans le prompt. L'ajout majeur ici est donc le composant de bibliothèque de prompt ou "Prompt lib" qui répertorie des prompts prêts à l'emploi et optimisent ainsi l'usage du LLM. Le rôle de prompt engineer prend toute son ampleur lors de cette phase. Il aura pour mission de formuler des prompts de haute qualité pour obtenir les meilleurs résultats, en veillant à ce que l'IA génère des réponses précises, pertinentes et adaptées aux besoins spécifiques.

En intégrant ce composant après avoir pu analyser les premiers usages de l'application par les utilisateurs, les prompt engineers pourront se concentrer sur les cas d'utilisation à plus forte valeur ajoutée.

# COMMENT LES TECHNOS SE POSITIONNENT ENTRE ELLES : NOTRE VISION D'UNE "ADVANCED LLM APP STACK"

MAJ - R5



# TIRER LES BÉNÉFICES DU PROMPT ENGINEERING AVEC UNE BASIC LLM STACK

MAJ - R6

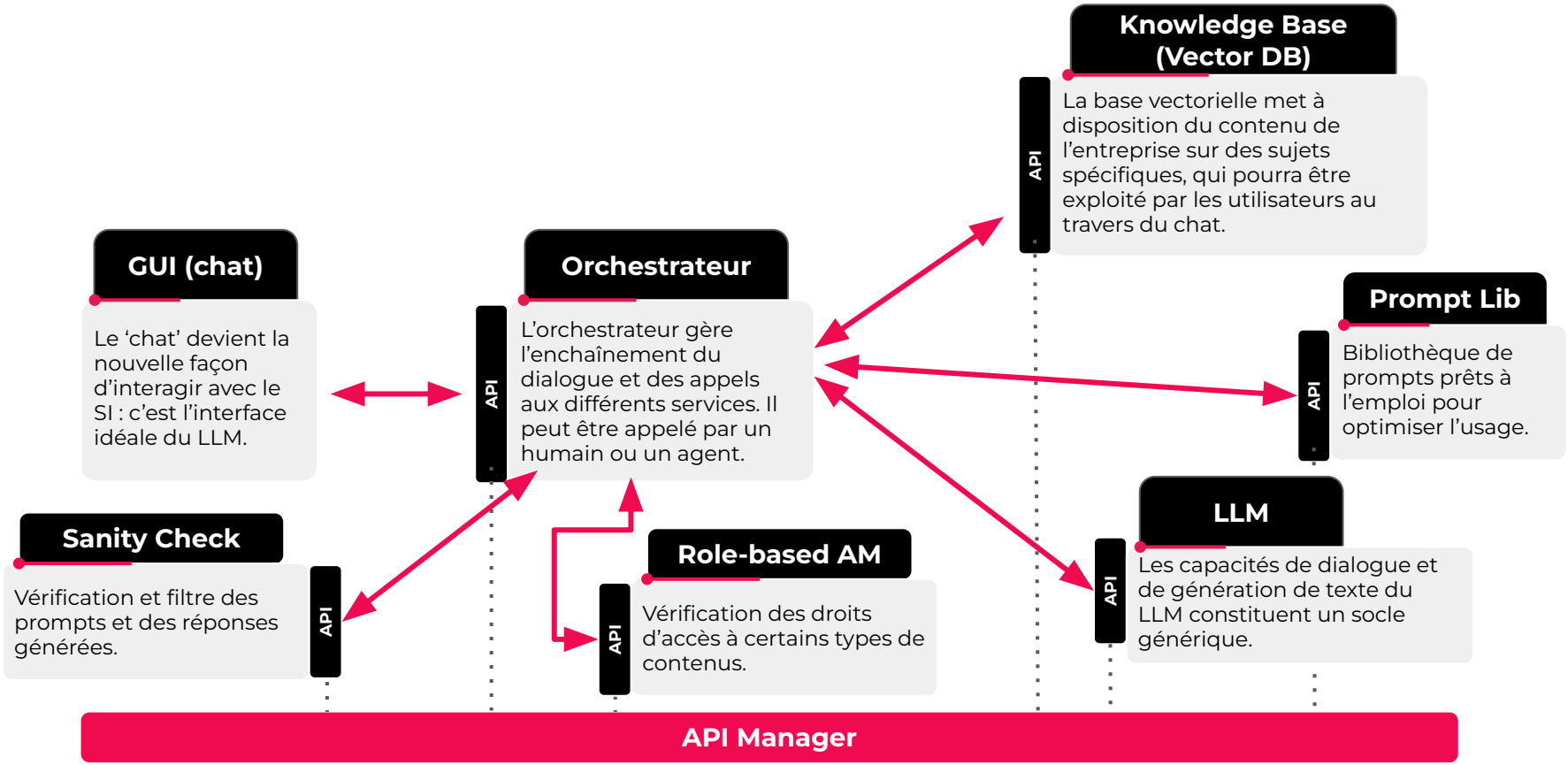
L'architecture d'une application basée sur des LLM doit pouvoir répondre aux défis posés par son industrialisation et son utilisation à l'échelle, en particulier en termes de gouvernance et de conformité.

Le contrôle de la diffusion des données ne doit pas s'arrêter à l'externe, il est impératif d'appliquer les règles de gestion des accès à la donnée pratiquées par l'entreprise au sein de l'application LLM. C'est pourquoi l'orchestrateur doit pouvoir interroger l'access management tool des données dans l'entreprise.

L'éthique et la conformité sont aussi des enjeux majeurs en phase d'industrialisation d'une telle application. Un composant type "sanity check" permet de s'assurer du respect des normes éthiques de l'entreprise, en entrée et en sortie du modèle. Il permet également de suivre les performances de qualité du modèle sur le long terme.

# COMMENT LES TECHNOS SE POSITIONNENT ENTRE ELLES : NOTRE VISION D'UNE "ENTREPRISE LLM APP STACK"

MAJ - R5



# Créer des Assistants pour encapsuler et maîtriser la complexité

MAJ - R8

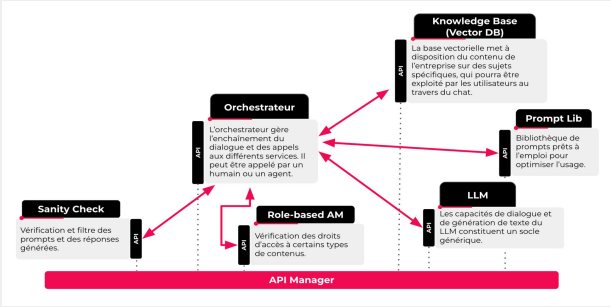
**GUI (chat)**

Le ‘chat’ devient la nouvelle façon d’interagir avec le SI : c’est l’interface idéale du LLM.

API

**Assistants**

L’assistant regroupe l’ensemble des paramètres à configurer et des données à utiliser pour répondre à un cas d’usage spécifique (ex: étudier des documents juridique, gérer la relation client, résoudre un incident, etc.).



MAJ - R6

# LA PLATEFORME GEN AI : 3 SERVICES DE BASE ET LA CAPACITÉ D'INDUSTRIALISER DES ASSISTANTS SPÉCIALISÉS

Une plateforme GenAI permet de répondre aux besoins en matière d'IA générative de la manière la plus efficace possible. Elle ouvre un monde de possibilités pour innover et améliorer les opérations de l'entreprise, en gardant ouvertes toutes les options possibles en matière de fournisseurs, de modèles, de cas d'usage ou encore d'intégration avec le SI et les processus de l'entreprise.

