Stage Développeur Full-Stack

Durée estimée : 6 à 10 heures (rendu sous 48h max)

Contexte

IALab développe des outils d'analyse documentaire alimentés par IA. Tu dois créer un prototype permettant d'extraire le contenu d'un PDF et d'en générer une analyse structurée via un LLM.

Objectif

Développer une application Python composée de :

1. **Un service d'extraction** qui prend un fichier PDF en entrée et retourne son contenu textuel
2. **Une intégration LLM** qui analyse le document et retourne un output structuré (API au choix : OpenAI, Anthropic, Mistral...)
3. **Une interface web** permettant d'uploader un PDF et d'afficher les résultats de l'analyse de manière claire et lisible

Spécifications de l'analyse

Le LLM doit retourner un JSON avec la structure suivante :

json

```
{
  "titre": "Titre du document ou titre suggéré",
  "resume": "Résumé du contenu en 2-3 phrases",
  "mots_cles": ["mot1", "mot2", "mot3"],
  "type_document": "facture | contrat | article | rapport | cv | autre",
  "langue": "fr | en | autre"
}
```

Ces informations doivent être affichées de manière structurée dans l'interface (pas un JSON brut).

Contraintes techniques

- Langage : Python
- Interface : libre (Streamlit, Gradio, FastAPI + front séparé, Flask...)
- **Docker obligatoire** : l'application doit se lancer via `docker compose up`
- Si architecture front/back séparée → docker-compose multi-services

Livrables attendus

- Repository Git (GitHub, GitLab) contenant :
  - Le code source
  - `Dockerfile` et `docker-compose.yml`
  - `README.md` avec : instructions de lancement, choix techniques expliqués, limites connues
  - Un exemple de PDF testé avec le résultat obtenu (screenshot ou copie du JSON)

Critères d'évaluation

Critère	Poids
Qualité du code — lisibilité, structure, nommage	30%
Documentation — README clair, code commenté si nécessaire	20%
Simplicité d'usage — `docker compose up` et ça marche	20%
UX de l'interface — affichage clair des résultats structurés	15%
Robustesse — gestion des erreurs, cas limites	15%

Bonus (non obligatoires)

- Gestion des PDFs scannés (OCR)
- Gestion de documents volumineux (chunking)
- Tests unitaires
- Déploiement en ligne