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

Restructuration de la Littérature Biomédicale dans une Architecture RAG pour la Génération de Réponse

Maël Lesavourey[1] Gilles Hubert[1]
(1) IRIT, Université de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 9, France
mael.lesavourey@irit.fr, gilles.hubert@irit.fr

RÉSUMÉ

Le Question Answering Biomédical (BQA) présente des défis spécifiques liés au vocabulaire spécialisé et aux structures sémantiques complexes de la littérature biomédicale. Les grands modèles de langage (LLMs) ont montré d'excellentes performances dans plusieurs tâches de compréhension et de génération du langage naturel. Cependant, leur efficacité tend à diminuer dans des domaines spécifiques, comme la biomédecine. Pour remédier à ce problème, les architectures de génération augmentée de récupération (RAG) sont devenues une approche prometteuse, combinant les avantages des méthodes de recherche d'information et des LLMs afin d'intégrer des connaissances spécifiques au domaine dans le processus de génération. Dans cet article, nous étudions le rôle du contexte dans l'amélioration des performances des pipelines RAG pour le BQA. Nous montrons que l'intégration d'un contexte basé sur une restructuration appropriée de la littérature influence positivement la qualité des réponses générées, en améliorant à la fois les métriques sémantiques et lexicales.

ABSTRACT

Here the title in English.

Biomedical Question Answering (BQA) poses specific challenges due to the specialized vocabulary and complex semantic structures of biomedical literature. Large Language Models (LLMs) have shown great performance in several Natural Language Understanding and Generation tasks. However, their effectiveness tends to drop in domain-specific contexts such as biomedicine. To address this issue, Retrieval-Augmented Generation (RAG) pipelines have become a promising approach, combining the strengths of retrieval methods with LLMs to incorporate domain-specific knowledge into the generation process. In this article, we investigate the role of context in enhancing the performance of RAG pipelines for BQA. We show that incorporating a context grounded on proper literature reshaping affects positively the quality of generated answers, improving both semantic and lexical metrics.

MOTS-CLÉS : Génération assistée de récupération, Réponse aux questions biomédicales, Recherche d'information, Génération de réponses.

KEYWORDS: Retrieval-Augmented Generation, Biomedical Question Answering, Information Retrieval, Answer Generation.

ARTICLE : **Accepté à** SCOLIA 2025.

Actes de la 20e Conférence en Recherche d'Information et Applications (CORIA), pages 183–200.

1 Introduction

Depuis leur apparition, les modèles de langage (LM) comme BERT (Devlin *et al.*, 2018) et GPT (Radford & Narasimhan, 2018) ont été largement adoptés pour traiter de nombreuses tâches de compréhension et de traitement automatique du langage naturel (TALN). Leur capacité à comprendre la relation sémantique entre les mots d'un document a transformé les approches traditionnelles dans divers domaines comme la recherche d'information (RI), remplaçant l'état de l'art dans de nombreuses tâches, telles que l'ordonnancement de documents, la classification et la génération de texte (Yates *et al.*, 2021; Zhu *et al.*, 2023). Cependant, ces modèles n'atteignent pas les mêmes performances lorsqu'ils sont appliqués à des corpus de domaines spécialisés, comme la littérature biomédicale et les documents juridiques (Zhang *et al.*, 2024; Chalkidis *et al.*, 2020). Les caractéristiques particulières de ces textes en sont les principales raisons. Elles amplifient l'écart sémantique entre les connaissances générales et les concepts spécialisés. La littérature biomédicale renferme des structures lexicales complexes, telles que des formules chimiques, des noms propres et des abréviations. De plus, la compréhension de cette littérature est d'autant plus difficile en raison de sa polysémie ; par exemple, les expressions « crise cardiaque », « infarctus du myocarde » et « accident cardiovasculaire » ont la même signification [1].