## Assistant Chat Généraliste

- Chat en langage naturel via une interface connectée à un LLM.
- Cet assistant permet aux collaborateurs de bénéficier d'un assistant très généraliste, non spécialisé au contexte de l'entreprise, pouvant répondre à des use cases tels que : rédiger un compte-rendu de réunion, formuler un email, réaliser une synthèse à partir d'un texte.

## Assistant documentaire dynamique

- Chat en langage naturel via une interface connectée à un LLM avec possibilité de charger un document dans la conversation. Le document ne sera pas conservé de manière persistante, l'assistant ne pourra interagir avec que pendant la durée du chat.
- Ce cas d'usage permet de bénéficier des capacités d'un LLM pour extraire et structurer de l'information textuelle issue d'un document.

## Assistant spécialisé, préconfiguré

- Assistant en langage naturel spécialisé sur un vertical métier, répondant à un ou plusieurs use cases spécifiques.
- Cet assistant sera pré-configuré par un administrateur via une interface dédiée. Cette interface permettra de sélectionner le LLM, de pré-rédiger des instructions et des prompts récurrents, et de charger un corpus documentaire tel qu'une base de connaissance ou de se connecter à une source telle qu'un CRM.

# APPLICATION DES PRINCIPES #GENAI AUX SGBD (MVP)

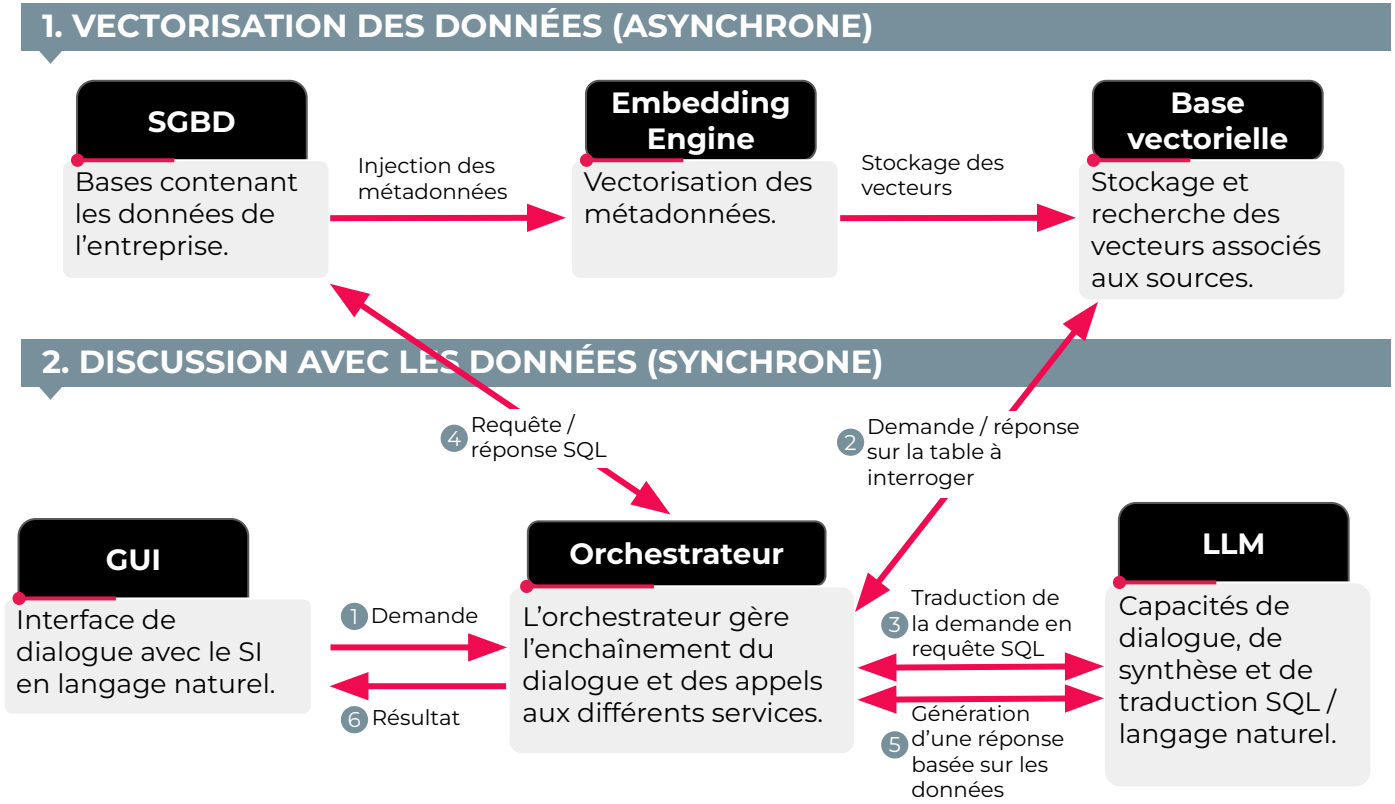
MAJ - R5

Le dialogue est **une forme idéale pour interroger les données de l'entreprise**. Il s'agit non plus de requêter des données, mais de collaborer avec un assistant qui va répondre à des demandes formulées en langage naturel.

Une telle architecture se décompose en 2 grands domaines : d'une part la préparation des données, qui s'effectue de manière asynchrone, d'autre part la gestion du dialogue, en temps réel.

**Attention cependant** à cette approche :

- elle demande que les bases de données soient correctement renseignées et les métadonnées intelligibles ;
- elle introduit des risques de sécurité (SQL injection).



# FINE-TUNING VS RETRIEVAL AUGMENTED GENERATION (RAG)

MAJ - R5

## FINE-TUNING : DÉFINITION

Le fine-tuning consiste à adapter un LLM à des tâches spécifiques ou à des domaines particuliers. Le processus implique de prendre un modèle pré-entraîné et de le réentraîner sur un ensemble de données plus restreint et ciblé, souvent constitué de textes spécifiques à la tâche envisagée. En ajustant les poids et les paramètres du modèle d'origine, le fine-tuning permet au LLM de mieux comprendre et de générer des réponses plus précises et adaptées à des contextes particuliers.

## CAS D'USAGE DU FINE-TUNING

- Adapter un modèle à un contexte particulier, par ex. :
  - aider un développeur,
  - adopter des tournures de phrases juridiques ou littéraires,
  - comprendre du vocabulaire propre à un domaine métier, une industrie…

## RAG: DÉFINITION

Le "RAG" tire parti des capacités de recherche intégrées aux LLM pour interroger une base de données vectorielle (voir fiche associée) ou des documents afin d'en extraire des informations pertinentes en réponse à une requête donnée. Ces informations récupérées sont ensuite utilisées pour enrichir la génération de texte, offrant ainsi une réponse plus contextuelle et informée, s'appuyant sur les connaissances de l'entreprise et plus seulement sur les connaissances générales du modèle.

## CAS D'USAGE DU RAG

- Enrichir le modèle avec les données de l'entreprise (tout en préservant le contrôle d'accès et la possibilité de supprimer ou mettre à jour le contenu), par ex. :
  - données financières,
  - descriptifs produits,
  - documentation,
  - framework technique…

## LES PLUS

- Capacité à adapter le modèle à son contexte (par ex. le retail, le BTP, le droit, etc.) ou à sa tâche (répondre à un client, donner des instructions, etc.)

