Datalchemy

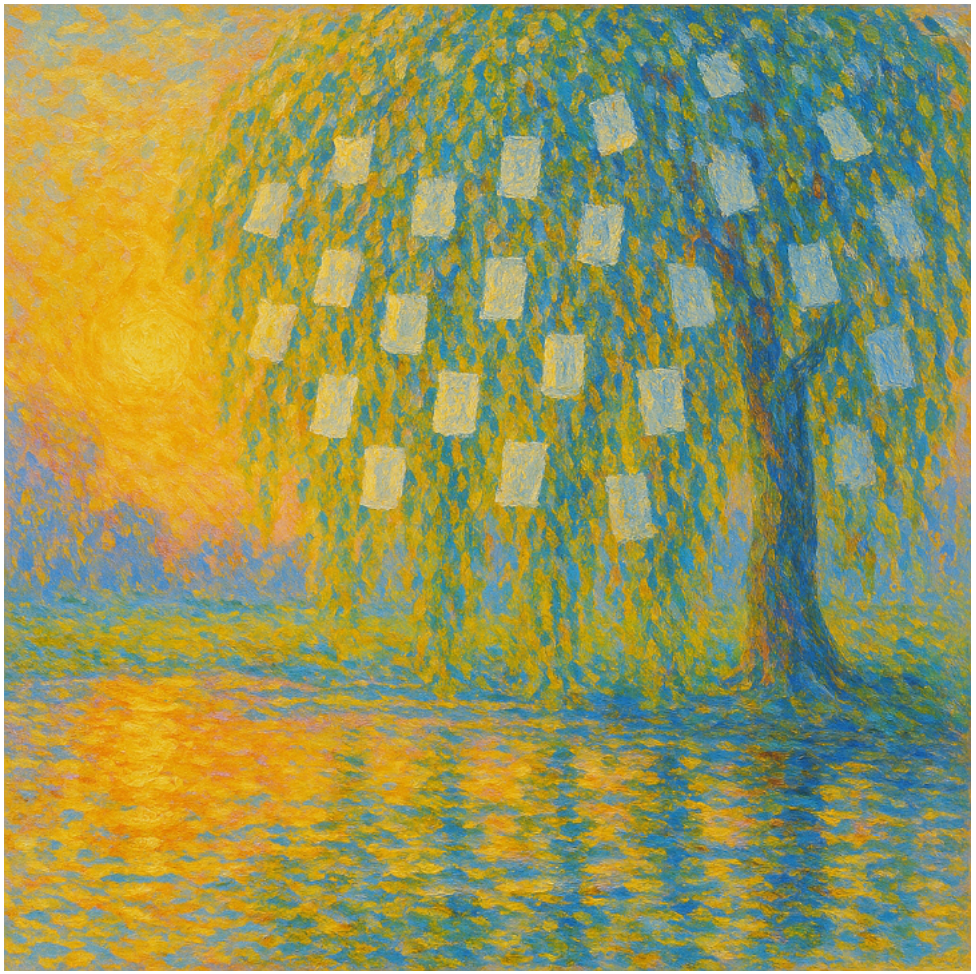
ECHOS

DE LA RECHERCHE #20

AVRIL 2025

GraphRAG is the new black

TL;DR ?



Cinq mot-clés de ces échos

#RAG, #GraphRAG, #documents, #connaissance, #graphes, #LLM

Pourquoi lire cette publi peut vous être utile concrètement ?

Vous vous apprêtez à tester le RAG pour « poser des questions à une base documentaire ». Ou plutôt : vous venez de tester ces solutions et êtes un peu déçu ? Voici l'occasion de revenir sur les défauts fondamentaux de ces approches et d'observer ce que la recherche académique récente peut proposer.

Si vous n’avez qu’une minute à consacrer à la lecture maintenant, voici le contenu essentiel en 7 phrases

1. Le RAG souffre toujours de défauts fondamentaux en industrialisation, observés par l’ensemble des ingénieurs se frottant au sujet.

2. Microsoft a proposé en 2024 une nouvelle approche baptisée GraphRAG promettant une meilleure structuration de l’information.

