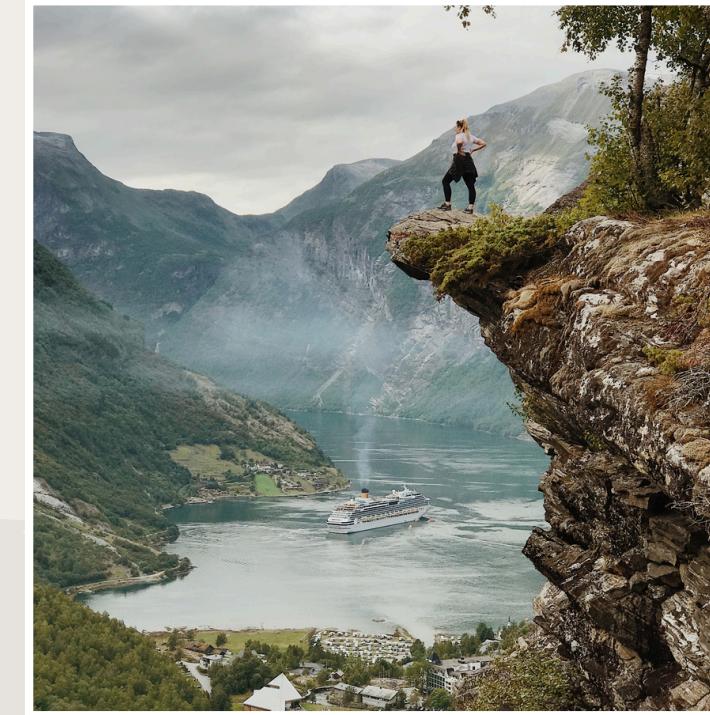


Samuel Beziat
Emma Choukroun
Jeanne Galet
Selima Khessairi

LO17
Projet 2 : RAG



GENOVA

guide touristique

Introduction

Contexte pédagogique

- Projet réalisé dans le cadre du cours LO17 – Indexation et Recherche d'Information

Choix du sujet

- Volonté de travailler sur un projet utile et engageant
- Réflexion collective : créer un guide touristique intelligent répondant à des questions pratiques

Présentation de GENOVA

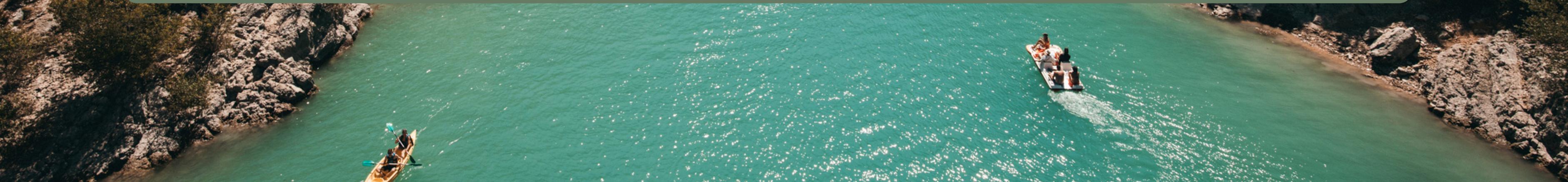
- GENOVA : Générateur de réponses touristiques avec navigation assistée
- Objectif : répondre à des requêtes comme :
 - "Que faire à Gênes en 2 jours ?"
 - "Où partir en septembre avec un petit budget ?"

Motivations personnelles

- Intérêt commun pour le voyage et les courts séjours
- Problèmes récurrents : information dispersée, difficilement accessible
- Objectif : centraliser les réponses touristiques simples et pertinentes

Objectifs techniques

- Mise en place d'un système RAG (Retrieval-Augmented Generation)
- Indexation dans ChromaDB
- Recherche sémantique + génération de réponse avec Gemini
- Améliorations du RAG naïf
- Interface utilisateur via Streamlit



Notre base de données



WIKIVOYAGE

Guide de voyage collaboratif et librement accessible.