## LES MOINS

- Processus relativement coûteux et lent
- Quasi-impossibilité de faire oublier à un modèle quelque chose qu'on lui a appris (sauf fine-tuning encore plus poussé et très coûteux)
- Pas de contrôle d'accès : toutes les informations apprises par le modèle deviennent accessibles à l'ensemble des utilisateurs du modèle

## LES PLUS

- Processus rapide (possibilité de mises à jour en continu : cf. les LLM qui s'appuient sur Google ou Bing pour ajouter du contenu d'actualité à leurs réponses)
- Accès aux documents et données de l'entreprise
- Possibilité d'un contrôle granulaire sur l'accès à l'information selon les rôles
- Peut se contenter d'un LLM générique

## LES MOINS

- N'a pas d'impact sur la manière qu'aura le LLM de répondre : ce sujet devra être pris en compte dans le prompt (besoin de prompt engineering)

## CONCLUSION

### LE RAG POUR LE FOND, PUIS LE FINE-TUNING POUR LA FORME ET LES SPÉCIFICITÉS DE L'ENTREPRISE

Le Fine-Tuning paraît séduisant, et il est poussé par les éditeurs, toutefois il s'avère complexe et dangereux à manipuler. S'il s'agit d'améliorer la pertinence des réponses, le RAG sera plus rapide et simple à mettre en œuvre et génèrera tout de suite de la valeur en complément de modèles légers, dont le rôle sera d'interagir en langage naturel. Le Fine-Tuning pourra intervenir plutôt sur la forme des interactions et sur le vocabulaire spécifique à l'entreprise.

# RETRIEVAL AUGMENTED GENERATION (RAG) : LES BONNES PRATIQUES

MAJ - R6

## NIVEAU DE BASE

**1. Optimisation du prompt et priorité à la simplicité**
Employez des techniques de prompt engineering efficaces pour conditionner le modèle, tout en commençant par des prompts simples.

**2. Détermination de la taille idéale des chunks**
Identifiez la taille optimale des morceaux de données (chunks) pour garantir une récupération et un traitement efficaces.

**3. Utilisation de résumés pour la concision**
Appliquez des techniques de résumé aux chunks de données afin de fournir au modèle une représentation concise de l'information.

**4. Gestion rigoureuse des données**
Portez une attention méticuleuse à la gestion, à la vérification, à la version et au nettoyage des sources de données et des pipelines. La qualité des données est essentielle pour obtenir un RAG de qualité.

**5. Évaluation régulière et adaptation continue**
Évaluez fréquemment la performance de la récupération et de la génération du contenu en utilisant des métriques spécifiques au cas d'utilisation. Restez ouvert à l'adaptation et à l'amélioration constante, tout en considérant des alternatives plus simples aux bases de données vectorielles, comme PGVECTOR.

## NIVEAU INTERMÉDIAIRE

**1. Filtrage métadonnées et pertinence**
Ajoutez des métadonnées aux chunks pour faciliter le traitement des résultats.

**2. Gestion des embeddings**
Élaborez des stratégies pour gérer les documents fréquemment mis à jour ou nouvellement ajoutés.

**3. Fiabilité et confiance**
Assurez l'exactitude et la fiabilité du contenu généré en utilisant des citations et en appliquant des techniques telles que l'estimation de la confiance, la quantification de l'incertitude et l'analyse des erreurs.

**4. Techniques de recherche hybride**
Intégrez différentes techniques de recherche, comme la recherche fondée sur des mots-clés et la recherche sémantique.

**5. Transformation de requête et arbitrages**
Explorez des stratégies telles que les embeddings de documents hypothétiques, la décomposition de requêtes en sous-requêtes, et l'évaluation itérative des requêtes pour combler les lacunes d'information. Tenez compte des compromis entre la précision, le rappel, le coût et le calcul pour optimiser le processus de récupération et de génération. Expérimentez également avec des stratégies de chunking avancées et le re-classement des documents récupérés pour améliorer les performances globales.

## NIVEAU AVANCÉ

**1. Affinage du modèle et des embeddings**
Affinez le modèle en continuant l'entraînement sur un ensemble de données plus spécifique ou en optimisant les embeddings pour mieux représenter les relations entre les données.

**2. Personnalisation des embeddings**
Personnalisez les embeddings en créant une matrice pour mettre l'accent sur les aspects pertinents du texte pour vos cas d'utilisation.

**3. Routage de requêtes**
Utilisez plus d'un index ou d'outil et dirigez les sous-requêtes vers l'index ou l'outil approprié.

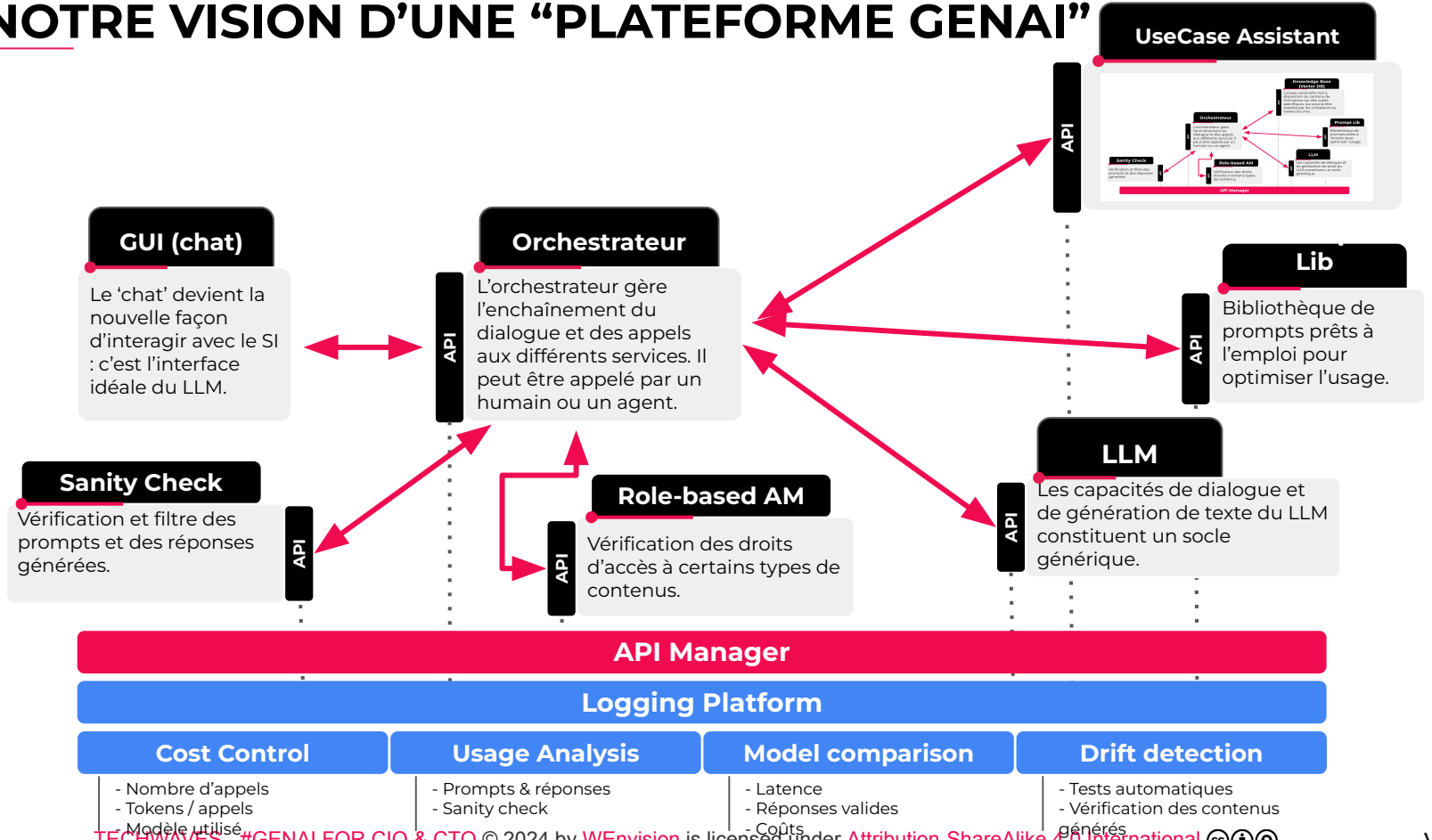
**4. Multi-récupération**
Combinez les résultats de plusieurs agents de récupération (et de génération) pour améliorer la qualité et la fidélité globales.

