Technologies et Infrastructures

LLM ChatBot Corporatif

Expérimentation avec l'intelligence artificielle générative

Mai 2024

Travailler avec l’IA générative

L'avènement récent de l'IA générative, propulsée par la technologie des transformateurs génératifs, marque une étape décisive vers une interaction intuitive homme-machine via le langage naturel, sans nécessiter d'expertise en informatique.

Cette technologie émergente, tout en soulevant des enjeux éthiques et sécuritaires, en est encore à ses balbutiements. Grâce à notre expertise de la solution infonuagique Azure, nous exploitons les solutions d'Azure OpenAI pour adresser ces défis et explorer le potentiel qu’offrent ces outils.

“

Comment pourrions-nous concrétiser l’utilisation de la technologie d’IA générative afin de la transformer en véritable outil pour nos collègues?

Les défis posés par l'IA générative

Confidentialité des données

L'interaction avec un ChatBot public tel que ChatGPT dans un contexte de travail pose un problème évident, sachant que les données sont persistées et peuvent être utilisé pour entraîner les modèles

Pertinence des résulats

L'utilisation d'un modèle public ne produit pas toujours des résultats pertinents en lien avec notre réalité d'affaire et risque parfois de générer des hallucinations

L'éthique de l'IA

Si les modèle d'IA générative continue de s'améliorer, une réflexion sur l'étendue des possibilités d'automatisation s'impose

Analyse d’affaire & recherche

Quelles sont les besoins à combler à travers cette preuve de concepte

- Certains collègues sont de plus en plus nombreux à utiliser des ChatBots publiques tel que ChatGPT
- D’autres collègues gagneraient à expérimenter avec l’IA générative en tant qu’assistant personnel
- Accroître la valeur d’une telle technologie avec des données pertinentes aux différents domaines d’affaires de nos collègues
- Utiliser l’opportunité d’approfondir nos connaissances sur le fonctionnement des technologies d’IA générative afin d’en comprendre ses avantages ainsi que ses limites
- Entamer une réflexion sur la nécessité de mettre en place des balises concernant les technologies d’IA générative

Plan de la preuve de concept

ChatBot corporatif

Modèle d'IA

Déterminer quel modèle utiliser ainsi que la méthode d'hébergement

Données internes

Déterminer comment permettre la contextualisation de données internes

Implémenter

Développer et implémenter la solution et la rendre disponible à l'interne

Tester

Impliquer nos collègues afin de recueillir des commentaires et des suggestions

Produit

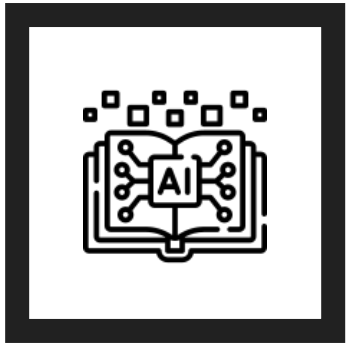
Transitionné la preuve de concept en véritable produit avec un cycle de développement

Services d'IA infonuagique

Les avantages d'utiliser Azure pour ses service d'IA

Propriété des données

Les entrées de données des utilisateurs ou les document téléversés restent privés et confidentiels



Données internes

Mettre à profit les services d'Azure nous permet d'utiliser des données internes pour améliorer la pertinence des résultats

