

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي و البحث 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



MASTER THESIS

Field: **Computer Science**
Specialty: **Artificial Intelligence and Data Science**

Retrieval-Augmented Generation for Multilingual Question Answering: Architectures, Challenges, and Future Directions

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
<i>Academic Year: 2024/2025</i>	

Abstract:

Large Language Models (LLMs) have achieved remarkable performance across a wide spectrum of natural language processing tasks. However, their tendency to generate factually incorrect information — commonly referred to as *hallucination* — and their inability to access knowledge beyond their training data represent fundamental limitations for knowledge-intensive applications. Retrieval-Augmented Generation (RAG) addresses these limitations by coupling a parametric generative model with a non-parametric retrieval mechanism over an external knowledge base.

This thesis presents a comprehensive survey of RAG systems with a particular focus on multilingual and cross-lingual question-answering scenarios, paying special attention to the challenges posed by morphologically rich and low-resource languages such as Arabic. We systematically review the theoretical foundations of text representation and transformer-based language models, analyse the core components of RAG pipelines — including chunking strategies, dense and sparse retrieval, neural reranking, and vector database indexing — and examine advanced architectures such as Modular RAG, Self-RAG, and Corrective RAG.

We further investigate the landscape of multilingual embedding models, Arabic NLP resources and models, language detection mechanisms, and evaluation frameworks for both retrieval and generation quality. The survey concludes with a synthesis of open research challenges and promising future directions, including multimodal RAG, agentic retrieval, and the ethical dimensions of deploying multilingual AI systems in under-resourced linguistic communities.

Keywords: Retrieval-Augmented Generation, Large Language Models, Multilingual NLP, Arabic NLP, Vector Databases, Question Answering, Hallucination, Reranking, Dense Retrieval.

Résumé:

Les grands modèles de langue (LLM) ont atteint des performances remarquables dans un large spectre de tâches de traitement automatique des langues. Cependant, leur tendance à générer des informations factuellement incorrectes — communément appelée *hallucination* — ainsi que leur incapacité à accéder à des connaissances au-delà de leurs données d’entraînement constituent des limitations fondamentales pour les applications à forte intensité de connaissances. La génération augmentée par récupération (RAG) répond à ces limitations en couplant un modèle génératif paramétrique à un mécanisme de récupération non paramétrique sur une base de connaissances externe.

Ce mémoire présente une revue complète des systèmes RAG avec un focus particulier sur les scénarios de questions-réponses multilingues et inter-langues, en accordant une attention spéciale aux défis posés par les langues morphologiquement riches et à faibles ressources telles que l’arabe. Nous passons en revue les fondements théoriques, les composants essentiels des pipelines RAG et les architectures avancées, avant de conclure par une synthèse des défis ouverts et des perspectives de recherche prometteuses.

Mots-clés : Génération augmentée par récupération, grands modèles de langue, TALN multilingue, TALN arabe, bases de données vectorielles, questions-réponses, hallucination.

ملخص:

حققت نماذج اللغة الكبيرة أداءً استثنائياً في طيف واسع من مهام معالجة اللغة الطبيعية. غير أن ميلها إلى توليد معلومات خاطئة من الناحية الواقعية — وهو ما يُعرف بظاهرة الهلوسة — وعجزها عن الوصول إلى المعرفة التي تتجاوز بيانات تدريبها يُمثّلان قيوداً جوهرية في التطبيقات كثيفة المعرفة. تعالج تقنية التوليد المعزّز بالاسترجاع (RAG) هذه القيود من خلال ربط نموذج توليدي ذي معاملات بآلية استرجاع غير معاملاتية تعمل على قاعدة معرفة خارجية.

تقدّم هذه الأطروحة مراجعة شاملة لأنظمة التوليد المعزّز بالاسترجاع، مع تركيز خاص على سيناريوهات الإجابة على الأسئلة متعددة اللغات وعبرها، مع الاهتمام البالغ بالتحديات التي تطرحها اللغات الغنية صرفياً وشحيحة الموارد كاللغة العربية. نستعرض بشكل منهجي الأسس النظرية لتمثيل النصوص ونماذج اللغة المبنية على بنية المحوّل، ونحلل المكونات الأساسية لقنوات معالجة التوليد المعزّز بالاسترجاع — بما فيها استراتيجيات التقطيع، والاسترجاع الكثيف والمتفرق، وإعادة الترتيب العصبي، وفهرسة قواعد البيانات المتجهية — ثم نفحص البنى المعمارية المتقدمة كـ RAG Modular و Self-RAG و RAG. Corrective.

نتناول كذلك واقع نماذج التضمين متعددة اللغات، وموارد ونماذج معالجة اللغة العربية، وآليات الكشف عن اللغة، وأطر تقييم جودة الاسترجاع والتوليد. تختتم المراجعة بتوليف للتحديات البحثية المفتوحة والاتجاهات المستقبلية الواعدة، بما فيها التوليد المعزّز بالاسترجاع متعدد الوسائط، والاسترجاع الوكيل، والأبعاد الأخلاقية لنشر أنظمة الذكاء الاصطناعي متعددة اللغات في المجتمعات اللغوية شحيحة الموارد.

الكلمات المفتاحية: التوليد المعزّز بالاسترجاع، نماذج اللغة الكبيرة، معالجة اللغة الطبيعية متعددة اللغات، معالجة اللغة العربية، قواعد البيانات المتجهية، الإجابة على الأسئلة، الهلوسة، إعادة الترتيب، الاسترجاع الكثيف.