**5. Compression contextuelle et auto-interrogation**
Appliquez des techniques de compression pour réduire la taille du contexte tout en maintenant sa pertinence. Utilisez l'auto-interrogation en utilisant la sortie du modèle comme une requête pour récupérer davantage d'informations et améliorer la réponse initiale.

Pour plus d'informations, voir le post X (anciennement Twitter) de Shyamal Anadkat, https://twitter.com/shyamalanadkat/status/1746645405276975581 [dernier accès le 17.01.2023]

# COMMENT LES TECHNOS SE POSITIONNENT ENTRE ELLES : NOTRE VISION D'UNE "PLATEFORME GENAI"

MAJ - R7



# TAKING LLMs TO PRODUCTION

MAJ - R7

## Sanity Check

Il est primordial, pour un usage professionnel des LLMs et de la GenAI en général, de surveiller le contenu généré. Il est essentiel de garantir que les utilisateurs ne sont pas confrontés à des contenus offensants ou qui ne respectent pas les règles de l'entreprise. **Des sanity checks sont intégrés dans certaines plateformes, comme Azure Open AI ou Google Vertex, qui proposent par défaut une analyse du contenu**. Il est également possible d'ajuster le niveau d'alerte en fonction de certains critères. Le niveau d'alerte doit être adaptable en fonction des uses case et être monitoré.

## Cost control and usages

La surveillance des algorithmes de GenAI, comme toute technologie, est essentielle pour comprendre pleinement les coûts impliqués. En outre, le logging des utilisations de l'application est cruciale. Les logs permettent une analyse rétrospective des KPI techniques (nombre de requêtes, tokens, latence, etc.) ainsi que business (types de demandes, fréquences d'utilisation, validité de la génération, etc.). **Pour choisir les modèles appropriés, il est nécessaire d'avoir des logs exhaustives pour calculer un rapport adapté entre coût, latence et pertinence pour chaque cas d'utilisation**.

## Model comparison

Avec la multiplication des modèles disponibles et la rapidité du développement de nouveaux modèles, il est crucial de maintenir une connaissance actualisée des capacités de ceux ci. En fonction des besoins spécifiques, **il est nécessaire de choisir judicieusement entre les outils disponibles, en tenant compte des latences attendues et de la pertinence des réponses**. Tous ces critères orientent les entreprises vers l'outil le plus adapté. Les logs permettent de prendre du recul par rapport à l'utilisation et aident à guider ce choix. **Les architectures doivent être indépendantes des modèles et permettre une transition facile vers de nouveaux modèles pour suivre l'évolution du domaine.**

## Drift detections

Une fois que les modèles sont déployés en production, il est impératif de surveiller leur utilisation et surtout la pertinence de leurs réponses. **Certains modèles peuvent évoluer avec le temps, il est donc essentiel de garantir leur cohérence par rapport à l'utilisation prévue**. Plusieurs approches peuvent être envisagées, telles que l'exécution régulière d'un jeu de tests avec une vérification des résultats, pouvant être réalisée à l'aide de différentes méthodes (embedding, autres LLM, température à 0, etc.). Une autre solution consiste à figer la version du modèle pour garantir qu'elle ne changera pas, bien que cette fonctionnalité ne soit pas disponible sur toutes les plateformes.

# FINOPS POUR LA GENAI

MAJ - R7

Les **coûts** liés à l'IT ont toujours été au **centre des préoccupations des entreprises,** mais la notion de **FinOps a émergé** avec l'arrivée du **modèle cloud**, où **l'usage *ET* l'architecture** des **solutions** ont un impact direct sur les coûts facturés, du fait de la variabilité de ceux-ci.
Lors du déploiement d'une plateforme GenAI, les aspects liés au coûts induits sont multiples, mais on peut imaginer un **framework** permettant d'évaluer non seulement l'impact des initiatives GenAI, leur viabilité financière, les options relatives à l'optimisation qui s'offrent aux équipes, mais aussi les coûts spécifiques à chaque type de déploiement. Il est notable, que de la même manière que l'IA générative est relativement récente, les **notions de finops pour GenAI sont, elles aussi, en cours d'établissement.**

### GEN AI Delivery

Dans ce segment du framework, on regroupe les aspects liés à la mise à disposition effective de modèles d'IA, aux campagnes d'enablement et de formation des équipes à l'IA.
**S'assurer que les collaborateurs, quel que soit leur usage de l'IA Generative (consommateur ou créateur de plateforme) ont les connaissance nécessaires pour en tirer partie de sereinement.**

### GEN AI Platform Cost Allocation

Ici l'accent est mis sur la manière dont les coûts associés à la plateforme GenAI sont répartis. Il s'agit notamment de cartographier précisément les dépenses et de répartir les coûts des services partagés. **De cette étape dépendra la capacité d'une entreprise à comprendre les coûts de sa plateforme genAI et à prendre ses décisions de manière éclairée.**

### GEN AI Cost Optimisation

Dans le cas où un modèle est entrainé ou amélioré (augmenté), **il est essentiel d'optimiser les coûts de manière continue, afin d'améliorer l'efficience du process, sans sacrifier la qualité**. De la même manière, on peut inclure dans cette section toutes les étapes annexes liées à l'amélioration de la qualité des données servant à nourrir l'implémentation.

### GEN AI Pricing model & Valeur Business

La connaissance approfondie des valeurs obtenues et actions menées dans une approche d'optimisation permet de **qualifier la viabilité financière des solutions de genAI** déployées, et permet de **valider (ou non) les investissements réalisés** dans ce domaine.

**Human Cost** + **Finops** = **Business Value**

# CINQ BONNES PRATIQUES À ADOPTER POUR ÉVITER LE SHADOW AI

MAJ - R6

**Le Shadow AI désigne l'utilisation d'outils et de technologies d'intelligence artificielle au sein d'une organisation sans l'autorisation ou le contrôle formel de l'IT ou de l'entreprise.** Cela se produit lorsque les employés ou les unités commerciales décident d'adopter des solutions d'IA en dehors du cadre établi par l'entreprise pour l'adoption de ces technologies.
**Cette utilisation non autorisée d'IA peut prendre différentes formes**, telles que l'utilisation de logiciels d'IA tiers non approuvés, l'accès à des services d'IA fondés sur le cloud sans l'approbation de la direction informatique, ou l'utilisation d'outils d'IA générative comme ChatGPT / Gemini / Perplexity pour accomplir diverses tâches professionnelles.
Les raisons de l'émergence du Shadow AI sont souvent liées à la recherche d'efficacité et à la nécessité de répondre rapidement à des besoins spécifiques. Cependant, **cette pratique peut présenter des risques importants pour l'entreprise**, notamment **en matière de sécurité des données, de conformité réglementaire et de gouvernance informatique.**

## 5 BONNES PRATIQUES À ADOPTER POUR ÉVITER LE SHADOW AI

### MISE EN PLACE D'UNE PLATEFORME GEN AI INTERNE

Investissez dans le développement d'une plateforme interne spécifiquement conçue pour répondre aux besoins de votre entreprise en matière d'IA générative.

