

Med-KAG, une approche de génération augmentée par des connaissances médicales : résultats préliminaires

Edouard HADDAG*, Gabriel H.A. MEDEIROS*, Lina F. SOUALMIA*

* Univ Rouen Normandie, INSA Rouen Normandie,
LITIS UR 4108, FR-76000 Rouen, France

Résumé. Nous présentons Med-KAG, une nouvelle architecture d'assistant en Intelligence Artificielle (IA) conçue pour améliorer la fiabilité de l'aide à la décision clinique. Elle étend le paradigme de Génération Augmentée par la Récupération (RAG) en intégrant un graphe de connaissances dans le but de réduire les hallucinations factuelles. Une étude préliminaire sur MedQA-US compare un grand modèle de langue de référence (Qwen3-235B-A22B) à sa variante enrichie par ces connaissances. Les résultats montrent une précision comparable entre les deux configurations et identifient le module de récupération comme principale source d'erreurs.

1 Introduction

L'intégration de l'IA en pratique clinique offre un potentiel majeur, notamment pour l'analyse des dossiers médicaux électroniques [Alghamdi et Mostafa (2025)]. Toutefois, l'adoption des grands modèles de langue (LLM) reste freinée par leur opacité et le risque d'hallucinations factuelles [Huang et al. (2025)]. Pour garantir la fiabilité nécessaire aux environnements critiques, nous proposons une nouvelle architecture d'assistant diagnostique. Notre approche enrichit la Génération Augmentée par la Récupération (RAG) [Gao et al. (2024)] avec un graphe de connaissances (KG) issu du Métathésaurus de l'UMLS [Bodenreider (2004)]. En ancrant le modèle dans cette base structurée, nous visons deux objectifs : **minimiser les hallucinations** en contraignant les réponses par des données factuelles, et **fournir une justification** en traçant l'origine du raisonnement. Cet article présente l'architecture proposée, baptisée Med-KAG¹, et démontre via une évaluation sur le jeu de données MedQA-US comment la synergie entre RAG et KG accroît la robustesse de l'aide au diagnostic.

2 Méthodes

Med-KAG vise à fiabiliser le diagnostic en substituant la récupération de texte non structuré par l'exploitation d'une base de connaissances structurée. Inspirée par le cadre MedRAG Xiong et al. (2024), notre architecture adopte une approche de Knowledge Augmented Generation (KAG). Bien que l'implémentation actuelle suive un pipeline linéaire (RAG Natif),

1. Disponible sur : <https://github.com/c2fc2f/Med-KAG>

Med-KAG, une approche de génération augmentée par des connaissances médicales

sa conception modulaire anticipe une évolution vers un RAG Modulaire [Gao et al. (2024)] permettant des flux complexes (auto-correction, itération).

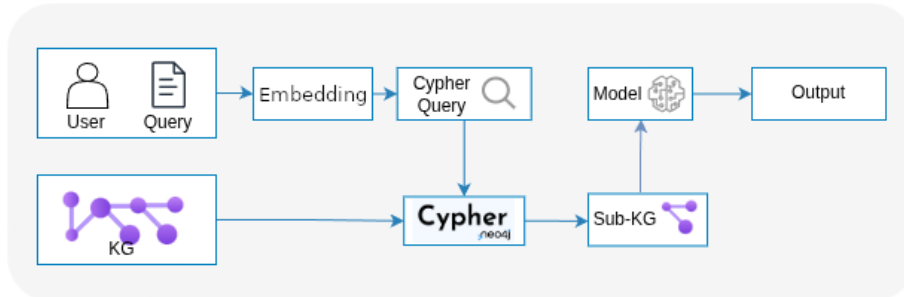


FIG. 1 – Schéma de l'architecture de Med-KAG.

Le cœur du système est un KG construit à partir du Métathésaurus de l'UMLS que nous avons développé [Medeiros et al. (2024)], garantissant des connexions vérifiables entre entités médicales. Contrairement aux approches vectorielles classiques, nous utilisons un récupérateur basé sur un LLM (Qwen3-235B-A22B [Yang et al. (2025)]). Ce module identifie les entités de la requête et extrait un sous-graphe pertinent. Ce contexte structuré est ensuite traité par le générateur (Qwen3), choisi pour sa capacité à verbaliser avec précision des données graphiques complexes, assurant ainsi cohérence computationnelle et exactitude factuelle.

