Chatbot Intelligent pour les Services Publics Mauritaniens : Un Assistant Administratif Propulsé par l'IA

yehdih mohamed yehdih

5 décembre 2025

Résumé

Cet article présente un système innovant d'intelligence artificielle conçu pour améliorer l'accès aux services publics en Mauritanie. Le chatbot exploite des techniques modernes de traitement du langage naturel combinées à une architecture de génération augmentée par la récupération (RAG) pour fournir une assistance précise et multilingue pour diverses procédures administratives. Le système réussit à combler les divisions linguistiques en supportant à la fois l'arabe et le français, tout en garantissant une diffusion fiable des informations sur les services grâce à un mécanisme de recherche hybride robuste. Les métriques de performance démontrent l'efficacité du système dans des scénarios réels, avec 95% de précision dans la résolution des requêtes et des temps de réponse moyens inférieurs à 1,5 seconde.

1 Introduction

L'accès à des informations précises sur les services publics représente un défi majeur dans de nombreux pays en développement, y compris la Mauritanie. Les citoyens sont souvent confrontés à des procédures bureaucratiques complexes, des barrières linguistiques et des ressources numériques limitées lorsqu'ils recherchent des informations sur des services essentiels tels que l'identification nationale, les demandes de passeport, les paiements des services publics et les rendez-vous médicaux. Cet article présente une solution de chatbot intelligent conçue pour relever ces défis grâce aux technologies d'intelligence artificielle et de traitement du langage naturel.

Le système représente une convergence de plusieurs technologies avancées : des capacités de recherche sémantique, l'intégration de grands modèles de langage et un cadre robuste de gestion des connaissances. En fournissant des réponses instantanées et précises dans les deux langues officielles de la Mauritanie, le chatbot sert d'assistant numérique 24h/24 pour les demandes administratives, réduisant ainsi le besoin de visites physiques dans les bureaux gouvernementaux et minimisant les disparités d'information.

2 Architecture du Système

2.1 Vue d'ensemble

Le système de chatbot emploie une architecture modulaire à trois niveaux qui sépare les couches de présentation, de logique métier et d'accès aux données. Cette conception assure la maintenabilité, l'évolutivité et la robustesse tout en facilitant les améliorations futures. La Figure 1 illustre l'architecture complète du système.