### COMMUNICATION OUVERTE

Encouragez la communication ouverte entre les départements pour comprendre leurs besoins. Les ressources disponibles doivent répondre aux exigences, réduisant ainsi le recours à des outils d'IA non approuvés.

### PROCESSUS D'APPROBATION RAPIDES

Simplifiez les processus d'approbation des nouveaux outils d'IA pour éviter que les employés ne se tournent vers le Shadow AI en raison de la lenteur des approbations.

### EDUCATION ET FORMATION

Sensibilisez les employés aux risques associés au Shadow AI grâce à des sessions de formation régulières, afin que chacun comprenne les implications de l'utilisation non autorisée d'IA.

### DÉFINIR DES POLITIQUES D'ACCÈS À L'IA GÉNÉRATIVE

Établissez des politiques claires sur l'accès des employés aux outils d'IA générative, déterminant où, quand et comment ils peuvent les utiliser, en tenant compte des avantages et des risques potentiels pour la sécurité et la confidentialité des données.

# QUELLE GOUVERNANCE POUR UNE IA RESPONSABLE ?

MAJ - R8

**Data Governance**

Assurer la qualité, la sécurité, la fraîcheur et l'utilisabilité des données à travers l'organisation

- Qualité des Données
- Accès et Confidentialité
- Connaissance du patrimoine
- Architecture des Données
- ...

&

**AI Governance**

Assurer que l'utilisation des systèmes d'IA soit éthiquement responsable, transparente, sécurisée, et conforme aux réglementations

- Sécurité des systèmes d'IA
- Transparence et explicabilité
- Éthique et responsabilité
- Supervision humaine
- ...

- Rôles et responsabilités
- Conformité Réglementaire
- Formation & sensibilisation
- Process & standards

# LA GOUVERNANCE DES DONNÉES JOUE UN RÔLE MAJEUR DANS L'AMÉLIORATION DE L'EFFICACITÉ

MAJ - R5

Une solide gouvernance des données n'est pas seulement un accessoire de la GenAI, mais un **rouage fondamental de sa fonctionnalité et de sa valeur.**

## QUALITÉ

**L'efficacité d'un modèle dépend directement de la qualité des données d'entrée**. Une gouvernance des données efficace peut maintenir et améliorer la qualité des données en veillant à ce que les données correctes, pertinentes et cohérentes soient récupérées et utilisées dans le processus de génération.

## PERTINENCE

Un framework de gouvernance de la donnée doit garantir que les données les plus pertinentes sont extraites pour la tâche à accomplir. Par exemple, **des données correctement cataloguées peuvent aider à identifier rapidement les bons ensembles de données** pour le fine tuning.

## COHÉRENCE

Il est essentiel que les données utilisées dans les modèles soient cohérentes. Les stratégies de gouvernance des données garantissent cette **cohérence entre les différentes sources de données**, ce qui rend les résultats plus fiables, notamment pour le processus de RAG.

## ACCÈS

Si un modèle accède à des données sensibles et les utilise, la gouvernance des données est essentielle pour garantir que ces informations soient traitées correctement, afin d'**éviter les violations, les fuites ou d'autres utilisations abusives de données sensibles.**

## DISTORSION

Une gouvernance efficace permet d'identifier et de réduire les biais dans les données utilisées dans les modèles. **Les biais peuvent conduire à des phénomènes d'hallucinations, où le modèle génère des informations qui peuvent être erronées ou inappropriées**. La gouvernance permet de s'assurer que les données sont équilibrées et représentative.

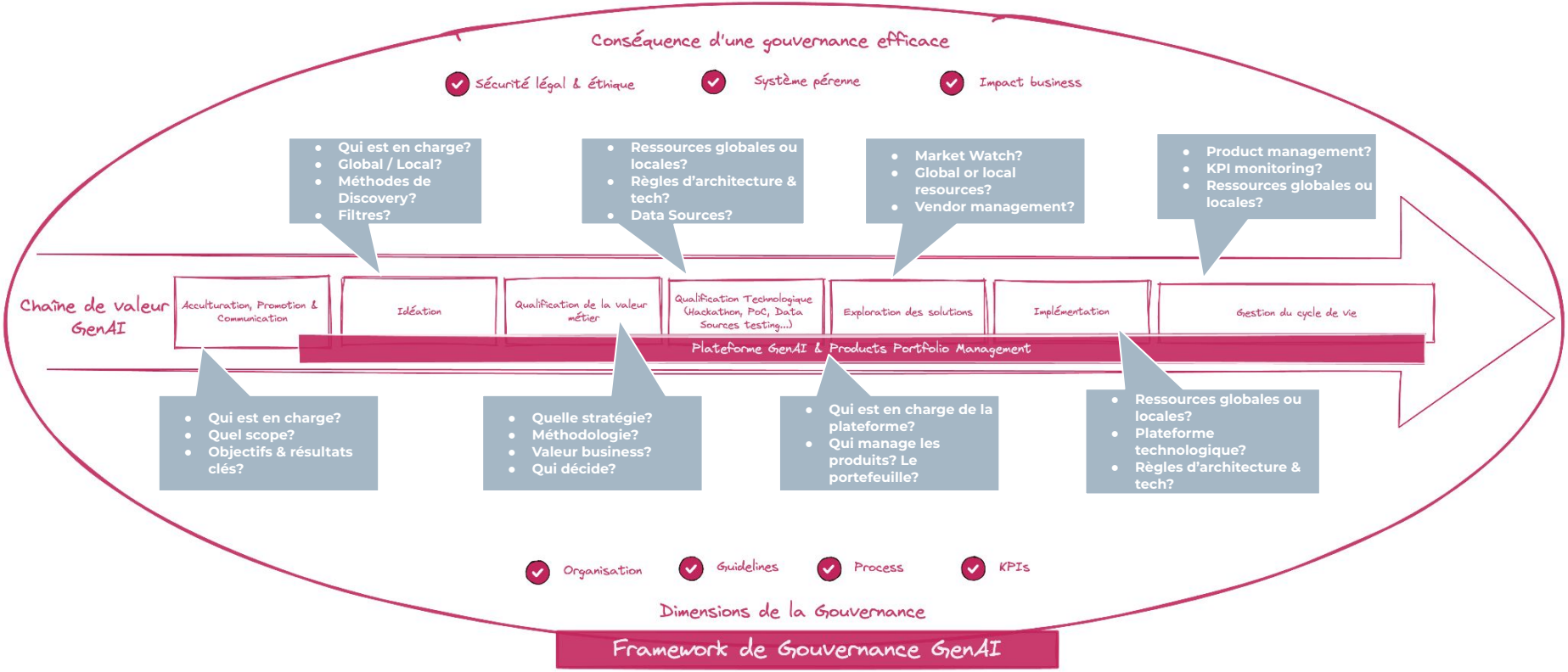
## CONFORMITÉ

La gouvernance des données garantit que les modèles sont **conformes à toutes les lois et réglementations** pertinentes en matière de confidentialité des données, évitant ainsi les complications juridiques.