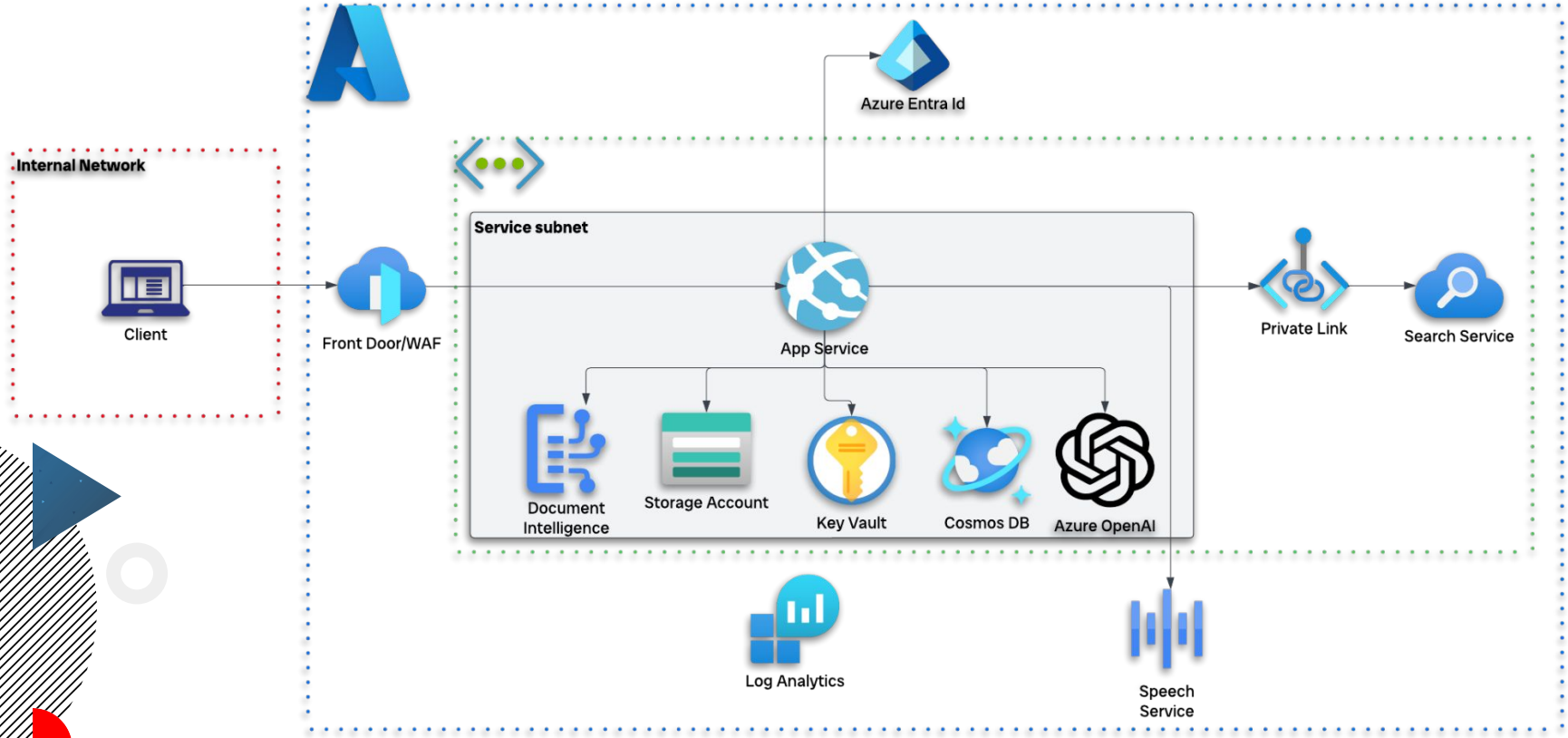
Expérience interne

Tirer parti de l'expérience de travail avec des fournisseurs infonuagiques déjà existante au sein de nos équipes

