Bouleverser l’accès à l’information en entreprise grâce aux LLM et au RAG avancé.

Comment accéder rapidement et efficacement aux connaissances dispersées parmi les ressources internes et se concentrer sur des tâches à haute valeur ajoutée ? Comment l'IA générative et les LLM peuvent aider à l'analyse, la décision et la création en entreprise ? Ce livre blanc décrypte comment connecter les Large Language Models (LLM) à la data interne, comment fonctionne le Retrieval Augmented Generation (RAG) ainsi que les bénéfices et utilisations possibles en entreprise.

Table des matières

Partie 1

LE MARCHÉ DES LLM

LE MARCHÉ DES LLM

IA générative et grands modèles linguistiques : quelle différence ?

Les LLM utilisent des données pour apprendre et développer une compréhension du langage. Ce sont des modèles qui traitent le langage naturel et génèrent du langage naturel. Toutes les IA génératives ne sont pas basées sur des LLM, mais l'ensemble des LLM sont une forme d'IA générative.

L'IA générative peut produire une variété de contenus : du texte, des images, des vidéos, de la musique, des voix, du code, etc... Elle regroupe une multitude d'outils conçus pour exploiter les données issues des LLM et d'autres modèles d'IA utilisant l'apprentissage automatique afin de créer de nouveaux contenus.

En revanche, un LLM est un type spécifique de modèle d'IA qui utilise un apprentissage automatique basé sur des milliards de paramètres pour comprendre et générer du texte.

L'histoire de l'IA générative

Source : Gartner.

2010

Traduction du langage naturel presque parfaite

Les chercheurs qui travaillent sur le langage naturel découvrent que des modèles exposés à une grande quantité de texte produisent de bien meilleurs résultats que les modèles qui utilisent juste des règles grammaticales.

2014

Maîtrise du sens des mots

Les modèles de langage commencent à comprendre le sens des mots dans un langage naturel en analysant le contexte dans lequel le mot est apparu.

2017-2022

Large language foundation models

La création de LLM est coûteuse, mais une fois créés, ils peuvent être personnalisés à l'aide d'une petite quantité de données supplémentaires afin d'obtenir des performances de pointe pour de nouvelles tâches.

2022

LLM conversationnels

2022 marque l'arrivée de ChatGPT, qui permet aux utilisateurs d'utiliser facilement les LLM. Une partie de son succès est due à son aspect conversationnel et sa capacité à traiter le langage naturel.

2023 : La prolifération de l'IA Générative

L'IA générative a été la grande révolution technologique de 2023, et nombreux sont ceux qui se sont amusés avec ChatGPT, ou Midjourney.

Alors qu'en 2022, le marché de l'IA générative représentait 23,2 milliards de dollars, il a atteint 44,9 milliards en 2023. C'est lors de cette même année qu'on a pu assister à la prolifération fulgurante de l'IA Générative.

Cette technologie est apparue comme une révolution et nombreux sont celles et ceux qui ont testé l'IA générative pour un usage personnel, avant de l'utiliser de manière plus professionnelle.

ChatGPT est arrivé comme figure de proue des générateurs de texte, mais la concurrence ne s'est pas fait attendre très longtemps. En 2023, ChatGPT détenait 19,7% des parts de marché, suivi par Jasper Chat (13,4%), YouChat (12,3%), DeepL (12,1%) et Simplifies (9,7%), laissant un tiers du gâteau à l'ensemble des autres acteurs. Côté génération d'images, Midjourney (26,9%), DALL-E (24,4%) et NightCafe (23,5%) étaient les plus connus.

Le succès rencontré par l'IA générative a rapidement permis de la monétiser, avec plus de 200 millions d'utilisateurs mensuels pour ChatGPT, 30 millions d'utilisateurs quotidiens et 25 à 50 millions d'utilisateurs payants.

On a rarement vu une technologie adoptée aussi rapidement, et l'étude menée par O'Reilly (une maison d'édition américaine spécialisée dans l'informatique) auprès de ses utilisateurs l'illustre bien :

67 % déclarent que leur entreprise utilise l'IA générative.

16 % des personnes travaillant avec l'IA utilisent des modèles open source.

54 % des utilisateurs d'IA s'attendent à ce que le plus grand avantage de l'IA soit une plus grande productivité.

Ce n'est que le début

De nombreux adeptes de l'IA en sont encore aux premiers stades : 26 % travaillent avec l'IA depuis moins d'un an.

Un obstacle persiste

La difficulté à trouver des cas d'utilisation appropriés constitue le principal obstacle à l'adoption, tant pour les utilisateurs que pour les non-utilisateurs.

2024 et demain : l'expérimentation

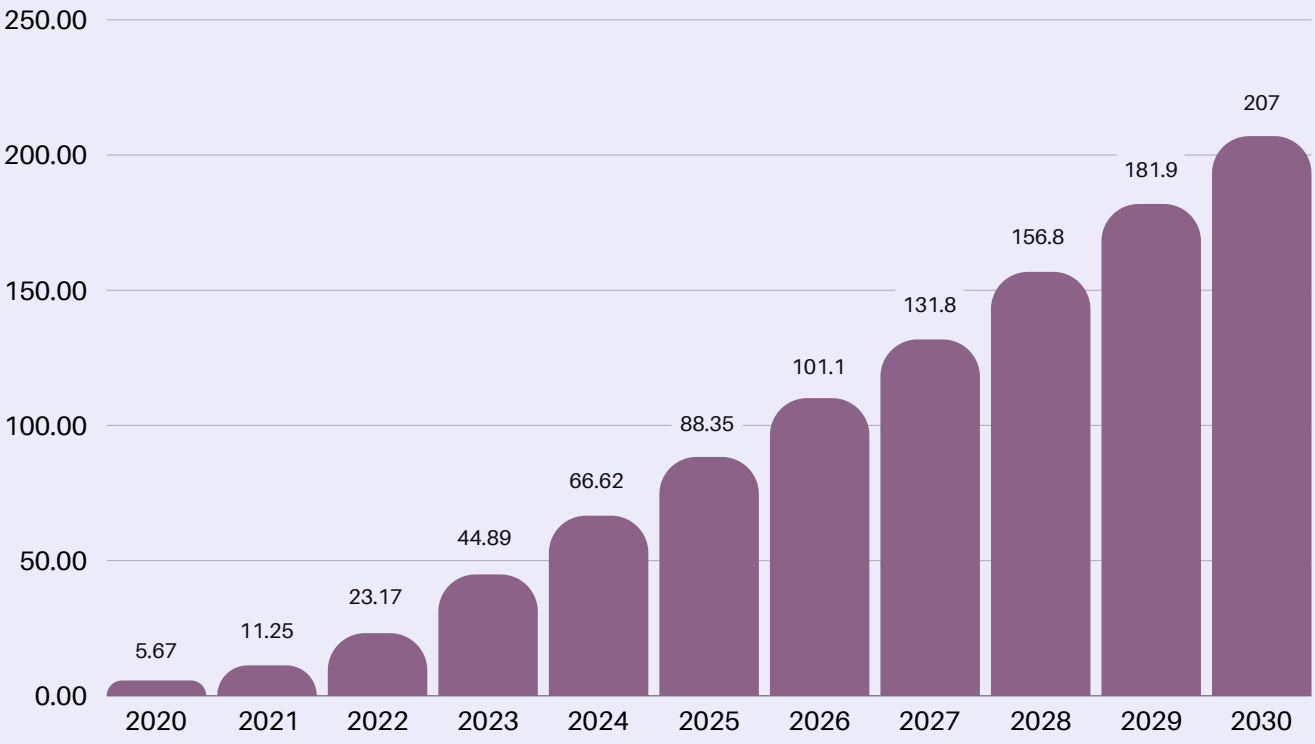
D'après les recherches menées par Sopra Steria Next, le marché de l'IA générative devrait être multiplié par dix d'ici 2028 et représenter environ 100 milliards de dollars, soit un taux de croissance annuelle de 65%.

Nous doublerons alors le temps passé sur les applications d'IA générative, passant de 30 minutes à 1h par jour, permettant ainsi d'augmenter le revenu moyen par utilisateur de 30 à 40$ par mois.

Projection du marché de l'IA générative

Taille du marché mondial en milliards de dollars, estimé en août 2023.

Source : Statista.