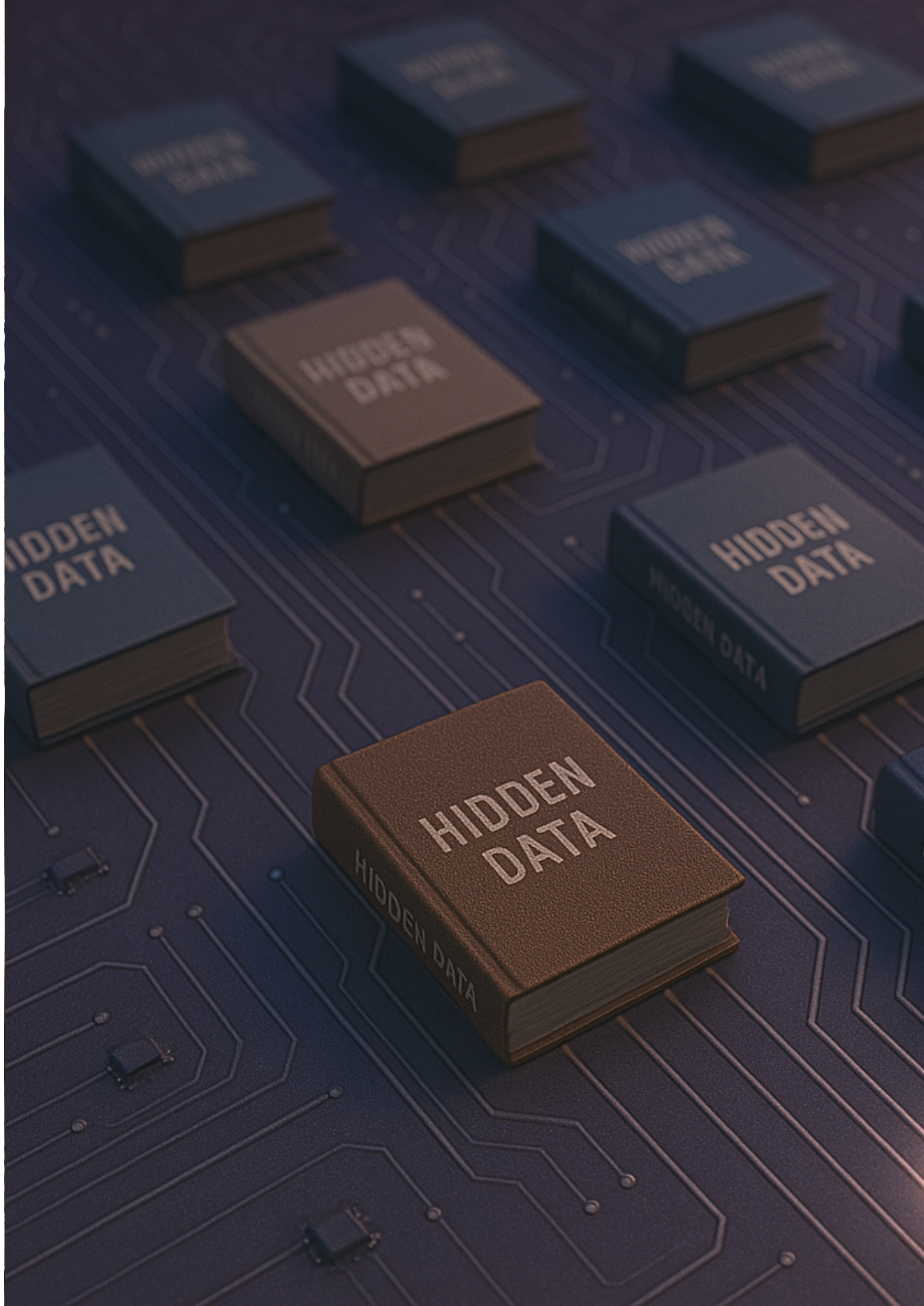
3. Le GraphRAG est aussi une approche massive, mais génère une hiérarchie de graphes sur l’information découverte et en permet une meilleure localisation.

4. Néanmoins, il reste difficile d’estimer la qualité de ces approches en l’absence de benchmarks réels et validés.

5. Le MediGraph RAG est une approche plus récente et très intéressante dans la mesure où elle est moins générique et spécialisée dans un domaine, ici, la littérature médicale.

6. Cette approche donne des clés intéressantes pour exploiter un graphe de connaissance disponible et construire un nouveau graphe issu de la documentation.

7. Nous donnons enfin un autre exemple d’approche utilisant un graphe déjà disponible reliant les éléments documentaires.



LE GRAPHRAG VA-T-IL SAUVER LE RAG ?

GraphRAG ?

Structurer et exploiter une base documentaire avec de l'IA ? "Poser des questions à une base documentaire", pour reprendre entre guillemets polis une promesse souvent affichées dans le monde onirique et marketing de Linkedin ? Le RAG (Retrieval Augmented Generation) est dans l'air du temps depuis plusieurs années, mais force est de reconnaître que cette approche reste critiquable. Depuis la publication originale de Meta AI, les tentatives d'implémentation dans l'industrie ont connu plus d'échecs que de succès, là où le monde académique continuait de proposer de nouvelles solutions plus ou moins originales...

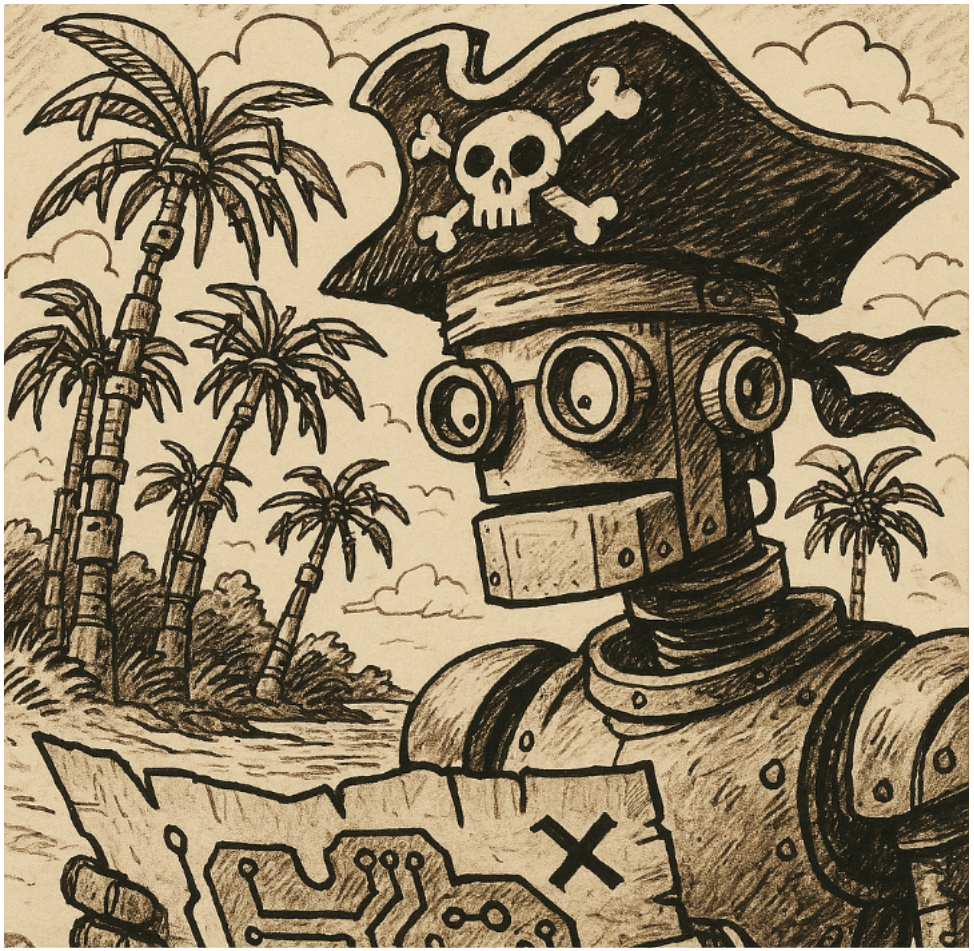
Nous avons déjà maintes fois signalé les problèmes de cette approche, mais considérant le sujet de cet article, il n'est pas superflu de reprendre :

