

# AI-Powered Document Search and Summarization System

</AI Detectives Team>

Catherine Maameri, Cristina Moussoungedi, Adel Zitouni



# Sommaire

**Introduction**

**Présentation du Sujet**

**Architecture & Fonctionnement**

**Conclusion & Perspectives**

**Appendix**

# Introduction

# Le défi : trop d'information, pas assez de temps



- ▶ Les documents longs sont difficiles à lire et analyser manuellement.
- ▶ La recherche d'informations pertinentes est lente et imprécise.
- ▶ Les résumés manuels demandent beaucoup de temps et d'effort.

Notre système RAG **retrouve et résume automatiquement** les informations pertinentes, rendant la recherche **plus rapide et plus claire**.

# Présentation du sujet



# Objectifs & Jeux de données

## Objectifs :

- Ingestion multi-formats (PDF, TXT)
- Recherche sémantique
- Résumé automatique
- Fonctionnement 100% local (CPU only)

## Jeux de données :

- Documents publics de l'UNESCO
- PDF + fichiers texte
- Contenu riche et structuré (définitions, analyses, politiques)
- Longueurs et styles différents → parfait pour évaluer le RAG

# Boîtes à outils utilisés

## BootCamp

- Python
- Jupyter / VS Code
- PyPDF2 (extraction PDF)
- Sentence-Transformers (embeddings)
- FAISS (similarity search)
- Transformers (t5-small)
- Vector store + metadata structure
- Chunking & preprocessing
- Evaluation methods (precision/recall)

## Appris par soi-même

- Human evaluation grid (clarity, relevance, coherence)
- Managing files and folder structure
- Colab integration in VS Code

- ▶ Documents collectés (PDF & TXT)
- ▶ "Text extraction" (PyPDF2)
- ▶ Cleaning & preprocessing
- ▶ Chunking of documents (200–500 tokens)
- ▶ Embedding generation (Sentence-Transformers)
- ▶ FAISS vector search for top-k relevant passages
- ▶ Custom retrieval function (get\_results)
- ▶ Summarization module (T5-small)
- ▶ Concatenation & summarization of top-k chunks
- ▶ Search evaluation (precision@k, recall@k)
- ▶ Human evaluation (clarity, relevance, fluency, coherence)
- ▶ Error handling for unsupported formats
- ▶ Metadata storage (source, chunk\_id, text)

## Fonctionnalités développées

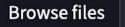


# Architecture et Fonctionnement

# Streamlit: Interface utilisateur

Déposez vos PDF/TXT (traitement immédiat)

 Drag and drop files here  
Limit 200MB per file • PDF, TXT



 UNESCO_Recommendation-on-Open-Scien...er2021.pdf	2.3MB	
 teachers_summary_1.txt	341.0B	
 teacher_shortages_notes.txt	269.0B	

Showing page 1 of 3  

2024\_TTF-UNESCO-Global-Report-on-Teachers\_EN.pdf ingéré (3 chunks)

open\_science\_summary\_1.txt ingéré (1 chunks)

preprocessing\_guide.txt ingéré (1 chunks)

semantic\_search\_intro.txt ingéré (1 chunks)

# Requête, Réponse et Résumé

## Recherche & Résumé

Votre requête

what is the link between teaching and open science

Lancer la recherche

Résultat 1 – UNESCO\_Recommendation-on-Open-Science\_november2021.pdf (chunk 6) | score=0.601