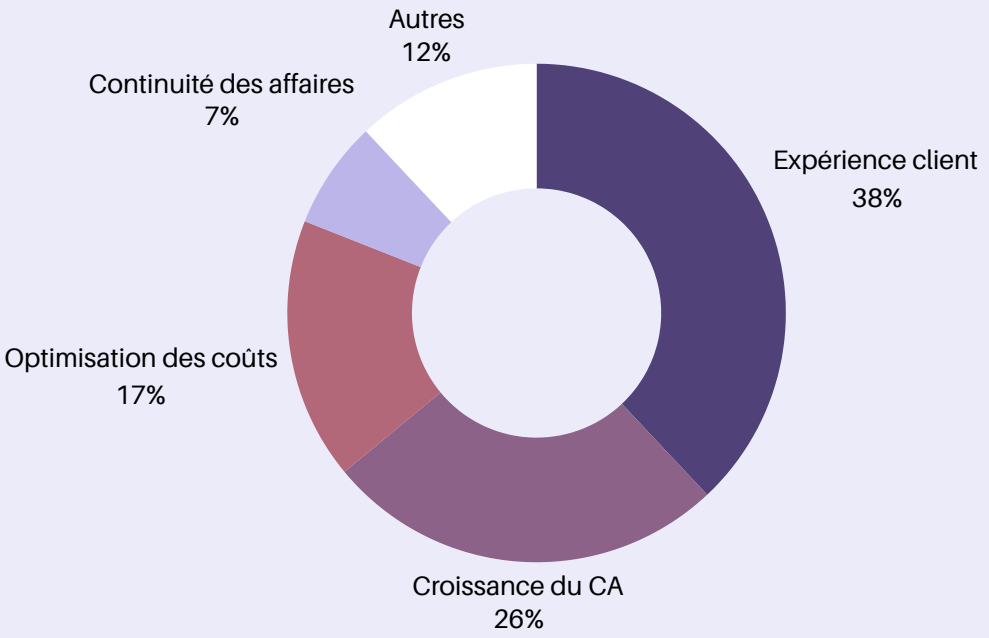


À horizon 3 ans, le potentiel gain de productivité en entreprise pourrait atteindre 10%, si l'IA générative se déploie sur 4 domaines d'application prioritaires : le service client, le marketing digital, l'ingénierie logicielle et la gestion des connaissances.

Dans un récent sondage Gartner mené auprès de plus de 2 500 dirigeants, 38 % ont indiqué que l'expérience clients est l'objectif principal de leurs investissements dans l'IA générative. Viennent ensuite la croissance des revenus (26 %), l'optimisation des coûts (17 %) et la continuité des activités (7 %).

Objectifs principaux des initiatives d'IA générative

Source : Gartner.



L'impact de l'IA générative sera significatif dans une multitude de secteurs, notamment la pharmacie, la fabrication, les médias, l'architecture, le design d'intérieur, l'ingénierie, l'automobile, l'aérospatiale, la défense, le domaine médical, l'électronique et l'énergie.

Médical

D'ici 2025, il est prévu que plus de 30 % des nouveaux médicaments et matériaux seront découverts de manière systématique grâce à des techniques d'IA générative.

Marketing

D'ici 2025, 30 % des messages marketing des grandes organisations seront générés par l'IA, une nette augmentation par rapport aux moins de 2 % enregistrés en 2022.

Et plus globalement, on promet à l'IA générative de beaux jours au sein de l'entreprise :

- D'ici 2024, 40% des applications d'entreprise intégreront une IA conversationnelle, contre moins de 5% en 2020.
- D'ici 2025, 30% des entreprises auront mis en œuvre une stratégie de développement et de test optimisée par l'IA, contre 5% en 2021.
- D'ici 2026, l'IA générative automatisera 60% des efforts de conception des nouveaux sites Web et applications mobiles.
- D'ici 2027, près de 15% des applications seront générées automatiquement par l'IA sans intervention humaine. En comparaison, cela ne se produit pas du tout aujourd'hui.

Étude de cas : exemples d'utilisations en entreprise

L'ORÉAL

Prenons l'exemple de L'Oréal, qui a récemment déployé son IA générative interne, à destination non seulement des développeurs du groupe, mais aussi de certains métiers.

Mis en ligne fin 2023, L'OréalGPT comptait déjà 19 000 utilisateurs distincts au bout de 3 mois. L'IA a généré 330 000 messages et 46 000 images. Dans le département marketing, cette IA permet de valider des idées. Mais L'Oréal ne compte pas s'arrêter là et souhaite multiplier les cas d'usages possibles.

Carrefour

Le groupe Carrefour utilise l'IA générative afin de créer les fiches produits de plus de 2000 références pour gagner du temps sur la production et la mise en ligne.

À terme, Carrefour prévoit de l'utiliser pour la totalité de ses fiches produits. En interne, l'entreprise utilise l'IA générative pour ses processus d'achats pour les accompagner dans leurs tâches quotidiennes, par exemple la rédaction d'appels d'offres ou encore l'analyse de devis.

Morgan Stanley

Morgan Stanley, banque d'investissement et un géant de la gestion de patrimoine, a annoncé en septembre 2023 qu'elle travaillait sur un assistant basé sur GPT-4. Baptisé AI @ Morgan Stanley Assistant, cet outil permet aux conseillers financiers d'accéder rapidement à une base de données d'environ 100 000 rapports de recherche et documents.

« Les conseillers financiers seront toujours au centre de l'univers de Morgan Stanley wealth management », a déclaré Andy Saperstein, coprésident de Morgan Stanley. « Nous croyons également que l'IA générative révolutionnera les interactions avec les clients, apportera de nouvelles efficacités aux pratiques des conseillers et, en fin de compte, aidera à libérer du temps pour faire ce que vous faites le mieux : servir vos clients. »

McKinsey & Company

McKinsey a étudié comment l'IA générative pourrait dupliquer les capacités de leurs collaborateurs, et ont alors lancé Lilli, leur propre solution d'IA générative destinée aux collaborateurs. Il s'agit d'une plate-forme qui fournit une recherche et une synthèse rationalisées et impartiales des vastes réserves de connaissances du cabinet afin d'apporter les meilleures informations, rapidement et efficacement, aux clients.

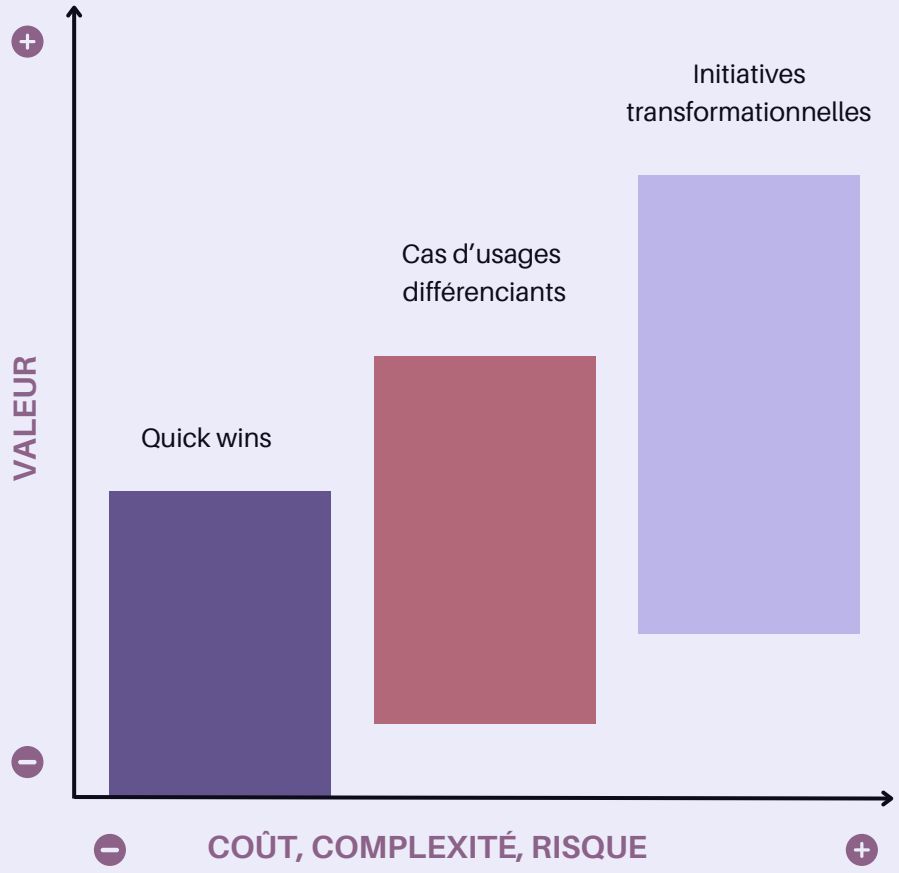
« Lilli regroupe pour la première fois nos connaissances et nos capacités en un seul endroit et nous permettra de passer plus de temps avec nos clients à activer ces informations et recommandations et à maximiser la valeur que nous pouvons créer », déclare Erik Roth, associé principal chez McKinsey qui dirige le développement de Lilli.

IA générative : quel ROI ?

Intégrer l'IA générative dans les entreprises est un défi technique d'une part, mais financier d'autre part. Le cabinet de conseil Gartner distingue 3 types de projets IA en entreprise : les "quick wins", les cas d'usage différenciants et les initiatives transformationnelles.