# QUESTIONS-CLÉS TOUT AU LONG DE LA CHAÎNE DE VALEUR

MAJ - R8

Affronter les défis posés par la GenAI nécessite l'établissement d'un cadre de **gouvernance robuste**.
Ce cadre doit, à chaque étape de la chaîne de valeur, fournir des réponses à des questions essentielles, afin de garantir la mise en œuvre d'**IA responsable**.



# ACCULTURER À LA GOUVERNANCE DES DONNÉES EST UNE DES CLÉS DU SUCCÈS DE L'IA GENERATIVE

MAJ - R8

La Gen AI promet des avancées révolutionnaires capables de générer du contenu innovant à partir de vastes quantités de données. Cependant, pour réaliser ce potentiel, un élément doit être pris en compte : une gouvernance des données efficace. La présence de données de mauvaise qualité ou comportant des biais peut compromettre la fiabilité des résultats et la pérennité des modèles algorithmiques. C'est ici que la formation à la gouvernance des données devient indispensable.

## FORMER À LA GOUVERNANCE DES DONNÉES POUR AMPLIFIER L'IMPACT DES IA GÉNÉRATIVES

La formation à la gouvernance des données permet d'impliquer chaque employé dans la collecte des données et le respect des standards de qualité. En rendant chacun acteur de cette démarche, des erreurs coûteuses peuvent être évitées et l'efficacité des technologies d'IA génératives maximisées.

## FORMER POUR PROMOUVOIR L'ÉTHIQUE ET LA RESPONSABILITÉ INDIVIDUELLE DANS L'USAGE DES DONNÉES

L'importance croissante des données souligne également des défis éthiques liés aux biais dans les ensembles de données qui peuvent influencer les résultats de l'IA. Former les employés à ces problématiques permet de promouvoir une gestion des données responsable en encourageant les utilisateurs à exercer un regard critique face aux outils et aux données qu'ils manipulent.

## BÂTIR UNE CULTURE DE LA DONNÉE

Enfin, instaurer une culture de la donnée à travers la formation est important pour que chaque acteur puisse comprendre le rôle que jouent les données dans le succès stratégique et opérationnel d'une entreprise. À travers ces formations, il s'agit aussi d'impulser une dynamique autour de la donnée et de ses enjeux afin de sensibiliser chacun de ses utilisateurs.

MAJ - R5

# COMPARER LES LLM : LES BENCHMARKS PUBLICS

## Définition

Les modèles de langage sont évalués sur diverses tâches pour vérifier leur **efficacité** et **faciliter leur comparaison**. En effet, les LLM peuvent exécuter un large éventail de tâches, il est donc crucial d'évaluer leur performance en fonction des exigences particulières de nos applications.

Pour ce faire, des benchmarks open source permettent de tester et de comparer les performances des modèles. Nous avons classé ces benchmarks en plusieurs catégories, en fonction du **type de problème que le modèle doit résoudre**. Dans nos applications, nous avons généralement besoin d'une utilisation particulière des LLM (génération de code, recherche augmentée, etc.). En fonction de cela, nous pouvons observer les performances des modèles pour choisir celui qui sera le mieux adapté.

## Notre œil d'expert : le Few-Shot Evaluation

Dans certains ensembles de données d'évaluation, les modèles ont été plus ou moins assistés dans la production de leurs réponses. C'est ce qu'on appelle le few-shot learning, une technique qui consiste à **fournir au modèle des exemples de questions et de réponses afin de l'orienter vers le type de réponse attendu.** Cela permet de guider le modèle vers la tâche qu'on souhaite qu'il accomplisse.

Par exemple, le benchmark **MMLU** a été évalué en mode 5-shot, ce qui signifie que le modèle a reçu 5 exemples de réponses avant d'être évalué. À privilégier donc pour tester la capacité d'un modèle à fournir le format de réponse attendu. À l'inverse, le benchmark **HumanEval** est en mode 0-shot, ce qui signifie que le modèle est simplement invité à résoudre le problème sans aucun exemple de solution. Ce test est rarement utilisé seul, car beaucoup plus versatile.

## DIFFÉRENTES CATÉGORIES DE BENCHMARK PERMETTENT D'IDENTIFIER LE MEILLEUR LLM SELON SON USAGE

### World of Knowledge

Les LLM peuvent être utilisés pour répondre à des questions sur des **connaissances générales ou sur des documents spécifiques.** Ces problèmes sont difficiles car les questions peuvent être complexes, les réponses implicites et les faits difficiles à identifier.
Pour cela, des benchmarks ont été créés, comme **TriviaQA (Joshi et al., 2017).** Ce benchmark regroupe un ensemble de questions permettant d'évaluer la capacité d'un modèle à répondre à des questions en langage naturel.

### Génération de code

Les LLM peuvent être utilisés pour générer du code. Pour évaluer la pertinence de cette application, un ensemble de benchmarks a été créé. Le benchmark **HumanEval (Chen et al., 2021)** regroupe un ensemble d'exercices de génération de code. Pour chaque exercice, le but est de tester si le modèle est capable de générer du code qui passe les tests unitaires. Les exercices sont variés et couvrent un large éventail de tâches de développement.

### Problèmes mathématiques

Les LLM ont fait des progrès significatifs dans de nombreuses tâches, mais ils ne sont pas encore capables de résoudre des problèmes mathématiques multi-étapes complexes. Des benchmarks comme **GSM8K (Cobbe et al., 2021)** spécifiques sont créés pour évaluer la pertinence des LLM sur des problèmes mathématiques simples. Ces benchmarks sont utiles pour évaluer la compréhension des modèles LLM et leurs capacités de "réflexion".

### Les benchmarks agrégés

Les LLM sont entrainés sur une quantité massive d'informations, y compris des sujets spécialisés. Cependant, il est encore difficile de savoir à quel point ces modèles sont capables d'apprendre et d'appliquer des connaissances de différents domaines. Pour les évaluer, des benchmarks généralisés ont été créés, comme le **MMLU (Hendrycks et al., 2020)**. Ce benchmark couvre 57 sujets, allant des sciences sociales à la philosophie en passant par les problèmes de physique.

# Fresh LLM

MAJ - R7

Un des obstacles majeurs des LLM réside dans leur incapacité à gérer et à intégrer des connaissances nouvelles et actualisées. **Les LLM, une fois entraînés, ne bénéficient pas de mises à jour en temps réel** et tendent à produire des réponses fausses ou obsolètes. Face à cette problématique, une avancée significative a été réalisée par plusieurs chercheurs chez Google. Présentée dans le papier ***FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation* (2023)**, leur solution repose sur le développement d'un nouveau benchmark intitulé FreshQA et d'une nouvelle méthodologie, FreshPrompt.

**FreshQA est un nouvel ensemble de données conçu pour tester la capacité des LLM à traiter des questions nécessitant des connaissances à jour**, y compris celles qui évoluent rapidement ou celles fondées sur des prémisses fausses. Parallèlement, **FreshPrompt est une méthode d'apprentissage en contexte qui intègre des informations pertinentes et récentes extraites directement de Google Search dans les invites des LLM.**

**Contrairement à Gemini ou à Copilot**, qui utilisent les résultats de moteurs de recherche pour répondre aux questions sans nécessairement intégrer ces informations de manière contextuelle ou mises à jour dans le modèle lui-même, **un Fresh LLM sait actualiser directement ses connaissances**. FreshPrompt permet une intégration plus sophistiquée et ciblée des données récentes, en extrayant et en incorporant des informations pertinentes issues de recherches web directement dans les prompts des LLM. Cela signifie que les LLM peuvent générer des réponses non seulement fondées sur leur formation initiale mais aussi enrichies par des données externes les plus récentes.

