ملخص

أدى التعقيد المتزايد في الأنظمة القانونية، إلى جانب الطبيعة غير المهيكلة والنطاق الإقليمي الخاص للوثائق القانونية، خاصة في الأنظمة متعددة اللغات، إلى ازدياد الاهتمام باستخدام تقنيات معالجة اللغة الطبيعية لتحسين الوصول إلى النصوص القانونية وفهمها. وقد أسهمت التطورات الحديثة في تقنيات الذكاء الاصطناعي في تقديم حلول واعدة من خلال الجمع بين توليد اللغة واسترجاع المستندات بشكل فوري. ومع ذلك، لا تزال الأنظمة الحالية تعاني من ضعف في عملية الاسترجاع، حيث تعتمد على عدد ثابت من المستندات، مما يؤدي إلى توليد معلومات غير دقيقة، أو تقديم محتوى غير كافٍ أو مشوش، بالإضافة إلى ضعف التوافق مع استفسارات المستخدمين. هذه التحديات تؤثر بشكل كبير على فعالية هذه الأنظمة في السياقات القانونية الحساسة. تعالج هذه الأطروحة هذه القيود من خلال اقتراح آلية جديدة لاختيار المستندات تعتمد على تكييف عددها حسب تعقيد الاستفسار، بدلاً من استخدام عدد ثابت. كما تسهم الأطروحة في تطوير مجموعة بيانات قانونية باللغة العربية مأخوذة من أحكام قضائية جزائرية، وتعرض واحدة من أولى المنظومات المتكاملة التي تجمع بين الاسترجاع والتوليد باللغة العربية. وقد تم تطوير وكيل حواري قادر على فهم الاستفسارات القانونية بالعربية وتقديم إجابات دقيقة ومرتكزة على السياق، مما يحسن من إمكانية الوصول إلى المعلومات القانونية في البيئات ذات الموارد المحدودة. ومن خلال معالجة التحديات الأساسية في الاسترجاع والتوليد، تسعى هذه الأطروحة إلى دعم البحث في المعالجة القانونية للغة العربية وتوفير أدوات عملية لتطوير مساعدين قانونيين ذكيين في العالم العربي.

كلمات مفتاحية: معالجة اللغة الطبيعية، الذكاء الاصطناعي القانوني، نماذج اللغة الضخمة، التوليد المعزز بالاسترجاع، المعالجة القانونية للغة العربية، استرجاع المعلومات القانونية، الوكلاء الحواريون، مجموعة بيانات قانونية جزائرية، الأنظمة القانونية متعددة اللغات.

Abstract

The growing complexity of legal systems, combined with the unstructured and region-specific nature of legal documents—particularly in multilingual jurisdictions—has sparked increasing interest in the use of Natural Language Processing (NLP) to enhance access to and comprehension of legal texts. Recent advances in Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) architectures offer promising solutions by combining language generation with real-time document retrieval. However, current RAG systems typically rely on fixed top-k document selection, which can result in hallucinations, insufficient or noisy context, and poor alignment with user queries. These retrieval inefficiencies significantly degrade the performance of LLMs in sensitive legal contexts. This thesis addresses these limitations by proposing a novel dynamic candidate selection mechanism that adapts retrieval thresholds based on query complexity. Furthermore, it contributes a new Arabic legal case dataset sourced from Algerian court rulings and introduces one of the first Arabic RAG pipelines that fully integrates retrieval and generation components. The developed conversational agent is capable of understanding legal Arabic queries and delivering grounded, contextually accurate responses—thus improving legal accessibility in under-resourced settings. By tackling key challenges in both retrieval and generation, this thesis advances research in Arabic legal NLP and provides practical tools for developing intelligent legal assistants in the Arab world.

Key words: *Natural Language Processing (NLP), Legal Artificial Intelligence, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Arabic Legal NLP, Dynamic Candidate Selection, Legal Information Retrieval, Conversational Agents, Algerian Legal Dataset, Multilingual Legal Systems.*