Catégories de cas d'utilisation de l'IA générative.

Source : Gartner.



Les quick wins, ou gains rapides, sont les projets qui peuvent être implémentés rapidement car ils ne présentent pas de grandes difficultés techniques. Le délai de valorisation est court (moins d'un an) et les améliorations en termes de productivité constituent un avantage concurrentiel qui décroît au fil du temps.

Les cas d'usage différenciants représentent les cas qui exploitent l'IA générative au sein d'une industrie spécifique avec des applications personnalisées, qui permettent d'exploiter les données internes à l'entreprise. Les coûts et risques sont plus élevés que pour les quick wins mais l'avantage concurrentiel est plus important.

Les coûts peuvent alors être compensés par la potentielle génération de revenus, mais le délai de valorisation est légèrement plus long : entre un et deux ans.

Les initiatives transformationnelles viennent avoir un fort impact sur les marchés et les modèles de revenu. Elles s'accompagnent de coûts et risques élevés et d'une grande complexité. Ces innovations peuvent mettre beaucoup plus de temps à être valorisées (plus de deux ans) mais elles offrent un très grand avantage concurrentiel. Les critères de décision d'investissement dans ce type de projets doivent donner la priorité aux avantages stratégiques qui peuvent être difficiles à quantifier plutôt qu'aux avantages financiers immédiatement identifiables.

Le point de vue d'Elqano

L'année 2023 a été marquée par l'essor de l'IA générative en entreprise pour des cas d'usage quick wins (résumé d'appels, tri d'emails...). L'année 2024 sera marquée par des cas d'usage plus complexes, à plus haute valeur ajoutée (travaux d'analyse, d'aide à la prise de décision, rédaction de documents complexes, etc).

60 à 70%

du temps des employés est constitué de tâche qui pourraient être automatisées par l'IA générative actuelle.

Source : MIT.

75%

de la valeur que l'IA générative pourraient apporter se répartissent dans 4 domaines : service client, marketing/sales, ingénierie logicielle et R&D.

Source : McKinsey.

0.1 à 0.6%

c'est la potentielle croissance de productivité au travail par an jusqu'à 2040, grâce à l'IA générative.

Source : McKinsey.

98%

des dirigeants dans le monde affirment que les modèles d'IA joueront un rôle important dans leurs stratégies organisationnelles au cours des 3 à 5 prochaines années.

Source : Accenture.

Partie 2

LLM en entreprise : comment ça marche et quelles limites ?

LLM en entreprise : comment ça marche et quelles limites ?

Comment fonctionnent les grands modèles de langage ?

Les Large Language Models fonctionnent grâce à une architecture complexe basée sur les réseaux de neurones artificiels, plus spécifiquement les Transformers.

Avant l'émergence des Transformers, la taille des modèles était limitée en raison de contraintes computationnelles et de difficultés techniques liées à l'apprentissage sur de grandes quantités de données. Cependant, les Transformers ont changé la donne en introduisant des mécanismes d'attention qui leur permettent de capturer les relations à longue distance dans les séquences de manière efficace.

Cette capacité à modéliser des dépendances complexes a ouvert la voie à la création de modèles de grande taille sans sacrifier la performance.

Les architectures des Transformers sont intrinsèquement parallélisables, ce qui permet de les entraîner efficacement sur de vastes ensembles de données à l'aide de matériel informatique moderne, tels que les GPU et les TPU.

En conséquence, les chercheurs ont pu développer des modèles de langage gigantesques comme GPT (Generative Pre-trained Transformer) et BERT (Bidirectional Encoder Representations from Transformers) avec des dizaines ou des centaines de milliards de paramètres.

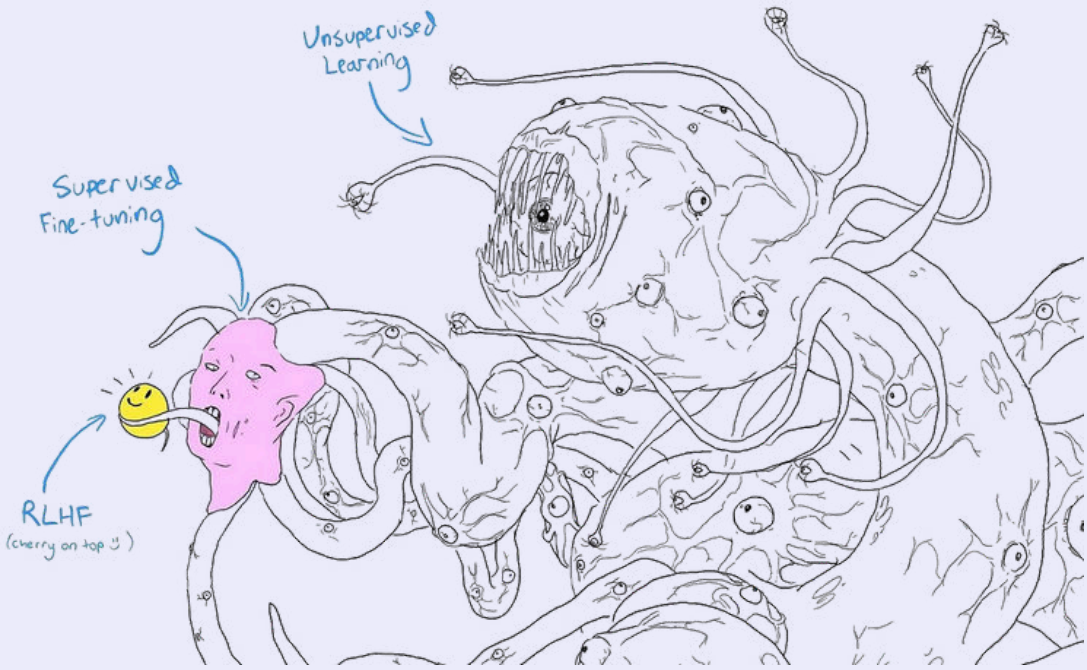
De façon simplifiée, les paramètres peuvent être des poids numériques dans les connexions entre les neurones d'un réseau de neurones, par exemple, un poids de 8 ou 3, qui détermine l'importance de la connexion entre deux neurones.

Avec ces milliards de paramètres, les Large Language Models (LLM) parviennent à saisir des structures et des nuances complexes dans les données textuelles : la machine arrive à représenter le sens d’un texte dans les paramètres.

Cycles de vie des LLM

1 L'entraînement

Les LLM sont pré-entraînés sur de vastes ensembles de données textuelles pour apprendre les modèles linguistiques de base, tels que la syntaxe, la sémantique et la structure des phrases. Pendant cette phase, le modèle apprend à prédire le prochain mot dans une séquence de mots donnée.



Réaliser ce pré-entraînement en entreprise est une lourde tâche : c'est très cher (plusieurs millions d'euros de dépense en puissance de calcul GPU), cela demande une très grande quantité de données ainsi qu'une importante expertise.

Puis vient le fine-tuning (SFT, RLHF) : le modèle pré-entraîné est conçu pour prédire le prochain mot. La plupart des modèles qui intéressent le marché sont ceux capables de répondre à des questions, avec (potentiellement) des instructions. Pour adapter le modèle on utilise des méthodes de fine-tuning (Supervised fine-tuning, et Reinforcement Learning with Human Feedback) qui permettent de l'entraîner sur des tâches de «question answering» et de prompt «instruction», des tâches spécifiques en utilisant des ensembles de données plus petits et spécialisés. Grâce à ces étapes, le modèle est maintenant capable de répondre correctement à des questions.

Cependant, pour apprendre de nouvelles notions et connaissances, il faut le ré-entraîner dans les couches inférieures pour "imprimer" de la connaissance dans les poids du modèle. Ce type de méthode n'a pas encore fait ses preuves et pousse généralement les modèles à halluciner car ils apprennent à parler sur des sujets moins connus.

Voici à quoi ressemble l'ensemble de données de réglage GPT4all dans la visionneuse d'ensembles de données Hugging Face. C'est assez simple : il se compose d'une série de prompts (première colonne) et de réponses/réponses attendues (deuxième colonne) pour affiner les réponses du modèle.

Prompt	Réponse
Bonjour J'ai une grille de données Wpf qui affiche une collection observable d'un...	Une solution possible consiste à utiliser une largeur fixe pour l'en-tête GroupItem et à aligner...
<h2>Bonjour, comment puis-je générer un PDF avec les données visuelles de l'écran, ou générer...	Pour générer un PDF avec les données visuelles de l'écran, vous pouvez utiliser une bibliothèque...
<pre><code>package com.kovair.omnibus.adapter.platform ; importer...	Le problème peut être lié au chargement des classes et au garbage collection. Lorsqu'un

Voir le dataset complet

2 L'inférence - utilisation du modèle

Une fois le modèle entraîné il peut être installé sur un ordinateur, souvent dans le cloud car les modèles peuvent être très gros : 7 milliards de paramètres pour les "petits" et jusqu'à des centaines de milliards de paramètres.