3 Résultats

Nous évaluons la pertinence de notre architecture via une étude comparative entre un modèle de référence (baseline) et notre système RAG enrichi. L'expérimentation repose sur le jeu de données MedQA-US. Bien que le benchmark MIRAGE [Xiong et al. (2024)] définisse le standard d'évaluation complet, nous nous concentrons ici spécifiquement sur la composante Question-Réponse médicale de MedQA-US pour isoler l'impact de l'ancrage des connaissances. La figure ci-dessous illustre un exemple typique tiré du jeu de données :

Question : *A junior orthopaedic surgery resident is completing a carpal tunnel repair with the department chairman as the attending physician. During the case, the resident inadvertently cuts a flexor tendon. The tendon is repaired without complication. The attending tells the resident that the patient will do fine, and there is no need to report this minor complication that will not harm the patient, as he does not want to make the patient worry unnecessarily. He tells the resident to leave this complication out of the operative report. Which of the following is the correct next action for the resident to take ?*

Options : (A) *Disclose the error to the patient and put it in the operative report, (B) Tell the attending that he cannot fail to disclose this mistake, (C) Report the physician to the ethics committee, (D) Refuse to dictate the operative report.*

Med-KAG Answer : B

Nous avons mesuré la précision diagnostique en comparant deux configurations : **Native**, le modèle Qwen3-235B-A22B seul, et **RAG**, le même modèle augmenté de notre récupérateur de KG UMLS. La précision, définie comme le taux de bonnes réponses, est reportée dans le Tableau 1.

Modèle	Réponses correctes	Taux de précision
Native (Qwen3-235B-A22B)	1167/1273	91,67%
RAG (Qwen3-235B-A22B + KG)	1162/1273	91,28%

Métrique	Moyenne	Médiane	Non nulle
Erreurs (Requête invalide)	0,22	0,00	-
Nœuds récupérés	23,92	0,00	502/1273 (39,4%)
Arêtes récupérées	27,92	0,00	254/1273 (20,0%)

TAB. 1 – Précision comparative sur MedQA-US et mesures de récupération (moyenne, médiane et non nulle).

Le modèle Qwen3-235B-A22B natif a atteint une précision de 91,67%. Notre architecture Med-KAG a obtenu une précision de 91,28% sur ce même ensemble de questions. Pour comprendre le comportement du récupérateur de notre système RAG, nous avons analysé les caractéristiques des sous-graphes qu'il renvoyait pour chaque requête. Nous avons également mesuré le taux d'échec du générateur de requêtes Cypher [Neo4j (2025)] basé sur un LLM. Dans ce contexte, une « erreur » désigne une requête Cypher syntaxiquement invalide générée par le récupérateur, ce qui entraîne par conséquent le renvoi d'un graphe vide. Nos résultats montrent un taux d'erreur moyen de 22% (une moyenne de 0,22).

Les données du Tableau 1 révèlent une distribution très asymétrique. Alors que les récupérations réussies ont généré des sous-graphes avec un nombre élevé de nœuds (moyenne 23,92) et d'arêtes (moyenne 27,92), la valeur médiane de 0,00 pour ces deux métriques indique qu'au moins la moitié de toutes les requêtes ont abouti à un graphe vide (39,4% d'entre eux ont des nœuds et 20,0% d'entre eux ont des arêtes). C'est l'effet combiné du taux d'erreur syntaxique de 22% et d'autres requêtes qui n'ont pas réussi à trouver d'entités pertinentes. Pour le cas de la question montrée précédemment (3), notre architecture Med-KAG a correctement répondu, contrairement au modèle natif. La Figure 2 illustre la transparence du système : le module de récupération a identifié l'entité clé (« flexor tendon ») et extrait son voisinage sémantique dans l'UMLS, fournissant au générateur un contexte biomédical vérifié.

Requête Cypher générée :

```

MATCH (c1:CUI)
WHERE c1.name CONTAINS "flexor tendon"
OPTIONAL MATCH (c1)-[r1:PAR|CHD|SY|RO]->(c2:CUI)
OPTIONAL MATCH (c2)-[r2:PAR|CHD|SY|RO]->(c3:CUI)
RETURN c1, c2, c3, r1, r2
LIMIT 250

```

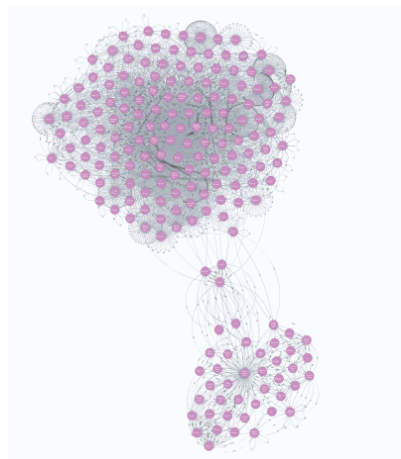


FIG. 2 – À gauche, la requête Cypher générée par le LLM; à droite, le sous-graphe UMLS récupéré servant de contexte ancré pour la réponse.

4 Discussion

Les résultats de notre évaluation mettent en évidence un goulot d'étranglement critique au sein de l'architecture actuelle. Avec une précision de 91,28 %, le système Med-KAG n'a pas surpassé, dans cette itération, le modèle de référence Qwen3-235B-A22B (91,67 %). Notre analyse indique que cet écart ne provient pas du générateur, mais des limites du composant de récupération. Comme le montre le Tableau 1, un taux d'erreur syntaxique de 22 % et une médiane de récupération nulle indiquent que le modèle peine à générer des requêtes Cypher valides. Le défi majeur identifié concerne la liaison d'entités (entity linking) : le LLM, probabiliste par nature, échoue à respecter la contrainte de correspondance exacte des identifiants UMLS, entraînant de fréquentes récupérations vides ou incomplètes.

