L'Intérêt des RAG dans la Gestion des Connaissances des Processus Administratifs Universitaires à l'ère des LLM

Luiz-Angelo Steffenel*, Laurent Lucas*

*Université de Reims Champagne-Ardenne
LICIIS - LRC CEA DIGIT
{prenom.nom}@univ-reims.fr

Résumé. Les modèles de langage à grande échelle (LLM) transforment profondément la gestion des connaissances dans les organisations. Cet article présente un projet co-porté par l'Université de Reims Champagne-Ardenne pour l'hébergement souverain de solutions LLM, et explore l'utilisation des systèmes de génération augmentée par récupération (RAG) pour répondre aux défis d'adaptation, d'accès sécurisé et de dissémination efficace de l'information au sein des administrations universitaires. Comme ce projet soulève plusieurs enjeux techniques, organisationnels et éthiques, cet article vise à lancer une session de discussion et d'échange afin de guider l'implémentation réussie du projet.

1 Introduction

Les modèles de langage à grande échelle (LLM) Brown et al. (2020) tels que GPT-4 révolutionnent la gestion des connaissances, en automatisant des tâches complexes et en améliorant l'accès à l'information. Dans un contexte universitaire, ces modèles peuvent simplifier la communication, centraliser les ressources, et répondre efficacement aux questions fréquentes.

En effet, des LLM comme GPT-4, Gemini, Mixtral ou Claude peuvent générer des rapports, synthèses, ou courriels en utilisant un langage professionnel et contextuellement adapté. Ils réduisent ainsi le temps nécessaire à la rédaction, tout en garantissant une qualité élevée. Des outils comme Grammarly ou des extensions intégrées aux plateformes d'entreprise illustrent l'intégration de ces modèles pour la rédaction collaborative Wu et al. (2023).

Un autre avantage des LLM souvent souligné est qu'ils permettent d'automatiser des tâches répétitives telles que le tri et la réponse aux courriels, l'organisation d'agendas complexes, et la mise à jour de bases de données. Des solutions commerciales, comme les intégrations de GPT dans Microsoft 365 Copilot, montrent l'adoption rapide de ces capacités dans les environnements professionnels.

Enfin, les LLM offrent une transcription précise de réunions et une traduction contextuelle de documents, surpassant souvent les outils de traduction basiques grâce à une meilleure intégration des nuances culturelles et linguistiques grâce à une plus importante évaluation du contexte Bender et al. (2021). Bien sûr, comme tout outil automatique, la performance dépend d'un corpus suffisamment développé, dont l'amplification et la diversification est l'un des fers de lance des principaux concurrents dans le marché des LLM.

Ces capacités positionnent les LLM non seulement comme des outils technologiques, mais aussi comme des facilitateurs de la gestion des connaissances, particulièrement dans les organisations où le volume et la diversité des données sont élevés.

2 Risques et Défis des LLM Hébergés

Malgré leurs avantages, les LLM posent des risques lorsqu'ils sont hébergés par des tiers :

— **Fuite de données sensibles** : Les données soumises à des plateformes tierces peuvent être exploitées à des fins d'entraînement et être enfin dévoilés par mégarde ou par des requêtes spécifiques visant l'obtention d'informations sur une entreprise ou structure ;

— **Conformité réglementaire** : Les données hébergées hors UE peuvent violer les normes du RGPD Voigt et Bussche (2017).

— **Attaques adversariales** : Les LLM sont vulnérables aux attaques visant à manipuler leurs sorties ou à extraire des informations sensibles.

En ce qui concerne l'administration universitaire, de nombreuses solutions commerciales ou "gratuites" sont utilisées de manière plus ou moins informelle, souvent sans une réelle réflexion sur les enjeux éthiques, la sécurité des données, les biais algorithmiques, et les limites liées à la dépendance aux données d'entraînement. Bien que des recommandations et directives invitent à la prudence et à la retenue, l'adoption de ces outils ne peut pas être ignoré ni arrêté.

Ces risques plaident en faveur d'une solution locale pour garantir la souveraineté technologique et la confidentialité des données.