Relever ces défis dans le cadre des tâches de question-réponse biomédicale (« biomedical question answering » - BQA) nécessite une prise en compte attentive des caractéristiques spécifiques du domaine. Il existe une grande variété de tâches de BQA (Jin *et al.*, 2022), chacune ayant ses propres particularités en termes de contenu de corpus, de format des réponses et de public ciblé. Nous considérons la littérature scientifique comme notre source d'information, tandis que la requête et sa réponse doivent s'adresser à des lecteurs spécialisés et être rédigées en langage naturel. Cette tâche se situe à l'intersection de la RI et de la génération de texte.

Une première méthode pour prendre en compte les caractéristiques spécifiques des corpus biomédicaux a été de pré-entraîner des LMs sur ces textes (Tinn *et al.*, 2021; Yasunaga *et al.*, 2022; Kanakarajan *et al.*, 2021). Cependant, plusieurs travaux (Dong *et al.*, 2022; Tan *et al.*, 2023; Xie *et al.*, 2024) ont montré que, malgré des performances améliorées, ces modèles manquent encore de compréhension sémantique. Avec l'émergence des grands modèles de langage (« Large Language Model » - LLM), cette méthode n'est plus envisageable, car entraîner un modèle depuis zéro devient trop coûteux. Ainsi, plusieurs approches ont été proposées pour intégrer des connaissances dans les LLMs. D'une part, la génération augmentée de récupération (« Retrieval-Augmented Generation », RAG) (Lewis *et al.*, 2020) combine la génération de texte avec des mécanismes de recherche de documents pertinents afin de contextualiser les réponses. D'autre part, l'apprentissage en contexte (« In-Context Learning », ICL) (Dong *et al.*, 2024) vise à aligner les réponses générées sur les attentes de l'utilisateur en fournissant directement des exemples parmi les entrées du modèle. L'efficacité de ces approches dépend du contexte extrait et de la manière dont il est structuré, par exemple sous forme de paires (requête, réponse), de texte brut issu de publications scientifiques ou de triplets sémantiques (sujet-prédicat-objet).

Dans cet article, nous étudions comment intégrer correctement des connaissances spécifiques à un domaine dans les LLMs en s'appuyant sur une tâche BQA.

Dans la suite de cet article, nous présentons d'abord les travaux connexes aux RAG et BQA. Ensuite, nous décrivons la méthode mise en œuvre pour traiter cette tâche, suivie d'une présentation détaillée

1. https://meshb.nlm.nih.gov/record/ui?ui=D009203

des modèles et technologies utilisés pour son implémentation et son évaluation. Nous analysons et discutons ensuite les résultats avant de conclure et introduire les perspectives de recherche future.

2 Travaux Connexes

Notre travail est lié à différents domaines, à savoir la RI, les LMs et la BQA, comme introduit dans les sections suivantes.

2.1 Recherche d'Information

Les premières approches en RI reposaient sur la correspondance lexicale, utilisant des statistiques mesurant les cooccurrences de mots entre plusieurs textes (par exemple, un document et une requête). Une méthode reconnue, BM25 (Robertson *et al.*, 2009), est basée sur le score TF-IDF et exploite divers concepts tels que la fréquence des termes, leur rareté et la longueur du texte pour calculer un score de similarité. Ces méthodes reposent sur des représentations éparses du texte, où chaque document est représenté sous forme d'un vecteur de grande dimension dans un espace défini par son vocabulaire. Leur principale limite réside dans leur incapacité à prendre en compte la signification sémantique du texte (par exemple, l'utilisation de synonymes ou de reformulations).

Pour y remédier, la proposition de représentations denses (Guo *et al.*, 2022) a permis de capturer des relations sémantiques qui dépassent la simple correspondance exacte des mots (Douze *et al.*, 2024; Karpukhin *et al.*, 2020).

2.2 Modèles de Langage

L'architecture Transformer a été introduite par (Vaswani *et al.*, 2017). Elle repose sur le mécanisme d'auto-attention, qui permet de capturer à la fois les dépendances locales et globales d'une séquence de jetons (tokens). On y distingue les modèles encodeurs et décodeurs. BERT (Devlin *et al.*, 2018) est l'encodeur le plus largement étudié depuis sa sortie. Son paradigme de pré-entraînement suivi d'un ajustement (« fine-tuning ») a conduit à des améliorations significatives dans des tâches telles que la classification de texte et la reconnaissance d'entités nommées.

