

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي
المدرسة العليا للإعلام الآلي 08 ماي 1945 بسيدي بلعباس

People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

Higher School of Computer Science

-08 May 1945- Sidi Bel Abbès



THESIS

To obtain the diploma of **State Engineer**

Field : **Computer Science**

Specialty : **Artificial Intelligence and Data Science**

Design and Development of a Multilingual AI Assistant for Public Administration (CASNOS) Based on Retrieval-Augmented Generation (RAG)

Presented by :

Alaa Eddine Merzouk

In front of the jury composed of :

Mr./Dr. Jury President President

Dr. Belkacem Khaldi Supervisor

Mr. Kallaci Amine Supervisor

Dr. Jury Examiner Examiner

Academic Year : 2024/2025

Abstract

Abstract (English)

This thesis presents the design, implementation, and deployment of a multilingual AI-powered conversational assistant for CASNOS (Caisse Nationale de Sécurité Sociale des Non-Salariés), the Algerian social security institution for self-employed workers. The system leverages Retrieval-Augmented Generation (RAG) to provide accurate, contextually relevant answers in French, Arabic, and English.

The architecture integrates a semantic chunker for document preprocessing, a multilingual embedding model (multilingual-E5-base) for vector representation, a Qdrant vector store for efficient similarity search, a cross-encoder reranker (BGE-v2-m3) for precision refinement, and a large language model (LLM) served via vLLM. The system features a modular pipeline with separate public and admin APIs, enabling citizen-facing interactions and administrative oversight respectively.

A key innovation of this work is the implementation of an intelligent **orchestrator** that dynamically routes user queries into three specialized categories : **contact** (agency addresses and phone numbers), **calculation** (contribution and retirement computations), and **general** (all other CASNOS-related inquiries). Each route is processed through an optimized pipeline with domain-specific prompts and logic. The system also incorporates robust **input validation** that filters malicious, off-topic, or gibberish queries using a hybrid approach combining rule-based fast-paths and LLM-based verification, and an **output validation** mechanism that ensures response quality through both lexical analysis and an LLM judge, with automatic fallback to safe, pre-defined messages when necessary.

A comprehensive evaluation framework was implemented to assess chunking strategies and embedding model performance using metrics such as MRR, Hit@N, and answer containment rate. The application is fully containerized using Docker, ensuring portability, scalability, and ease of deployment across different environments. The system architecture includes a FastAPI-based public endpoint for stateless chat interactions with Server-Sent Events (SSE) streaming, an admin dashboard for conversation review and document management, and a complete re-indexing mechanism for the knowledge base.

Evaluation results demonstrate strong retrieval precision and multilingual coverage, with significant improvements over baseline keyword search. The project successfully

showcases the practical application of RAG systems in public administration, providing a scalable, maintainable, and production-ready solution for citizen information services.

Keywords : RAG, LLM, NLP, Chatbot, Multilingual, Qdrant, Sentence-Transformers, Orchestrator, Input Validation, Output Validation, Public Administration, CASNOS, Docker, Containerization.

Résumé (Français)

Ce mémoire présente la conception, l'implémentation et le déploiement d'un assistant conversationnel multilingue basé sur l'intelligence artificielle pour la CASNOS (Caisse Nationale de Sécurité Sociale des Non-Salariés). Le système utilise la génération augmentée par récupération (RAG) pour fournir des réponses précises et contextuellement pertinentes en français, arabe et anglais.

L'architecture intègre un découpeur sémantique pour le prétraitement des documents, un modèle d'embedding multilingue (multilingual-E5-base) pour la représentation vectorielle, une base de données vectorielle Qdrant pour la recherche par similarité, un re-classeur cross-encodeur (BGE-v2-m3) pour l'affinement de la précision, et un grand modèle de langue (LLM) servi via vLLM. Le système propose une architecture modulaire avec des API publiques et administratives distinctes, permettant respectivement les interactions avec les citoyens et la supervision administrative.

Une innovation clé de ce travail est l'implémentation d'un **orchestrateur** intelligent qui achemine dynamiquement les requêtes utilisateur vers trois catégories spécialisées : **contact** (adresses et numéros de téléphone des agences), **calcul** (calculs de cotisations et de retraites), et **général** (toutes les autres demandes liées à la CASNOS). Chaque route est traitée par un pipeline optimisé avec des prompts et une logique spécifiques au domaine. Le système intègre également une **validation d'entrée** robuste qui filtre les requêtes malveillantes, hors sujet ou incompréhensibles en utilisant une approche hybride combinant des chemins rapides basés sur des règles et une vérification par LLM, ainsi qu'une **validation de sortie** qui garantit la qualité des réponses à travers une analyse lexicale et un juge LLM, avec un repli automatique vers des messages sûrs et prédéfinis si nécessaire.