En mode inférence, un Large Language Model (LLM) utilise les connaissances qu'il a acquises lors de son entraînement pour accomplir diverses tâches liées au langage. Lorsqu'une requête est soumise au modèle, il parcourt la séquence de mots en entrée, en attribuant une attention variable à chaque mot en fonction de son contexte et de sa signification.

En utilisant cette attention, le modèle évalue la probabilité des mots suivants dans la séquence, générant ainsi une prédiction ou une réponse cohérente avec la requête donnée.

Point lexique

Token

Un token représente une unité de base du langage, comme un mot, un sous-mot ou même un caractère. Lorsque le texte est traité par un LLM, il est d'abord divisé en tokens, ce qui permet au modèle de comprendre et de manipuler le langage de manière plus granulaire.

Embedding

Un embedding est la représentation numérique d'un token. Les tokens sont introduits dans le modèle puis convertis afin d'être traités de manière mathématique. Chaque token est associé à un vecteur numérique qui capture ses propriétés sémantiques et syntaxiques, permettant ainsi au modèle de comprendre les relations entre les mots dans le langage.

Sentence Embedding

Ou embedding de phrase : la représentation numérique d'une phrase ou d'un texte dans un espace vectoriel continu. Cela permet de capturer les caractéristiques sémantiques et syntaxiques de la phrase, ce qui facilite la comparaison et l'analyse de la similarité entre différentes phrases.

Les embeddings de phrases sont généralement obtenus à partir de modèles de langage pré-entraînés, tels que les Sentence Transformers, qui convertissent les mots individuels de la phrase en vecteurs numériques et les combinent ensuite pour former un embedding global représentatif de la phrase entière.

À garder en tête

Les LLM sont des modèles d'IA qui arrivent à prédire les réponses les plus probables, grâce aux technologies de Transformers, la machine arrive à traduire le sens sémantique d'un mot et d'une phrase en quelque chose qu'elle maîtrise bien : des nombres (vecteurs).

À quoi servent les LLM ?

De façon générale, les LLM sont utiles dans toutes les tâches de langage naturel. Voici certaines des utilisations principales des LLM :

Génération de texte

Les LLM peuvent être utilisés pour générer du texte de manière automatique, que ce soit pour la création de contenu, la rédaction automatique de mails, la production de rapports ou la génération de code.

Personnalisation

Ils peuvent être utilisés pour personnaliser les expériences en ligne, telles que la recommandation de produits, l'adaptation de contenus ou la personnalisation des interactions avec les utilisateurs.

Résumé automatique

Ils peuvent résumer de grands volumes de texte de manière concise et précise, aidant ainsi à extraire les informations pertinentes d'un document.

Traduction automatique

Les LLM peuvent être employés pour améliorer les systèmes de traduction automatique en fournissant des traductions plus précises et fluides dans différentes langues.

Analyse des sentiments

Les LLM peuvent analyser de grandes quantités de texte pour déterminer les sentiments, les opinions ou les tendances à partir de données telles que les réseaux sociaux, les critiques de produits ou les commentaires en ligne.

Prise de décision

Les LLM peuvent fournir des analyses et des insights basés sur le langage naturel pour aider à la prise de décision dans divers domaines, tels que la finance, la médecine, le droit ou la gestion d'entreprise.

Recherche d'informations

Ils peuvent aider à la recherche d'informations en extrayant des informations pertinentes à partir de grandes bases de données ou de textes.

À garder en tête

Les LLM sont les experts du langage naturel, ils sont capables de faire le rapprochement entre les mots et le sens de mots, ils peuvent comprendre des instructions, et ont, grâce à leur entraînement sur des très grands jeux de données, appris une très grande partie de la connaissance humaine (publique).

Les fournisseurs et technologies disponibles

OpenAI

Le plus connu et grand leader du marché, produit toute la suite GPT (GPT-3.5, GPT-4, et bientôt GPT-5). Ces modèles sont utilisés pour diverses applications telles que la génération de texte, la traduction, la réponse à des questions...

Google

Propose plusieurs grands modèles linguistiques, notamment son dernier model Gemini 1.5 Pro, un modèle multimodal (pour des entreprises sélectionnées : jusqu'à 1M tokens, 10M destiné à la recherche).

Microsoft

Dès les débuts de l'IA générative, Microsoft s'est positionné comme un acteur principal de l'écosystème, dans un premier temps avec des partenariats privilégiés avec OpenAI.

Microsoft a ensuite mis en place ces IA (GPT-3.5, GPT-4), sur son service pour le cloud, mais aussi très rapidement dans l'intégration de l'IA générative dans sa suite Office avec Copilot.

amazon

Amazon aussi se positionne comme un acteur majeur du service cloud pour l'utilisation des IA avec des outils qui facilitent le travail des développeurs pour la mise en place d'IA.

Amazon a investi largement dans Anthropic qui travaille sur un LLM tout aussi puissant que celui d'OpenAI : Claude (1,2,3).

MISTRAL AI_

La start-up française qui s'est fait figure de proue de l'open source, propose des modèles tels que 8x22b, 8x7b, 7b, large.

Le dernier sorti : 8x22b est "un modèle de mélange d'experts peu dense (SMoE) qui n'utilise que 39 milliards de paramètres actifs sur 141 milliards, ce qui offre une rentabilité inégalée pour sa taille".

Meta

Meta propose entre autres le modèle Llama 3, dernière version de sa famille de modèles Llama en open source également.

Deux modèles pré-entraînés et fine-tuned avec des paramètres de 8B et 70B ont été publiés avec la capacité à prendre en charge un large éventail de cas d'usage.

Les autres types d'acteurs

LangChain

Puissant framework open source pour développer des applications alimentées par des modèles de langage. Il se connecte aux modèles d'IA que vous souhaitez utiliser et les relie à des sources externes.

LangSmith donne une visibilité sur ce qui se passe avec votre application LLM, qu'elle soit construite avec LangChain ou non, afin que vous sachiez comment agir et améliorer la qualité.

Enfin, LangServe rend clé en main la fourniture d'une API pour votre application LangChain.

LlamaIndex

Framework de données pour les applications LLM permettant d'intégrer, de structurer et d'accéder à des données privées ou spécifiques à un domaine.

LangSmith

Plate-forme de développement tout-en-un pour chaque étape du cycle de vie des applications basée sur LLM, que vous construisez avec LangChain ou non.

ragas

Framework qui aide à évaluer les pipelines de récupération augmentée (RAG).

À garder en tête

L'engouement autour de la puissance des LLM induit un écosystème qui évolue très vite. Tous les mois les cartes sont rebattues et les leaders sont déclassés. Cependant, très peu d'entreprises sont capables de développer "from scratch" leur propre LLM du fait de la complexité et du coût d'entraînement.

Les limites des LLM

Malgré leur puissance, les LLM ont aussi des limites. Voici quelques limites importantes qu'il faut connaître :

Les LLM ne connaissent pas les données privées des entreprises

Les informations connues par l'IA sont celles apprises pendant son entraînement, ses connaissances sont donc statiques. Les LLMs sont entraînés sur de vastes ensembles de données textuelles provenant d'Internet (entre autres), mais ils ne sont pas nativement intégrés aux systèmes internes et aux données spécifiques de chaque entreprise.

De nombreuses entreprises cherchent des solutions pour faire marcher ces IA sur leur base de connaissances. Nous verrons par la suite les difficultés et les solutions envisageables pour y parvenir.

Un tel modèle peut donc formuler des réponses obsolètes ou génériques lorsque vous attendez une réponse spécifique et actuelle, s'il n'a pas intégré les données les plus récentes.

Par exemple, les tendances, les normes industrielles, les préférences des consommateurs et les pratiques commerciales peuvent évoluer rapidement, rendant les données sur lesquelles les modèles ont été formés moins pertinentes pour les tâches actuelles.

Cela peut être particulièrement préoccupant dans les domaines où la précision et la pertinence des informations sont cruciales, comme la prise de décision commerciale, la recherche et développement, ou la personnalisation des services client.

Lorsque l'on souhaite utiliser l'IA générative en entreprise, il est crucial que le modèle fournisse des réponses fiables et à jour. C'est là que le modèle Retrieval Augmented Generation (RAG) entre en jeu.

Le point de vue d'Elqano

Dans la très grande majorité des cas, les techniques de fine-tuning ou de grande fenêtres contextuelles ne sont pas les plus recommandées pour avoir des résultats de qualité, notamment à cause de leur coût, leur complexité et leurs performances.