LE ROUTARD

Guide de voyage pratique axé sur les bons plans et les conseils d'expérience.



**MINISTÈRE
DE L'EUROPE
ET DES AFFAIRES
ÉTRANGÈRES**

*Liberté
Égalité
Fraternité*

DIPLOMATIE.GOUV

Site officiel pour les conseils aux voyageurs et les informations sur la situation internationale.

Extraction, nettoyage et indexation du corpus touristique

1. Collecte ciblée de sources fiables

- **Objectif** : constituer une base documentaire exploitable par le système RAG
- **Sources choisies** :
 - Wikivoyage : contenu structuré (dump XML, 11554 pages)
 - Le Routard : retours d'expérience via crawling (28 032 pages)
 - Diplomatie.gouv.fr : fiches pays officielles (192 pages)
- **URLs extraites, filtrées et sauvegardées en CSV**

2. Chargement et nettoyage du contenu

- **URLs chargées** avec WebBaseLoader (LangChain)
- **Documents enrichis** de métadonnées (source, type, date)
- **Pas de nettoyage** du bruit des pages web
- **Traitement parallélisé** pour optimiser le temps
- **Fusion dans un fichier unique** (all_docs_lo17.pkl) : **39 000** documents prêts à vectoriser

3. Indexation vectorielle et résilience

- **Embeddings** générés par Gemini (embedding-001)
- **Indexation** dans ChromaDB (base vectorielle persistante)
- **Pas de chunking** car documents de petite taille
- **Filtrage** des documents vides
- **Gestion fine des quotas via** :
 - Batchs de 10 documents
 - Pause de 1 seconde
 - Checkpoints automatiques pour reprise en cas d'interruption

Améliorations effectuées sur notre RAG

prompt engineering renforcé

définition d'un rôle explicite, interdiction de suggestions ou de généralisations, exigence d'une formulation claire et directe, obligation de citer les URLs exactes des documents

transparence et traçabilité

sources explicites

mémoire conversationnelle

historique : intégration d'un module de mémoire (ConversationBufferMemory)

détection des hallucinations

repérage automatique de réponses potentiellement inventées

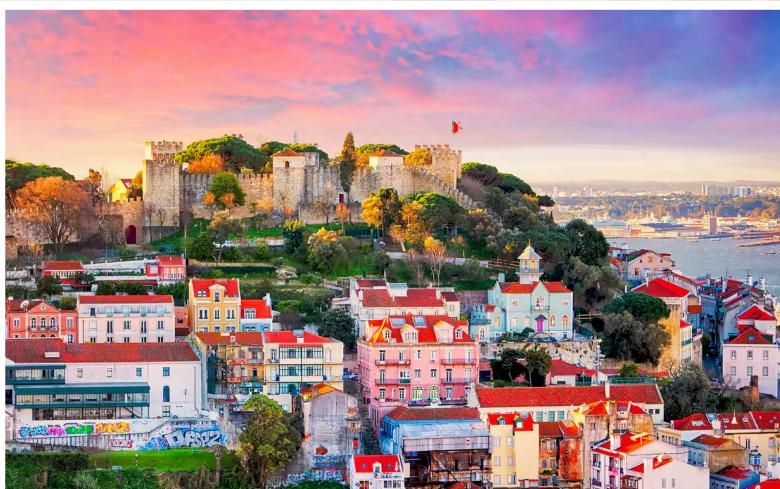
Exemple de questions



Quels sont les plats typiques de Naples ?



Quel est le climat en Albanie ?



Propose moi un itinéraire à Lisbonne



Que faire à Tokyo ?

Démonstration

Application déployée : [Genova - Guide de Voyage IA · Streamlit](#)

Interface

Nouveau chat

Historique

1. quels musées visiter à Rome ?

Exemples de questions

Que faire à Tokyo ?

Quels sont les plats typiques de Naples ?

Propose moi un itinéraire à Lisbonne

Quel est le climat en Albanie ?