## FreshQA

Conçu autour de **600 questions réparties en quatre catégories**, FreshQA est un **benchmark dynamique** qui met à l'épreuve les LLM sur des sujets nécessitant des connaissances évolutives. **Le modèle a été entraîné sur plus de 50 000 questions posées à des utilisateurs**. Deux modes d'évaluation étaient possibles : un mode "*relaxed*" qui se concentre sur la justesse de la réponse principale, et un mode "*strict*" qui vérifie la factualité de toutes les affirmations faites dans la réponse. **Les résultats révèlent les limitations des LLM face aux connaissances changeantes, aux prémisses fausses, et au raisonnement multi-étapes, indépendamment de la taille du modèle.**

## FreshPrompt

FreshPrompt permet d'améliorer les LLM en intégrant les dernières informations du web, contournant ainsi leurs limites sur l'actualisation des connaissances. **Il débute par la collecte de données variées via Google Search.** Cette collecte inclut non seulement les réponses directes mais aussi des éléments plus nuancés comme les interrogations connexes des utilisateurs. **Les informations-clefs de ces données sont ensuite extraites** (sources, dates, titres et mots-clés saillants, etc.), **puis organisées par ordre chronologique dans l'invite du LLM** et enrichies par des démonstrations ciblées. **Ce processus guide ainsi le modèle vers des réponses logiques, précises et actualisées**, même pour des questions avec des prémisses erronées.

## Quelles perspectives pour les entreprises ?

L'introduction du concept de FreshPrompt invite à repenser l'utilisation des LLM au sein des entreprises. **Pour rester compétitives, elles doivent s'assurer de fournir à leurs LLM des informations internes "fraîches"** – qu'il s'agisse des chiffres de ventes de la veille ou même des données de ventes en temps réel, ainsi que des actualités et des événements internes récents. **Cette nécessité de disposer de données internes actualisées dépasse largement le cadre de l'utilisation des recherches Google comme source d'information pour les LLM.** En actualisant leurs données, les entreprises pourraient non seulement enrichir la qualité des analyses produites par les LLM mais aussi accroître leur capacité à répondre de manière proactive aux évolutions du marché.

# TECHWAVES “GEN AI FOR CIO & CTO”

[sfeir]

WENVISION

# MÉTHODOLOGIE DE NOS TECHWAVES

Les Techwaves sont le résultat d'une méthodologie éprouvée chez SFEIR et WEnvision, pour sélectionner les technologies les plus pertinentes au sein d'un domaine. Celles qui permettront de faire émerger les ambitions des entreprises sans accroître leur dette technique.

Pour ce faire, nous évaluons divers critères, comme la facilité de prise en main, la constitution d'une communauté, la disponibilité de modules de formation voire de livres, etc.

Un exercice possible grâce au crowdsourcing interne au groupe, qui nourrit aussi bien notre veille technologique que nos analyses.

Le résultat de ces analyses permet de positionner les technologies sur notre courbe, découpée en 6 secteurs →

**1.**

**EXPÉRIMENTATION**

Concerne les développeurs et entreprises innovantes, start-up, qui cherchent des moyens d'obtenir un avantage décisif grâce à la technologie. Ils testent beaucoup, réalisent des "proofs of concept".

**2.**

**DÉCOLLAGE**

La technologie est adoptée par les start-up et DSI qui sont prêts à miser beaucoup dessus afin d'obtenir un avantage différenciant. Pour eux, il est vital que le projet aboutisse.

**3.**

**CROISSANCE**

Les entreprises un peu plus prudentes, mais qui souhaitent tout de même pouvoir s'appuyer rapidement sur des technologies innovantes, adoptent à leur tour la technologie.

**4.**

**PLATEAU**

La technologie a trouvé son marché. Les plus conservateurs, qui souhaitent partir sur des technologies éprouvées, l'ont adoptée. Le marché est capable de répondre à la demande.

**5.**

**DÉCROISSANCE**

De moins en moins de nouveaux projets adoptent cette technologie. L'usage est en décroissance. L'utilisation de cette technologie peut être assimilée à de la dette.

**6.**

**MENACE**

L'utilisation de cette technologie sur un nouveau projet est à proscrire. Peu d'acteurs sont encore impliqués dessus.

# LLM GÉNÉRALISTES

MAJ - R8

## Définition

Un LLM (Large Language Model) est un type de modèle de machine learning qui a été entraîné sur un gros volume de données de type texte. Cela permet aux LLM de "comprendre" des contenus existants et de générer de nouveaux contenus, tels que des textes, du code, des scripts, des paroles de musique, des emails, des lettres, etc.

Les LLM sont encore en cours de développement, mais ils ont le potentiel de révolutionner un large éventail d'industries, en particulier sur les fonctions de développement logiciel, marketing, ventes, service client et développement de nouveaux produits.

## Cas d'usage

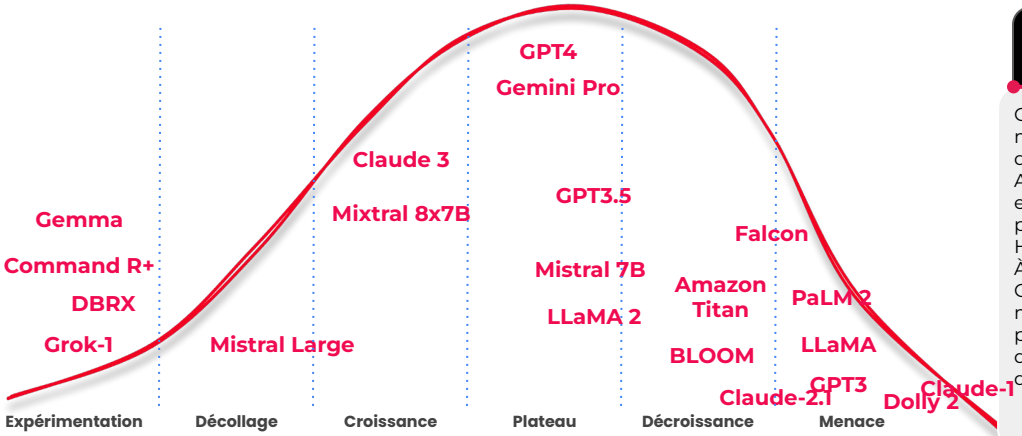
Les LLM peuvent être utilisés dans divers domaines, tels que le traitement du langage naturel (classification de texte et l'analyse de sentiment), la traduction automatique, l'écriture créative et la réponse aux questions. Ils sont également utiles pour générer du code ou du texte, aider à l'analyse de données, détecter des fraudes et segmenter des clients.

## Notre œil d'expert

Le domaine des LLM est en constante évolution, avec des leaders tels que GPT-4 d'OpenAI, Claude d'Anthropic et Gemini de Google, qui dominent actuellement le marché. Cependant, les modèles open source ont leurs propres avantages, notamment en matière de personnalisation et de déploiement *on premises*. Ces technologies évoluent rapidement, et il est crucial de rester à jour avec les dernières tendances et régulations, notamment l'AI Act de l'Union Européenne.