Construire un bon prompt n'est pas toujours facile

Écrire un prompt efficace n'est pas donné à tout le monde, et les LLM y sont très sensibles. La même question, posée de manière différente, peut générer des réponses elles-mêmes totalement différentes. C'est notamment pour ça que de nombreuses entreprises recrutent des Prompt Engineers. Voici quelques-unes des limites courantes associées à cette tâche :

La complexité du langage professionnel : les entreprises opèrent souvent dans des domaines spécialisés avec un jargon spécifique et des conventions linguistiques particulières. Formuler des prompts qui reflètent précisément ces nuances peut être difficile, surtout si les modèles de langage ne sont pas pré-entraînés sur des données sectorielles spécifiques.

Le besoin d'expertise : la construction de prompts efficaces nécessite une compréhension approfondie du domaine d'activité de l'entreprise et des besoins spécifiques de la tâche à accomplir. Cette tâche peut donc exiger l'implication d'experts métier pour formuler des requêtes pertinentes.

La sensibilité aux biais : les prompts mal formulés peuvent introduire des biais dans les résultats générés par les LLM. Par exemple, des formulations ambiguës ou tendancieuses peuvent conduire à des réponses inappropriées ou partiales, ce qui peut compromettre la qualité et l'objectivité des résultats obtenus.

L'optimisation du rendement : trouver le bon équilibre entre la spécificité de la requête et la diversité des réponses peut être un défi. Une formulation trop restrictive peut limiter la capacité du modèle à générer des réponses variées, tandis qu'une formulation trop vague peut produire des résultats imprécis ou non pertinents.

De façon générale, pour écrire un bon prompt il faut respecter les mêmes codes de base de la communication : à qui suis-je en train de parler, que sait-il sur le sujet... Voici une liste (non exhaustive) des éléments utiles pour écrire un prompt : persona, contexte, tâches, format attendu, audience, information additionnelle, ton.

La taille de la fenêtre contextuelle

La Fenêtre contextuelle (ou Context Window) est une notion importante dans le domaine du traitement automatique du langage naturel. Elle fait référence à la portée des informations (plage de mots) prises en compte par un modèle de langage pour générer une prédiction ou une réponse... Les modèles avec de grandes fenêtres contextuelles ont la capacité de prendre en compte plus de contexte autour de chaque mot lors de la génération de texte ou de la prise de décision.

Plus la quantité de texte à prendre en compte augmente, plus la complexité informatique de la tâche s'accroît.

GPT-4 a étendu sa fenêtre à 128K tokens pour gpt-4-turbo/gpt4-vision, et bien que supérieure à la plupart des concurrents, cette limite reste contraignante pour les tâches plus complexes comme par exemple pour examiner de nombreux documents.

La taille de la fenêtre contextuelle est cruciale dans l'exploitation des LLM car elle permet de fournir toute l'information nécessaire pour répondre à une question portant sur des informations non connues par le LLM (données d'entreprise ou information récente par exemple).

La taille de la fenêtre de contexte fait débat dans le monde de la recherche, d'un côté Google présente des LLM avec des fenêtres de 1M+ tokens, de l'autre côté des chercheurs montrent que la pertinence des réponses diminue après un certain seuil.

" Grâce à notre étude, nous avons découvert que si les LLMs montrent une performance prometteuse sur des entrées jusqu'à 20K tokens, leur capacité à traiter et comprendre des séquences plus longues diminue de manière significative."

Source : Arxiv.

Il est important de noter que le principal mode de rémunération des fournisseurs de LLM est la facturation par token (par mot) envoyés au LLM et les mots générés par le LLM.

À garder en tête

Pour une utilisation en entreprise, les LLM comportent des limites. Et même si certaines méthodes permettent de les contourner, elles sont très coûteuses et complexes à mettre en place. Il existe tout de même des solutions prometteuses : en combinant la génération de texte avec des systèmes de récupération d'informations, on peut améliorer la précision, la pertinence et l'efficacité des réponses des LLM. Cette technique, appelée RAG, ouvre la voie à des applications plus performantes et adaptées aux besoins spécifiques des entreprises.

Partie 3

Faire fonctionner le LLM avec les données internes grâce au RAG

Faire fonctionner le LLM avec les données internes grâce au RAG

C'est quoi le RAG ?

Le RAG étend les capacités des LLM en intégrant des informations spécifiques à des domaines ou à une base de connaissances interne. C'est l'action de fournir au LLM de l'information supplémentaire (via un prompt) pour l'aider à répondre à une question. Aujourd'hui le RAG est la méthode la plus efficace pour enrichir le LLM avec de l'information nouvelle, sur laquelle il n'a pas été entraîné.

Comment ça marche ?

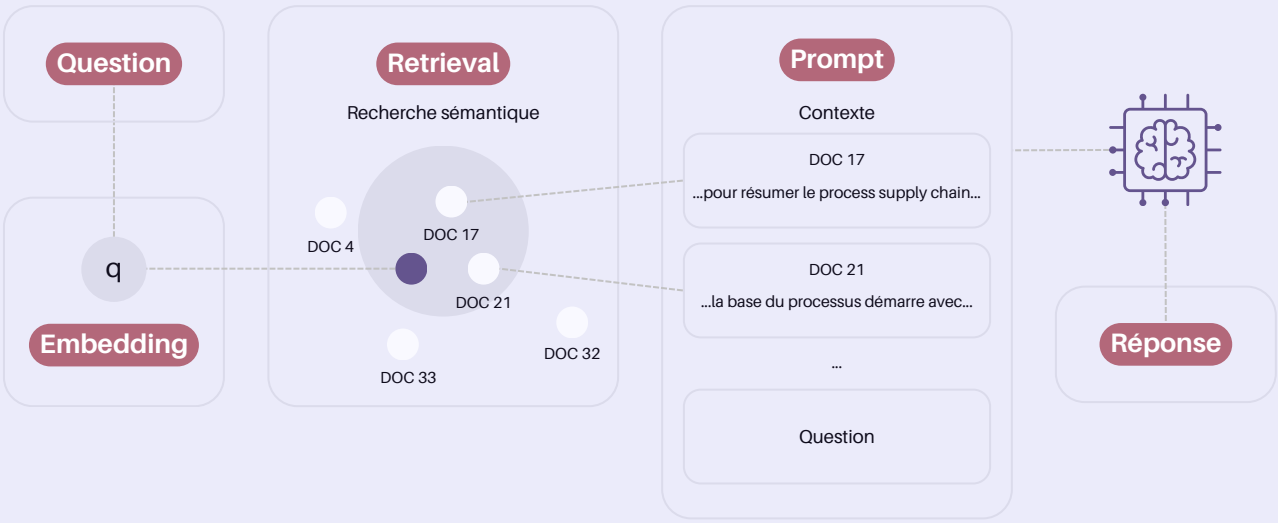
Les RAG combinent le prompt de l'utilisateur avec des bases de données ou documents pour enrichir le contexte et suivent deux étapes principales.

La première, la phase de récupération (retrieval), consiste à explorer les données afin de collecter les informations pertinentes à la question de l'utilisateur.

Ces informations, accompagnées de la question, sont ensuite intégrées au prompt envoyé au LLM. La seconde étape, la génération, permet au LLM de répondre naturellement à la question, à l'image de ce que fait ChatGPT.

Bien que le RAG puisse paraître simple, son impact en entreprise est considérable. Contrairement aux modèles linguistiques traditionnels qui se restreignent à des données parfois obsolètes de plusieurs mois ou années, le RAG permet au modèle de rester à jour.

Le fonctionnement des RAG peut être simplifié de la sorte :



Préparation du RAG : indexation de la base de connaissance

L'objectif est de transformer la base de connaissance de formats variés (pdf, word, pptx, etc), en une base de connaissance facilement et rapidement exploitable par des outils de recherche d'information.

Généralement elle se traduit par la récupération du texte contenu dans les documents choisis (pdf, words etc), puis l'enregistrement dans une base de données spécifiquement conçue pour la recherche d'information.

Avec l'arrivée des transformers, les technologies de recherche d'information ont aussi été améliorées, on parle aujourd'hui de semantic search et de base vectorielle.

Pour créer une base de donnée vectorielle il suffit de convertir notre texte en embeddings grâce aux modèles de "sentence transformers". Il est conseillé de découper le texte en paragraphes pour des raisons de performances (et limitation) des modèles d'embeddings.

Utilisation du RAG

Étape 1 - Retrieval : récupération de l'information pertinente

Pour trouver les extraits de textes les plus pertinents pour répondre à la requête utilisateur, on effectue une comparaison sémantique entre la requête et un morceau de texte de notre index, pour cela il faut :

- transformer la requête utilisateur en un vecteur (embedding) grâce au même modèle de sentence transformer
- comparer le vecteur issu de la requête utilisateur avec les vecteurs dans notre base de connaissance et prendre ceux qui sont les plus proches (distance entre vecteurs)

En sortie on obtient les extraits de textes les plus proches sémantiquement de la requête. Ils serviront comme contexte fourni au LLM.