- Les approches originelles entraînaient un modèle sur la base documentaire, ce qui va à l'encontre des approches applicatives : personne ne veut réentraîner un modèle chaque fois qu'un nouveau document apparaît.

- L'approche RAG conduit souvent à des résultats très disparates... Dans certains cas, cela marche, dans d'autres cas non. Et le fier mais néanmoins conscient data-scientist sait trop bien qu'il n'existe pas de solution pour corriger quoi que ce soit.

- L'approche RAG vise, à partir d'une question, à identifier les éléments (chunks) pertinents dans la base documentaire via un pré-filtrage et une similarité de cosinus, pour ensuite les injecter dans le contexte du LLM. Découper, structurer ces éléments est fondamental (au-delà du simple découpage ligne à ligne), mais les solutions ne sont pas évidentes.

Pertinent pour sa promesse de valeur (tout le monde a une base documentaire aussi précieuse qu'ignorée), mais défaillant dans ses résultats, le RAG a donc connu de très nombreuses propositions d'améliorations techniques. Et nous vous proposons, aujourd'hui, de nous intéresser à un courant qui fait de plus en plus de bruit : le GraphRAG. Combiner ces objets mathématiques agréables que sont les graphes avec les approches du RAG débloque-t-il enfin la situation ? Petit point de recherche depuis trois publications fondamentales sorties ces deux dernières années.



FROM LOCAL TO GLOBAL: A GRAPHRAG APPROACH TO QUERY-FOCUSED SUMMARIZATION, EDGE ET AL, MICROSOFT

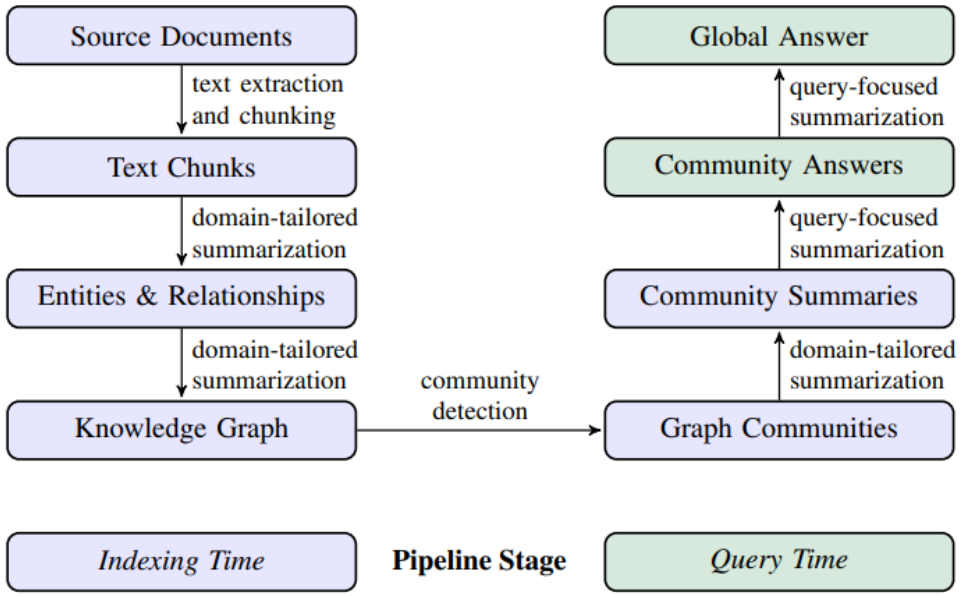
Cette publication scientifique, de Microsoft Research, est considérée comme fondatrice du mouvement du GraphRAG. L'argument central donné par les auteurs porte sur les cas de figure où on veut poser une question nécessitant une compréhension "complète" du dataset. Par exemple, si nous imaginons un corpus documentaire de publications sur dix ans, nous pourrions vouloir extraire les principaux thèmes sur l'ensemble de ces publications. Et dans ces cas-là, les approches RAG classiques sont incapables de travailler, dans la mesure où elles commencent par extraire un sous-ensemble du dataset documentaire qui, de fait, ne pourra contenir toute l'information nécessaire pour répondre.

D'une manière plus générale, le GraphRAG s'inscrit ici dans une lignée de travaux qui cherchent à exploiter une structure de graphe pour appréhender un ensemble d'informations. Faut-il le rappeler, le graphe est un objet mathématique très simple et très riche pour modéliser des éléments et les liens qui existent entre ces éléments. On peut supposer que là où le RAG est fortement limité, ayant accès à une base documentaire via une base de vecteurs brute, la découverte et l'exploitation de liens entre les informations modélisées en graphe pourrait enrichir fortement l'utilisation de cette connaissance.

ci, l'approche se distingue par l'idée qu'un graphe est une entité assez facile à séparer en sous-graphes ou à agrége. Cette approche va être particulièrement importante pour disposer d'une vision hiérarchique sur l'information : à haut niveau, une vision globale et sommaire, mais la capacité de descendre à bas niveau pour observer les relations fines entre chaque entité.

