Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Approche méthodologique pour la génération de question-réponse portant sur plusieurs documents

Hui Huang[1, 2] Julien Velcin[1] Yacine Kessaci[2]
(1) Université Lumière Lyon 2, Université Claude Bernard Lyon 1, ERIC, 69007, Lyon, France
(2) Worldline S.A., 92059, Paris La Défense Cedex, France
hui.huang@univ-lyon2.fr, julien.velcin@univ-lyon2.fr,
yacine.kessaci@worldline.com

RÉSUMÉ

Les systèmes de questions-réponses (QA pour *Question Answering*) actuels ont du mal à synthétiser les preuves dispersées dans les documents. Alors que les jeux de données QA scientifiques existants se concentrent sur le raisonnement portant sur un document seul, la tâche de recherche peut exiger l'intégration de contenus provenant de plusieurs articles. Pour répondre à cette limitation, nous proposons un cadre pour créer un jeu de données QA multi-documents qui s'appuie sur l'analyse de graphes de citations afin de regrouper des articles connexes et utilise un grand modèle de langage (LLM) pour générer des questions complexes. Des expériences préliminaires réalisées sur 23 882 articles démontrent la faisabilité de ce cadre, produisant 238 paires QA qui nécessitent une synthèse sur plusieurs articles. D'autres expériences indiquent que la recherche d'information dense actuelle obtient un rappel limité pour ces questions multi-documents, soulignant le besoin de mécanismes de recherche d'information et de raisonnement plus avancés. Il s'agit d'un projet en cours d'élaboration. Nous visons à terme à fournir un jeu de données QA robuste qui capture la complexité et la nature interconnectée des publications scientifiques, ouvrant la voie à des évaluations plus réalistes des systèmes de QA.

ABSTRACT

Methodological approach for generating question-answers on multiple documents

Current question-answering (QA) systems struggle to synthesize evidence scattered across scientific documents. While existing scientific QA datasets focus on reasoning from single documents, real-world research often demands integrating content from multiple articles. To address this limitation, we propose a framework to construct multi-document QA datasets by leveraging citation-graph community detection to cluster related papers and employing a large language model (LLM) to generate multi-document questions. Preliminary experiments on 23,882 machine-learning papers demonstrate the framework's feasibility, yielding 238 QA pairs requiring inter-article synthesis. Further experiments indicate that current dense retrieval has limited recall for these multi-document questions, highlighting the need for more advanced retrieval and reasoning mechanisms. This is an ongoing project, we aim to provide a robust QA dataset that captures the complexity and interlinked nature of scientific publications, paving the way for more realistic evaluations of QA systems.

MOTS-CLÉS : Questions-réponses multi-documents, Réseau de citations, Détection de communautés, Grande modèles de langage.

KEYWORDS: Multi-document Question Answering, Citation Network, Community Detection, Large Language Models.

Actes de la 20e Conférence en Recherche d'Information et Applications (CORIA), pages 21–30.

1 Introduction

Les systèmes de questions-réponses (QA pour *Question Answering*) ont pris de plus en plus d'importance pour aider les chercheurs et les étudiants à accéder rapidement à des informations complexes et à les comprendre (Hermann *et al.*, 2015). Cependant, à mesure que la littérature scientifique se développe, extraire efficacement et précisément les contenus clés d'une vaste bibliothèque et fournir des réponses pertinentes, est devenu un défi important (Chen *et al.*, 2017). En effet, de nombreuses questions de recherche du monde réel nécessitent la réalisation d'une synthèse d'éléments (preuves) provenant de plusieurs sources. Se baser sur une unique source (un document) ne suffit pas.

La plupart des jeux de données QA traditionnels se concentrent sur des documents uniques ou effectuent un raisonnement multi-hop (Kwiatkowski *et al.*, 2019; Yang *et al.*, 2018; Kočiskỳ *et al.*, 2018). Bien que les tâches QA multi-hop précédentes nécessitent plusieurs documents, elles se basent sur un raisonnement qui suit des chaînes logiques ou déductives. In fine, la réponse finale à la question se trouve dans un seul document (Zhu *et al.*, 2024). Or, il arrive souvent que la réponse soit dispersée à travers des articles car elle concerne des méthodologies partagées ou contrastées, ou même des concepts qui se chevauchent mais ne sont pas identiques. Les jeux de données QA existants pour la littérature scientifique capturent rarement ces relations inter-documents ou ne proposent pas des questions-réponses nécessitant la consultation de plusieurs documents (Jin *et al.*, 2019; Tsatsaronis *et al.*, 2015). Par conséquent, nous pensons qu'il existe une opportunité de créer des nouvelles QA plus complexes qui concernent *plusieurs* documents.