Pour dépasser cette limitation, nous avons identifié deux axes d'amélioration complémentaires. Le premier repose sur la **récupération par embeddings** : l'utilisation exclusive d'une recherche vectorielle permettrait de contourner la contrainte d'exact-match et d'augmenter le taux de récupération. Toutefois, cette approche peut générer des sous-graphes volumineux et « bruités », diluant le contexte relationnel précis requis pour le raisonnement médical. Le second axe correspond à une **récupération hybride (recommandée)** : une stratégie en deux étapes apparaît plus adaptée, dans laquelle une première recherche vectorielle identifierait des nœuds « graines » pertinents, suivie d'une construction guidée par LLM de la logique de traversée du graphe. Cela permettrait au modèle de se concentrer sur le raisonnement relationnel plutôt que sur la génération syntaxique de requêtes.

En résumé, bien que l'implémentation actuelle demeure perfectible, ces premiers résultats confirment la faisabilité de l'approche et isolent clairement les verrous technologiques à lever. Les travaux futurs se concentreront sur le développement d'un récupérateur hybride plus robuste afin d'exploiter pleinement l'ancrage dans les connaissances biomédicales et d'améliorer la fiabilité globale de l'assistance clinique.

5 Conclusion

Med-KAG est une nouvelle architecture d'assistant IA conçue pour améliorer la fiabilité du diagnostic médical en ancrant un système RAG dans un KG basé sur l'UMLS. L'évaluation préliminaire sur MedQA-US a montré que notre implémentation actuelle (91,28% de précision) n'a pas surpassé le modèle de référence natif (91,67%). Cependant, nous avons réussi à identifier le principal goulot d'étranglement : le récupérateur basé sur le LLM, qui peine à générer des requêtes Cypher valides et pertinentes, entraînant un taux d'erreur de 22% et de fréquents échecs de récupération. Ce constat n'invalide pas le concept de RAG basé sur les connaissances ; il isole plutôt clairement le composant qui nécessite une amélioration. Nos travaux immédiats se concentrent sur le développement d'un récupérateur hybride plus robuste. En combinant la recherche sémantique basée sur les embeddings avec le parcours de graphe basée sur Cypher, nous visons à surmonter les limites actuelles et à exploiter pleinement les connaissances structurées pour construire un assistant IA plus précis et explicable pour les cliniciens. La validation future impliquera des évaluations plus étendues utilisant des données réelles et l'intégration avec des entrepôts de données cliniques.

Références

- Alghamdi, H. et A. Mostafa (2025). Advancing ehr analysis : Predictive medication modeling using llms. *Information Systems 131*, 102528, doi: <https://doi.org/10.1016/j.is.2025.102528>.
- Bodenreider, O. (2004). The Unified Medical Language System (UMLS) : integrating biomedical terminology. *Nucleic Acids Research 32*(Database issue), D267–D270, doi: 10.1093/nar/gkh061.
- Gao, Y., Y. Xiong, M. Wang, et H. Wang (2024). Modular rag : Transforming rag systems into lego-like reconfigurable frameworks.

- Huang, L., W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et T. Liu (2025). A survey on hallucination in large language models : Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* 43(2), 42, doi: 10.1145/3703155.
- Medeiros, G. H. A., L. F. Soualmia, et C. Zanni-Merk (2024). Harnessing the core propagation phenomenon ontology to develop a knowledge graph for tracking health-related phenomena. *Studies in Health Technology and Informatics* 316, 1933–1937, doi: 10.3233/SHTI240811.
- Neo4j (2025). Neo4j cypher manual : Overview. <https://neo4j.com/docs/cypher-manual/current/introduction/cypher-overview/>. Accessed : 2025-11-03.
- Xiong, G., Q. Jin, Z. Lu, et A. Zhang (2024). Benchmarking retrieval-augmented generation for medicine.
- Yang, A., A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, et Z. Qiu (2025). Qwen3 technical report.

Summary

We present Med-KAG, a new Artificial Intelligence (AI) assistant architecture designed to improve reliability in clinical decision support. Med-KAG extends the Retrieval-Augmented Generation (RAG) paradigm by integrating a biomedical Knowledge Graph derived from the Unified Medical Language System (UMLS) Metathesaurus. This approach aims to reduce factual hallucinations and enhance transparency by grounding responses in verified medical relationships between diseases, symptoms, and treatments. A preliminary evaluation on MedQA-US compares a baseline large language model (Qwen3-235B-A22B) with its knowledge-enhanced variant. The results show comparable accuracy between both configurations and identify the retriever as the primary source of errors. This work highlights the potential of combining structured medical knowledge with generative AI to achieve more explainable clinical assistance, while emphasizing key technological bottlenecks that remain to be addressed.