Assistant Académique IA – HEC / Polytechnique

Document de travail – Discussion technique & stratégique

1. Vision & Philosophie

L’objectif n’est pas de créer un simple chatbot conversationnel, mais un agent académique institutionnel, structuré, gouverné, traçable et évolutif.

Principes directeurs :

- Architecture LLM-agnostique (pas de dépendance fournisseur, possibilité d’A/B testing)

- Approche RAG prioritaire (réponses fondées sur des sources)

- Validation humaine (human-in-the-loop)

- Conception institution-ready (logs, permissions, gouvernance)

2. Démo actuelle – Fonctionnalités existantes

Architecture :

- Pipeline RAG

- Base vectorielle

- Interface web

- Données publiques institutionnelles

- Score de confiance

- Système de feedback utilisateur

Fonctionnalités :

- Questions sur règlements académiques

- Deadlines publiques

- Structure des programmes

- Recherche sémantique multi-documents

- Citation des sources

- Détection des réponses à faible confiance

3. Stack technique cible

Frontend :

- Next.js (App Router)

- Tailwind CSS

Backend :

- Node.js (API routes)

- Validation stricte (Zod)

- Middleware de rate limiting

Base de données :

- PostgreSQL

- Extension pgvector

- Row-Level Security (RLS)

Couche RAG :

- LlamaIndex

- Recherche hybride (BM25 + vector search)

Couche LLM :

- Interface abstraite LLMProvider

- Implémentations OpenAIProvider / GeminiProvider

- Provider unique pour embeddings

Authentification :

- Phase 1 : Magic link

- Phase 2 : Intégration Azure AD / Microsoft Entra ID

Observabilité :

- Logs structurés

- Dashboard analytics

- Monitoring latence et coût

4. Fonctionnalités futures

Court terme :

- Back-office administrateur

- Import documents (PDF/DOCX/URL)

- Workflow Draft → Publish

- Dashboard analytics

Moyen terme :

- Authentification Microsoft

- Table deadlines structurée (SQL)

- Synchronisation calendrier (ICS)

Avancé :

- Agents autonomes supervisés

- Moteur de suggestion FAQ

- Clustering des questions récurrentes

- Digest hebdomadaire pour l’administration

5. Blocages identifiés

- Validation intégration Azure AD

- Accès aux API internes

- Gouvernance éditoriale

- Clarification conformité RGPD

- Positionnement institutionnel (officiel ou expérimental)

6. Besoins de collaboration

Technique :

- Optimisation avancée du RAG

- Sécurisation production-grade

Institutionnel :

- Accès API

- Intégration Azure AD

- Validation gouvernance des données

Stratégique :

- Cadre Hi! PARIS

- Structuration inter-écoles