Pour combler cette lacune dans la littérature, nous proposons un cadre pour construire un jeu de données QA spécialisé nécessitant la référence à plusieurs documents. L'idée clé est de générer des paires QA à travers de petits groupes d'articles qui sont sémantiquement liés (par ex. via des liens de citation), tout en garantissant que chaque élément QA implique plusieurs sources. Plus précisément, nous proposons d'exploiter la détection de communautés sur les graphes de citations pour identifier des groupes d'articles, puis d'utiliser un grand modèle de langage (LLM) pour générer des paires questions-réponses nécessitant des informations provenant de plusieurs articles. Nous utilisons un jeu de données en libre accès SPIQA (Pramanick *et al.*, 2025), que nous enrichissons avec des informations de structure (ici, les liens de citation), pour générer des paires QA et nous menons des expériences préliminaires sur ces QA. Bien que le travail soit en cours, les expériences préliminaires démontrent qu'un tel cadre peut produire des paires QA multi-document difficiles, allant au-delà du périmètre des tâches typiques nécessitant un seul document, même dans un cadre multi-hop.

Dans ce papier, nos contributions se présentent comme suit :

1. Nous identifions clairement le manque de données QA multi-documents dans les domaines scientifiques et présentons différentes méthodes de génération QA présentes dans l'état de l'art (section 2).

2. Nous détaillons notre méthodologie de génération QA multi-documents dans la section 3.

3. Nous rapportons des données et des expériences préliminaires dans la section 4 afin d'illustrer la faisabilité de notre approche.

2 Travaux Connexes

2.1 Jeux de données pour la QA sur les articles scientifiques

Au cours des dernières années, les chercheurs ont construit divers jeux de données de différentes échelles et formes pour les tâches de question-réponse centrées sur la littérature scientifique. Les recherches passées utilisaient des méthodes automatisées pour extraire des entités nommées et leurs relations à partir de documents afin de générer des paires de questions-réponses (Rajpurkar *et al.*, 2016; Jin *et al.*, 2019). Par la suite, des jeux de données tels que PubmedQA (Jin *et al.*, 2019), BIOASQ (Tsatsaronis *et al.*, 2015) et Qasper (Dasigi *et al.*, 2021) ont progressivement élargi la construction de questions et réponses pour inclure des résumés et des parties du texte entier, mettant l'accent sur une information sémantique plus riche. De nouveaux jeux de données, tels que QASA et SPIQA, ont été proposés pour encore élargir les frontières des tâches de QA. QASA (Lee *et al.*, 2023) met l'accent sur la compréhension du contenu du texte intégral, en créant des paires de QA à travers la lecture manuelle de documents complets, et couvre plus de détails sur le document. SPIQA (Pramanick *et al.*, 2025) introduit le traitement d'informations multimodales, en utilisant des grands modèles de langage visuels (VLLMs) pour générer des paires de QA de haute qualité basées sur les figures et les citations correspondantes dans les articles. Néanmoins, ces jeux de données restent généralement centrés sur des réponses qui apparaissent dans un document unique. Cela limite les requêtes comparatives ou synthétiques plus larges qui nécessitent des éléments répartis sur plusieurs articles.

2.2 Problématique des Questions-Réponses Complexes

Plusieurs travaux ont traité de méthodes de compréhension textuelle et de raisonnement basé sur la connaissance pour augmenter la complexité du raisonnement. Des jeux de données QA classiques sur un seul document comme SQuAD (Rajpurkar *et al.*, 2016) et Natural Question (Kwiatkowski *et al.*, 2019) aux tâches de raisonnement à multi-hop telles que HotpotQA (Yang *et al.*, 2018) ou aux jeux de données hybrides intégrant des données structurées comme Hybrid QA (Chen *et al.*, 2020b) et OTT-QA (Chen *et al.*, 2020a), il y a eu une tendance continue vers des configurations de QA plus diversifiées et complexes. Cependant, la plupart d'entre eux se limitent encore à des scénarios de document unique ou à des passages multi-hop soigneusement sélectionnés.