Un cadre d'évaluation complet a été mis en place pour évaluer les stratégies de découpage et les performances des modèles d'embedding à l'aide de métriques telles que MRR, Hit@N et le taux de réponse. L'application est entièrement conteneurisée avec Docker, garantissant la portabilité, l'évolutivité et la facilité de déploiement dans différents environnements. L'architecture comprend un endpoint public basé sur FastAPI pour les interactions sans état avec streaming Server-Sent Events (SSE), un tableau

de bord administrateur pour la révision des conversations et la gestion des documents, ainsi qu'un mécanisme complet de ré-indexation de la base de connaissances.

Les résultats d'évaluation démontrent une forte précision de recherche et une couverture multilingue efficace, avec des améliorations significatives par rapport à la recherche par mots-clés. Ce projet illustre avec succès l'application pratique des systèmes RAG dans l'administration publique, offrant une solution évolutive, maintenable et prête pour la production.

Mots-clés : RAG, LLM, TAL, Chatbot, Multilingue, Qdrant, Sentence-Transformers, Orchestrateur, Validation d'entrée, Validation de sortie, Administration publique, CASNOS, Docker, Conteneurisation.

ملخص (العربية)

يقدم هذا المذكرة تصميم وتطوير ونشر مساعد محادثة متعدد اللغات قائم على الذكاء الاصطناعي لصندوق الضمان الاجتماعي للعمال غير الأجراء، (CASNOS) المؤسسة الجزائرية للضمان الاجتماعي للعمال المستقلين. يستخدم النظام تقنية التوليد المعزز بالاسترجاع (RAG) لتقديم إجابات دقيقة وذات صلة بالسياق باللغات الفرنسية والعربية والإنجليزية.

تدمج البنية المعمارية نموذج تقطيع دلالي لمعالجة الوثائق مسبقاً، ونموذج تضمين متعدد اللغات (*multilingual-E5-base*) للتمثيل المتجهي، وقاعدة بيانات متجهية Qdrant للبحث بالتشابه، ومُعِد ترتيب من نوع cross-encoder (BGE-v2-m3) لتحسين الدقة، ونموذج لغوي كبير (LLM) يعمل عبر vLLM. يتميز النظام ببنية معيارية بواجهات برمجة تطبيقات (API) عامة وإدارية منفصلة، تتيح على التوالي التفاعل مع المواطنين والإشراف الإداري.

من الابتكارات الرئيسية لهذا العمل تطبيق منسق ذكي يقوم بتوجيه استفسارات المستخدم ديناميكياً إلى ثلاث فئات متخصصة: الاتصال (عناوين وأرقام هواتف الوكالات)، الحساب (حسابات الاشتراكات والتقاعد)، وعام (جميع الاستفسارات الأخرى المتعلقة بالصندوق). تتم معالجة كل مسار عبر خط أنابيب محسّن مع تعليمات ومنطق خاص بالمجال. يتضمن النظام أيضاً التحقق من صحة الإدخال القوي الذي يقوم بتصفية الاستفسارات الخبيثة أو غير ذات الصلة أو غير المفهومة باستخدام نهج هجين يجمع بين المسارات السريعة القائمة على القواعد والتحقق بواسطة LLM، بالإضافة إلى التحقق من صحة الإخراج الذي يضمن جودة الردود من خلال التحليل المعجمي وحكم LLM، مع العودة التلقائية إلى رسائل آمنة ومحددة مسبقاً عند الضرورة.

تم تنفيذ إطار تقييم شامل لتقييم استراتيجيات التقطيع وأداء نماذج التضمين باستخدام مقاييس مثل MRR و Hit@N ومعدل احتواء الإجابة. التطبيق معبأ بالكامل باستخدام Docker، مما

يضمن قابلية النقل والتوسع وسهولة النشر في بيئات مختلفة. تشمل البنية نقطة نهاية عامة تعتمد على FastAPI للتفاعلات عديمة الحالة مع بث أحداث الخادم، (SSE) ولوحة تحكم إدارية لمراجعة المحادثات وإدارة الوثائق، وآلية كاملة لإعادة فهرسة قاعدة المعرفة. تُظهر نتائج التقييم دقة استرجاع عالية وتغطية متعددة اللغات، مع تحسينات ملحوظة مقارنةً بالبحث بالكلمات المفتاحية. ينجح هذا المشروع في عرض التطبيق العملي لأنظمة RAG في الإدارة العامة، مقدماً حلاً قابلاً للتوسع والصيانة وجاهزاً للإنتاج لخدمات معلومات المواطنين.

الكلمات المفتاحية: RAG، LLM، معالجة اللغة الطبيعية، روبوت المحادثة، متعدد اللغات، Sentence-Transformers، Qdrant، التحقق من صحة الإدخال، التحقق من صحة الإخراج، الإدارة العامة، CASNOS، الحوسبة السحابية، Docker.