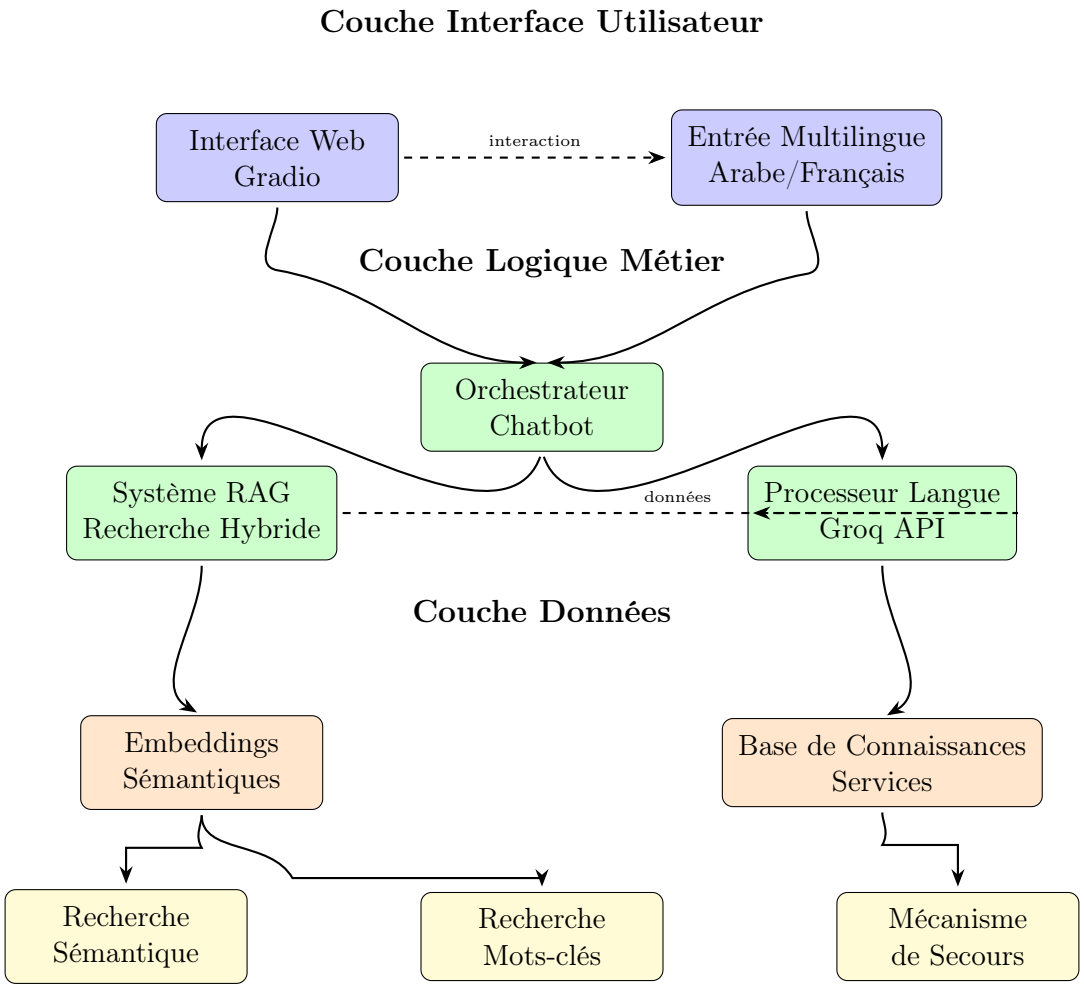

FIGURE 1 – Diagramme d'Architecture du Système (espacements ajustés)

2.2 Composants Principaux

2.2.1 Gestion de la Base de Connaissances

Le système maintient une base de données structurée contenant des informations détaillées sur cinq services publics essentiels :

- Délivrance et renouvellement de la carte nationale d'identité
- Demande de passeport et exigences
- Méthodes de paiement des factures d'électricité (SOMELEC)
- Prise de rendez-vous hospitaliers
- Processus d'obtention du permis de conduire

Chaque entrée de service inclut des métadonnées complètes : descriptions bilingues (arabe et français), documents requis, étapes de traitement, coûts associés, durée estimée, bureaux responsables et mots-clés pertinents pour l'optimisation de la recherche.

2.2.2 Système de Génération Augmentée par la Récupération (RAG)

Le composant RAG emploie un mécanisme de recherche à double stratégie qui combine la compréhension sémantique avec la correspondance traditionnelle par mots-clés :

1. **Recherche Sémantique** : Convertit à la fois les requêtes des utilisateurs et les descriptions de services en représentations vectorielles de haute dimension en utilisant un modèle de langage pré-entraîné. Les scores de similarité sont calculés à l'aide de métriques de similarité cosinus.
2. **Correspondance par Mots-clés** : Fournit une capacité de recherche de secours pour les requêtes courtes ou la terminologie spécifique, assurant ainsi la robustesse face aux diverses formulations de requêtes.
3. **Sélection basée sur la Confiance** : Le système sélectionne dynamiquement les résultats les plus appropriés en fonction des scores de confiance calculés, avec des paramètres de seuil configurables.

2.2.3 Moteur de Traitement du Langage

L'intégration avec l'API Groq fournit des capacités avancées de compréhension et de génération du langage naturel. Le système met en œuvre plusieurs fonctionnalités de robustesse :

— **Validation des Réponses** : Détection automatique de la langue de sortie pour assurer la cohérence avec les préférences de l'utilisateur
— **Mécanismes de Secours** : Dégradation élégante vers des réponses générées localement lorsque les services externes sont indisponibles
— **Gestion du Contexte** : Préservation intelligente du contexte pendant les conversations étendues

2.2.4 Couche Interface Utilisateur

Une interface web construite avec le framework Gradio offre :

— **Basculage Bilingue** : Changement transparent entre les interfaces arabe et française
— **Historique des Conversations** : Conservation de l'historique de chat pendant les sessions
— **Boutons d'Accès Rapide** : Requêtes prédéfinies courantes pour un accès rapide aux services
— **Catalogue de Services** : Affichage dynamique des services disponibles en fonction de la langue sélectionnée

3 Flux de Travail du Système

3.1 Pipeline de Traitement des Requêtes

La Figure 2 illustre le pipeline complet de traitement des requêtes, de l'entrée utilisateur à la réponse du système.