Les plus beaux quartiers de Buenos Aires

Wednesday 18 June 2025

Genova

Votre compagnon de route, guide IA des voyageurs curieux

Posez votre question sur une destination

Où voulez-vous aller ? Que voulez-vous savoir ?

Vous : quels musées visiter à Rome ?

Genova : Voici quelques musées à visiter à Rome, selon le contexte fourni :

- **Musées du Vatican:** Ils abritent 10 musées, des galeries, et près de 1 400 salles. Les collections incluent de l'art antique, étrusque, égyptien, des œuvres de la Renaissance et de l'âge baroque, ainsi que de l'art contemporain. La chapelle Sixtine est un incontournable.
- **Musées capitolins:** Situés sur la colline du Capitole, ils sont répartis dans deux palais reliés par une galerie souterraine. Le Museo del palazzo dei Conservatori se trouve dans un palais du XVe siècle, tandis que le Museo del palazzo Nuovo a été inauguré en 1654.
- **Galleria Borghese:** Située au nord-est du parc de la Villa Borghese, elle abrite l'une des plus belles collections privées d'Italie.
- **Museo nazionale etrusco (dans la Villa Borghese)**
- **Galleria nazionale d'Arte moderna (dans la Villa Borghese)**
- **Palazzo Altemps**
- **Palazzo Massimo**
- **Centrale Montemartini:** Un musée de statues antiques situé dans une ancienne centrale électrique, dans le quartier de la Garbatella.
- **MAXXI:** Musée d'art contemporain, situé au nord de Rome.

Liens utiles :

- <https://www.routard.com/fr/guide/a/que-voir-que-visiter/europe/italie/rome>
- <https://www.routard.com/fr/guide/a/itineraires-conseilles/europe/italie/rome>
- <https://www.routard.com/fr/guide/a/liens-utiles/europe/italie/rome>
- <https://www.routard.com/fr/guide/europe/italie/rome/rome-antique/capitole-campidoglio/musei-capitolini-musees-capitolins>
- <https://www.routard.com/fr/guide/europe/italie/rome/vatican-et-environs/musei-del-vaticano>

Partie Evaluation

Dataset avec 6 questions:

- **user_input** : question évaluées,
- **response** : réponse générée,
- **retrieved_contexts** : contenu des sources,
- **reference** : réponse de référence.

