ERAG : le RAG fonctionne-t-il pour la recherche d'experts sur le TREC Enterprise Track ?

Sarah Nouali Ismail Badache Patrice Bellot
Aix-Marseille University, Université de Toulon, CNRS, LIS, Marseille, France
{sarah.nouali, ismail.badache, patrice.bellot}@lis-lab.fr

RÉSUMÉ

La recherche d'experts est depuis longtemps une application courante de la recherche d'information (RI). Les premiers travaux se concentraient principalement sur des modèles probabilistes basés principalement sur l'analyse de documents. Cependant, l'émergence des grands modèles de langage (LLM) a ouvert de nouvelles possibilités. Dans cet article, nous revisitons la piste TREC Enterprise en appliquant des méthodes récentes basées sur les LLM. Nous visons à évaluer leur efficacité dans la recherche d'experts, en examinant dans quelle mesure les techniques de génération à enrichissement contextuel (RAG) aident à identifier correctement des experts, système que nous nommons dans cet article : ERAG (Expert Retrieval with Augmented Generation). Nos expériences ont été menées sur la collection de tests CERC (CSIRO), l'un des premiers benchmarks dans ce domaine. Il nous a permis d'évaluer les deux phases d'un RAG : la recherche documentaire et le classement des experts. Nos résultats montrent que le RAG égale et surpasse les performances les plus élevées rapportées dans la présentation originale de TREC 2007 Enterprise Track, soulignant le potentiel de ces méthodes pour la recherche d'experts.

ABSTRACT

ERAG : Does RAG work for expert search on the TREC Enterprise Track ?

Expert retrieval has long been a common problem in information retrieval (IR). Early works focused primarily on document-based probabilistic models. However, the emergence of large language models (LLM) has opened up new possibilities. In this paper, we revisit the TREC Enterprise track by applying recent LLM-based methods. We aim to evaluate their effectiveness in expert retrieval by examining the extent to which retrieval-augmented generation (RAG) techniques help identify experts, a system that we call in this paper : ERAG (Expert Retrieval with Augmented Generation). Our experiments were conducted on the CERC test collection (CSIRO), one of the first benchmarks in this field. It allowed us to evaluate document retrieval and expert ranking. Our results show that RAG matches and surpasses the highest performance reported in the original TREC 2007 Enterprise Track overview, highlighting the potential of these methods for expert retrieval.

MOTS-CLÉS : Recherche d'information, Grands modèles de langage, Recherche d'experts, Génération à enrichissement contextuel, TREC Enterprise, RAG.

KEYWORDS: Information Retrieval, Large Language Models, Expert Search, Retrieval-Augmented Generation, TREC Enterprise, RAG.

1 Introduction

La croissance rapide et exponentielle de l'information numérique a fait du Web une source centrale de connaissances, mais il est également devenu de plus en plus difficile d'y naviguer pour trouver des informations fiables et identifier les bonnes personnes possédant l'expertise appropriée (Balog *et al.*, 2012). Ces personnes seraient en mesure de répondre à des questions, de collaborer entre elles sur des tâches ou de monter des projets. La recherche d'experts, contrairement à la recherche de documents qui fait référence à la capacité de trouver des textes pertinents, consiste à identifier des personnes correspondant aux critères recherchés à partir de ce que l'on peut savoir d'elles sur le Web (sites Web personnels, publications...).

En 2005, la conférence Text REtrieval Conference (TREC) [1] a lancé la piste Enterprise, axée à la fois sur la recherche documentaire et la recherche d'experts (Craswell *et al.*, 2005). Active jusqu'en 2008 (Craswell *et al.*, 2005; Soboroff *et al.*, 2006; Bailey *et al.*, 2007b; Balog *et al.*, 2008), elle a fourni les premiers benchmarks pour évaluer les approches de recherche d'experts. De nombreux modèles fondamentaux ont été développés et ces benchmarks demeurent une référence essentielle, même s'ils ne sont plus distribués par les organisateurs (NIST [2]).

Dans cet article, nous revisitons cette piste en appliquant des méthodes récentes basées sur les grands modèles de langage (LLM). Notre objectif est d'évaluer l'efficacité des LLM dans la recherche d'experts sur l'un des jeux de données TREC Enterprise, en examinant si les techniques de génération à enrichissement contextuel (RAG)(Lewis *et al.*, 2020) peuvent identifier efficacement les experts du domaine. À cette fin, nous proposons ERAG (*Expert Retrieval with Augmented Generation*), une approche combinant LLM et RAG pour améliorer l'identification et la qualification des experts à partir de données anciennes. Cet ensemble de données est particulièrement difficile à traiter, en raison de son âge, de la présence de données HTML bruitées et non nettoyées, et de sa structure complexe, car il a été initialement construit à partir de collections d'e-mails. Notre principale question de recherche est la suivante : quelle contribution les techniques modernes basées sur les LLM et le RAG peuvent-elles apporter à la recherche d'experts en les appliquant à un ensemble de données vieux de près de 20 ans ?

