

# Réunion 29 avril 2026

- Administratif et organisationnel
- Extraction des données

# Administratif et organisationnel

- Cumul CAPES - Doctorat
  - Experience
  - Présence aux oraux
- ollama sur CaSciModOT
- Préparation de DeepLearn lun 11/05 (heure à venir)

# Extraction des données - documents

- librairie choisie : PyMuPDF4LLM
- extraction au format Markdown
- Chunking : [langchain\\_text\\_splitters.MarkdownTextSplitter](#)
  - perfectible (cas du pdf journal\_officiel...pdf)
- Données extraites au format CSV
  - doc\_id
  - doc\_name
  - page\_number
  - chunk\_number (dans l'ordre du pdf)
  - text



# TODO

- Création de la base de données ChromaDB
- Récupération des documents extraits par le modèle lors d'une execution
- création de la sortie produite au format pour le challenge
- liste des perturbations possibles
  - récupération des mots répandus
  - déduction de perturbation
- execution de la pipeline sur les requêtes de test
-