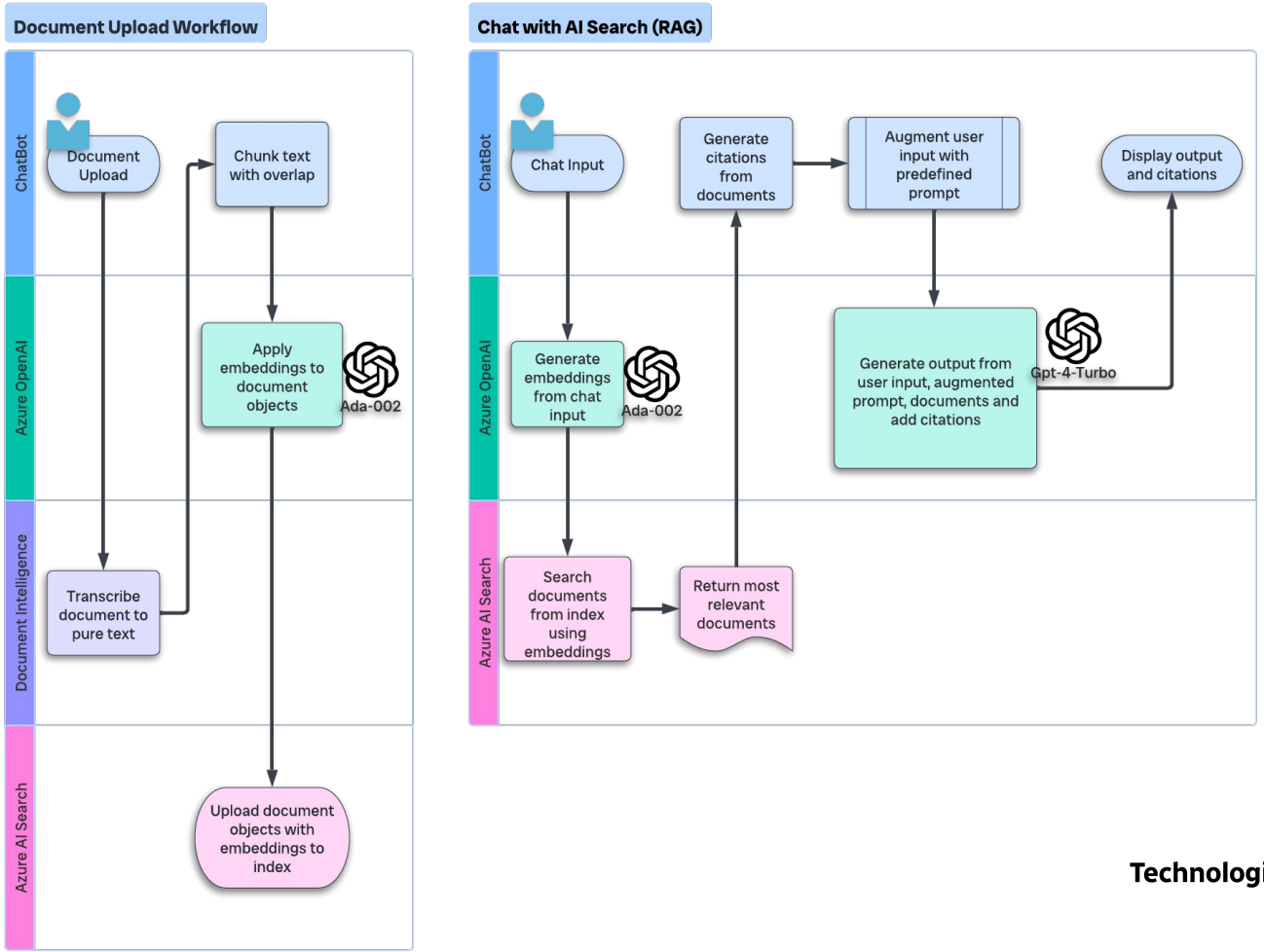
Gestion de serveurs

La solution infonuagique évite des investissements en matériel ainsi que de la maintenance à long terme

ChatBot Architecture



ChatBot Retrieval-Augmented Generation (RAG)



Évaluation du projet

Les critères d'évaluation de la preuve de concept

Faisabilité technique

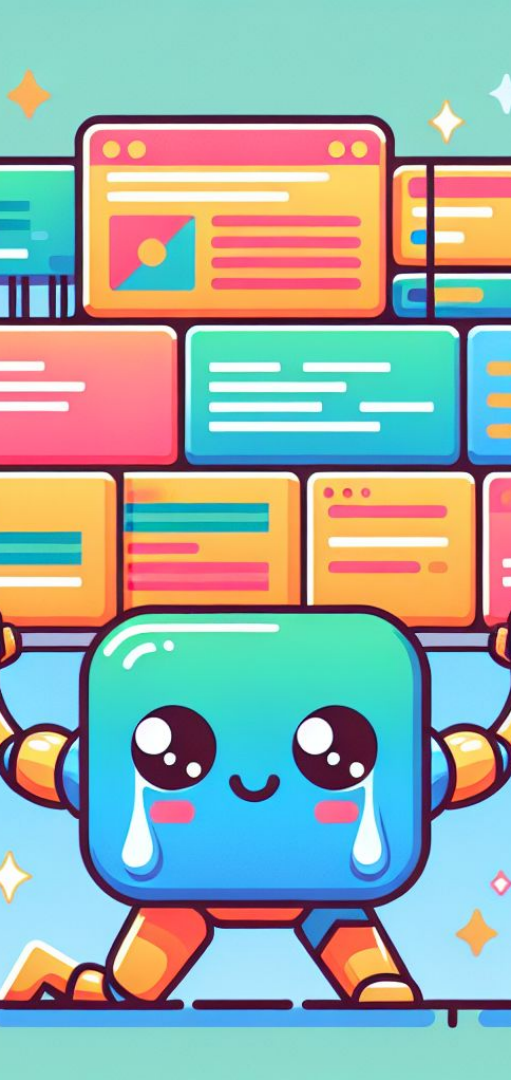
Est-ce que le plan d'architecture peut effectivement être déployé et utilisé?