Le reste de cet article est organisé comme suit. La section 2 passe en revue le contexte et les travaux connexes. La section 3 décrit notre approche de recherche d'experts basée sur le RAG. Les sections 4 et 5 présentent le protocole d'évaluation ainsi que les résultats expérimentaux. Enfin, la section 6 conclut l'article et présente quelques perspectives.

2 Contexte et état de l'art

Cette section présente : i) des informations générales sur les ensembles de données TREC Enterprise, et ii) quelques travaux de recherche d'experts spécifiquement réalisés sur TREC Enterprise.

2.1 Jeux de données TREC Enterprise

La piste TREC Enterprise a fourni deux collections de test : la collection Wide Web Consortium (W3C) pour les éditions 2005 et 2006, et la collection CSIRO Enterprise Search test Collection (CERC) (Bailey *et al.*, 2007a) pour les éditions 2007 et 2008 qui est celle employée ici. Le premier

1. https://trec.nist.gov/
2. https://trec.nist.gov/data/enterprise.html

ensemble de données W3C a été le premier benchmark pour la recherche d'experts. Le corpus comprenait la documentation interne du World Wide Consortium W3C, récupérée sur les sites publics « w3.org » en juin 2004. Cependant, malgré sa valeur intrinsèque, les évaluations présentaient des limites en termes de précision en raison de l'absence d'évaluateurs experts et des jugements de pertinence restreints. Ces limites ont incité l'équipe TREC à constituer un autre ensemble de données, mais cette fois en utilisant les jugements d'experts. L'ensemble de données CERC a été constitué à partir des pages accessibles au public du domaine web « csiro.au » de l'agence scientifique nationale australienne Commonwealth Scientific and Industrial Research Organization (CSIRO)[3], extraites en mars 2007.

2.2 Recherche d'experts sur les ensembles de données TREC Enterprise

Les principales approches explorées sur le corpus TREC durant la période pré-LLM étaient les suivantes : i) les approches basées sur le contenu ; ii) les approches basées sur les réseaux et iii) les approches hybrides.

Tout d'abord, les approches basées sur le contenu(Fang & Zhai, 2007; Balog, 2007; Shen *et al.*, 2007; Petkova & Croft, 2008) ont été étudiées de manière approfondie au cours des éditions TREC. Elles comprennent deux types de modèles : les modèles basés sur les candidats et les modèles basés sur les documents. L'approche basée sur les candidats, ou approche basée sur les profils, ou approche indépendante de la requête, se concentre sur la création de profils d'experts à l'aide de représentations textuelles, puis sur le classement de la liste des experts pour chaque requête à l'aide de modèles de recherche ad hoc traditionnels. L'approche basée sur les documents, ou approche dépendante de la requête, se concentre quant à elle sur la recherche des documents pertinents pour la requête, puis sur l'extraction de l'expert associé à ces documents. Une combinaison de ces deux modèles a également été explorée.

Deuxièmement, les approches basées sur les réseaux (Shen *et al.*, 2007) s'appuient sur une représentation graphique de la relation entre les experts et les documents. Ces modèles utilisent des méthodes d'analyse graphique, telles que la propagation d'autorité ou des algorithmes inspirés du PageRank, pour exploiter la structure relationnelle sous-jacente aux collections de documents.

Enfin, les approches hybrides (Duan *et al.*, 2007) combinent des informations basées sur le contenu et sur la structure du réseau afin d'exploiter simultanément les indices textuels et les relations entre entités. Ce type d'approche vise à tirer parti de la complémentarité entre le contenu et la structure et s'est avéré particulièrement utile dans les contextes où la diversité des sources et des liens offre un potentiel accru pour la détection fine de l'expertise.

Dans cet article, notre contribution s'inscrit dans le même objectif des travaux réalisés dans le cadre du TREC Enterprise Track. Cependant, notre approche ERAG explore des techniques récentes basées sur les LLM et le RAG pour accomplir simultanément des tâches de recherche documentaire et de recherche d'experts. À notre connaissance, aucune publication ne présente de résultats récents sur la tâche TREC Enterprise, ni sur une tâche équivalente en termes de nature des données et de complexité. De plus, le paradigme d'évaluation de cette tâche exige que la proposition d'experts en réponse à une requête soit justifiée par des documents. Il n'est donc pas envisageable d'utiliser uniquement des LLM sans mécanisme de ciblage documentaire et de citation de sources, sous peine de compromettre la validité et la vérifiabilité des résultats. Le RAG, qui renforce le contenu généré par le LLM en fournissant un contexte à la requête est la réponse naturelle à cette exigence. Il permet de réduire les

3. https://www.csiro.au

hallucinations tout en permettant de restreindre la recherche d'experts à la temporalité de la collection.