Un point d'attention important avant d'aborder la méthode : ces domaines de recherche sont tellement récents qu'il existe très peu de benchmarks valables pour mesurer leur qualité. Ce point est déjà une alerte sur le suivi de ces travaux où un score affiché peut être totalement décorrélé des résultats réels d'un outil. Mais c'est aussi une raison pour laquelle les auteurs ont généré leur propre benchmark en générant des cas de figure avec des LLMs. Cela encourage une certaine prudence sur la généralisation de cette méthode.

La méthode est représentée dans le schéma ci-dessous, schéma que nous allons ensuite détailler :



- Source Documents → Text Chunks

Le découpage des documents en "chunks" est classique par rapport aux approches usuelles.

- **Text Chunks → Entities & Relationships**

Une fois ces "chunks" extraits, un LLM va être interrogé spécifiquement pour extraire, depuis chaque élément, les entités découvertes et les relations entre ces entités. On retrouve des approches utilisées en recherche de causalité (cf dernier article), approches que l'on sait faillibles. Il est d'ailleurs intéressant d'observer que le nombre d'entité extraites évolue beaucoup en fonction du nombre d'appels du LLM et de la taille du chunk :

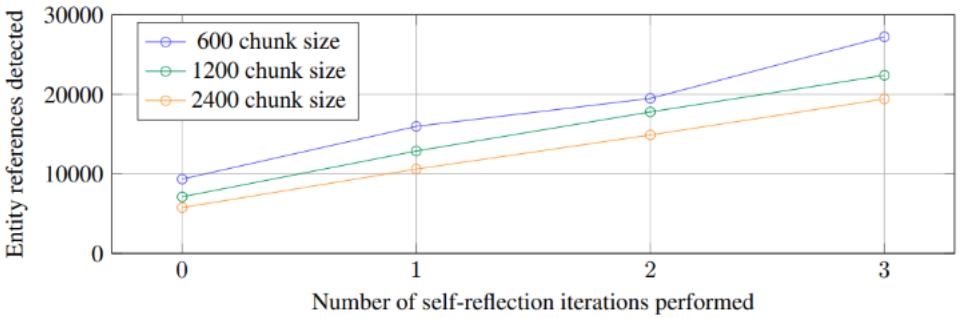


Figure 3: How the entity references detected in the HotPotQA dataset (Yang et al., 2018) varies with chunk size and self-reflection iterations for our generic entity extraction prompt with `gpt-4-turbo`.

- **Entities & Relationships → Knowledge Graph**

Les extractions réalisées ont vocation à s'intégrer dans un graphe global. Évidemment, un même "node" du graphe sera extrait plusieurs fois de la base documentaire. Chaque entité est représentée par un contexte et un résumé de définition de l'entité.

- **Knowledge Graph → Graph Communities**

Ce point est particulièrement intéressant. Le graphe généré est immense, et nous ne pourrons travailler directement dessus. Des algorithmes de similarité vont donc viser à identifier des sous-graphes d'éléments très proches les uns des autres, ici les "Graph Communities". Cette approche est récursive et donne à une extrémité le graphe complet, et à l'autre une vision très globale de l'ensemble des éléments. Cette approche génère donc une hiérarchie sur l'information. Dans le schéma ci-dessous, on voit à gauche une partition (couleurs) très globale avec des nœuds (ronds visibles) qui représentent chacun un point d'entrée dans une "communauté". A droite, nous descendons d'un niveau et gagnons en précision, ou chaque "communauté" est elle-même un graphe de sous-communautés.

(a) Root communities at level 0 (b) Sub-communities at level 1

Figure 4: Graph communities detected using the Leiden algorithm (Traag et al., 2019) over the MultiHop-RAG (Tang and Yang, 2024) dataset as indexed. Circles represent entity nodes with size proportional to their degree. Node layout was performed via OpenORD (Martin et al., 2011) and Force Atlas 2 (Jacomy et al., 2014). Node colors represent entity communities, shown at two levels of hierarchical clustering: (a) Level 0, corresponding to the hierarchical partition with maximum modularity, and (b) Level 1, which reveals internal structure within these root-level communities.

- **Graph Communities → Community Summaries**

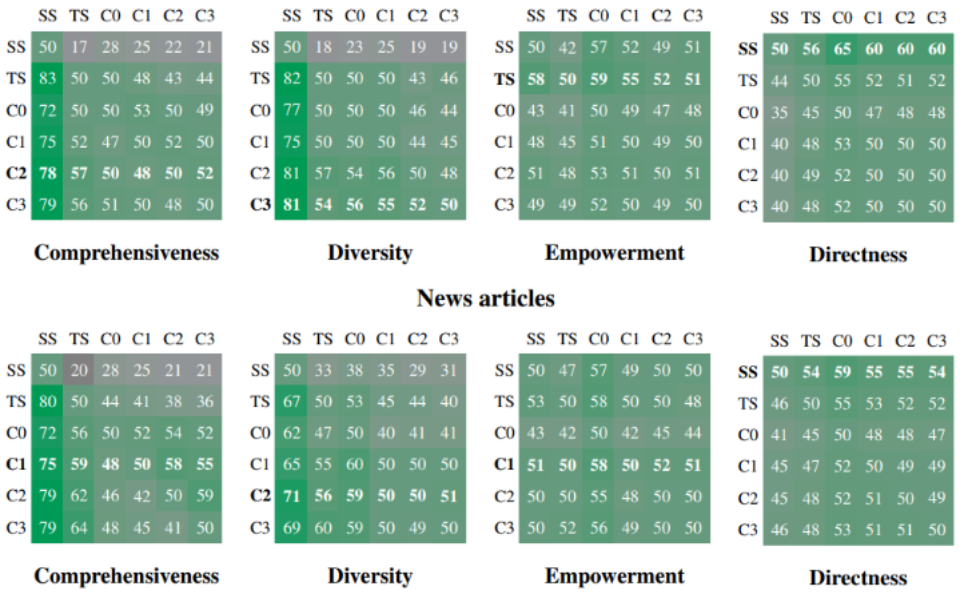
Chaque communauté (à chaque niveau hiérarchique) sera représentée par un résumé. A un niveau fin, ce résumé sera une représentation de l'information du nœud. A plus haut niveau, les résumés des enfants seront agrégés.