Étape 2 : prompt final

Dans cette étape finale du processus de RAG, le système prend en compte les extraits pertinents lors de l'étape précédente pour formuler le prompt final.

C'est la requête finale ou la description donnée au modèle de langage pour générer le contenu souhaité. Il s'agit de la dernière étape avant que le modèle ne commence à générer le texte final en réponse à cette requête.

Ce prompt peut inclure des phrases-clés, des résumés, des extraits pertinents, des questions spécifiques, ou tout autre élément qui guide efficacement le modèle vers la production du contenu souhaité.

Le processus de RAG simple comme présenté ci-dessus est très performant dans le cas idéal : lorsque l'information est similaire à celle d'un dictionnaire, les questions sont factuelles, et la réponse figure directement dans le corpus.

Le RAG n'est plus adapté pour les données déstructurées ou visuelles comme des slides de présentation ou des documents par exemple.

C'est là que le RAG avancé intervient, afin de démultiplier les bénéfices du RAG en le rendant plus performant et plus adapté aux besoins des entreprises.

Voici la différence de réponse entre une IA générative sans RAG, et une qui utilise le modèle RAG :

IA générative sans RAG

12:55PM
Avons-nous déjà travaillé sur l'IA générative ?

LLM *12:56PM*
Je suis désolé, mais je ne dispose pas des informations nécessaires pour répondre à cette question.

IA générative avec RAG

12:55PM
Avons-nous déjà travaillé sur l'IA générative ?

LLM *12:56PM*
Il existe en effet deux projets en cours sur l'IA générative :

Generative_AI_Project.docx

GenAI_Consulting.docx

Limite du RAG simple dans le cadre de son utilisation en entreprise

L'approche de parsing de texte de documents, leur découpage en paragraphes, et l'insertion de ces segments dans une base vectorielle pour faire du "Retriever-Answer Generation" (RAG) est une méthode courante en traitement automatique du langage pour faciliter la recherche d'information et la génération de réponses.

Cependant, cette méthode présente certaines limites en matière de conservation et de représentation complète de l'information, principalement pour les raisons suivantes :

- Perte de contexte global :

Lorsque les documents sont découpés en paragraphes individuels, chaque paragraphe est souvent traité comme une unité d'information indépendante.

Cette segmentation peut entraîner la perte du contexte global ou de la continuité narrative qui existe dans le document complet. Les paragraphes pris de manière isolée peuvent ne pas contenir toutes les informations nécessaires pour comprendre le sujet en profondeur, ou peuvent être mal interprétés sans le contexte fourni par les sections adjacentes.

- Perte d'informations structurales et multimodales :

Les documents d'entreprise peuvent inclure non seulement du texte, mais aussi des graphiques, des tableaux, et d'autres éléments visuels ou structurés qui sont difficiles à encoder efficacement dans une base vectorielle purement textuelle.

La segmentation en paragraphes et l'encodage textuel standard ignorent ces éléments, qui peuvent être essentiels pour comprendre l'information complète.

- Problèmes de granularité :

Le choix de la granularité des segments de texte (paragraphes, phrases, ou documents entiers) a un impact significatif sur les performances du système RAG.

Les paragraphes peuvent ne pas être de bonne granularité pour certains types d'informations ou de questions. Comme lorsqu'une réponse adéquate nécessite une intégration d'informations provenant de multiples paragraphes ou même de différents documents.

- Défis liés à la cohérence et à la continuité :

Dans le cas où des réponses ou des informations doivent être générées ou récupérées à partir de multiples entrées vectorielles (paragraphes), il peut être difficile de maintenir la cohérence et la continuité dans les réponses générées, surtout si les segments proviennent de différentes parties du document ou de différents documents.

Le RAG avancé

Il existe un grand nombre de stratégies possibles pour optimiser le RAG.

Le RAG avancé est une évolution du concept de RAG simple, le concept a été introduit pour pallier aux limitations du RAG dont nous avons parlé précédemment.

L'objectif est à la fois d'être capable de conserver 100% de l'information dans son outil de recherche lors de l'indexation, de la transformer de façon à ce qu'un LLM puisse mieux comprendre le contexte de l'information, mais aussi d'être capable de répondre à un plus large panel de questions.

Le RAG avancé permet également de trouver de façon plus précise les informations au sein de la base de connaissance.

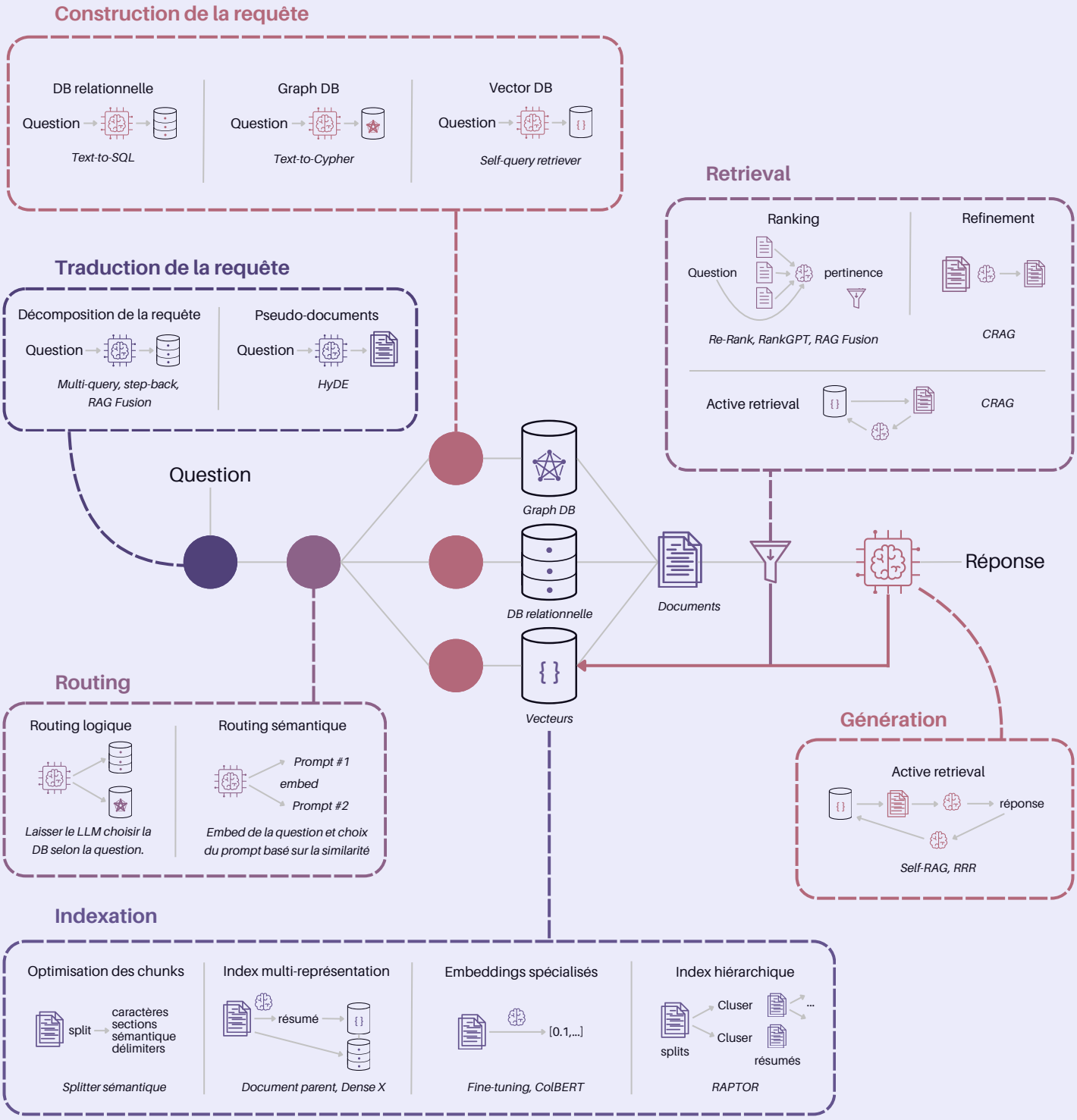
Les nouvelles technologies de traitement du langage permettent de tirer pleinement profit de la base de connaissance en la manipulant et la transformant pour extraire le maximum d'informations.

Les propriétés principales du RAG avancé sont :
1 - la transformation d'une base de données hétérogène en une base de connaissance structurée avec un index vectoriel multi-niveaux.

2 - l'utilisation d'un raisonnement multi-étapes pour parcourir efficacement la base de données et répondre à des questions complexes.

Les stratégies de RAG avancé peuvent être regroupées en 2 grandes catégories : la stratégie d'indexation et la stratégie de recherche d'information.