Bien que les études récentes en recherche d'experts s'appuient sur d'autres corpus (Azimi *et al.*, 2025), ce jeu de données reste pertinent et intéressant à analyser, notamment en comparaison avec des approches modernes telles que l'utilisation de LLMs.

3 ERAG : Recherche d'experts basée sur le RAG

Dans cette section, nous présentons notre approche ERAG, qui repose sur le paradigme RAG (Retrieval-Augmented Generation). Le RAG est largement étudié dans la littérature et se décline en plusieurs variantes (Singh *et al.*, 2025) : le RAG naïf, le RAG avancé, le RAG modulaire, le graph RAG et le RAG agentique. Dans le cadre de cette étude, nous avons mis en œuvre différentes variantes basées sur les approches RAG naïf et RAG avancé(Gao *et al.*, 2023), qui constituent le fondement de notre proposition ERAG (*Expert Retrieval with Augmented Generation*).

RAG Naïf : Il s'agit de l'architecture fondamentale de l'approche RAG, basée sur une simple recherche par mot-clé des données nécessaires pour alimenter le LLM. Son principal avantage est sa simplicité de mise en œuvre, au prix de quelques limitations inhérentes aux approches sacs de mots classiques (polysémie, homonymie, ordre des mots...).

RAG Avancé : Il pallie les limites du RAG naïf en ajoutant une meilleure compréhension sémantique grâce à l'exploitation de représentations sémantiques dans la phase de récupération, au reclassement des modèles ou à la récupération itérative.

La figure 1 illustre le processus de l'approche ERAG, incluant l'ensemble des méthodes et des modèles implémentées, ainsi qu'un exemple concret du déroulement du processus.

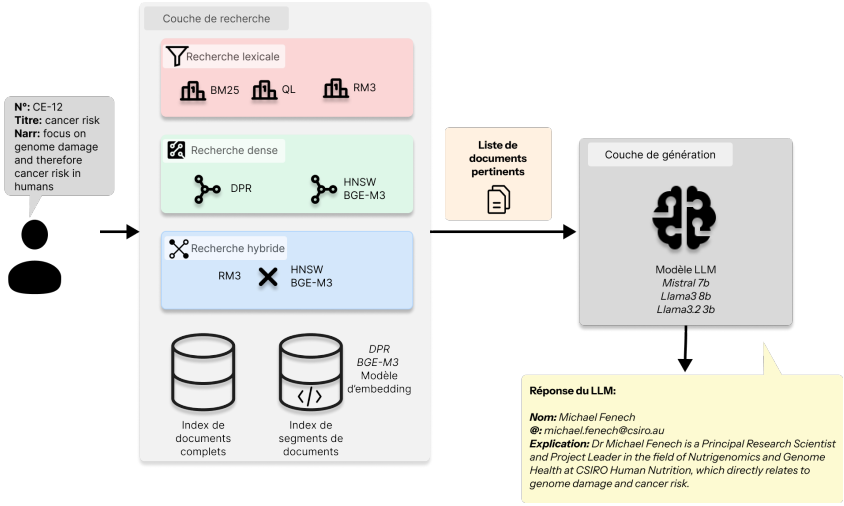

FIGURE 1 – ERAG : Une approche RAG pour la recherche d'experts

3.1 Ingestion des données

Nous avons créé deux index : un index de documents complets pour la recherche clairsemée (*sparse retrieval*) et un index de documents segmentés (*chunks*) pour la recherche dense (*dense retrieval*). Après avoir testé différentes valeurs, nous avons fixé le nombre maximal de jetons (*tokens*) à 250 par segment, tout en veillant à ne pas couper les phrases. Cette taille de segment a été choisie pour s'adapter aux modèles denses et optimiser la précision de la recherche.

3.2 Méthodes de recherche

Afin de récupérer les documents pertinents pour les requêtes, nous avons mis en œuvre trois types de récupération. Pour la récupération clairsemée, nous avons mis en œuvre BM25 (Best Match 25) (Robertson *et al.*, 1994), une fonction de classement basée sur la fréquence des termes, QL (Query Likelihood) (ChengXiang, 2008), un modèle probabiliste basé sur des modèles de langage classiques, et RM3 (Relevance Model 3) (Lv & Zhai, 2010), une méthode de pseudo-rétroaction de pertinence qui élargit la requête initiale en utilisant BM25 comme modèle de récupération initial. Pour chaque méthode, nous avons testé différents paramètres afin d'optimiser les performances. Pour la recherche dense, nous avons effectué une recherche sémantique à l'aide des modèles DPR[4](Dense Passage Retriever) (Karpukhin *et al.*, 2020) et du modèle BGE-M3[5] (BAAI[6] General Embeddings-Multi-Linguality, Multi-Functionality, Multi-Granularity) (Chen *et al.*, 2024). De plus, nous avons évalué une approche de recherche hybride en combinant notre meilleur résultat RM3 avec BGE-M3.