Notre travail vise à soutenir les scénarios de QA nécessitant la synthèse d'éléments distincts et complémentaires provenant de plusieurs documents. Cela diffère de la QA multi-hop en ce que la réponse finale n'est pas entièrement contenue dans une seule source et ne peut pas être obtenue simplement en effectuant des "sauts" à l'intérieur d'un texte unique ou même de plusieurs documents composant un corpus.

Par ailleurs, il existe également des travaux portant sur de multiples documents, par exemple dans le cadre de la création de revues de littérature ou de résumés multi-documents (Ma *et al.*, 2022; Wan & Yang, 2008; Ma, 2024). Ils mettent l'accent sur la fusion et le résumé d'informations issues de diverses sources, dans le but d'en produire une vue d'ensemble. Dans notre travail, nous nous distinguons de ces approches de tâches multi-documents plus globales, en ciblant explicitement la résolution de questions qui nécessitent la mise en commun des preuves issues de plusieurs articles distincts.

3 Méthodologie proposée

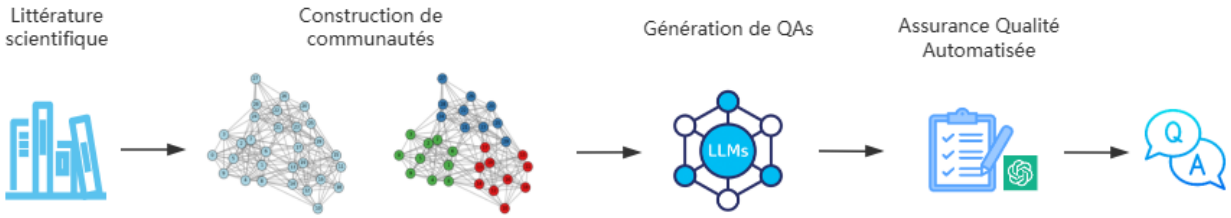

FIGURE 1 – Séquence des modules composant notre méthodologie pour générer des questions-réponses à partir de la littérature scientifique.

Nous présentons un processus pour la construction d'un jeu de données de question-réponse à partir de la littérature scientifique. Comme illustré dans la figure 1, ce cadre comprend trois modules principaux : Construction de Communauté (cf. §3.1), Génération de Questions-Réponses (cf. §3.2), et Assurance Qualité Automatisée (cf. §3.3). L'idée principale de ce cadre est de construire des communautés de haute qualité en capturant les structures relationnelles entre documents, de générer des questions et réponses basées sur plusieurs documents reliés entre eux, et d'effectuer une vérification systématique de la qualité et une évaluation des paires de questions-réponses générées.

3.1 Construction de Communauté

Dans la littérature scientifique, les articles présentent souvent des connexions relationnelles complexes et diverses, telles que des citations, des similitudes méthodologiques, des similarités thématiques, des chevauchements de mots-clés ou des réseaux de co-auteurs (Qiu *et al.*, 2014). Nous nous concentrons sur les liens de citation comme point de départ : les citations étant par nature orientées, il est pertinent de modéliser chaque article comme un nœud dans un graphe dirigé. Dans ce contexte, la direction de l'arête aide à analyser la propagation d'idées ou les dépendances bibliographiques.

Un algorithme de détection de communautés est ensuite appliqué pour produire de petits groupes d'articles connexes. Dans l'implémentation spécifique, nous adoptons l'algorithme de Speaker-listener Label Propagation (SLPA) (Kuzmin *et al.*, 2013), qui conserve plusieurs étiquettes lors des itérations répétées. Comparé aux algorithmes traditionnels de partitionnement strict ou de détection de communautés à relation unique comme Leiden (Traag *et al.*, 2019) et (Blondel *et al.*, 2008), le SLPA permet à un article d'apparaître dans plusieurs communautés s'il couvre plusieurs domaines de recherche. Cet alignement nous semble mieux correspondre aux phénomènes observés dans le monde réel.