Dans le même temps, des décodeurs axés sur la génération de nouveaux « tokens » ont été développés. GPT-1 (Radford & Narasimhan, 2018) a démontré que l'entraînement sur de grands corpus pouvait produire un modèle génératif capable de gérer plusieurs tâches de compréhension et de génération du langage. GPT-2 (Radford *et al.*, 2019) a marqué une avancée majeure en augmentant considérablement le nombre de paramètres des LLMs ainsi que la taille du corpus d'entraînement. Sa capacité à effectuer différentes tâches a également constitué un tournant, ouvrant la voie au ICL, qui permet d'orienter le comportement des LLMs sans nécessiter de « fine-tuning ».

Plus récemment, de nouveaux LLMs, notamment GPT-3 (Brown *et al.*, 2020), LLaMA (Touvron *et al.*, 2023a,b; Dubey *et al.*, 2024) et Mixtral (Jiang *et al.*, 2024), ont repoussé les limites de ce qui peut être accompli avec les Transformers. Ils ont également mis en lumière certaines limites des LLMs, notamment en termes de biais (comme les hallucinations (Ji *et al.*, 2023)) et de contraintes computationnelles liées à leur taille.

L'architecture RAG combine les atouts des méthodes de RI et des LLMs génératifs (Lewis *et al.*, 2020), créant ainsi un pont entre la RI et la génération de texte. Dans cette approche, un module récupère des documents ou des passages pertinents en fonction d'une requête, et un modèle génératif utilise les informations récupérées pour produire une réponse contextualisée (Cuconasu *et al.*, 2024; Ram *et al.*, 2023). En créant un contexte dynamique et spécifique à la requête, le RAG permet aux LLMs de concentrer leur attention sur les informations les plus pertinentes, améliorant ainsi la précision et réduisant les hallucinations (Ayala & Bechard, 2024). Cette méthode constitue une alternative pertinente aux LLMs basés exclusivement sur des paramètres statiques.

2.3 Question Réponse Biomédicale

Au cours des douze dernières années, les campagnes d'évaluation BioASQ (Nentidis *et al.*, 2024a) ont permis le développement de diverses méthodes pour le BQA. Elles ont suivi l'évolution de la RI et de la génération de réponses. Ainsi, les premières approches se concentraient sur des techniques de résumé extractif basées sur la correspondance lexicale, telles que TF-IDF ou LexRank (Erkan & Radev, 2004). Au fil des années, des méthodes basées sur l'apprentissage supervisé et profond sont apparues, surpassant ainsi les travaux précédents.

Plus récemment, l'effort s'est focalisé sur l'apprentissage par transfert, avec des modèles pré-entraînés sur des jeux de données généraux ou biomédicaux(Lee *et al.*, 2019; Tinn *et al.*, 2021), puis ajustés sur le dataset BioASQ (Krithara *et al.*, 2023).

Avec l'émergence des LLMs, les architectures RAG ont suscité un intérêt particulier. Différentes méthodes de récupération de documents ont été proposées : éparses, denses ou hybrides (Almeida *et al.*, 2024; Ateia & Kruschwitz, 2024; Chih *et al.*, 2024; Gao *et al.*, 2024; Merker *et al.*, 2024; Panou *et al.*, 2024). La plupart d'entre elles exploitent un modèle de type BERT avec une architecture d'encodeur croisé pour réordonner les publications et créer un contexte adapté à la requête.

Concernant la génération de réponses, les approches proposées diffèrent essentiellement dans le choix des modèles et leurs paramètres (par exemple, le format des prompts, l'utilisation de ICL, l'optimisation des hyperparamètres). Pour plus de détails sur les approches utilisées dans les tâches de BQA, nous invitons le lecteur à consulter le « survey » (Jin *et al.*, 2022).