3.3 Modèles de génération

Pour identifier les experts, nous avons tiré parti de la puissance des LLM afin d'extraire les entités et leurs expertises à partir des documents pertinents que nous avons récupérés et ajoutés à la requête initiale. Nous avons demandé au modèle de générer une liste d'experts avec leurs noms, leurs rôles ou leurs fonctions, leurs adresses e-mail et la raison pour laquelle ils ont été choisis comme experts. Nous avons expérimenté différents volumes de documents afin d'identifier les experts les plus pertinents. L'extraction des experts est réalisée sur n documents, traités séquentiellement : chaque document est transmis individuellement au LLM. Nous avons utilisé Mistral 7b[7], Llama3 8b[8] et Llama3.2 3b[9]. Les trois modèles ont une quantification de 4 bits et une température fixée à 0 afin de garantir des générations déterministes et reproductibles entre les modèles.

4 Évaluation expérimentale

Dans cette section, nous commençons par présenter l'ensemble de données utilisé, puis nous décrivons le protocole d'évaluation de notre approche ERAG.

4. https://huggingface.co/facebook/dpr-ctx_encoder-single-nq-base,https://huggingface.co/facebook/dpr-question_encoder-single-nq-base
5. https://huggingface.co/BAAI/bge-m3
6. BAAI : Beijing Academy of Artificial Intelligence
7. https://ollama.com/library/mistral
8. https://ollama.com/library/llama3
9. https://ollama.com/library/llama3.2

4.1 Description de l'ensemble de données

Nous avons utilisé la collection de tests CERC de 2007 TREC Enterprise Track. L'ensemble de données CERC contient 370 715 documents, dont 89 % sont des pages HTML, 4 % sont des documents PDF, Word ou Excel, et le reste est un mélange de fichiers multimédia, de scripts et de fichiers journaux (Jiang *et al.*, 2007).

Nous utilisons la liste des sujets ou requêtes et des experts de l'édition 2007, car elle a été créée par des experts, des scientifiques du CSIRO, qui avaient pour tâche de trouver des employés experts pour répondre aux différentes demandes du public et aider les cellules de communication à rédiger des pages pertinentes sur les activités scientifiques (Bailey *et al.*, 2007a). Dans les jugements de pertinence *(qrels)*, les experts sont représentés par leurs adresses électroniques : c'est sous cette forme que les réponses aux requêtes doivent ainsi être fournies.

Les requêtes *(topics)* du jeu de données sont constituées de : numéro de la requête (N°), le titre de la requête (Titre) et une description plus détaillée de la requête (Narr.). Un exemple est illustré sur la figure 1.

4.2 Modèles de référence

Extraction des adresses e-mail et BM25. Un système simple mis en œuvre consiste à récupérer les documents pertinents à l'aide d'un score BM25, puis à renvoyer la liste des adresses e-mail extraits des documents trouvés à l'aide d'une expression régulière qui respecte le format « first-name.lastname@csiro.au » fourni par les organisateurs aux participants.

Modèle LLM seul. L'une des nouvelles approches que nous avons mises en œuvre consiste à interroger un LLM pour voir s'il peut répondre aux requêtes et fournir les experts en fonction de ses propres données d'entraînement. Cela permet non seulement d'obtenir une première base de référence, mais aussi de voir si le modèle a été entraîné sur l'ensemble de données CERC. Nous avons testé le modèle open source Mistral 7b et Llama3 8b.

4.3 Métriques d'évaluation

Nous évaluons les performances à l'aide des métriques TREC standards, notamment la précision moyenne (MAP), le rang réciproque moyen (MRR), la précision à k (P@k) et le gain cumulé normalisé actualisé (nDCG). Ces mesures nous permettent d'évaluer à la fois la qualité du classement et la couverture des documents et des experts récupérés. Nous les comparons aux meilleurs systèmes TREC 2007 (TREC 2007) de l'évaluation officielle (Bailey *et al.*, 2007b).

5 Résultats expérimentaux

Dans cette section, nous présentons les résultats obtenus par notre approche ERAG pour les deux tâches évaluées dans TREC Enterprise, correspondant aux étapes du RAG : la recherche de documents (équivalent RAG : récupération) et la recherche d'experts (presque équivalent RAG : experts présents dans la réponse générée).

