Data battle 2025

Lilian Michel-Dansac, Pablo Canella, Mathis Firmino

Données et preprocess

- Structure interne des pdf -> certains étaient “cassés”
- pdf -> json avec un ocr
- faiss pour les chunks

Choix du modèle

Via Huggingface : mistralai/Mistral-7B-Instruct-v0.3

Autre modèles testés :

- meta-llama/Llama-3.2-3B-Instruct
- ministral/Ministral-3b-instruct

CodeCarbon preprocess

Ajout pdf : 4.2759177160106695e-06 kgCO2e

- CPU : 0.00050826098 kWh
- GPU : 0 kWh
- Total : 0.00052103089 kWh

Preprocess data : 0.010542363909032614 kgCO2e

- CPU : 1.237948753 kWh
- GPU : 0.06600149169003999 kWh
- Total : 1.34187343392 kWh

Démonstration