On peut résumer le RAG avancé avec ce schéma :



L’étape la plus importante, c’est la quatrième étape : l’indexation, représentée le plus en bas sur le schéma précédent. À l’aide de vos sources de données internes, le RAG avancé vient ensuite l’enrichir de la sorte :

- L’indexation multi-level :

Cette étape permet de résumer et contextualiser les documents. Sur la base de la requête utilisateur, le chunk (morceau d’un document) le plus pertinent est extrait et transmis au LLM avec le document dont il fait partie. Cela permet une compréhension plus approfondie du contexte, une extraction plus précise des caractéristiques des données et une meilleure adaptabilité à diverses tâches et domaines.

- RAG vision :

On en parle plus en détail dans la partie suivante, mais cette étape permet de réaliser une description textuelle des slides ou document visuels. C’est tout simplement l'intégration de capacités de traitement d'images dans le modèle RAG.

Cette extension permet au modèle de comprendre et d'utiliser des informations visuelles en plus des données textuelles pour améliorer ses performances.

- Metadata :

Assignation automatique de metadata par l’IA pour enrichir la compréhension du contenu et améliorer la génération de texte. Les metadata sont des informations structurées qui décrivent, identifient, et permettent d'organiser les données. Dans le contexte du RAG, elles sont utilisées pour enrichir la compréhension et la recherche d'informations dans les différentes sources de données (textes, pdf, visuels...).

- Indexation séquentielle :

Création des résumés de page d’un document. Afin de conserver le contexte des pages précédentes nous effectuons une indexation séquentielle, c’est à dire qu’en plus de la page du document nous allons fournir les informations sur la pages précédentes

Stratégie de recherche d’informations

Afin de rechercher la bonne information, le RAG avancé vient tout d’abord modifier la requête initiale de l’utilisateur. Elle est en effet enrichie et/ou traduite en une requête riche sémantiquement. Les métadonnées sont extraites et la requête est décomposée en plusieurs requêtes.

Multi-index/multi level search : selon la requête utilisateur, la recherche va ensuite être effectuée dans différents index.

Active retrieval : lors de la recherche d’information, une étape de vérification de l’information est ajoutée. Si l’information n’est pas pertinente, l’algorithme recommence en modifiant les paramètres de la recherche.

Voici un exemple de requête et les étapes par lesquelles le RAG avancé va passer :

Requête : “Quel est le secteur d’activité de nos clients ?”.
Le RAG avancé va détecter l’intention utilisateur d’un point de vue macro sur sa base de connaissances, faire la recherche dans l’index de résumés de documents. Il récupère donc une grande quantité de résumés de documents qui concernent les clients de l’entreprise. Enfin, il synthétise l’information dans une réponse finale qui catégorise les clients par type d’industrie.

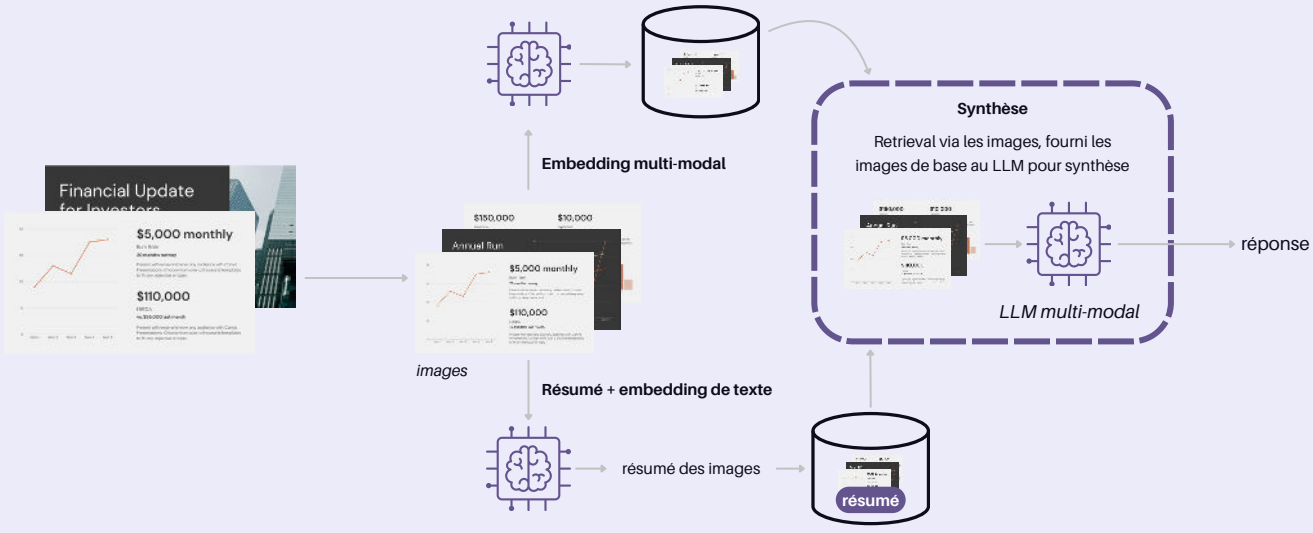
Zoom sur le RAG vision

Au sein d’une entreprise, de nombreuses données et connaissances se trouvent sur des présentations sous forme de slides ou autres contenus visuels.

La méthodologie RAG vision permet d’incorporer ce type de données à la base de connaissances.

Ce processus extrait les slides sous forme d'images, utilise un LLM pour résumer chaque image, intègre les résumés d'images avec un lien vers les images originales, récupère l'image pertinente sur la base de la similarité entre le résumé d'image et la question de l'utilisateur, et enfin transmet le tout au LLM pour la synthèse de la réponse. Il existe cependant deux approches :

Approche 1 : intégrations multi-modales



Approche 2 : récupération multi-vecteur avec résumé d’images

Entre ces deux options, la seconde, qui permet de résumer l'image, est bien plus performante. Elle est cependant plus complexe, et plus coûteuse. Voici les différents scores de précision résultants d'une étude Langchain en décembre 2023 :

Approche	Score (précision CoT)
Haut k RAG (texte uniquement)	20%
Approche 1 : intégrations multi-modales	60%
Approche 2 : récupération multi-vecteurs avec résumé d'images	90%

*Voir l'*étude complète

Le point de vue d'Elqano

Pour un RAG avancé performant, l'étape la plus critique est l'enrichissement et la transformation de la base de données grâce à l'IA. Cette étape est absolument nécessaire pour répondre efficacement à des questions complexes et permettre des raisonnements multi-étapes.

Un RAG puissant pour ouvrir les portes à la Knowledge Intelligence

Le RAG représente une avancée majeure dans la manière dont les entreprises peuvent accéder et utiliser leurs données pour générer des connaissances exploitables. Il va au-delà de la simple extraction de données et fournit des réponses et des insights pertinents, transformant ainsi la manière dont les entreprises exploitent leur capital de connaissances.

Le RAG favorise la collaboration et le partage des connaissances au sein des entreprises. En permettant aux utilisateurs de poser des questions naturelles et d'obtenir des réponses instantanées, il encourage l'échange d'informations et l'intelligence collective.

Les fonctionnalités de génération de contenu favorisent également la création de bases de connaissances internes, alimentées par les contributions de l'ensemble des collaborateurs, ce qui renforce la culture de partage et d'apprentissage continu au sein de l'entreprise.

On peut ainsi dire que le RAG ouvre la voie à une "knowledge intelligence" plus avancée. En combinant des capacités de recherche, d'analyse et de génération, il permet aux entreprises de prendre des décisions éclairées, anticiper les tendances du marché et innover plus rapidement.

En fournissant un accès facile et intuitif à une mine de données et de connaissances, le RAG devient un atout stratégique pour les entreprises cherchant à rester compétitives et innovantes.

		RAG simple	RAG avancé
Indexation	Chunk de documents	✓	✓
	Résumé de documents	✗	✓
	Description des images	✗	✓
Retriever/search	Sur le chunks	✓	✓
	Sur les résumés et descriptions	✗	✓
Génération	À partir de textes	✓	✓
	À partir d'images	✓	✓

		RAG simple	RAG avancé
Questions simples	Nombre de documents faible (moins de 20)	✓	✓
	Nombre de documents élevé (plus de 20)	✓	✓
	Nombre de documents très élevé (plus de 10 000)	✓	✓
Questions complexes	Nombre de documents faible (moins de 20)	✓	✓
	Nombre de documents élevé (plus de 20)	✗	✓
	Nombre de documents très élevé (plus de 10 000)	✗	✗

À garder en tête

L'utilisation de la puissance des LLM sur les données d'entreprise est possible. En revanche, l'environnement hostile de la base de connaissances des entreprises et les besoins des utilisateurs par métier nécessitent une mise en place de techniques robustes et élaborées afin d'assurer le bon fonctionnement. La bonne connexion entre les données des entreprises et les LLM ouvre les portes à la Knowledge intelligence.

