Cahier des charges

Projet de Piscine – Intelligence Lab

Agent Vocal IA Agentique Multimatière en Temps Réel (100 % local)

Auteur : Romain Mallet
Référent Intelligence Lab : Martial ROBERGE
Durée estimée : 3 à 6 semaines
Environnement principal : Google Colab (local, GPU activé)

1. Contexte et objectif du projet

Ce projet a pour objectif la conception d'un **agent vocal d'intelligence artificielle en temps réel**, capable d'interagir oralement avec un utilisateur et de l'accompagner dans l'apprentissage de plusieurs matières (mathématiques, physique,français, anglais, etc.).

L'agent devra :

- Comprendre la voix de l'utilisateur via un module de reconnaissance vocale local (ASR).
- Exploiter un **RAG Agentique** propre à chaque matière, alimenté par des supports pédagogiques (cours, exercices et corrigés).
- Générer une réponse contextualisée en moins d'une seconde.
- Restituer la réponse simultanément à l'écrit et/ou à l'oral.
- Guider l'utilisateur par indices progressifs, sans jamais fournir la solution complète.

Le projet se déroule **entièrement en local**, sans utilisation d'API externes. L'ensemble des modèles (ASR, RAG, LLM/SLM, TTS) sera exécuté directement sur Google Colab, en exploitant le GPU lorsque disponible.

2. Objectifs pédagogiques

Ce projet vise à :

- Démontrer la maîtrise d'un pipeline complet d'intelligence artificielle temps réel (voix → texte → raisonnement avec RAG → voix).
- Explorer la mise en place d'un **système RAG agentique**, avec un module distinct pour chaque matière.
- Comprendre et appliquer les principes de la **génération augmentée par la recherche** sans recourir à des solutions cloud.
- Mettre en œuvre une approche d'apprentissage progressif où l'intelligence artificielle agit comme un tuteur plutôt qu'un fournisseur de réponses.
- Évaluer les limites de performances et de latence des modèles locaux (SLM).

3. Description fonctionnelle

3.1. Fonctionnalités principales

Fonctionnalité	Description
Reconnaissance vocale (ASR)	Transcription en temps réel de la voix de l'utilisateur à l'aide d'un modèle local (Whisper ou Faster-Whisper).
Sélection de la matière	L'utilisateur choisit la matière d'étude. L'agent utilise alors le RAG et le SLM correspondants.
RAG par matière	Chaque matière dispose d'un index FAISS alimenté par des documents spécifiques (cours, exercices, corrigés).
SLM spécialisé	Chaque RAG est associé à un petit modèle de langage local (SLM) adapté à la matière.
Génération d'indices progressifs	L'agent accompagne l'utilisateur avec des indices graduels au lieu de donner la solution.
Synthèse vocale (TTS)	Restitution vocale locale via un moteur open source (Piper ou Coqui TTS).
Réponse textuelle	Affichage simultané du texte généré pour assurer la compréhension complète.

3.2. Fonctionnalités optionnelles en fonction du temps restant

Fonctionnalité	Description
Interface graphique	Interface ergonomique (Gradio ou Streamlit) affichant les états d'écoute, de réflexion et de réponse.
Module de vision	Intégration d'un modèle multimodal permettant d'analyser des images d'exercices via la caméra l'utilisateur en temps réel.
Personnalisation des voix	Attribution d'une voix distincte à chaque matière afin de renforcer l'immersion pédagogique.

4. Architecture technique

4.1. Vue d’ensemble

```
Microphone
    ↓
Reconnaissance vocale (ASR)
    ↓
Texte de la question
    ↓
Sélection du RAG selon la matière
    ↓
Recherche des passages pertinents (FAISS)
    ↓
Réponse générée par un modèle SLM local
    ↓
Production d’indices progressifs
    ↓
Synthèse vocale (TTS) et affichage texte
```

4.2. Modèles utilisés

Type de modèle	Modèles envisagés	Taille / Dimensions	Utilisation principale
SLM (LLM local)	Phi-3.1 Mini / Qwen2.5-3B / Mistral 7B Instruct (selon GPU)	1B à 7B paramètres	Génération des réponses selon la matière
Moteur RAG	Sentence-Transformers + FAISS	N/A	Indexation et recherche de passages pertinents dans les documents
Embeddings	all-MiniLM-L6-v2 / bge-small	384–768 dimensions	Création des vecteurs pour l’index FAISS
Reranker (optionnel)	cross-encoder/ms-marco-MiniLM-L6-v2	N/A	Amélioration de la pertinence des résultats RAG

ASR (Speech-to-Text)	Faster-Whisper / Whisper.cpp	N/A	Transcription de la voix utilisateur en texte

5. Données et structure du projet

Chaque matière disposera de son propre répertoire contenant :

```
/data/{matiere}/
   ├── cours/
   ├── exercices/
   ├── corriges/
   ├── index.faiss
   └── meta.json
```

Les données d'entrée seront issues de ressources pédagogiques publiques ou de cours internes à l'ECE.
Les documents seront traités localement : extraction de texte via PyMuPDF, segmentation en chunks, embeddings, et indexation dans FAISS.

6. Indicateurs de performance

Critère	Objectif
Temps de réponse global	< 1 seconde pour la réponse textuelle et vocale
Précision RAG	≥ 80 % de passages pertinents
Taux d'erreurs ASR	< 10 % sur phrases courtes
Cohérence des indices	Validation sur un jeu d'exercices
Stabilité du système	10 interactions consécutives sans crash

7. Contraintes techniques

- Aucune dépendance à une API ou service externe.
- Exécution exclusivement en local sous Google Colab.
- Nécessité d'un GPU pour le chargement simultané du modèle LLM et du modèle ASR.
- Taille des modèles limitée pour garantir une latence inférieure à une seconde.
- Environnement Python 3.10 ou supérieur.

8. Livrables attendus

1. **Code source complet** du projet (fichiers Python, données et modèles) via un repo GitHub.
2. **Fichier README** détaillant l'installation, la configuration et le lancement.
3. **Documentation technique** expliquant le pipeline et les choix technologiques.
4. **Vidéo de démonstration** (1 à 2 minutes) présentant le fonctionnement en temps réel.
5. **Rapport synthétique** (3 à 4 pages) sur les résultats, difficultés rencontrées et enseignements.

9. Perspectives d'évolution

- Intégration d'un module de **vision locale** (analyse d'exercices en temps réel via la caméra de l'utilisateur).
- Création d'une **mémoire de session** pour un suivi personnalisé de l'élève.
- Déploiement sur un serveur cloud privé (GCP ou Hugging Face Spaces).
- Enrichissement des profils vocaux et comportementaux pour chaque "professeur".