Plusieurs études sur des corpus généraux ont montré que l'organisation du contexte au niveau des documents a un impact significatif sur les performances des LLMs sur des tâches de question-réponse (Cuconasu *et al.*, 2024). Notre travail portant sur des articles scientifiques en biomédecine, donc des documents longs, nous nous intéressons à la structuration du contexte à un niveau de granularité plus fin, à savoir la phrase.

3 Méthode

Comme mentionné précédemment, ce travail vise à traiter la problématique de la structuration du contexte dans une architecture RAG appliquée au BQA. Cette section formalise le problème et détaille les étapes de la méthode que nous proposons.

3.1 Question Réponse

Cette tâche peut être définie comme une génération de réponse en fonction d'un contexte construit à partir de publications biomédicales. Soit q une question biomédicale exprimée en langage naturel, et $D = \{d_1, d_2, ..., d_n\}$ un ensemble de publications biomédicales. Le système a pour objectif de générer une réponse a, avec :

$$a = LLM(I,\, q,\, C), \tag{1}$$

où LLM est une implémentation d'un LLM, C est le contexte extrait et potentiellement restructuré à partir d'un sous-ensemble de D afin de maximiser la pertinence de a, et I représente les instructions données au modèle, c-à-d une chaîne de caractères qui contient un ensemble de directives explicites destinées à guider le comportement du modèle lors de la génération.

3.2 Module de RI

Un module de RI a pour objectif de créer un sous-ensemble $D' \subset D$ contenant les m articles les plus pertinents pour la requête q. Formellement :

$$D' = Retriever(q,\, D,\, m), \tag{2}$$

où $Retriever$ est une instance d'un module de RI qui cherche à maximiser le Rappel (Recall) parmi les m articles récupérés afin de contenir l'information recherchée.

3.3 Structuration du Contexte

Les documents de D' sont décomposés en unités textuelles de base, c'est-à-dire en phrases. L'ensemble obtenu $S = \{s_1, s_2, ..., s_k\}$ est constitué de toutes les phrases extraites de D'.

Chaque phrase s_i est encodée dans un espace vectoriel à l'aide d'un encodeur. L'embedding d'une phrase est obtenu en effectuant un $mean\_pooling$ sur l'embedding de chaque token de la phrase (moyenne de l'ensenble des vecteurs sur chacune des dimensions). Pour simplifier les notations, nous noterons simplement :

$$E = \{SEncoder(s_i) \mid s_i \in S\}, \tag{3}$$

$SEncoder(s_i)$ étant le $mean\_pooling$ appliqué aux tokens encodés de s_i .

Pour guider l'attention du LLM lors du traitement du contexte, les éléments sémantiquement proches sont regroupés. L'intuition derrière cette approche est qu'un contexte structuré aidera le modèle à « comprendre » l'information fournie en entrée, plutôt que d'avoir des informations dispersées dans S. Les plongements (embeddings) dans E sont regroupés (en clusters) de façon disjointe en utilisant la similarité cosinus, $E = \{E_1, E_2, ..., E_t\}$, où E_i désigne un groupe de plongements. Notons $T = \{T_1, T_2, ..., T_t\}$ les groupes correspondants de phrases, où chaque T_i est un groupe de phrases traitant d'un même « sujet » :

$$T_i = \{s \in S \mid \forall e_1,\, e_2 \in E_i,\, cos\_sim(e_1,\, e_2) \geq seuil\} \tag{4}$$

Pour chaque groupe T_i, un algorithme de classement est appliqué afin d'identifier les phrases informatives, notées $T'_i \subset T_i$, et :

$$T'_i = sentenceRank(T_i,\, l), \tag{5}$$

où $sentenceRank$ est une implémentation d'une méthode d'ordonnancement et l représente le nombre de phrases sélectionnées.