Faithfulness

Fidélité au sens du contexte

Answer Relevancy

Alignement entre réponse générée et attendues

Context recall

Rappel de la réponse par rapport au contexte

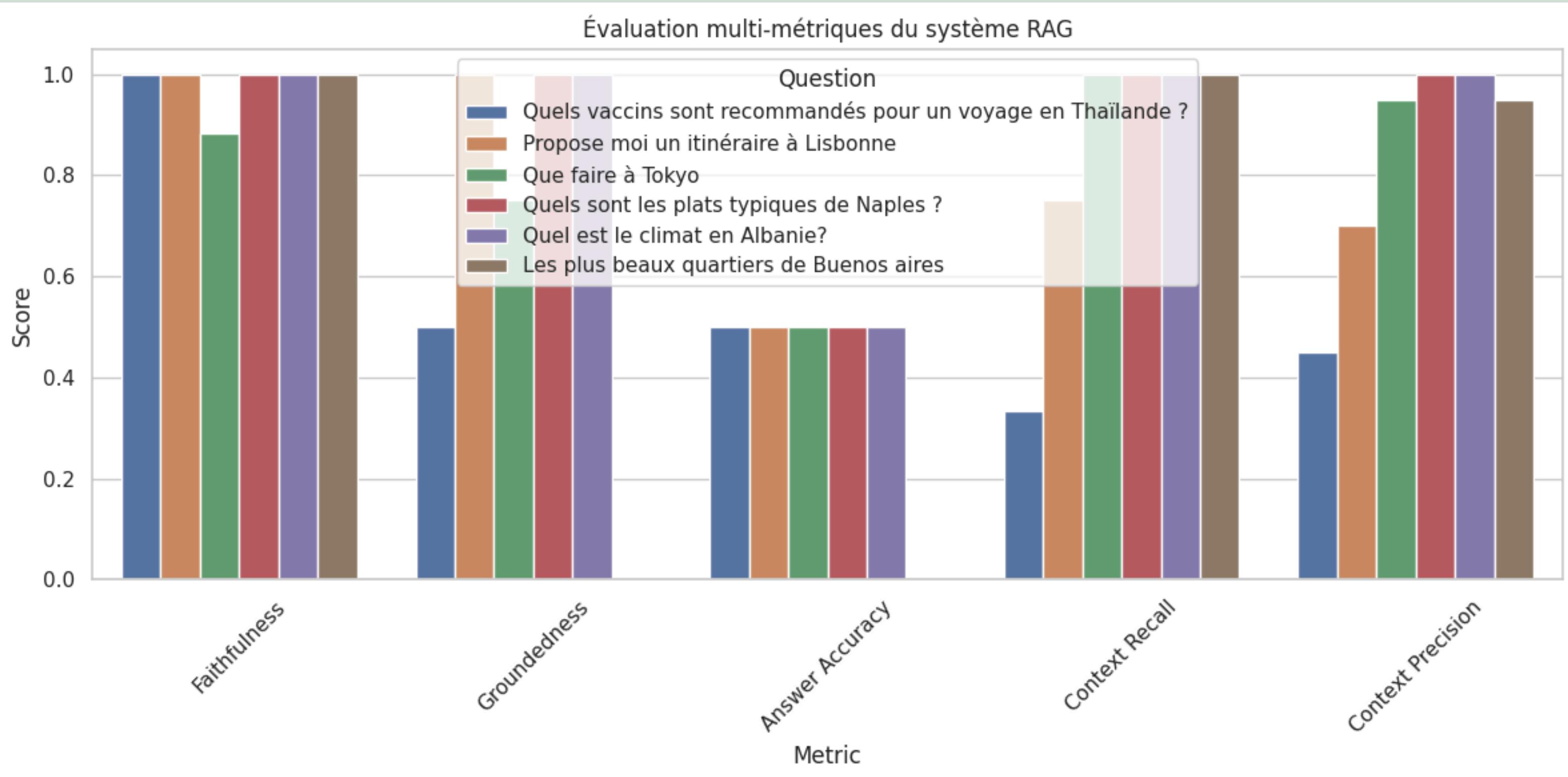
Groundedness

Conformité de la réponse au contexte

Context precision

Précision de la réponse par rapport au contexte

Partie Evaluation ★★★★★



Bilan de l'expérience

- Travail en groupe de 4 : répartition efficace, entraide constante
- Défis techniques : quotas API, gestion des appels, ajustement des workflows
- Utilisation de LangChain, embeddings, prompts sur-mesure
- Structuration manuelle du corpus : choix, nettoyage, vectorisation
- Collaboration agile : revues croisées, itérations rapides, tests réguliers
- Montée en compétences : pipeline RAG, intégration API, UI/UX Streamlit
- Projet formateur, ancré dans une problématique concrète et réaliste