Remplacer ChatGpt

Est-ce que la démarche fait en sorte que nos collègues cessent d'utiliser les AI qui ne garantissent pas la confidentialité des données?

Ajout de valeur

Est-ce que l'application est utilisée par nos collègues pour les assistés dans leurs tâches de tous les jours



Écueils du projet

Obstacles et difficultés rencontrées

Instabilité des resources Azure

Les ressources IA d'Azure sont en très grande demande, ce qui a parfois causé des problèmes de disponibilités dans certaines régions.

Une implémentation robuste de "Infrastructure as Code" est recommandée pour palier à ce problème.

Compétition avec ChatGPT

Bien que les modèles d'Azure soient très performants, les fonctionnalités de l'interface de ChatGPT sont très affinées, notamment "My GPT".

Il se doit de rapidement développer des alternatives avec agilité.

Réussites du projet

Succès rencontrés dans la démarche

Technologie impressionnante

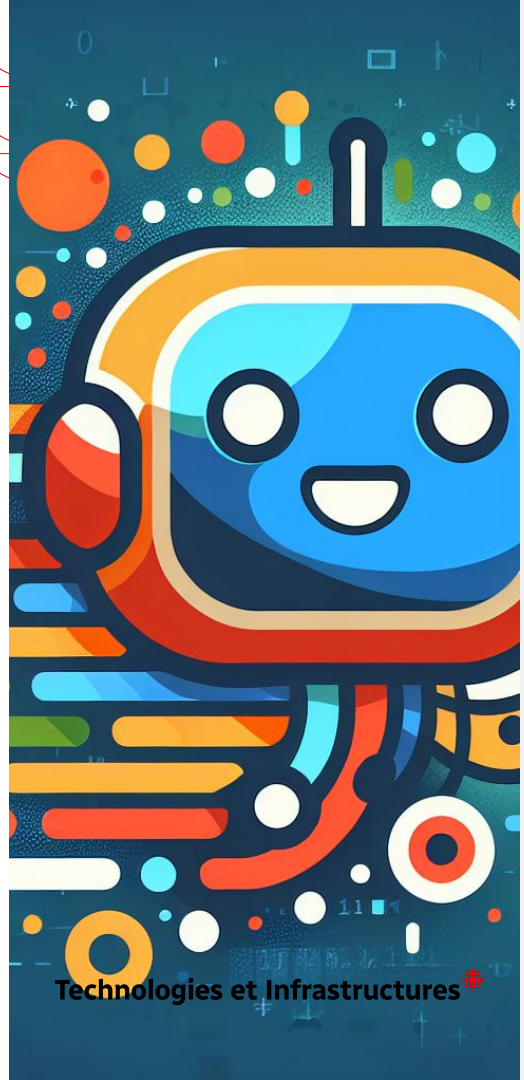
Utiliser un ChatBot est une expérience relativement impressionnante, surtout lorsqu'on tire parti des fonctionnalités RAG avec des documents internes.

En sachant que nos données restent confidentielles, nos collègues ont confiance lorsqu'ils utilisent l'application.

Flexibilité de l'infrastructure infonuagique

L'hébergement des modèles en mode "SaaS" permet de rapidement mettre sur pied des projets d'IA générative.

Il est aisé d'expérimenter avec différents produits d'IA et de tester les interactions entre ceux-ci.