5.1 Recherche de documents

La première tâche évaluée durant TREC Enterprise track est la recherche de documents. L'objectif était de répertorier les pages et les documents clés qui aideraient à créer une page de présentation autour d'un certain sujet.

Afin de comparer nos résultats avec ceux du TREC 2007 (Bailey *et al.*, 2007b), nos modèles ont récupéré 1000 documents pertinents par sujet (requête).

Nous avons testé plusieurs approches de recherche, notamment des méthodes clairsemées, denses et hybrides. Le tableau 1 présente les résultats les plus performants pour chaque approche, classés par MAP, conformément à la présentation officielle de la piste TREC Enterprise. **Domination du**

Recherche	Index	Méthode	Requête	MAP	P@5	P@20	nDCG
Sparse	Complet	BM25	Titre	**0.5007**	**0.8040**	**0.7160**	**0.7523**
	Complet	QL	Titre	0.4957	0.80	0.6880	0.7502
	Complet	RM3	Titre	0.4645	0.7880	0.6660	0.7331
Dense	Segments	BGE-M3	Complète	0.2738	0.7160	0.5710	0.5579
	Segments	DPR	Complète	0.0445	0.1560	0.1450	0.1856
Hybrid	Segments	RM3 & BGE-M3	Titre & Complète	0.4017	0.7640	0.6620	0.6925
TREC 2007	/	/	/	0.422	/	0.743	0.527

TABLE 1 – Meilleurs scores pour l'étape de récupération - recherche documentaire

***Sparse Retrieval* :** les méthodes lexicales classiques surpassent nettement les approches neuronales modernes. Notre configuration la plus performante, BM25 ($k1$=1.8 et b=0.7) appliqué à l'index des documents complets, a surpassé toutes autres méthodes sur toutes les métriques (MAP= 0.5007, nDCG= 0.7523), dépassant même la meilleure exécution automatique de TREC 2007 (MAP= 0.422). Pour cette exécution, nous avons utilisé uniquement le titre de la requête pour récupérer tous les documents pertinents. Cela indique que la correspondance exacte par mots-clés reste le signal le plus fiable pour ce corpus.

Limites des modèles denses : les modèles denses affichent une efficacité moindre, particulièrement DPR (MAP= 0.0445). Bien que BGE-M3 soit plus robuste, l'écart de performance avec BM25 souligne le défi du "décalage de domaine" (*domain shift*) : sans ajustement spécifique, les plongements sémantiques peinent à égaler la seule précision lexicale.

Approche Hybride : la combinaison de RM3 et BGE-M3 en utilisant la stratégie RRF (MAP= 0.4017) ne parvient pas à surpasser la solution BM25. Le "bruit" introduit par le moteur dense semble l'emporter sur les gains sémantiques, suggérant la nécessité d'une pondération plus fine ou encore d'une stratégie de fusion plus sophistiquée.

5.2 Recherche d'experts

La deuxième tâche du volet TREC Enterprise est la recherche d'experts. L'objectif était de trouver les contacts clés à répertorier dans les pages de présentation d'un sujet donné. La liste d'experts doit être fournie sous forme d'adresses électroniques respectant le format « firstname.lastname@csiro.au ». Dans cette section, nous présentons les performances de notre approche basé sur le RAG par rapport aux résultats de TREC 2007 (Bailey *et al.*, 2007b) et de nos modèles de référence *(baselines)*.

Pour la tâche de recherche d'experts, nous avons pris nos meilleurs résultats de recherche et avons essayé de varier le nombre de documents fournis aux LLM Mistral 7b, Llama3 8b et Llama3.2 :3b.

L'extraction des experts est réalisée sur n documents, traités séquentiellement : chaque document (représentant une itération) est transmis individuellement au LLM. La métrique de précision est ici définie comme le ratio entre les experts pertinents identifiés et le nombre total d'experts retournés par le modèle.

Comparaison des Meilleures Performances : Le tableau 2 présente les meilleures exécutions obtenues selon la MAP, pour chaque méthode de recherche, quels que soient le modèle et le nombre de documents injectés dans le prompt. Pour cette expérimentation, nous avons utilisé les deux modèles de taille comparable, Mistral 7b et Llama3 8b. Dans le cas de la variante "Index=Segments*", nous avons effectué la recherche sur les segments (chunks), puis nous avons fourni les documents complets au LLM pour l'étape de génération.