Notre méthode permet également de raffiner ou d'étendre ces communautés en ajoutant ou en substituant d'autres mesures de similarité, telles que les embeddings sémantiques ou les graphes de co-auteur. En fin de compte, chaque sous-ensemble d'articles émerge avec un cohérence thématique ou méthodologique, servant de groupe candidat pour générer des paires de QA inter-articles.

3.2 Génération de QA

Une fois les communautés formées, nous sollicitons un grand modèle de langage pour créer des paires de questions-réponses impliquant directement plusieurs articles au sein de chaque communauté. Plus précisément, nous avons conçu un prompt structuré pour le LLM, lui demandant explicitement de générer des questions qui croisent au minimum deux documents au sein de chaque communauté, et de fournir une réponse concise intégrant ces multiples sources.

Plusieurs LLMs ont été évalués dans nos expérimentations préliminaires, notamment Claude-3.5-Sonnet [1], GPT-4o [2], et Gemini-1.5-Pro [3]. Nous avons appliqué le même prompt à 50 communautés extraites aléatoirement, puis évalué, pour chaque modèle, sa capacité à générer des questions multi-documents. Claude-3.5-sonnet s'est montré le plus performant, produisant des questions et réponses plus cohérentes et précises, ce qui a motivé notre choix de le retenir pour la suite de notre travail. Cependant, tous les LLMs avancés futurs capables de gérer un contexte multi-documents peuvent êtres employés. Nous veillons à ce que les QA générées contiennent des références spécifiques au contenu de *tous* les articles de support, augmentant ainsi leur complexité et leur nature multi-documents.

3.3 Assurance Qualité Automatisée

Bien que les LLMs soient capables de générer des questions complexes à partir de plusieurs documents, il arrive que certains résultats puissent encore être répondus par un seul document, ce qui compromet notre objectif d'un jeu de données multi-documents. Pour y remédier, nous concevons un processus d'assurance qualité automatisé.

Pour chaque paire question-réponse (q_i, a_i) et son ensemble de documents correspondant $S = \{s_1, s_2, ...\}$, nous fournissons individuellement au modèle un document s_j de la communauté, permettant au modèle de répondre à q_i, et enregistrons cette réponse comme a_i^m. Nous laissons ensuite le LLM comparer a_i et a_i^m pour déterminer quelle réponse est plus précise ou complète. En parcourant tous les documents de S, si seul a_i peut couvrir toutes les informations ou intégrer des conclusions contradictoires, la paire question-réponse est conservée, éliminant les problèmes de faible complexité auxquels on peut répondre par un seul document.

Ce processus d'assurance qualité automatisé permet d'éliminer les paires question-réponse qui semblent complexes mais qui, en réalité, peuvent être résolues par un seul document dans le cadre de la génération de données à grande échelle. Ainsi, il renforce la dépendance multi-documents de l'ensemble du jeu de données.

Notre procédure automatisée a été confrontée à une vérification manuelle structurée (échantillon de 30 paires QA vérifiées par trois experts du domaine). Elle a confirmé que toutes les questions conservées après filtrage nécessitaient réellement la consultation de plusieurs documents support.

1. https ://claude.ai/
2. https ://openai.com/index/gpt-4o-system-card/
3. https ://deepmind.google/models/gemini/

4 Expériences Préliminaires

Dans cette section, nous utilisons le cadre décrit précédemment pour construire un petit jeu de données de questions-réponses multi-documents et réaliser des expériences préliminaires. Ces expériences nous aident à vérifier la faisabilité et les limitations de la méthode proposée dans un environnement de littérature scientifique réel, fournissant une référence pour la construction de jeux de données de plus grande échelle et l'optimisation des algorithmes à venir.

4.1 Préparation des QA multi-documents

Notre étude est basée sur des publications académiques en libre accès collectées par SPIQA (Pramanick *et al.*, 2025). SPIQA a collecté 25 859 articles évalués lors de 19 conférences de haut niveau sur l'apprentissage automatique entre 2018 et 2023, en se concentrant sur les fichiers pdf accessibles au public et les fichiers source TeX correspondants pour extraire des textes d'articles originaux de haute qualité. Les articles utilisés dans notre corpus sont rédigés en anglais. Nous excluons par ailleurs 1 977 documents de l'ensemble SPIQA original en raison de leur illisibilité et incohérence causées par un contenu manquant, des anomalies de formatage ou des segments TeX non traitables. Nous utilisons l'API de Semantic Scholar [4] pour obtenir des informations de citation et construisons un graphe de citation en utilisant le système de gestion de bases de données graphes publique Neo4j [5], en considérant chaque document comme un nœud et chaque citation comme un arc.