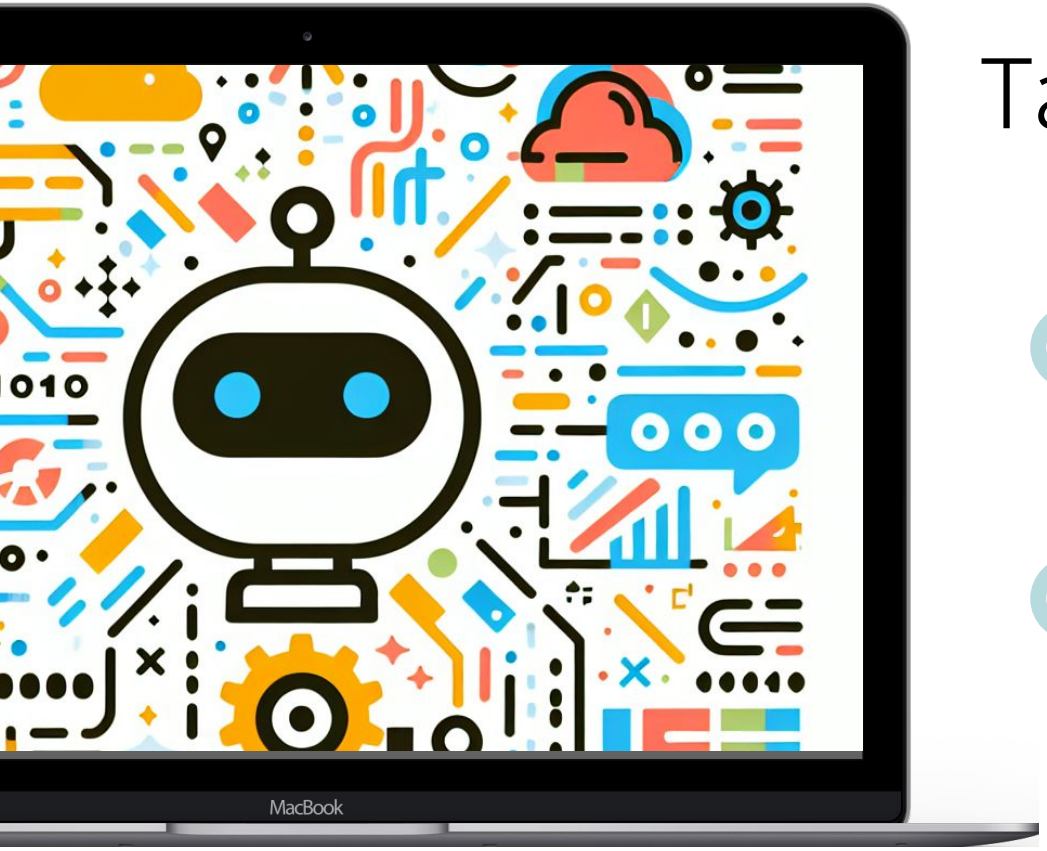


Table ronde

01

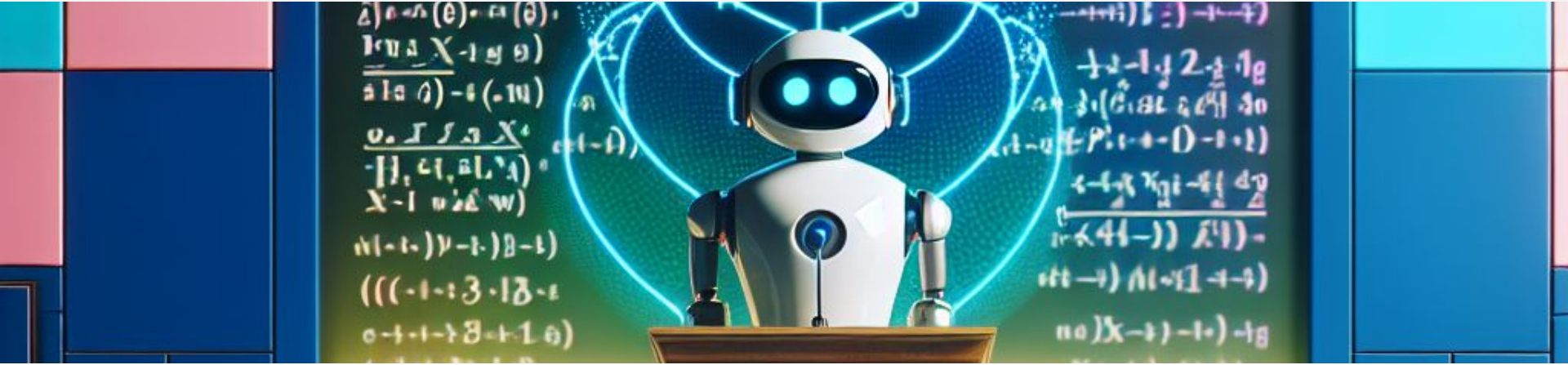
Questions

Avez-vous des questions par rapport avec la présentation?