Recherche	Index	Méthode	Modèle	Itérations	MAP	P@5	P@20	Précision
Baselines	Complet	BM25	/	/	0.1457	0.1240	0.0510	**107**/45275
	/	LLM	Mistral 7b	/	0.0521	0.0250	0.0063	3/69
	/	LLM	Llama3 8b	/	0.0102	0.0041	0.0010	1/68
Sparse	Complet	BM25	Llama3 8b	30	**0.4728**	**0.2760**	**0.0970**	101/1264
	Complet	QL	Llama3 8b	30	0.3997	0.2440	0.0870	92/1560
	Complet	RM3	Llama3 8b	30	0.4108	0.2640	0.0890	94/1539
Dense	Segments*	BGE-RM3	Mistral 7b	50	0.3530	0.2000	0.0780	91/2737
TREC 2007	/	/	/	/	0.4632	0.2280	0.0910	106/4688

TABLE 2 – Résultats globaux des meilleures exécutions par méthode de recherche

L'approche RAG "Sparse + LLM" (en particulier BM25 et Llama3 8b) s'impose comme la meilleure configuration, dépassant même les méthodes originelles de TREC 2007.

Les solutions utilisant les LLM seuls (Mistral 7b et Llama3 8b), sans contexte injecté dans le prompt, affichent des résultats quasi nuls (MAP < 0.06).

L'approche exhaustive consistant à faire une extraction des adresses e-mail d'experts après la recherche BM25, donne des améliorations mais elle reste très faible comparée à TREC 2007.

Le système RAG basé sur une recherche Sparse (BM25) surpasse la recherche Dense (BGE-M3) avec la combinaison BM25 + Llama3 atteignant une MAP de 0.4728, contre 0.3530 pour l'approche dense. Dans un contexte de recherche d'experts, les mots-clés spécifiques (noms, termes techniques rares) capturés par le Sparse semblent fournir un contexte plus qualitatif au LLM que les vecteurs sémantiques de l'approche dense.

Nous avons remarqué lors de nos expérimentations que bien que Mistral 7b et Llama3 8b soient de tailles comparables, Llama3 8b donne des performances globalement meilleures.

Au final, notre système RAG obtient une performance équivalente à la référence TREC 2007 à partir de 3.5 fois moins de documents (101/1264, comparé à 106/4688). Cette fiabilité bien supérieure démontre que le LLM agit comme un filtre capable d'identifier avec une grande précision l'expertise contenue dans le contexte injecté.

Impact de la taille des LLM : Le tableau 3 permet d'analyser à quel point la taille et l'architecture du LLM influencent la précision du RAG, à paramètres de recherche constants (50 documents analysés séquentiellement, méthode de recherche BM25 ($k1$=2, b=0.9) et index des documents complets).

Llama3 8b s'impose comme le modèle le plus performant (MAP 0.4555), frôlant les performances du meilleur système TREC 2007, tout en conservant une précision supérieure (108/2083).

Recherche	Méthode	MAP	P@5	P@20	Précision
Mistral 7b	BM25	0.4069	0.2440	0.0910	**110**/2744
Llama3 8b	BM25	0.4555	**0.2680**	**0.1000**	108/2083
Llama3,2 3b	BM25	0.3213	0.2120	0.0910	103/2194
TREC 2007	/	**0.4632**	0.2280	0.0910	106/4688

TABLE 3 – Scores de la recherche d'experts pour 50 documents

Llama3.2 3b affiche les scores les plus faibles (MAP 0.3213). Le modèle de 3b de paramètres, bien que plus rapide, entraîne une chute de MAP de près de 30 % par rapport au modèle 8b, ce qui se traduit par un bruit plus important (103/2194).

Nous remarquons que Llama3 8b est plus sélectif. Il commet moins d'erreurs d'attribution (faux positifs) que Mistral 7b (108/2083 comparé à 110/2744). Cette capacité de filtrage est cruciale dans un pipeline RAG pour éviter la pollution des résultats finaux.

Analyse de la fenêtre de contexte : le Tableau 4 présente l'évolution des performances du modèle Llama3 8b en fonction du nombre d'itérations pour la tâche de recherche d'experts.

Index	Modèle	Itérations	MAP	P@5	P@20	Précision
Complet	Llama3 8b	1	0.3156	0.1348	0.0337	31/80
Complet	Llama3 8b	5	0.4156	0.2160	0.0630	63/259
Complet	Llama3 8b	10	0.4437	**0.2760**	0.0850	85/521
Complet	Llama3 8b	30	**0.4728**	**0.2760**	0.0970	101/1264
Complet	Llama3 8b	50	0.4555	0.2680	**0.1000**	108/2083
Complet	Llama3 8b	60	0.4327	0.2560	0.0980	113/2468
Complet	Llama3 8b	80	0.3821	0.2680	**0.1000**	108/2083
Complet	Llama3 8b	100	0.3768	0.2120	0.0970	**121**/4054
TREC 2007	/	/	0.4632	0.2280	0.0910	106/4688

TABLE 4 – Scores de la recherche d'experts avec BM25, Llama3 et pour différents nombres de documents

L'augmentation du nombre d'itérations de 1 à 30 montre une progression constante de la MAP, passant de 0.3156 à 0.4728. À ce stade (30 itérations), notre méthode surpasse les performances de l'état de l'art (TREC 2007 : 0.4632). Cela démontre que des signaux d'expertise sont présents au-delà des premiers documents renvoyés par le moteur de recherche, et que le LLM parvient à les isoler efficacement malgré une baisse de la pertinence intrinsèque des documents de rang inférieur. Elle montre aussi que l'analyse séquentielle par document permet de reconstruire une liste d'experts plus précise que les méthodes de classement globales de l'état de l'art associé à TREC Enterprise.