- **Community Summaries → Community Answers**

face à une question, un LLM va être utilisé pour comparer les résumés de chaque communauté à la question. Une agrégation va ensuite être mise en place pour générer le contexte de réponse du LLM.

Concernant les résultats que nous allons présenter ensuite, rappelons que le dataset de test est généré par appel à un LLM via l'usage de "Personas". Mais plus dangereux, les résultats numériques viennent aussi d'un appel à un LLM qui comparera différentes réponses pour évaluer un score. Ces critères portent sur la compréhension (combien de détails sont donnés en réponse et ces détails couvrent-ils l'étendue de la question), la diversité (richesse de la réponse) et l'efficacité. Gardons un peu de recul face aux résultats qui, s'ils sont intéressants, restent dangereux à évaluer.

Le schéma ci-dessous montre, pour chaque couple d'approches, le nombre de fois qu'une approche a eu un meilleur score que la deuxième. Par exemple, ci-dessous, en diversité (Diversity), l'approche "TS" a un meilleur score que l'approche "SS" dans 82% des cas. Les approches ici sont C0 à C3 : différentes déclinaisons du GraphRag où on utilise uniquement un niveau hiérarchique du graphe sélectionné, TS une approche simplifiée et SS l'approche RAG "classique" :



Enfin, à titre d'illustration, vous trouverez ci-après une question, la réponse GraphRAG, la réponse RAG classique, ainsi que la "décision" du LLM :

Question	Which public figures are repeatedly mentioned across various entertainment articles?
Graph RAG	The entertainment industry is vast and diverse, encompassing film, television, music, sports, and digital media. Certain public figures stand out due to their significant contributions and influence across these sectors. The following summary highlights key individuals who are repeatedly mentioned in various entertainment articles, reflecting their impact and presence within the industry. **Actors and Directors** [...] **Public Figures in Controversy** [...] **Musicians and Executives** [...] **Athletes and Coaches** [...] **Influencers and Entrepreneurs** [...] The repeated mention of these figures in entertainment articles signifies their ongoing relevance and the public's interest in their work. Their influence spans across various aspects of entertainment, from shaping cultural narratives in film and television to driving trends in music and digital media. These individuals not only contribute to their respective fields but also influence the broader cultural landscape, often becoming central figures in social discussions and public discourse.
Naïve RAG	Public figures who are repeatedly mentioned across various entertainment articles include Taylor Swift, Travis Kelce, Britney Spears, and Justin Timberlake. These individuals have been highlighted for various reasons, ranging from their professional achievements to their personal lives. **Taylor Swift** [...] **Travis Kelce** [...] **Britney Spears** [...] **Justin Timberlake** [...] These figures are frequently covered due to their high-profile status and the public's interest in their careers and personal lives. Their activities, whether in music, sports, or personal relationships, have significant cultural and economic impacts, as evidenced by the media coverage and public reactions.
LLM Decision	**Comprehensiveness: Winner=1 (Graph RAG)**. Answer 1 is better because it provides a more comprehensive and detailed list of public figures from a wider range of entertainment sectors, including film, television, music, sports, gaming, and digital media. It also includes examples of their contributions and the impact they have on their respective fields, as well as mentions of controversies and their implications. Answer 2, while detailed in its coverage of a few individuals, is limited to a smaller number of public figures and focuses primarily on their personal lives and relationships rather than a broad spectrum of their professional influence across the entertainment industry. **Diversity: Winner=1 (Graph RAG)**. Answer 1 is better because it provides a more varied and rich response by covering a wide range of public figures from different sectors of the entertainment industry, including film, television, music, sports, gaming, and digital media. It offers insights into the contributions and influence of these figures, as well as controversies and their impact on public discourse. The answer also cites specific data sources for each mentioned figure, indicating a diverse range of evidence to support the claims. In contrast, Answer 2 focuses on a smaller group of public figures, primarily from the music industry and sports, and relies heavily on a single source for data, which makes it less diverse in perspectives and insights.

MEDICAL GRAPH RAG: TOWARDS SAFE MEDICAL LARGE LANGUAGE MODEL VIA GRAPH RETRIEVAL-AUGMENTED GENERATION

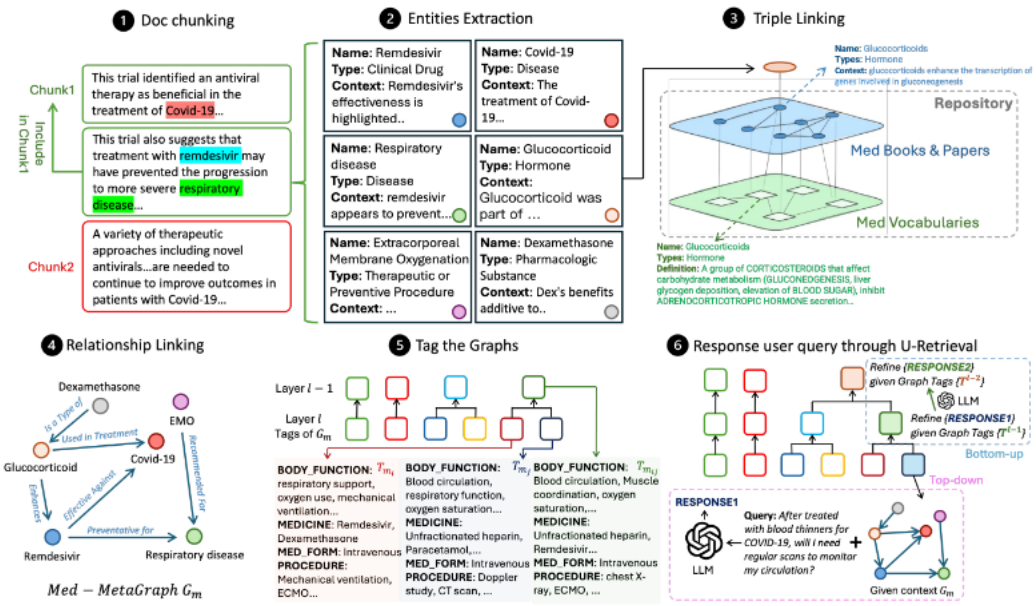
Le GraphRAG de Microsoft a lancé tout un nouveau domaine de recherche, avec une promesse savoureuse pour les galériens du RAG : découvrir et exploiter des relations entre les informations découvertes pour accompagner la réponse. Plus d'une centaine de travaux sont sortis à la suite pour proposer adaptations et évolutions. Nous vous proposons maintenant de nous concentrer sur une spécialisation de l'approche particulièrement intéressante, ce pour deux raisons :

