

Partie 2 Architecture et déploiement en production

1. Déploiement d'un Agent IA (FastAPI + React) sur Azure avec CI/CD

Prérequis :

- Compte GitHub actif
- Projet backend avec main.py, requirements.txt (+ Dockerfile si déploiement via conteneur)
- Projet frontend avec package.json
- Compte Azure actif
- Abonnement active pour App Services
- Groupe de ressources créé dans Azure

Backend (FastAPI)

- Pousser le code sur GitHub
- Créer une Web App sur Azure
(Azure Portal → App Services → Créer une application Web)
 - Choisir pile d'exécution : Python 3.13
 - Choisir méthode de déploiement : GitHub Actions
- Connecter le repo GitHub
 - Lien le compte GitHub
 - Organisation
 - Dépôt
 - Branche
- Configuration automatique
 - crée un pipeline CI/CD avec GitHub Actions pour déployer automatiquement à chaque push
 - installe les dépendances à partir de requirements.txt
 - livre un lien public d'API
- Configurer les clés API / variables d'environnement
(Azure App Settings / GitHub Secrets)

Frontend (React)

- Pousser le code sur GitHub
- Créer une application web statique
(Azure Portal → App Services → Créer une application Web statique)
 - Choisir un plan d'hébergement: Free / Standard / Dedicated
- Connecter le repo GitHub :
 - Lien le compte GitHub
 - Organisation
 - Dépôt
 - Branche
- Configuration automatique
 - crée un pipeline CI/CD avec GitHub Actions
 - installe les dépendances à partir de package.json
 - livre un lien public d'API interface

Connexion Frontend ↔ Backend

Monitoring post-deployment : Azure Monitor pour gérer les logs, les metriques, les alertes.

Coût mensuel : 500 requêtes/jour = 25€/mois

Partie 3 Vision Language Model

1. Pas de configuration avec GPU
2. Pour traiter des documents contenant à la fois du texte et des éléments visuels (images, graphiques, schémas) dans un pipeline RAG, on peut convertir chaque page ou bloc du document en image, en extraire le texte, puis générer une requête représentative (query) pour chaque unité de contenu.

Ces données (image, texte, requête) sont stockées dans un fichier Parquet structuré. Ensuite, pour répondre à une question utilisateur, on transforme ce dataset en une base vectorielle multimodale : on encode à la fois le texte et l'image avec un modèle VLM, ce qui permet de faire une recherche sémantique efficace sur les documents complexes.

Demo :

Assistant IA Recherche Web

Pose-moi ta question 🗡️ Je cherche pour toi les infos les plus récentes sur Google!

Assistant IA Recherche Web

annonces cette semaine avec des sources fiables. Les trois documents fournis ne mentionnent pas de développements récents spécifiques. Pour répondre à votre question, j'aurais besoin d'avoir accès à des sources plus récentes et pertinentes sur les dernières avancées en IA générative.

Nouvelles Russie et Ukraine du juillet 2025 avec les sources.

1. La Russie a mené de nouvelles attaques nocturnes ciblant la ligne Soumy-Kharkiv-Dnipro, selon Volodymyr Zelensky. Source: https://www.lemonde.fr/international/live/2025/07/26/en-direct-guerre-en-ukraine-les-dernieres-informations_6622319_3210.html 2. La troisième session de pourparlers de paix entre Russie et Ukraine s'est terminée sans avancées significatives, tandis que la Russie poursuit ses avancées territoriales. Source: <https://www.tf1info.fr/international/direct-guerre-ukraine-russie-zelensky-poutine-les-dernieres-informations-aujourd-hui-dimanche-27-juillet-2025-2385342.html> 3. Un centre commercial a été bombardé par les forces russes à Kamianske, dans l'oblast de Dnipropetrovsk, en Ukraine, le 26 juillet 2025. Source: https://www.lemonde.fr/international/live/2025/07/27/en-direct-guerre-en-ukraine-les-dernieres-informations_6624528_3210.html