III. OPEN SCIENCE CORE VALUES AND GUIDING PRINCIPLES 13. The core values of open science stem from the rights-based, ethical, epistemological, economic, legal, political, social, multi-stakeholder and technological implications of opening science to society and broadening the principles of openness to the whole cycle of scientific research. They include the following: a. Quality and integrity: open science should respect academic freedom and human rights and support high-quality research by bringing together multiple sources of knowledge and making research methods and outputs widely available for rigorous review and scrutiny, and transparent evaluation processes. b. Collective benefit: as a global public good, open science should belong to humanity in common and benefit humanity as a whole. To this end, scientific knowledge should be openly available and its benefits universally shared. The practice of science should be inclusive, sustainable and equitable, also in opportunities for scientific education and capacity development. c. Equity and fairness: open science should play a significant role in ensuring equity among researchers from developed and developing countries, enabling fair and reciprocal sharing of scientific inputs and outputs and equal access to scientific knowledge to both producers and consumers of knowledge regardless of location, nationality, race, age, gender, income, socio-economic circumstances, career stage, discipline, language, religion, disability, ethnicity or migratory status, or any other grounds. d. Diversity and inclusiveness: open science should embrace a diversity of knowledge, practices, workflows, languages, research outputs and research topics that support the needs and epistemic pluralism of the scientific community as a whole, diverse research communities and scholars, as well as the wider public and knowledge holders beyond the traditional scientific community, including indigenous peoples and local communities, and social actors from different countries and regions, as appropriate. 17 14. The following guiding principles for open science provide a framework for enabling conditions and practices within which the above values are upheld, and the ideals of open science are made a reality: a. Transparency, scrutiny, critique and reproducibility: increased openness should be promoted in all stages of the scientific endeavour, with the view to reinforcing the strength and rigour of scientific results, enhancing the societal impact of science and increasing the capacity of society as a whole to solve complex

Résumé :

open science should respect academic freedom and human rights and support high-quality research by bringing together multiple sources of knowledge and making research methods and outputs widely available for rigorous review and scrutiny . the practice of science should be inclusive, sustainable and equitable, also in opportunities for scientific education and capacity development.

# Requête, Réponse et Résumé

## Recherche & Résumé

Votre requête

what is the recommendation on teaching

Lancer la recherche

Résultat 1 — teacher\_shortages\_notes.txt (chunk 0) | score=0.438

Notes on Teacher Shortages

- Growing global demand for qualified teachers
- High attrition rates in low-income countries
- Limited professional support reduces retention
- UNESCO calls for better teacher policies
- Strengthening training pipelines is essential

Résultat 2 — teachers\_summary\_1.txt (chunk 0) | score=0.416

UNESCO Global Report on Teachers (2024) – Summary The report highlights the global teacher shortage, especially in low-income countries. Key issues include lack of training, poor working conditions, and insufficient career development. UNESCO recommends increasing investment in teacher education and improving professional support.

Résumé :

UNESCO recommends increasing investment in teacher education and improving professional support . key issues include lack of training, poor working conditions, and insufficient career development .