Le domaine concerné est ici le domaine médical, particulièrement exigeant en termes de qualité et de précision des réponses.

Cette spécialisation nous informe d'une manière très intéressante sur les moyens de sortir de l'approche générique proposée par Microsoft. Or, en intelligence artificielle, il est rarissime de pouvoir appliquer directement une approche académique sans l'adapter au problème cible. Cet exemple est donc précieux :)

Un autre élément important est ici d'utiliser des connaissances externes déjà disponibles, conjointement à la base documentaire cible. Le monde médical regorge de définitions exactes et précises, de taxonomies et autres pouvant accompagner l'approche RAG. Enfin, les auteurs soulignent que l'approche GraphRAG, notamment la génération des différentes communautés hiérarchiques dans le graphe, est particulièrement coûteuse en temps de calcul.

Mais alors, quels mécanismes se cachent dans cette nouvelle approche spécialisée ? Le schéma ci-dessous donne une vision d'ensemble, que nous allons prendre le temps de décortiquer :)



- **Semantic Document Chunking**

Toujours incontournable, le chunking de la base documentaire est ici un poil plus avancé. Au-delà du chunking "classique" (basé sur une longueur maximale), les auteurs exploitent ici un LLM pour mesurer la cohérence sémantique entre une ligne et la suivante. L'idée est d'avoir des chunks qui soient le plus cohérents possibles à l'intérieur. Une fenêtre glissante est exploitée pour éviter les coupures trop brutales.

- **Entities Extraction**

L'extraction d'entités reste centrale dans ces approches. Cette extraction se fait toujours via un LLM instrumenté pour l'occasion. Le texte extrait représente une entité en agrégeant son nom, son type (tel que déterminé par le LLM) ainsi qu'une description du contexte dans lequel l'entité a été découverte.

- **Triple Linking**

C'est ici que le spécifique prend sa place. Trois graphes différents vont exister, les deux premiers étant génériques quelle que soit la documentation cible, et le troisième généré depuis la base documentaire. Le premier graphe exploite une base de publications scientifiques médicales, et le second un dictionnaire médical spécialisé. Via des recherches de similarité, les éléments des deux graphes vont être liés entre eux, puis liés au graphe contenant les entités extraites dans la documentation en cours d'exploitation. L'idée est bien ici d'exploiter des bases de connaissances transversales et validées, que l'on condense en "Knowledge Graph" pour pouvoir ensuite y appliquer le graphe découvert dans la documentation.

- **Relationship Linking**

Chaque couple d'éléments du graphe global va être questionné (via un LLM, une nouvelle fois), pour qualifier le lien (s'il existe) entre les deux éléments.