Nous appliquons ensuite l'algorithme SLPA pour la détection de communautés. Cet algorithme prend en charge les graphes orientés et non orientés. Nous choisissons de l'appliquer sur un graphe orienté après des expérimentations. Cela permet de produire davantage de petites communautés, tandis que le graphe non orienté génère des communautés beaucoup plus grandes. Nous avons besoin des communautés plus petites, ce qui permet de formuler des questions plus précises lors du processus de génération. En tenant compte de la granularité et de la stabilité de la division des communautés, nous fixons le nombre d'itérations à 50 et le paramètre de filtrage à 0,1. Nous obtenons initialement 11 373 communautés.

Nous affinons ensuite les communautés pour garantir que les tâches de QA suivantes reflètent des scénarios multi-documents sans surcharger le contexte pour le LLM. Nous écartons 6 787 communautés contenant un seul nœud, issues de citations isolées ou d'articles plus récents non cités. Nous éliminons également 414 communautés de grande taille contenant plus de 13 documents car elles risqueraient de dépasser la limite de jetons en entrée du LLM lors de la génération des questions-réponses. Après filtrage, nous conservons 4 172 communautés de petite à moyenne taille, chacune comprenant de 2 à 7 documents. Les articles dans ces communautés couvrent 14 698 documents de la collection originale.

Nous sélectionnons ensuite aléatoirement 100 communautés pour lancer la génération de QA basée sur LLM, ce qui donne un ensemble initial de 298 paires question-réponse. Cette étape nous permet de valider la faisabilité de notre approche à petite échelle. À terme, nous prévoyons d'étendre ce processus à l'ensemble des communautés identifiées afin de construire un jeu de données multi-documents plus complet pour de futures recherches. Nous les filtrons à travers le filtre de QA présenté à la section 3.3, écartant environ 20% qui étaient jugées pouvant être répondues par un seul article.

4. https ://www.semanticscholar.org/product/api
5. https ://neo4j.com/

Enfin, nous obtenons 238 paires question-réponse constituant notre jeu de données préliminaire. Le tableau 4.1 présente deux exemples représentatifs de QA retenues après filtrage. Chaque exemple répertorie plusieurs articles sources de la même communauté, illustrant la nature inter-articles de la question. Nous constatons que de nombreuses questions impliquent des configurations expérimentales contrastées, une comparaison des résultats différente ou la fusion de méthodes issues de différents articles.

Exemple 1 : **Question :** How does point supervision strategy differ between infrared small target detection and crowd counting tasks in terms of network design and performance ? **Réponse :** In infrared target detection, point supervision uses single points per target and evolves them into masks through label evolution, achieving 70% IoU of fully supervised methods. For crowd counting, point supervision uses multiple points (typically 5-20) per image and employs multiscale fusion networks to handle varying crowd densities, reaching 95% performance of fully supervised approaches. The key difference lies in the supervision density - infrared targets require minimal points due to target simplicity, while crowd counting needs denser point annotations to capture crowd distribution variations. **Sources :** [2304.01484], [2202.09942]
Exemple 2 : **Question :** What are the different approaches to achieving differential privacy in language models and their relative trade-offs ? **Réponse :** There are three main approaches : 1) Local DP that anonymizes each text sample individually but requires high noise leading to incoherent language, 2) Global DP that trains language models privately on the full dataset allowing better utility but needs careful privacy accounting, and 3) Selective DP that protects only sensitive portions allowing better utility-privacy trade-offs. The choice depends on trust assumptions and utility requirements. **Sources :** [2210.13918], [2108.12944], [1712.05888]

TABLE 1 – Exemples de paires QA multi-documents provenant du jeu de données préliminaire. Chaque paire nécessite l'utilisation d'informations provenant de deux articles ou plus (cf. Sources).