Il est essentiel de développer des solutions flexibles, permettant de remplacer facilement un LLM par un autre. Les LLM ne sont pas universels, il peut donc être judicieux d'utiliser différents modèles pour différentes tâches.

Pour tirer le meilleur parti des LLM, les entreprises doivent se tenir informées des meilleures pratiques et des avancées technologiques tout en respectant les régulations en vigueur. La clé du succès réside dans l'adaptabilité, la compatibilité, et le choix du modèle approprié pour chaque tâche spécifique.




## Focus sur...

### ANTHROP\C

Claude 3 est le dernier modèle de langage développé par Anthropic. Il se décline en 3 versions de performance croissante : Haiku, Sonnet et Opus. À l'instar de GPT-4 ou Gemini Pro Claude 3 est multimodal, ce qui lui permet de traiter à la fois du texte et des images dans ses interactions.

### Google

Google a annoncé en décembre le lancement de Gemini, de nouveaux modèles d'IA multimodaux capables de comprendre et de générer du texte, mais aussi de comprendre des images et vidéos. Ils existent en 3 tailles : Nano, Pro et Ultra avec des compétences de plus en plus avancées. NB : le modèle Ultra n'est pas encore disponible

### MISTRAL AI_

Mistral AI est une startup française spécialisée en IA, proposant les modèles Mistral-7B et Mixtral-8x7B. Il s'agit actuellement des modèles open source les plus performant et ont permis de démontrer la force de l'approche Mixture of Experts (MoE). Leur dernier né, Mistral Medium promet d'encore meilleures performances.

# ORCHESTRATEURS

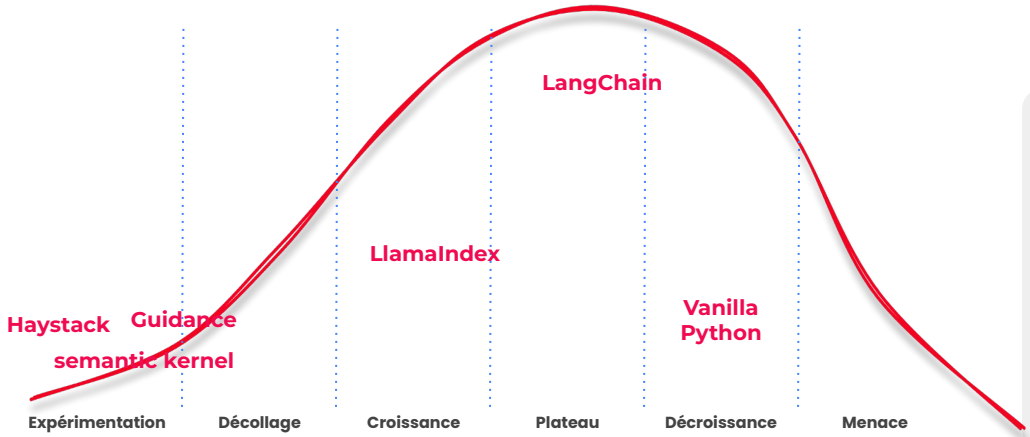
MAJ - R8

## Définition

Les orchestrateurs LLM sont conçus pour faciliter l'interaction entre les utilisateurs et les LLM en gérant des tâches complexes qui nécessitent généralement plusieurs appels d'API. Ils aident à structurer les conversations, à gérer l'état du dialogue, à gérer les entrées et sorties et à contrôler les échanges avec le modèle.
Ces orchestrateurs sont des outils précieux pour exploiter pleinement les capacités des LLM, en facilitant l'interaction avec les modèles et en offrant des fonctionnalités supplémentaires pour une utilisation plus efficace et adaptée aux besoins spécifiques des utilisateurs.

## Cas d'usage

Ces outils peuvent être utilisés pour construire des chatbots capables de générer du texte, de traduire des langues, de répondre à des questions... Ils permettent également d'automatiser des tâches actuellement effectuées manuellement et d'intégrer des LLM avec des bases de données vectorielles pour créer des applications plus puissantes. Les orchestrateurs permettent in fine de créer des agents autonomes, capables de planifier des actions et de les effectuer.




## Notre œil d'expert

Ces outils ne sont disponibles que depuis peu de temps. Ils sont très vite devenus indispensables pour construire des applications basées sur les LLM. Ils permettent non seulement d'accélérer le développement, mais également d'ajouter des fonctionnalités supplémentaires aux LLM, ce qui les rend incontournables lors de la création d'un outil. Que ce soit pour la gestion des prompts, l'optimisation du contexte, le chaînage des appels API ou encore la vectorisation des documents en vue de leur intégration dans le modèle, les orchestrateurs simplifient et accélèrent l'intégration de ces fonctionnalités.

L'un des orchestrateurs les plus populaires est Langchain, une initiative open source qui a connu une progression remarquable en passant de 0 à plus de 49 000 étoiles sur GitHub en seulement quelques mois. Ce framework permet de se connecter aux modèles les plus couramment utilisés et de réaliser des chaînages d'appels avec des API externes. Grâce à une communauté active et engagée, de nouveaux composants sont régulièrement ajoutés, enrichissant ainsi les fonctionnalités de l'outil de manière quasi quotidienne.

Concernant les hyperscalers, Microsoft a sorti Guidance en open source. Bien que moins utilisé que Langchain, il n'en reste pas moins un solide concurrent dans l'avenir.

## Focus sur...

### LangChain

LangChain est un framework open source de création d'applications utilisant des LLM. Il offre une API de haut niveau, prend en charge plusieurs LLM pour faciliter le développement d'applications telles que des chatbots, des outils de synthèse de documents, des outils d'analyse de code, des assistants personnels....

### Microsoft

Microsoft Guidance est un cadre ouvert pour contrôler les LLM, avec une syntaxe simple. Il prend en charge diverses structures de sortie, le prototypage et la mise en cache. Les applications potentielles incluent la génération de texte, la traduction, et l'interaction utilisateur dans les chatbots.

### Google

Vertex AI Search et Conversation (Anciennement Gen App Builder) de Google Cloud permet aux de créer rapidement des applications d'IA générative d'entreprise en utilisant des modèles, des outils de recherche et des technologies d'IA conversationnelle. Elle permet de créer des chatbots, des assistants numériques, etc. en quelques minutes.

## UN GRAND MERCI !

**Les TechWaves sont un exercice collectif. Nous adressons tous nos remerciements aux équipes de SFEIR et WENVISION qui ont contribué à les mettre en musique.**

**Un merci tout particulier à Adrien Lasselle, Arnaud Domard, Bertrand Mondolot, Julia Wabant, Michaël Sherding, Pierre-Alexandre Picard, Romain Viau, Salim Elakoui, Vincent Matthys.**


WENVISION

**TECHWAVES - #GENAI FOR CIO & CTO**

sfeir.com

wenvision.com

# Techwaves GenAI

**L'analyse des experts de SFEIR et WEnvision sur les solutions de GenAI**

Vous êtes perdus et ne savez par où commencer sur le sujet GenAI ?
Les équipes de SFEIR et WEnvision ont compilé les solutions du marché pour vous offrir un condensé et un classement sur une courbe de tendances technologiques des solutions de GenAI.

Les Techwaves sont mises à jour toutes les 6 semaines.

A retrouver sur :
consulting.wenvision.com et sfeir.com