Partie 4

Bénéfices et cas d'usage en entreprise

Bénéfices et cas d'usage en entreprise

Bénéfices du RAG en entreprise

Lorsqu'il est correctement appliqué, le modèle RAG peut être un allié de taille en entreprise, voici les principaux avantages de l'amélioration de l'IA générative grâce à la technologie RAG :

Bénéfices techniques

- Des réponses à jour et précises : les RAG garantissent que les réponses sont fondées sur des informations actuelles.
- Des réponses transparentes : les RAG sont capables de citer les sources.
- Efficacité des ressources : les modèles n'ont pas besoin d'être ré-entraînés avec de nouvelles données.

Bénéfices stratégiques

- Trouver des informations internes : le RAG permet d'extraire rapidement des informations pertinentes à partir de grandes bases de données internes de l'entreprise.
- Gain de temps : au lieu de passer des heures à rechercher des données manuellement, les collaborateurs peuvent obtenir des réponses précises et pertinentes en un temps record.
- Amélioration de la qualité des productions : les collaborateurs peuvent créer des documents, des rapports ou des présentations de meilleure qualité et plus rapidement.
- Meilleure prise de décision : le RAG fournit aux décideurs des informations clés nécessaires pour éclairer leurs décisions.

À garder en tête

En plus de répondre à des requêtes simples, le RAG avancé peut être utilisé pour la résolution de problèmes, la prise de décision et la réponse à des questions plus complexes.

Cas d’usage par industrie

Comment le modèle RAG peut-il vous servir en entreprise ? Voici quelques cas concrets d’usages possibles pour différents secteurs.

Cabinets de conseil

- Analyse de marché et recherche de tendances :

En utilisant le RAG pour explorer des bases de données externes, des rapports sectoriels et des articles pertinents, une entreprise de conseil peut obtenir des informations précieuses sur les tendances du marché, les comportements des consommateurs et les innovations sectorielles.

Ces informations peuvent ensuite être utilisées pour formuler les recommandations stratégiques aux clients. Un gain de temps important pour les consultants.

- Élaboration de rapports et de documents :

Le RAG peut être utilisé pour générer des résumés, des analyses et des rapports à partir de grandes quantités de données textuelles. Une entreprise de conseil pourrait utiliser le RAG pour rédiger automatiquement des rapports personnalisés en agrégeant et en synthétisant des informations provenant de diverses sources.

Exemple de requête : “réalise un résumé de toutes les recommandations que nous avons faites pour le client X et décrypte les résultats pour identifier les meilleures recommandations.”

Agences de publicité/digital :

- Génération de contenu créatif :

L'agence peut utiliser le RAG pour générer des idées de campagnes publicitaires en intégrant des informations provenant de bases de données externes, telles que les tendances du marché, les comportements des consommateurs et les succès passés de l'entreprise.

Elle peut également utiliser le RAG pour générer des scripts publicitaires en se basant sur des exemples de publicités efficaces dans le même secteur ou en utilisant des données sur les préférences et les intérêts du public cible.

- Identification des meilleures pratiques et des succès passés :

Dans le cadre de réponse à un appel d’offre, le RAG peut aider l'agence à extraire des exemples de meilleures pratiques et de succès passés dans des projets similaires, en utilisant des données provenant de son propre historique de projets ou en récupérant des informations sur les réalisations d'autres agences dans le même domaine.

Exemple de requête : “analyse nos réponses aux appels d’offres remportés dans le secteur des FMCG et formule la meilleure réponse possible à un nouvel appel d’offre du secteur.”

Fonds d’investissement :

- Sélection de portefeuille :

En intégrant des données provenant de diverses sources, telles que les rapports annuels, les communiqués de presse et les analyses financières, le RAG pourrait aider à identifier les entreprises ou les secteurs qui présentent le meilleur potentiel de rendement pour le portefeuille du fonds. Il pourrait également être utilisé pour évaluer les risques associés à chaque investissement potentiel.

- Due diligence :

Lors de la réalisation d'une due diligence , le fond pourrait utiliser le RAG pour récupérer et analyser des informations clés, telles que les antécédents financiers, les performances passées, les relations avec les parties prenantes, etc.

Cela permettrait d'obtenir une vue d'ensemble plus complète et d'identifier les éventuels risques ou opportunités.

- Surveillance des investissements :

Une fois que des investissements ont été réalisés, le RAG pourrait être utilisé pour surveiller en continu les performances des entreprises du portefeuille, les tendances du marché et les événements susceptibles d'avoir un impact sur les investissements. Cela permettrait de prendre des décisions informées et réactives en temps réel.

Exemple de requête : “réalise un résumé des 10 derniers boards de la startup X” ou “analyse ces deux startups du même secteur et formule une comparaison approfondie de leurs performances”

Cabinets juridiques :

- Recherche juridique améliorée :

L'un des cas d'utilisation les plus évidents du RAG pour un cabinet légal est d'améliorer la recherche juridique. Le RAG pourrait être utilisé pour récupérer des informations juridiques pertinentes à partir de vastes bases de données de jurisprudence, de lois et de précédents juridiques. Ces informations pourraient ensuite être utilisées pour générer des résumés, des analyses ou des réponses à des questions spécifiques des clients.

- Analyse de contrats et de documents juridiques :

Le RAG pourrait être utilisé pour analyser et interpréter des contrats et d'autres documents juridiques.

Le RAG pourrait aider à identifier les risques potentiels, à repérer les clauses ambiguës et à fournir des conseils sur la manière de rédiger ou de négocier des contrats de manière plus efficace.

- Assistance à la rédaction de documents juridiques :

Le RAG pourrait être utilisé pour fournir une assistance à la rédaction de documents juridiques tels que des motions, des requêtes ou des avis juridiques. En utilisant des modèles de langage génératifs, le RAG pourrait aider les avocats à rédiger des documents plus rapidement en fournissant des suggestions de formulation, des exemples pertinents et des références juridiques.

Partie 5

Solution qui utilise le RAG avancé : le chat d'entreprise "Playground" connecté aux données internes

Solution qui utilise le RAG avancé : le chat d'entreprise "Playground" connecté aux données internes

À quoi ça sert ?

Playground est notre dernière innovation chez Elqano, cette solution intègre les Grands Modèles de Langage (LLM) avec de la recherche interne avancée sur vos données pour améliorer l'efficacité du travail. Par l'intermédiaire d'une interface type ChatGPT, Playground permet de générer du contenu en temps réel, discuter avec ses données, automatiser des tâches et accéder rapidement aux informations internes.

Playground utilise le concept de RAG avancé en combinant une indexation séquentielle multi-level vision (cf partie RAG avancé), et les stratégies présentées également dans la partie pour gérer les requêtes utilisateurs complexes.

La fonction Intelligent Search de Playground permet d'accéder rapidement aux informations internes. Conçue pour être efficace, elle aide les employés à trouver rapidement les données dont ils ont besoin.

Quelques exemples d'usage de la solution Playground

	RH	Finance	IT
REQUÊTE SIMPLE	Peux-tu me montrer les résultats du dernier sondage de satisfaction ?	Quel est le statut du budget du département marketing ce mois ?	Quels sont les derniers protocoles de sécurité pour l'accès à distance ?

	Sales	Légal	Marketing
REQUÊTE SIMPLE	Où puis-je trouver la liste des leads du dernier salon ?	Quel est le statut des derniers dépôts de brevets de l'entreprise ?	Quels sont les retours sur la dernière campagne d'emailing ?

	RH	Finance	IT
REQUÊTE COMPLEXE	Peux-tu analyser les tendances des taux de roulement du personnel sur les cinq dernières années et identifier les départements à risque ?	Peux-tu comparer les dépenses actuelles avec le budget alloué et identifier les écarts significatifs pour chaque département ?	Quel est l'impact des récentes mises à jour de sécurité sur la performance du réseau de l'entreprise ?

	Sales	Légal	Marketing
REQUÊTE COMPLEXE	Peux-tu effectuer une analyse de rentabilité des différents segments de clients pour identifier les plus rentables et ceux à risque ?	Peux-tu analyser les délais de traitement des cas juridiques l'an passé et identifier les processus qui nécessitent une optimisation ?	Quels ont été les retours sur investissement (ROI) pour chaque canal de marketing utilisé au cours de l'année passée ?

Et après ?

Vous souhaitez approfondir ces sujets avec nos experts en RAG avancé ? Contactez-nous !

Pour approfondir

RAG avancé : l'indexation

RAG multi-modal

Principales questions sur l'IA générative

Mesurer le ROI du RAG

L'IA générative dans le consulting

Cas d'usages de l'IA générative

elqano
WHEN KNOWLEDGE FINDS PEOPLE