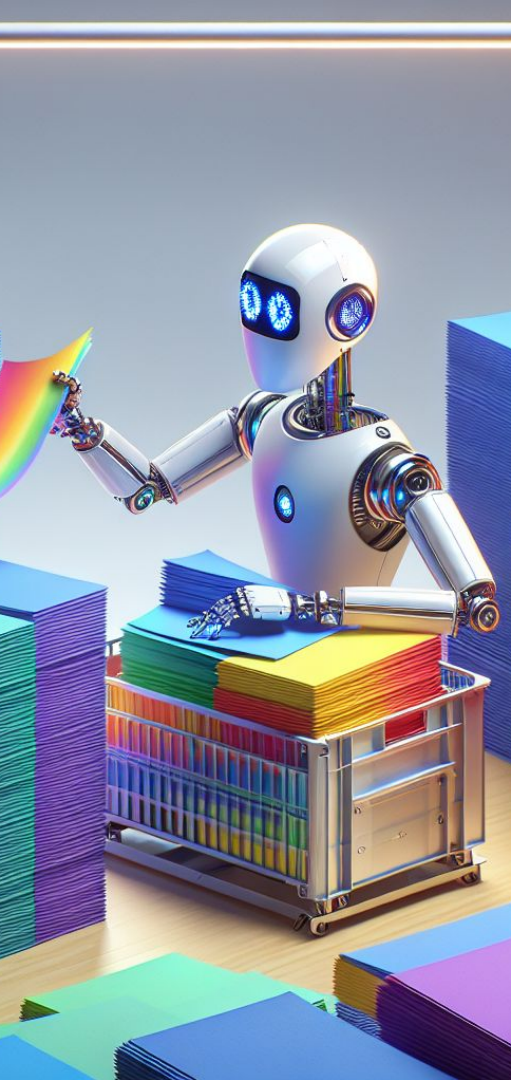
02

Présentation

Souhaitez-vous voir l'application à l'oeuvre si le temps le permet?



Information supplémentaire



Les modèles derrière l'IA générative

Modèle de langage à grande échelle

Un modèle de langage à grande échelle (LLM pour Large Language Model) est un système d'intelligence artificielle conçu pour interpréter et générer du texte en langage naturel.

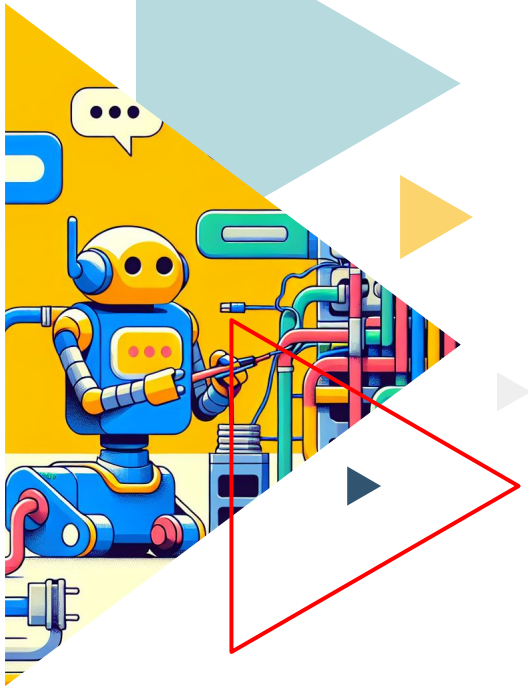
Les LLMs utilisent l'architecture transformateur générative pré-entraînée (GPT pour Generative Pre-trained Transformer) pour prédire la probabilité que chaque mot se succède à un autre **tout en accordant de l'importance aux mots qui dictent le context d'une phrase**, ce qui permet de générer des réponses cohérentes et contextuellement pertinentes.