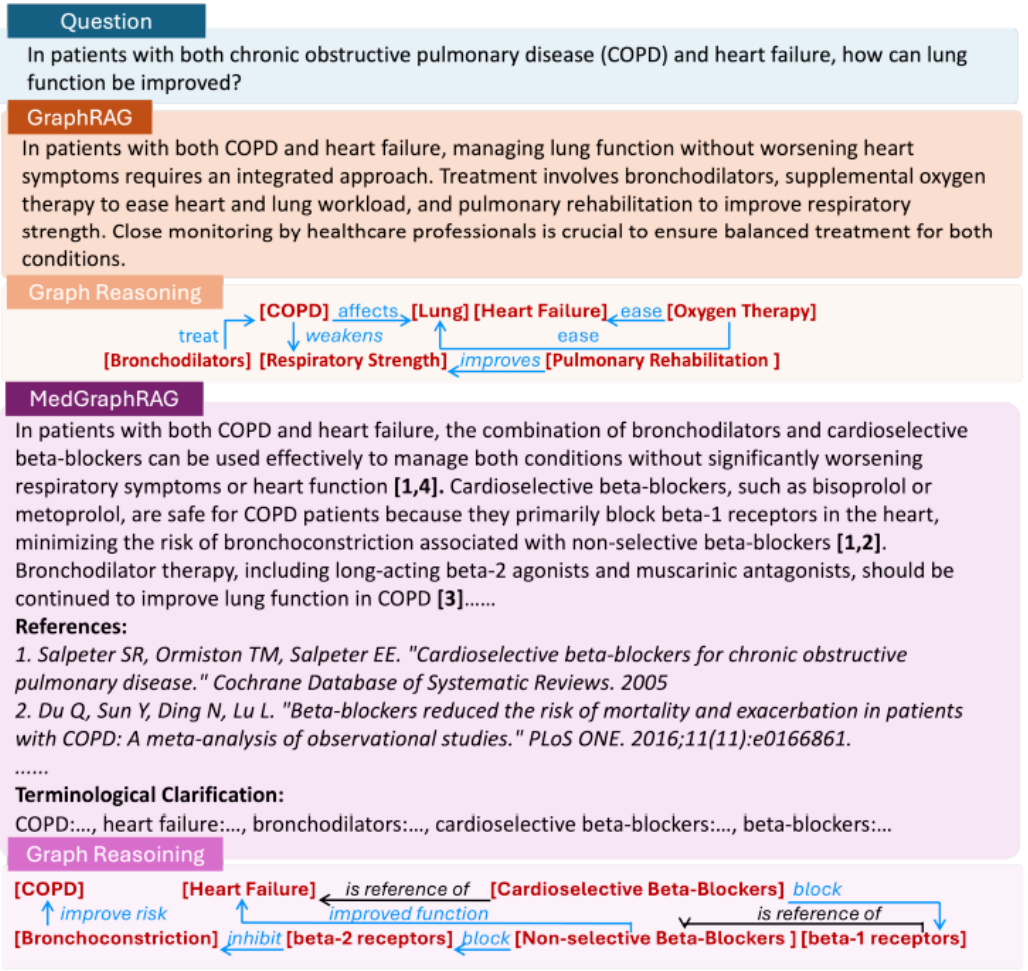
- **Tag the Graphs**

L'enjeu ici est de ne pas reproduire la découverte des "Graph Communities" générées par le GraphRAG originel. Les auteurs exploitent une base de tags (classifications) déjà existantes dans le domaine médical, et commencent par appliquer le tag correspondant à chaque nœud du graphe. Un algorithme va ensuite remonter vers des sous-graphes de plus en plus conséquents, en générant des tags "résumés" pour les sous-groupes de nœuds ayant des tags proches.

- **Réponse à une question**

Face à une demande, l'approche va déjà identifier les tags les plus importants. Ces derniers vont servir à sélectionner le sous-graphe le plus pertinent. Les éléments retenus seront utilisés comme contexte de réponse au LLM.

Le schéma ci-après montre deux exemples de questions et de réponses avec comparaison entre le GraphRAG original et cette version spécialisée.



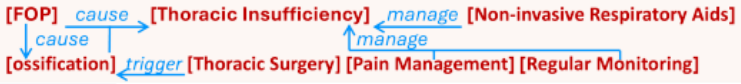
Question

What are the most effective strategies for managing progressive thoracic insufficiency in patients with fibrodysplasia ossificans progressiva (FOP)?

GraphRAG

Managing progressive thoracic insufficiency in patients with fibrodysplasia ossificans progressiva (FOP) focuses on avoiding thoracic surgeries that can trigger further ossification. Supportive care includes non-invasive respiratory aids, effective pain management, and regular monitoring of lung function and chest mobility to maintain quality of life and manage symptoms conservatively. Additionally, physical therapy tailored to individual needs can help improve mobility and function without exacerbating the condition.

Graph Reasoning



MedGraphRAG

Managing progressive thoracic insufficiency in FOP patients requires a careful balance between effective respiratory support and preventing heterotopic ossification (HO) **[2]**. Early use of non-invasive ventilation (NIV) **[1]**, such as BiPAP, particularly during sleep or in response to respiratory infections, can improve respiratory function without invasive procedures **[2]**, minimizing the risk of trauma **[3]**...Gentle, supervised physical therapy helps maintain lung capacity while avoiding trauma that could trigger HO **[4]**...

References:

1. Shulmanl, et al. "Maternal genetic skeletal disorders: lessons learned from cases of maternal osteogenesis imperfecta and fibrodysplasia ossificans progressiva." Journal of Clinical ..

2. Pignolo RJ, et al. "The natural history of flare-ups in fibrodysplasia ossificans progressiva (FOP): a comprehensive global assessment." Journal of Bone and Mineral Research. 2016

.....

Terminological Clarification:

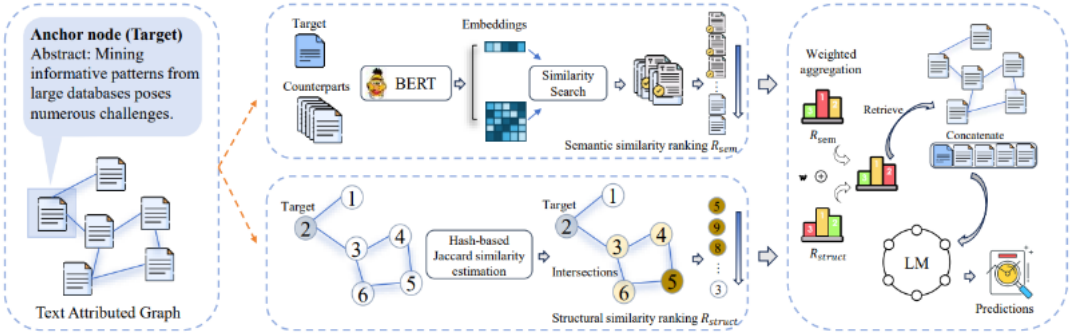
heterotopic ossification (HO):..., non-invasive ventilation (NIV):..., BiPAP:...

Graph Reasoning

Les résultats ci-dessous présentent les résultats sur différentes approches, depuis la simple interrogation brute de LLM jusqu'à cette approche spécifique. Ces résultats sont aussi l'occasion de montrer qu'en termes de résultats "bruts", ces approches ne sont pas si révolutionnaires. C'est plus sur la capacité à exposer le graphe à un utilisateur tiers que ces approches gagnent en pertinence, en exposant partiellement le fonctionnement de la recherche.