Perspectives

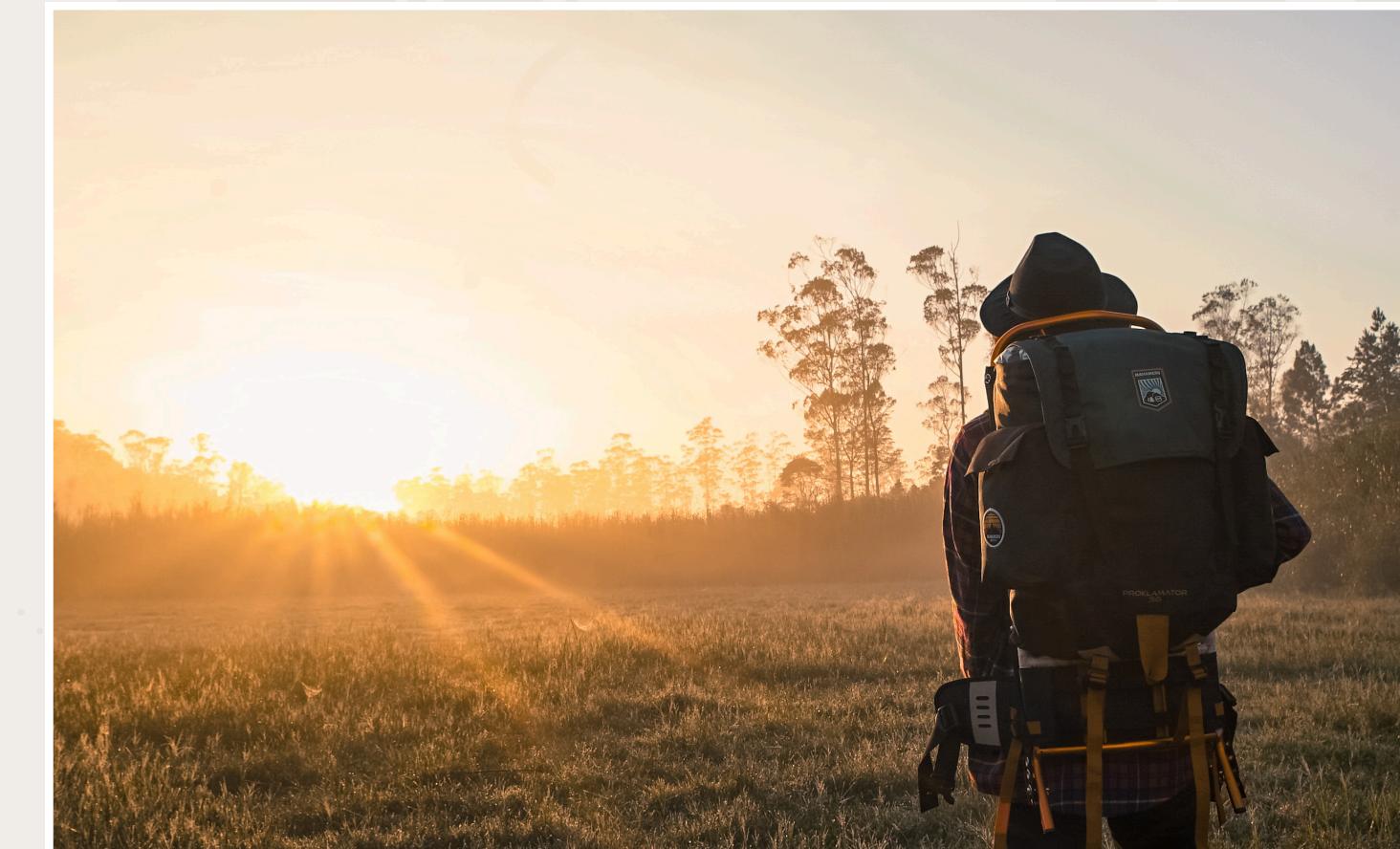
- Améliorer la synthèse (itinéraires, programmes de séjour)
- Intégrer un modèle open-source local (plus d'autonomie)
- Ajouter un module de fact-checking automatique
- Développer une version multilingue
- Expérimenter l'interaction vocale (assistant parlé)
- Enrichir l'interface avec du retour utilisateur
- Ouvrir à des sources plus variées (blogs, forums, avis voyageurs)
- Ajouter un système de personnalisation des réponses (profil utilisateur, préférences)

**LAISSEZ GENOVA
VOUS PLANNIFIER**

**VOTRE PROCHAINE
AVENTURE**



<https://genova-rag.streamlit.app/>



**MERCI POUR VOTRE ÉCOUTE,
N'HÉSITEZ PAS POUR TOUTE(S)
QUESTION(S) !**