La comparaison de la précision (fiabilité) entre notre méthode et les méthodes traditionnelles met en lumière une différence assez importante : la meilleure exécution de TREC 2007 affiche une précision de 2.26 % (106 experts retenus sur 4688 suggestions). Cela indique un système qui génère énormément de bruit pour capturer le signal pertinent. Contrairement à notre méthode, où au rang 30 l'on atteint une précision de 7.99 % (101/1264). Le modèle est ainsi 3,5 fois plus fiable : pour chaque expert suggéré, la probabilité qu'il soit réellement pertinent est largement supérieure.

L'analyse des itérations révèle un phénomène de dilution de la précision à mesure que l'on s'éloigne

des documents de tête. À une itération et avec une précision de 38,7 %, le modèle est extrêmement sélectif. Le document classé en premier par BM25 offre un contexte pur qui minimise les erreurs d'interprétation du LLM. À partir du rang 50, bien que le nombre d'experts pertinents trouvés continue de croître (de 101 à 121), la précision s'effondre pour atteindre 2,98 % au rang 100, mais qui reste meilleure que TREC 2007.

À mesure que nous augmentons le rang, le LLM traite des documents de moins en moins pertinents. Sa propension à identifier des experts dans des contextes de faible qualité augmente le nombre de faux positifs. Le point de rupture se situe à 30 itérations : c'est le moment où le gain en rappel (nouveaux experts trouvés) ne compense plus la perte de précision (experts erronés introduits).

Par ailleurs, nous avons remarqué que les meilleures méthodes de recherche documentaires ne sont pas forcément celles qui donnent les meilleurs résultats de la recherche d'experts. En effet, un document considéré comme pertinent dans l'étape d'évaluation de la recherche documentaire ne contient pas forcément d'experts pertinents ce qui impact le classement d'experts.

6 Conclusion

Cet article a visé à examiner l'efficacité et l'impact des approches récentes de recherche d'information (RI) basées sur des LLM et le RAG. Nous avons évalué leur contribution à travers le processus de notre approche ERAG, présenté dans la figure 1, lorsqu'elles sont appliquées à un jeu de données à la fois ancien, structurellement complexe et bruité, tel que celui de TREC Enterprise, pour la recherche d'experts à partir de requêtes exprimées en langage naturel.

Nos résultats démontrent que l'utilisation d'un LLM comme extracteur séquentiel (par itération sur une liste de documents trouvés) permet non seulement de surpasser les benchmarks TREC officiels en termes de MAP, mais assure également une fiabilité (précision) jusqu'à trois fois supérieure. Ils démontrent que les techniques RAG basées sur les LLM peuvent contribuer de manière significative à la recherche d'experts, même sur un ensemble de données vieux de près de 20 ans. Alors qu'un LLM génératif seul donne des résultats extrêmement médiocres, le fait d'ancrer le modèle dans des documents récupérés, en particulier avec BM25, améliore considérablement la MAP et la précision initiale, ce qui montre que la principale contribution du LLM réside dans le raisonnement sur des preuves pertinentes plutôt que dans la mémorisation d'experts. La recherche lexicale reste essentielle dans les corpus d'entreprise hétérogènes, où les approches basées uniquement sur l'intégration ont du mal à saisir les distinctions thématiques fines.

La plupart des solutions de 2007 utilisaient un modèle en deux étapes ou un modèle de vote, dans lequel le score d'un expert était une simple agrégation ou somme des scores de pertinence des documents dans lesquels il apparaissait. Le RAG remplace cette approche par un raisonnement sémantique, permettant au LLM d'évaluer la manière dont une personne est mentionnée (par exemple, en tant que responsable de projet ou simple mention secondaire), ce qui constituait une difficulté majeure pour les systèmes traditionnels.

En perspective, l'évolution du système vers des approches non séquentielles ainsi que l'implémentation d'un critère d'arrêt dynamique basé sur le score de confiance du modèle, permettrait d'optimiser les coûts computationnels tout en affinant la précision. Par ailleurs, une évaluation qualitative des explications générées par le LLM pour justifier chaque expertise, constituerait une étape importante pour garantir la transparence et la pertinence du système final.