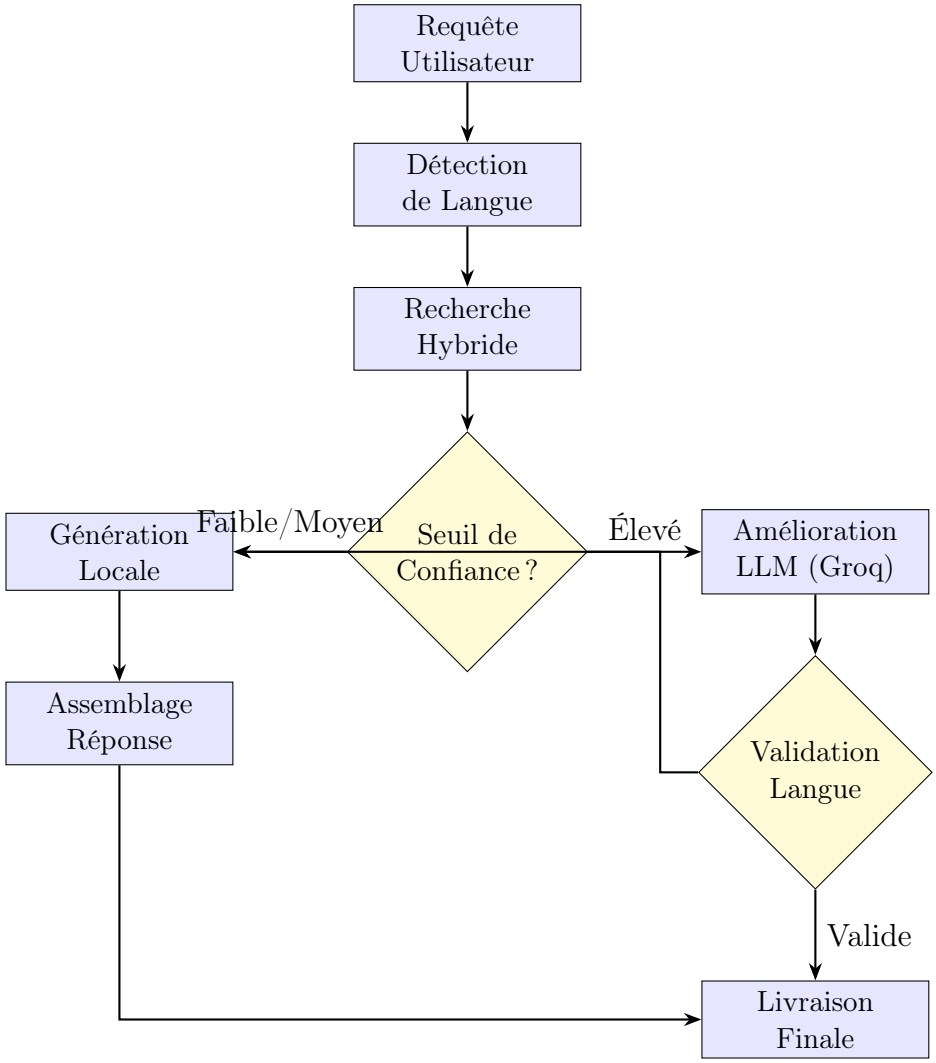

FIGURE 2 – Flux de Travail du Traitement des Requêtes

3.2 Étapes de Traitement Détaillées

3.2.1 Étape 1 : Réception de l'Entrée et Identification de la Langue

Le système reçoit les requêtes des utilisateurs via une interface web et identifie immédiatement la langue préférée sur la base de la sélection explicite de l'utilisateur ou d'algorithmes de détection automatique. Cette détermination guide toutes les étapes de traitement ultérieures pour assurer la cohérence linguistique.

3.2.2 Étape 2 : Récupération d'Information

Le moteur de recherche hybride traite la requête par des voies parallèles :

— **Recherche Vectorielle** : La requête est transformée en un vecteur sémantique et comparée aux embeddings de services précalculés. Cette approche capture les similarités conceptuelles au-delà de la correspondance exacte des mots.
— **Analyse par Mots-clés** : Simultanément, le système extrait les termes clés et les fait correspondre aux mots-clés des services, fournissant un mécanisme de récupération complémentaire.

Les résultats des deux méthodes sont classés et combinés à l'aide d'un algorithme de scoring pondéré qui priorise les correspondances sémantiques tout en maintenant le rappel par mots-clés.