Spécifications du ChatBot

Modèles et fonctionnalités de l'application

- Modèles d'IA Générative
  - GPT 4 turbo - LLM
  - Ada-002 - Plongements (Embeddings)
  - GPT 4 Vision - Multimodale avec Reconnaissance d'image
  - Dall E 3 - Génération d'image

- Services IA Azure
  - Speech Services - Transcription voix à texte
  - Document Intelligence - Transcription de documents et images à texte pure
  - Azure AI Search - Génération augmentée de récupération (RAG)



IA & contexte d’affaire

Génération augmentée de récupération (RAG)

Interagir avec des données internes

L’architecture RAG permet aux LLMs de consulter de l’information présélectionnée spécifique à un domaine. Ceci réduit non seulement le risque d’hallucinations, mais permet aussi de chercher, de résumer ou d’analyser des documents internes.

Les avantages de l’architecture RAG

L’architecture RAG est relativement rapide à implémenter. En contrepartie, l'entraînement ou l’affinement des modèles est plus dispendieux et requiert un effort humain significatif.

Au delà du ChatBot

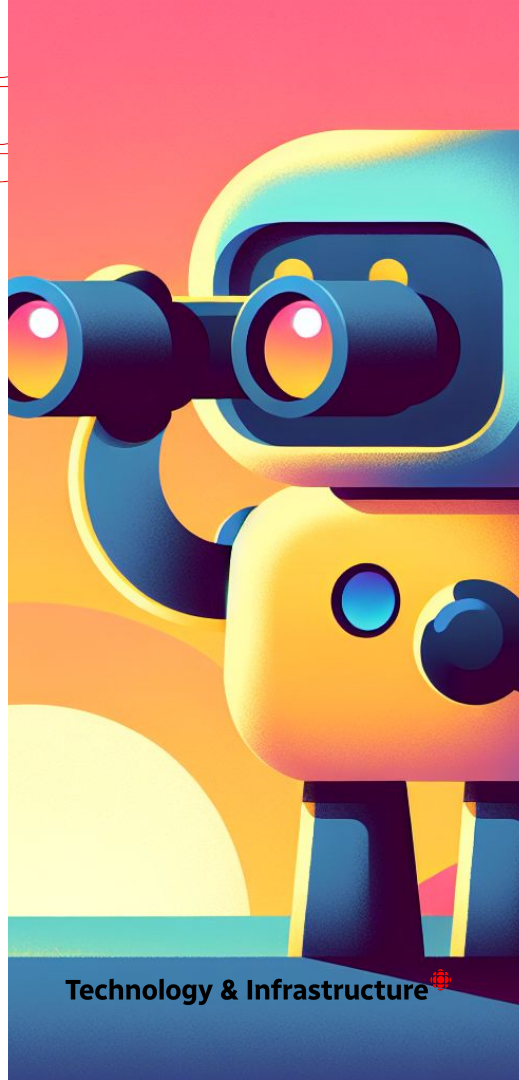
Apprivoiser le potentiel de l'IA générative pour l'automatisation des processus

Reconnaissance d'images

Les modèles de reconnaissance d'images et de reconnaissance vocale peuvent potentiellement être utilisés pour analyser des vidéos et appliquer automatiquement des actions prédéfinies

Automatisation des flux de travail

Le modèle GPT 4 Turbo comprend une fonctionnalité de création de JSON, ce qui pourrait potentiellement permettre d'utiliser les LLMs pour automatiser des processus et obtenir des résultats standardisés



Affinement des modèles

Les modèles de langage à grande échelle peuvent être affinés afin de les adapter à nos besoins et pour contourner certaines limitations ou filtres des modèles de base.

Afin d'affiner un modèle, il est nécessaire d’ingérer une quantité d’objets contenant un exemple de *prompt* et un résultat attendu.

Documentation additionnelle

- Service Azure OpenAI
- Article du Financial Times sur l'IA générative (Anglais)
- Guide de "Prompt Engineering"
- Raffinage Azure OpenAI
- Article sur les plongements vectoriels (Embeddings)
- Documentation Azure sur le RAG