Références

AZIMI M., MOFFAT A. & ZOBEL J. (2025). Expert finding revisited : A uniform exploration of methods. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, p. 478–487.

BAILEY P., CRASWELL N., SOBOROFF I. & DE VRIES A. P. (2007a). The CSIRO enterprise search test collection. *SIGIR Forum*, **41**(2), 42–45. DOI : 10.1145/1328964.1328969.

BAILEY P., DE VRIES A. P., CRASWELL N. & SOBOROFF I. (2007b). Overview of the TREC 2007 enterprise track. In E. M. VOORHEES & L. P. BUCKLAND, Éds., *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007*, volume 500-274 de *NIST Special Publication* : National Institute of Standards and Technology (NIST).

BALOG K. (2007). People search in the enterprise. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 916–916.

BALOG K., FANG Y., DE RIJKE M., SERDYUKOV P. & SI L. (2012). Expertise retrieval. *Foundations and Trends® in Information Retrieval*, **6**(2–3), 127–256. DOI : 10.1561/1500000024.

BALOG K., SOBOROFF I., THOMAS P., BAILEY P., CRASWELL N. & DE VRIES A. P. (2008). Overview of the TREC 2008 enterprise track. In E. M. VOORHEES & L. P. BUCKLAND, Éds., *Proceedings of The Seventeenth Text REtrieval Conference, TREC 2008, Gaithersburg, Maryland, USA, November 18-21, 2008*, volume 500-277 de *NIST Special Publication* : National Institute of Standards and Technology (NIST).

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

CHEN J., XIAO S., ZHANG P., LUO K., LIAN D. & LIU Z. (2024). Bge m3-embedding : Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

CHENGXIANG Z. (2008). Statistical language models for information retrieval a critical review. *Foundations and Trends$\hat{W}$ in Accounting*, **2**(1–2), 137–213.

CRASWELL N., DE VRIES A. P. & SOBOROFF I. (2005). Overview of the TREC 2005 enterprise track. In E. M. VOORHEES & L. P. BUCKLAND, Éds., *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005*, volume 500-266 de *NIST Special Publication* : National Institute of Standards and Technology (NIST).

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DUAN H., ZHOU Q., LU Z., JIN O., BAO S., CAO Y. & YU Y. (2007). Research on enterprise track of trec 2007 at sjtu apex lab. In *TREC*.

FANG H. & ZHAI C. (2007). Probabilistic models for expert finding. In *European conference on information retrieval*, p. 418–430 : Springer.

GAO Y., XIONG Y., GAO X., JIA K., PAN J., BI Y., DAI Y., SUN J., WANG H., WANG H. *et al.* (2023). Retrieval-augmented generation for large language models : A survey. *arXiv preprint arXiv :2312.10997*, **2**(1), 32.

JIANG J., LU W. & LIU D. (2007). Csir at trec 2007 expert search task. In *TREC*.

KARPUKHIN V., OGUZ B., MIN S., LEWIS P., WU L., EDUNOV S., CHEN D. & YIH W.-T. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6769–6781, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.550.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W.-T., ROCKTÄSCHEL T. *et al.* (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, **33**, 9459–9474.

LV Y. & ZHAI C. (2010). Positional relevance model for pseudo-relevance feedback. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, p. 579–586.

PETKOVA D. & CROFT W. B. (2008). Hierarchical language models for expert finding in enterprise corpora. *International Journal on Artificial Intelligence Tools*, **17**(01), 5–18.

ROBERTSON S. E., WALKER S., JONES S., HANCOCK-BEAULIEU M. & GATFORD M. (1994). Okapi at TREC-3. In D. K. HARMAN, Éd., *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 de *NIST Special Publication*, p. 109–126 : National Institute of Standards and Technology (NIST).

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SHEN H., CHEN G., CHEN H., LIU Y. & CHENG X. (2007). Research on enterprise track of TREC 2007. In E. M. VOORHEES & L. P. BUCKLAND, Éds., *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007*, volume 500-274 de *NIST Special Publication* : National Institute of Standards and Technology (NIST).

SINGH A., EHTESHAM A., KUMAR S. & KHOEI T. T. (2025). Agentic retrieval-augmented generation : A survey on agentic rag. *arXiv preprint arXiv :2501.09136*.

SOBOROFF I., DE VRIES A. P. & CRASWELL N. (2006). Overview of the TREC 2006 enterprise track. In E. M. VOORHEES & L. P. BUCKLAND, Éds., *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006, Gaithersburg, Maryland, USA, November 14-17, 2006*, volume 500-272 de *NIST Special Publication* : National Institute of Standards and Technology (NIST).