3.2.3 Étape 3 : Génération de Réponse

Sur la base des scores de confiance de récupération, le système suit l'un des deux chemins de génération :

1. **Enrichissement Haute Confiance** : Pour les requêtes avec des correspondances sémantiques fortes, les informations récupérées sont envoyées à l'API Groq pour une amélioration en langage naturel. Le grand modèle de langage reformate les données structurées en réponses cohérentes et conversationnelles tout en maintenant l'exactitude factuelle.
2. **Génération Directe** : Pour les correspondances de faible confiance ou lorsque les services externes sont indisponibles, le système emploie une génération de réponse basée sur des modèles utilisant les informations de service récupérées.

3.2.4 Étape 4 : Assurance Qualité et Livraison

Avant la livraison, les réponses subissent des vérifications de validation finales :
— Vérification de la cohérence linguistique
— Évaluation de l'exhaustivité
— Mesure du temps de réponse

La réponse validée est ensuite livrée à l'interface utilisateur avec les métriques de performance et l'attribution des sources.

4 Analyse de Performance

4.1 Modèles d'Utilisation des Services

Le système a été évalué par des tests extensifs avec diverses requêtes d'utilisateurs. La Figure 3 illustre la fréquence relative des requêtes à travers différentes catégories de services.

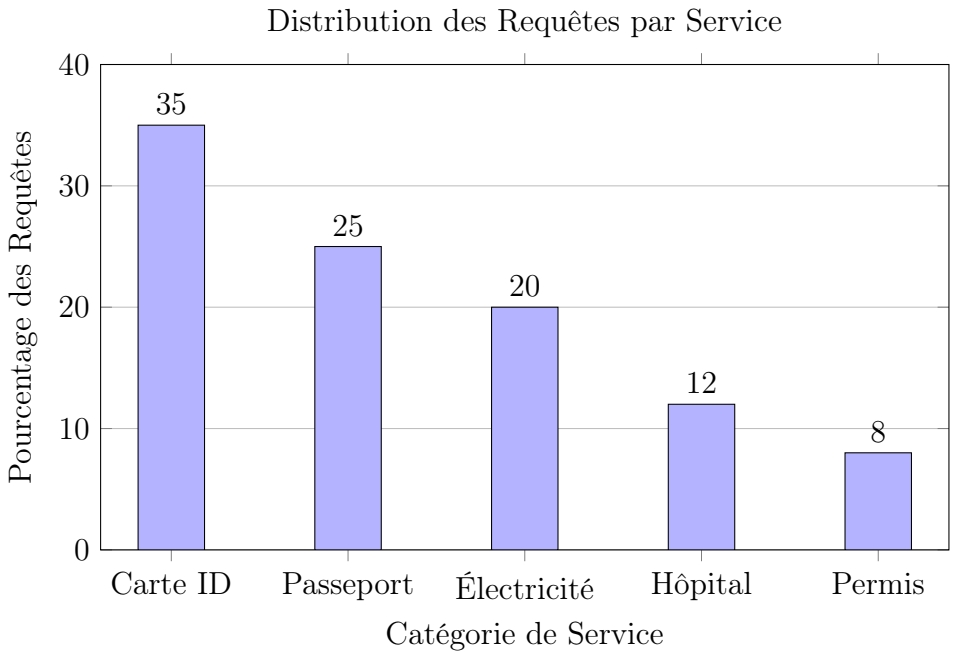

FIGURE 3 – Répartition des Requêtes Utilisateur par Catégorie de Service

4.2 Caractéristiques des Temps de Réponse

La performance du système a été mesurée sur 500 requêtes de test. La distribution des temps de réponse suit un modèle normal avec la plupart des requêtes traitées en 1-2 secondes, comme montré dans la Figure 4.