3 Le projet Fédération ILaaS (Inférence LLM as a Service)

Afin de contrer ces inconvénients de l'utilisation de LLM hébergés sur des sites tiers, plusieurs universités dont l'Université de Reims Champagne-Ardenne, Rennes, Lille, Paris 1 Panthéon-Sorbonne et CentraleSupélec Paris-Saclay ainsi que la DGRI se sont lancées dans un projet fédérateur pour la mutualisation de serveurs d'inférence hébergés dans la fédération de datacenters labellisés [1] Cette mutualisation vise offrir des services LLM à la communauté universitaire tout en renforçant trois piliers :

— **La Confiance :** le niveau de confiance est mitigé envers des fournisseurs de services LLM tels que OpenAI, qui proposent des offres « éducation ». Or, en raison des activités de calcul scientifique menés par plusieurs des institutions partenaires de ce projet, nous avons tout ou partie de l'infrastructure nécessaire pour garantir un meilleur niveau de confidentialité à nos usagers ;

— **La Souveraineté :** le choix des LLM et de leurs variantes (quantification, fine-tuning, etc.) est important dans certains contextes (biomédical, informatique, etc.). La capacité d'offrir différentes variantes et de les adapter aux différents cas d'usage nécessite une indépendance technologique et une plus grande proximité avec le public cible. Dans le cas de ce projet, cette versatilité sera proposée en s'appuyant sur les bases du projet Aristote [2] ;

1. https://www.aefinfo.fr/depeche/721518-creation-dilaas-une-nouvelle-federation-detablissements-du-superieur-pour-aller-vers-plus-de-sobriete-numerique
2. https://github.com/CentraleSupelec/aristote-dispatcher

— **La Soutenabilité :** le coût au token reste élevé en raison des investissements onéreux et de la (sur)réservation de ressources GPU nécessaires pour garantir la réactivité (token/s) attendue par les usagers. Il apparaît pertinent d'optimiser l'usage des ressources matérielles en mixant judicieusement les besoins (pédagogie, administration, recherche) selon les périodes d'activité des usagers et la nature des tâches.

En plus de la mutualisation de ressources et compétences, des évolutions de projets numériques sont envisagés pour gérer le découplage entre l'infrastructure d'inférence et les services applicatifs, tels que RAGaRenn [3] (Université de Rennes) ou le projet CRISalid [4] (porté par les universités Paris 1 Panthéon-Sorbonne, Toulon, Paris-Saclay, Claude Bernard Lyon 1, Montpellier, l'université Polytechnique Hauts-de-France, Nantes Université et l'EHESS). À terme, l'objectif est de rendre accessibles ces services via une API unique et standard facilitant l'utilisation de frameworks courants dans le domaine des LLM, exposés par la fédération de serveurs d'inférence LLM avec priorisation inter-datacenter.

4 RAG : un Catalyseur pour la Gestion des Connaissances

Parmi les services LLM qui pourront être hébergés par les partenaires du projet Fédération ILaaS, les RAG (Retrieval-Augmented Generation) Lewis et al. (2020) font partie des services qui pourront apporter le plus de bénéfices à la communauté universitaire. En effet, les systèmes RAG combinent des bases de données locales avec des LLM pour centraliser les informations organisationnelles et améliorer leur accessibilité. Ils réduisent l'éparpillement des données et simplifient les recherches grâce à des interfaces intuitives basées sur le langage naturel.

Contrairement aux systèmes de recherche traditionnels, les RAG utilisent un LLM pour interpréter les requêtes en langage naturel. Cela permet d'extraire plus de réponses pertinentes, même lorsque les utilisateurs ne connaissent pas les mots-clés exacts ou le format du document recherché. Par exemple, un employé pourrait poser une question complexe comme "Comment soumettre une demande de congé exceptionnelle ?" ou "Puis-je pour effectuer des vacations d'enseignement dans une autre université ?", et recevoir directement une réponse synthétique accompagnée des documents nécessaires, voire les services ou personnes à contacter.

Les RAG permettraient, d'ailleurs, de résoudre un problème typique des approches utilisées actuellement (site intranet, par exemple). En effet, dans une intranet, l'indexation des documents dépend de l'effort de plusieurs acteurs (service concerné, webmaster/service communication, etc.) et dont la visibilité n'est pas forcément adéquate vis-à-vis de l'utilisateur. Il n'est pas rare de tomber sur des informations périmées, alors que les RAG peuvent être connectés à des bases de données dynamiques, garantissant que les utilisateurs consultent toujours la version la plus récente d'un document. Cela est essentiel dans des environnements où les politiques ou procédures évoluent fréquemment.

Enfin, un avantage clé des RAG réside dans leur capacité à gérer les niveaux d'accès aux informations en fonction des droits des utilisateurs. Cette fonctionnalité est particulièrement cruciale pour les organisations qui manipulent des données sensibles ou réglementées :

— **Contrôle basé sur les rôles :** Les RAG permettent d'intégrer des mécanismes d'authentification et d'autorisation. Par exemple, un membre du personnel administratif

3. https://ragarenn.eskemm-numerique.fr/
4. https://crisalid.org

peut accéder à des informations générales sur les politiques RH, tandis qu'un gestionnaire aura également accès aux détails confidentiels des dossiers des employés. Ces restrictions s'appliquent au niveau des documents, mais aussi aux réponses générées par le système, évitant ainsi toute fuite accidentelle d'informations.