4.2 Recherche basée sur des vecteurs sémantiques

Configuration Expérimentale. Pour valider que ces QA nécessitent effectivement une recherche d'information multi-documents, nous testons un modèle de recherche sémantique standard qui indexe le texte de chaque article en embeddings vectoriels. Nous appliquons le modèle d'embedding BGE (Xiao *et al.*, 2023) pour générer des représentations vectorielles à la fois pour le texte des questions et pour les paragraphes de chaque article. Nous indexons ces embeddings de paragraphes en utilisant la bibliothèque Faiss [6] pour permettre une récupération efficace basée sur la similarité. Au moment du test, un embedding de question est comparé à tous les embeddings de paragraphes dans l'index, et les top-k paragraphes sont retournés en fonction de la similarité cosinus. Nous évaluons la performance en termes de rappel@10, rappel@20 et rappel@50, où rappel@k mesure à quelle fréquence les paragraphes nécessaires apparaissent parmi les top-k résultats récupérés. Cela reflète la couverture et la capacité de rappel du système de récupération dans les scénarios multi-documents.

Résultat. Comme montré dans le tableau 2, nous observons que le rappel@10 n'est que de 0,44, et même au rappel@50, il reste en 0.63. Ces scores de rappel relativement faibles suggèrent que

6. https ://github.com/facebookresearch/faiss

Méthode	Rappel@10	Rappel@20	Rappel@50
BGE	0.44 $\pm$ 0.34	0.53 $\pm$ 0.35	0.63 $\pm$ 0.37

TABLE 2 – Performance de la récupération par vecteurs denses avec le modèle d'embedding BGE.

les méthodes de récupération basées sur les vecteurs existantes sont insuffisantes pour les questions nécessitant des preuves multi-documents. Cette performance limitée s'explique également par le fait que les questions-réponses portent sur des articles liés de manière structurelle, et non uniquement par la similarité sémantique des mots employés. En effet, le système identifie souvent un document avec une pertinence partielle mais ne parvient pas à récupérer d'autres documents contenant des résultats complémentaires ou contradictoires essentiels pour répondre pleinement à la question. Cette constatation montre que la similarité sémantique ne garantit pas à elle seule le succès pour résoudre la tâche QA multi-documents. Elle justifie ainsi le projet d'élaborer des QA plus complexes.

4.3 Discussion et Perspectives

Bien que ces premiers tests confirment la faisabilité du processus, de nombreuses questions restent ouvertes. Par exemple, certaines communautés peuvent être trop larges, entraînant des questions trop générales qui manquent de "réponses correctes et précises" bien définies, tandis que des communautés très petites ne produisent que quelques paires QA, limitant ainsi la couverture. Les travaux futurs exploreront différents algorithmes de clustering (par exemple, en définissant différentes tailles de communautés), affineront ou élargiront les prompts de génération pour des requêtes plus riches et plus diversifiées et proposerons de nouveaux algorithmes qui peuvent exploiter les informations structurelles entre les articles pour répondre aux QA multi-documents.

Nous visons à publier un jeu de données QA multi-documents à plus grande échelle dérivé de ce cadre, accompagné de résultats expérimentaux de référence, afin d'encourager une participation plus large de la communauté QA à la réponse aux besoins d'information concernant plusieurs articles liés. En effet, le projet ne se limite pas à un échantillon de 100 communautés, mais vise à traiter l'intégralité des communautés identifiées. Cette démarche permettra d'obtenir un volume plus grand de paires QA, renforçant ainsi la robustesse et la représentativité du benchmark multi-documents.

5 Conclusion

Ce travail présente un effort continu pour construire des paires question-réponse qui nécessitent véritablement des preuves provenant de plusieurs publications scientifiques. Nous nous concentrons sur un cadre flexible qui regroupe des articles connexes, génère automatiquement des questions multi-documents et applique un filtrage de qualité pour éliminer les éléments triviaux nécessitant un seul document. Notre jeu de données préliminaire montre la viabilité de cette approche et les limites de la méthode de récupération basée uniquement sur la sémantique. À l'avenir, nous prévoyons d'élargir la génération de données pour encourager le développement de systèmes QA multi-documents reflétant mieux la complexité de la littérature scientifique.