# Conclusion et Perspectives



# Pourquoi notre solution fait la différence

**Notre solution transforme totalement la manière de rechercher et comprendre des documents volumineux.**

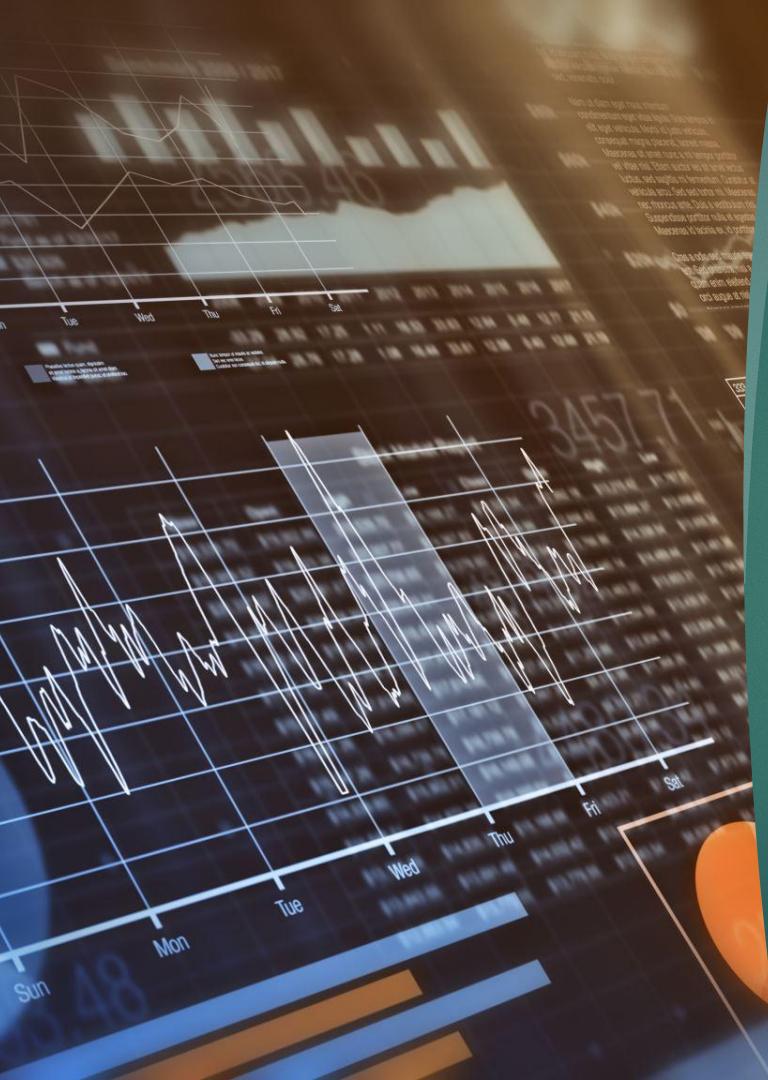
**Elle permet en quelques secondes :**

- **une recherche intelligente**, bien plus précise qu'un simple CTRL+F
- **des résumés clairs**, même pour des textes longs, denses, ou techniques
- **un accès instantané à l'information utile**, sans perdre du temps dans des centaines de pages
- **une interface simple**, directement utilisable par n'importe quel étudiant, chercheur ou professionnel

**En résumé : notre outil fait gagner du temps, apporte de la clarté, et améliore la prise de décision.**

# Limitations

- ▶ Données hétérogènes
- ▶ Test terrain non réalisé
- ▶ Petit Modèle (T5-small) → résumés parfois génériques
- ▶ Pas de fine-tuning



# Améliorations Futures

- ▶ Support DOCX
- ▶ Embeddings plus puissants
- ▶ Chunking dynamique
- ▶ Modèle de résumé plus grand
- ▶ Interface web complète
- ▶ Gestion du Multi-Langues
- ▶ Chatbot intelligent pour dialogue multi-questions

# 5. Appendix

# Difficultés surmontées

## Techniques

- Gestion de différents formats de fichiers (PDF, TXT)
- Extraction de texte parfois bruyante ou incomplète
- Découpage optimal des documents en chunks
- Création et interrogation de l'index FAISS
- Problèmes de compatibilité (VS Code ↔ Colab)
- Avertissements / limites du modèle T5-small
- Mise en place de l'évaluation (precision@k, recall@k)

## Humaines & Organisationnelles

- Coordination efficace dans l'équipe
- Répartition des tâches et synchronisation des versions
- Gestion du stress et du temps pendant le hackathon
- Adaptation rapide aux imprévus techniques
- Communication claire pour aligner les décisions
- Apprentissage de nouveaux outils en temps limité

# Ce dont nous sommes fiers

---

Construire un pipeline RAG complet en si peu de temps

---

Réussir à faire fonctionner la recherche FAISS et le résumé automatique

---

Transformer des documents bruts en réponses claires et structurées

---

Répartir les tâches et coordonner efficacement les tâches dans l'équipe

---

Surmonter les problèmes techniques (environnement, kernels, formats)

---

Communication et entraide constantes, malgré la pression du hackathon

---

Apprentissage rapide de nouveaux outils (Colab dans VS Code, FAISS, T5-small)

# Links

Github link (PUBLIC REPO):  
[HACKATHON-2](#)

2-mn video link :

<https://www.loom.com/share/4f35f8a7e97b45e28a883d3857b751f0>



Merci pour votre Attention!

Q&R