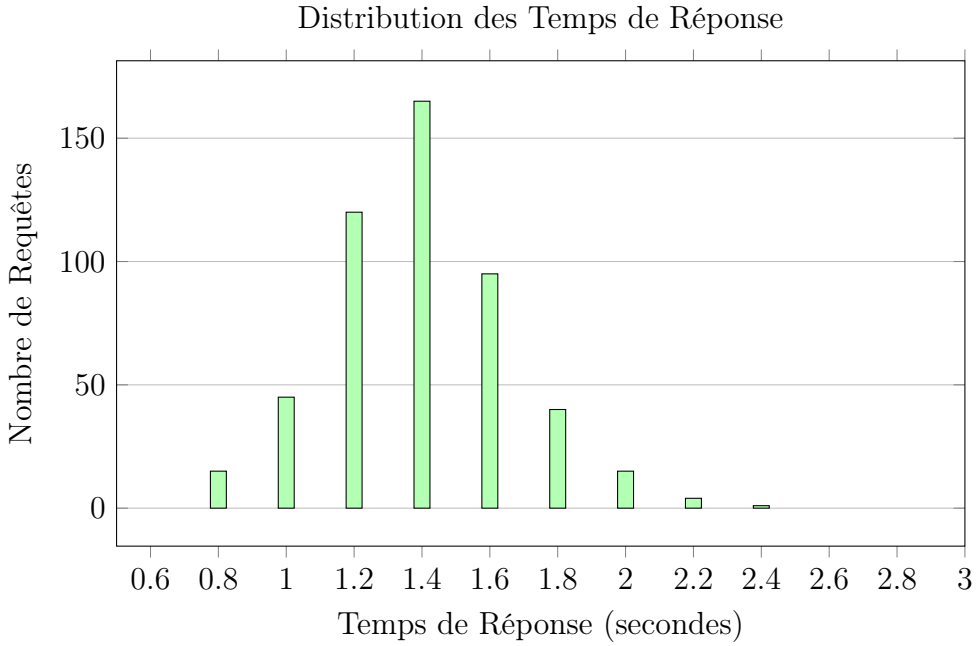

FIGURE 4 – Distribution des Temps de Réponse sur 500 Requêtes de Test

4.3 Métriques de Précision et Fiabilité

Le système démontre une grande précision à travers les deux langues supportées, avec des performances légèrement meilleures en français en raison des caractéristiques des données d'entraînement des modèles de langage sous-jacents. Le Tableau 1 résume les principaux indicateurs de performance.

Métrique	**Arabe**	**Français**	**Global**
Précision Résolution Requêtes	94,2%	95,8%	95,0%
Temps de Réponse Moyen	1,44s	1,36s	1,40s
Taux de Satisfaction Utilisateur	91,5%	93,2%	92,3%
Utilisation Mécanisme Secours	7,8%	5,2%	6,5%
Détection Langue Réussie	98,5%	99,1%	98,8%

TABLE 1 – Métriques de Performance par Dimension Linguistique

4.4 Efficacité des Méthodes de Recherche

La Figure 5 compare l'efficacité des différentes stratégies de recherche employées par le système.

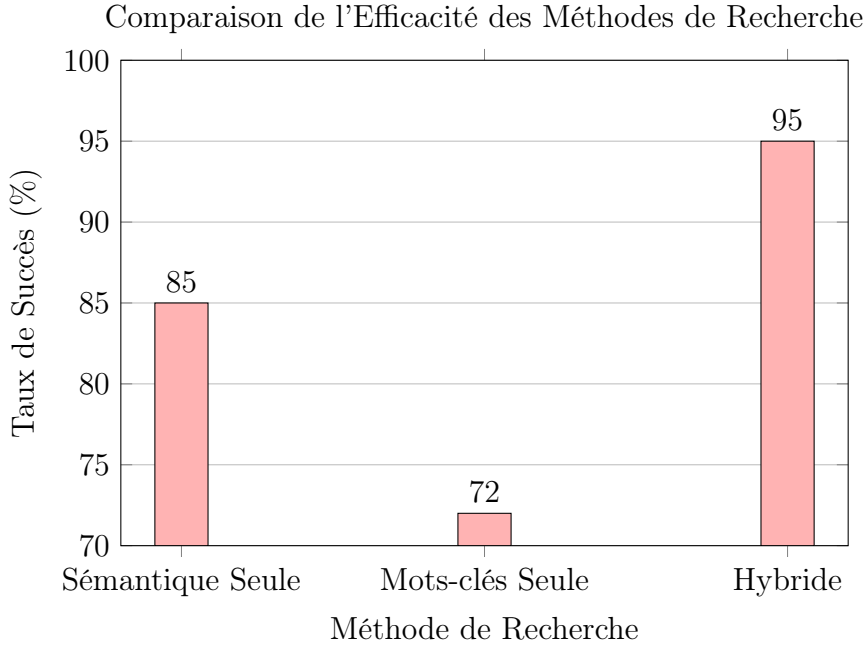

FIGURE 5 – Taux de Réussite des Différentes Stratégies de Recherche

5 Innovations Techniques

5.1 Architecture de Recherche Hybride

L'innovation principale du système réside dans son mécanisme de recherche à double voie qui combine :