Remerciements

Nous remercions Worldline pour son soutien, qui nous a permis d'utiliser des ressources internes afin de mener à bien nos expérimentations. Ces travaux ont bénéficié d'un accès aux ressources de calcul en IA et de stockage au IDRIS au travers de l'allocation de ressources 2024-AD011014704R1 attribuée par GENCI sur la partition V100 du calculateur Jean Zay.

Références

BLONDEL V. D., GUILLAUME J.-L., LAMBIOTTE R. & LEFEBVRE E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics : theory and experiment*, **2008**(10), P10008.

CHEN D., FISCH A., WESTON J. & BORDES A. (2017). Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv :1704.00051*.

CHEN W., CHANG M.-W., SCHLINGER E., WANG W. Y. & COHEN W. W. (2020a). Open question answering over tables and text. In *International Conference on Learning Representations*.

CHEN W., ZHA H., CHEN Z., XIONG W., WANG H. & WANG W. Y. (2020b). Hybridqa : A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 1026–1036.

DASIGI P., LO K., BELTAGY I., COHAN A., SMITH N. A. & GARDNER M. (2021). A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4599–4610.

DOUZE M., GUZHVA A., DENG C., JOHNSON J., SZILVASY G., MAZARÉ P.-E., LOMELI M., HOSSEINI L. & JÉGOU H. (2024). The faiss library.

HERMANN K. M., KOCISKY T., GREFENSTETTE E., ESPEHOLT L., KAY W., SULEYMAN M. & BLUNSOM P. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems*, **28**.

JIN Q., DHINGRA B., LIU Z., COHEN W. & LU X. (2019). Pubmedqa : A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* : Association for Computational Linguistics.

KOČISKỲ T., SCHWARZ J., BLUNSOM P., DYER C., HERMANN K. M., MELIS G. & GREFENSTETTE E. (2018). The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, **6**, 317–328.

KUZMIN K., SHAH S. Y. & SZYMANSKI B. K. (2013). Parallel overlapping community detection with slpa. In *2013 International Conference on Social Computing*, p. 204–212 : IEEE.

KWIATKOWSKI T., PALOMAKI J., REDFIELD O., COLLINS M., PARIKH A., ALBERTI C., EPSTEIN D., POLOSUKHIN I., DEVLIN J., LEE K. *et al.* (2019). Natural questions : a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, **7**, 453–466.

LEE Y., LEE K., PARK S., HWANG D., KIM J., LEE H.-I. & LEE M. (2023). Qasa : advanced question answering on scientific articles. In *International Conference on Machine Learning*, p. 19036–19052 : PMLR.

MA B. (2024). Mining both commonality and specificity from multiple documents for multi-document summarization. *IEEE Access*.

MA C., ZHANG W. E., GUO M., WANG H. & SHENG Q. Z. (2022). Multi-document summarization via deep learning techniques : A survey. *ACM Computing Surveys*, **55**(5), 1–37.

PRAMANICK S., CHELLAPPA R. & VENUGOPALAN S. (2025). Spiqa : A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, **37**, 118807–118833.

QIU J.-P., DONG K. & YU H.-Q. (2014). Comparative study on structure and correlation among author co-occurrence networks in bibliometrics. *Scientometrics*, **101**, 1345–1360.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). Squad : 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 2383–2392.

TRAAG V. A., WALTMAN L. & VAN ECK N. J. (2019). From louvain to leiden : guaranteeing well-connected communities. *Scientific reports*, **9**(1), 1–12.

TSATSARONIS G., BALIKAS G., MALAKASIOTIS P., PARTALAS I., ZSCHUNKE M., ALVERS M. R., WEISSENBORN D., KRITHARA A., PETRIDIS S., POLYCHRONOPOULOS D. *et al.* (2015). An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, **16**, 1–28.

WAN X. & YANG J. (2008). Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, p. 299–306.

XIAO S., LIU Z., ZHANG P. & MUENNIGHOFF N. (2023). C-pack : Packaged resources to advance general chinese embedding.

YANG Z., QI P., ZHANG S., BENGIO Y., COHEN W. W., SALAKHUTDINOV R. & MANNING C. D. (2018). Hotpotqa : A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv :1809.09600*.

ZHU A., HWANG A., DUGAN L. & CALLISON-BURCH C. (2024). Fanoutqa : A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 18–37.