— **Traçabilité des consultations :** Les systèmes RAG peuvent enregistrer les interactions des utilisateurs avec la base documentaire, offrant une visibilité sur qui a accédé à quelles informations. Cela est utile pour garantir la conformité avec des régulations telles que le RGPD.

5 Défis de l'Implémentation d'un RAG Local

La mise en place d'un système RAG au sein du représente une avancée majeure pour optimiser la gestion des connaissances. Cependant, plusieurs défis doivent être pris en compte pour assurer son succès.

Tout d'abord, la complexité technique. En effet, l'intégration d'un RAG dans l'écosystème existant nécessite des efforts importants pour connecter des sources de données hétérogènes (dossiers partagés, bases de données, intranet, etc.). De plus, l'entraînement ou l'adaptation d'un LLM pour interpréter efficacement les spécificités des documents universitaires peut s'avérer complexe et coûteux. L'utilisation d'un framework tel que celui du projet Aristote, ainsi que les ressources de calcul mutualisés (serveurs GPU, notamment) permet de diluer la charge technique et de lisser le coût de l'adaptation aux différentes pratiques des universités.

Les enjeux liés aux ressources matérielles et humaines sont aussi à prendre en charge. Bien que le projet Fédération ILaaS vise l'acquisition de ressources matérielles et un partage d'expérience entre les universités partenaires, la gestion et la mise à jour du système exigent une expertise technique pas forcément présente au sein des équipes de l'université. Une équipe technique "centrale" qui coordonne les équipes locales est prévue afin d'optimiser le déploiement de la fédération de serveurs LLM.

Aussi important que les ressources matérielles et humaines, il sera nécessaire de garantir la confidentialité des données et leur conformité aux réglementations, comme le RGPD. Les mécanismes de restriction d'accès doivent être robustes, et des audits réguliers doivent être prévus pour prévenir les risques de fuite ou d'accès non autorisé.

Enfin, l'adoption et formation des utilisateurs est essentielle. Même avec une solution techniquement aboutie, l'adoption par les utilisateurs reste un défi. La formation des employés et des enseignants sur les usages du système, ainsi que la mise en place d'un support technique efficace, seront essentielles pour maximiser les bénéfices du RAG. Des événements tels que l'AI Week [5] seront également promus afin de faire connaître les outils mis à disposition de la communauté. Enfin, il est important d'évaluer le taux d'adoption revient à estimer le retour sur investissement, notamment en termes de gain de temps et de réduction des erreurs, ce qui peut s'avérer difficile à court terme, bien que les bénéfices soient clairs sur le long terme.

5. https ://tinyurl.com/mrsdr8f7

6 Conclusion

La mise en œuvre d'un système RAG représente une opportunité stratégique pour les universités souhaitant moderniser leur gestion des connaissances. En centralisant les ressources, en simplifiant leur recherche et en garantissant leur sécurité, un RAG local peut devenir un levier clé de la transformation numérique. Le projet Fédération ILaaS vise à procurer des ressources permettant aux universités d'héberger des services LLM souverains, dont le RAG. Toutefois, pour réussir ce projet, il est crucial de surmonter les défis techniques, organisationnels et humains associés à son déploiement.

Références

Bender, E. M., T. Gebru, A. McMillan-Major, et S. Shmitchell (2021). On the dangers of stochastic parrots : Can language models be too big ? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, New York, NY, USA, pp. 610–623. Association for Computing Machinery, doi: 10.1145/3442188.3445922.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, et D. Amodei (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, et H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Volume 33, pp. 1877–1901. Curran Associates, Inc., doi: 10.48550/arXiv.2005.14165.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, et D. Kiela (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, et H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Volume 33, pp. 9459–9474. Curran Associates, Inc.

Voigt, P. et A. Bussche (2017). *The EU General Data Protection Regulation (GDPR) : A Practical Guide*. Springer, doi: 10.1007/978-3-319-57959-7.

Wu, H., W. Wang, Y. Wan, W. Jiao, et M. Lyu (2023). ChatGPT or Grammarly ? Evaluating ChatGPT on grammatical error correction benchmark, doi: 10.48550/arXiv.2303.13648.

Summary

Large-scale language models (LLM) are profoundly transforming knowledge management in organizations. This article presents a project co-led by the University of Reims Champagne-Ardenne for hosting sovereign LLM solutions, and explores the use of retrieval-augmented generation (RAG) systems to address the challenges of adaption, secure access and efficient dissemination of information within university administrations. As this project raises technical, organizational and ethical issues, this article aims to establish a discussion and exchange session to guide the successful implementation of the project.