Model	Fake Health	Pub Health	MedQA	Med MCQA	Pub MedQA	MMLU Col-Med	MMLU Col-Bio	MMLU Pro-Med	MMLU Anatomy	MMLU Gene	MMLU Clinic
Baselines without retrieval											
Llama2-13B	53.8	49.4	42.7	37.4	68.0	60.7	69.4	60.3	52.6	66.0	63.8
Gemini-pro	60.6	63.7	59.0	54.8	69.8	69.2	88.0	77.7	66.7	75.8	76.7
GPT-4	71.4	70.9	78.2	72.6	75.3	76.7	95.3	93.8	81.3	90.4	86.2
Baselines with RAG											
Llama2-13B	56.2	54.3	48.1	42.0	68.6	62.5	68.3	63.7	51.0	64.5	67.4
Gemini-pro	72.5	68.4	64.5	57.3	76.9	79.0	91.3	86.4	79.5	80.4	83.9
GPT-4	78.6	77.3	88.1	76.3	77.6	81.2	95.5	94.3	83.1	92.9	93.1
Baselines with GraphRAG											
Llama2-13B	58.7	57.5	52.3	44.6	72.8	64.1	73.0	64.6	52.1	66.2	67.9
Gemini-pro	73.8	70.6	65.1	59.1	75.2	79.8	90.8	85.8	80.7	81.5	84.7
GPT-4	78.4	77.8	88.9	77.2	77.9	82.1	95.1	94.8	82.6	92.5	94.0
Baselines with MedGraphRAG											
Llama2-13B	64.1	61.2	65.5	51.4	73.2	68.4	76.5	67.2	56.0	67.3	69.5
Gemini-pro	79.2	76.4	71.8	62.0	76.2	86.3	92.9	89.7	85.0	87.1	89.3
GPT-4	**86.5**	**83.4**	**91.3**	**81.5**	83.3	**91.5**	**98.1**	**95.8**	**93.2**	**98.5**	**96.4**

LARGE LANGUAGE MODELS BASED GRAPH CONVOLUTION FOR TEXT-ATTRIBUTED NETWORKS, ZHOU ET AL



Ce dernier travail, présenté à l'ICLR 2025, est l'occasion d'observer une approche assez différente mais néanmoins classique dans le domaine. En effet, quitte à travailler sur des graphes, ne pourrait-on utiliser les modèles Deep Learning qui sont spécialisés sur la gestion des graphes, les bien nommés Graph Neural Networks (GNNs) ? Si l'arrivée des LLMs a eu l'effet d'une lame de fond sur l'ensemble de la recherche en Deep Learning, avec le temps, les approches différentes (et dans de nombreux cas, plus pertinentes) se rappellent au souvenir des chercheurs, ne serait-ce que pour explorer notre capacité à comparer, voir à mixer ces techniques entre elles.

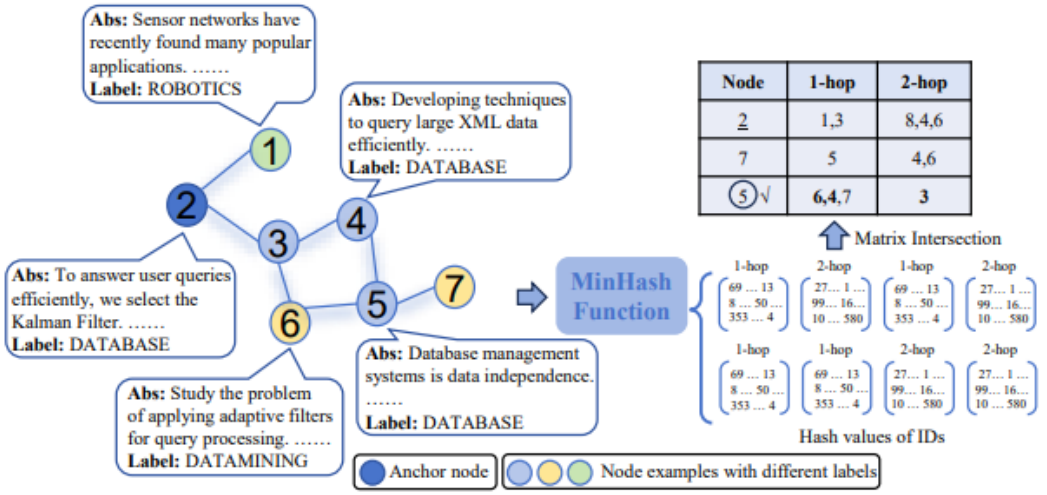
Le problème originel (il y a toujours un problème) est que les approches GNNs ont une vision sur les liaisons entre éléments mais restent assez aveugles sur le contenu textuel de chaque élément.

Ici, les chercheurs vont donc mettre en concurrence les deux approches, en ayant initialement un graphe composé de ressources textuelles (Text Attributed Graph) :

- L'approche sémantique (en haut du graphe) ou via des embeddings et un calcul de similarité, les éléments les plus intéressants seront conservés.

- L'approche structurelle, cherchant à partir d'un noeud source les noeuds du graphes les plus pertinents à retenir.

L'approche structurelle compare des nœuds en comparant les voisins communs de deux nœuds pour estimer leur proximité. On estime que deux nœuds du graphe sont d'autant plus proches qu'ils ont beaucoup de nœuds voisins en commun (voisins directs ou indirects). Ci-dessous, on estime que le nœud source (nœud 2) a pour plus proche voisin le nœud 5.



CONCLUSIONS

Le GraphRAG est encore un domaine très récent, voire trop récent. L'approche originale est extrêmement lourde, et quand il faut générer synthétiquement un benchmark de test et utiliser des LLMs comme métriques, on peut partir du principe que l'approche est trop jeune pour être qualifiée correctement.

Le GraphRAG ne règle pas tous les défauts du RAG : nous restons sur une approche massive avec de vraies difficultés de mesure et de test, sans solution de correction immédiate. En revanche, l'utilisation ou la génération de graphes de connaissances est une approche très pertinente pour modéliser la connaissance (et sortir un meilleur parcours de l'information que la simple liste de vecteurs stockés) comme pour exposer l'information à l'utilisateur.

Nous continuons de suivre le domaine et, déjà aujourd'hui, pouvons récupérer de ces travaux des pistes pertinentes pour améliorer nos solutions 😊

Datalchemy

contact@datalchemy.net

Images extraites des articles respectifs ou générées par IA