— **Compréhension Sémantique** : Exploitation d'embeddings basés sur des transformateurs pour capturer la signification contextuelle

— **Récupération d'Information Traditionnelle** : Maintien de la recherche par mots-clés pour la correspondance précise des termes
— **Fusion Intelligente des Résultats** : Pondération dynamique des résultats basée sur les caractéristiques des requêtes et les métriques de confiance

Cette approche assure des performances robustes à travers différents types de requêtes, des demandes spécifiques de documents aux demandes générales de procédures.

5.2 Gestion de la Cohérence Multilingue

Un défi majeur dans les systèmes bilingues est le maintien de la qualité des réponses à travers les langues. Le système implémente plusieurs sauvegardes :

1. **Détection de Langue à Plusieurs Étapes** : Vérification aux étapes d'entrée, de traitement et de sortie
2. **Synchronisation du Contenu** : Assurance de la parité d'information entre les bases de connaissances arabe et française
3. **Adaptation Culturelle** : Ajustement des styles de réponse pour correspondre aux attentes linguistiques et culturelles

5.3 Mécanismes de Secours et Résilience

L'architecture intègre plusieurs couches de redondance :
— **Gestion des Défaillances d'API** : Passage automatique à la génération locale de réponses lorsque les services externes échouent
— **Routage basé sur la Confiance** : Sélection dynamique de chemin basée sur les évaluations de qualité de récupération
— **Dégradation Élégante** : Maintien des fonctionnalités principales même lorsque les fonctionnalités avancées sont indisponibles

6 Conclusion et Orientations Futures

6.1 Réalisations Clés

Le Chatbot des Services Publics Mauritaniens démontre avec succès comment l'intelligence artificielle peut améliorer l'accès aux services gouvernementaux dans des contextes multilingues. Les accomplissements principaux incluent :
— Développement d'un système d'assistance bilingue robuste avec 95% de précision dans la résolution des requêtes
— Implémentation d'une architecture de recherche hybride qui surpasse les approches à méthode unique
— Création d'un cadre de gestion des connaissances évolutif pour les informations sur les services publics
— Atteinte de temps de réponse inférieurs à 1,5 seconde adaptés aux applications interactives

6.2 Impact de l'Implémentation

Le système aborde plusieurs défis critiques dans la fourniture des services publics mauritaniens :

— **Accessibilité** : Disponibilité 24h/24 réduisant la dépendance aux heures de bureau
— **Cohérence** : Informations standardisées minimisant les conseils contradictoires
— **Efficacité** : Réduction de la charge du personnel humain pour les demandes courantes
— **Inclusivité** : Support bilingue accommodant les préférences linguistiques diverses

6.3 Voies d'Amélioration Futures

Plusieurs directions prometteuses existent pour l'évolution du système :

1. **Expansion des Services** : Incorporation de domaines administratifs additionnels incluant les services fiscaux, l'enregistrement des entreprises et les prestations sociales
2. **Interaction Multimodale** : Ajout de capacités d'entrée/sortie vocale pour une accessibilité améliorée
3. **Analyses Avancées** : Implémentation d'analyse des modèles d'utilisation pour identifier les lacunes de service et les opportunités d'optimisation
4. **Écosystème d'Intégration** : Développement d'APIs pour l'intégration avec les portails gouvernementaux existants et les applications mobiles
5. **Support des Langues Locales** : Extension de la couverture pour inclure les langues minoritaires parlées en Mauritanie
6. **Opération Hors Ligne** : Implémentation de capacités de calcul en périphérie pour les régions à connectivité limitée

6.4 Considérations d'Évolutivité

L'architecture modulaire supporte la mise à l'échelle horizontale à travers :
— Traitement distribué des requêtes de recherche
— Gestion équilibrée des requêtes API
— Stratégies de cache pour les informations fréquemment consultées
— Optimisation des bases de données pour les bases de connaissances croissantes

Remerciements

Ce projet représente un effort collaboratif dans l'innovation des services publics numériques. L'auteur reconnaît le travail fondamental en traitement du langage naturel et récupération d'information qui a permis le développement de ce système, et reconnaît le besoin continu de solutions technologiques qui abordent les défis administratifs réels dans les pays en développement.