Pour créer notre contexte final C, les groupes de $T' = \{T'_1, T'_2, ..., T'_t\}$ sont classés en fonction de leur pertinence par rapport à la requête q. Pour un cluster T'_i, un encodeur croisé produit une probabilité p_i indiquant sa pertinence vis-à-vis de q :

$$p_i = cross\_encoder(q,\, T'_i) \tag{6}$$

Les groupes les plus pertinents sont ensuite ordonnés pour construire C :

$$C = \{T'_{i_1}, T'_{i_2}, ..., T'_{i_c} \mid p_{i_1} \geq p_{i_2} \geq ... \geq p_{i_c}\}, \tag{7}$$

où $c < t$ est le nombre de groupes sélectionnés.

3.4 Génération de Réponse

Le contexte C, combiné avec les instructions I, et la requête q sont donnés en entrée d'un LLM pour générer une réponse a :

$$a = LLM(I,\, q,\, C) \tag{8}$$

Il est possible d'enrichir I en fournissant des exemples de paires (question, réponse). Ce mécanisme, le ICL, permet au LLM d'apprendre une tâche spécifique à partir d'exemples fournis dans I, sans mise à jour explicite de ses poids.

Cette méthode a pour but de générer des réponses hautement contextualisées et pertinentes pour q en exploitant des documents spécialisés, tout en minimisant le bruit et les informations non pertinentes.

4 Expérimentations

Dans cette section, nous présentons les jeux de données et les métriques utilisés pour mener nos évaluations. Nous détaillons également les paramètres d'implémentation pour le module de RI, la sélection de phrases, le classement des sujets et la génération de réponses.

4.1 Jeux de Données

Comme le montre (Jin *et al.*, 2022), il existe peu de jeux de données abordant directement la tâche spécifique que nous traitons. Nous avons choisi de travailler sur le jeu de données BioASQ-TaskB (Nentidis *et al.*, 2024b), car il correspond à nos spécifications telles qu'explicitées en introduction.

BioASQ-TaskB se compose de deux phases. La Phase A vise à récupérer les 10 publications les plus pertinentes pour une requête donnée (à partir de la base de données biomédicale PubMed[2]) et à en extraire les passages pertinents. La Phase B se concentre sur l'extraction et la génération de réponses en proposant une « réponse exacte » et une « réponse idéale ».

Les « réponses exactes » suivent un format spécifique selon le type de question (« Oui/Non », « Résumé », « Factuel », « Liste »). Les « réponses idéales » sont des textes en langage naturel rédigés comme le ferait un expert biomédical. Pour générer ces réponses, les équipes participantes disposent des vérités terrain de la Phase A, c'est-à-dire des articles pertinents et des extraits correspondants. Depuis BioASQ 12, la Phase A+ a été introduite. Son objectif est identique à celui de la Phase B, mais sans disposer des vérités terrain de la Phase A. Toutes les données de BioASQ-TaskB sont annotées manuellement par des experts biomédicaux, fournissant ainsi des références de qualité pour diverses tâches de TALN biomédical.

Nous avons isolé les requêtes et leurs « réponses idéales » issues des campagnes BioASQ 11 et 12, ce qui nous a permis d'évaluer notre travail sur deux collections distinctes composées respectivement de 327 et 340 requêtes biomédicales.

4.2 Métriques

L'équipe organisatrice de BioASQ propose une évaluation manuelle des réponses générées par les systèmes participants. Chaque annotateur attribue une note sur 5 selon les critères de précision, rappel, lisibilité et répétition. Les scores ROUGE2 et ROUGE-SU4 (Rappel, F1) (Lin, 2004) sont également fournis. Cependant, les scores manuels ne sont calculés que pendant la durée de la campagne d'évaluation. Notre étude étant menée a posteriori, nous n'avons pas pu y avoir recours.

Afin d'évaluer notre travail et de comparer les performances de nos modèles aux méthodes proposées lors de la campagne d'évaluation, nous avons choisi d'utiliser les métriques ROUGE2 Rappel, Précision et F1, notées respectivement R2-R, R2-P et R2-F1.

Toutefois, ces métriques lexicales présentent une limite intrinsèque lorsqu'elles sont appliquées à des tâches de génération de texte. En effet, ROUGE2 évalue le chevauchement des bi-grammes entre un texte de référence et une réponse candidate. Une réponse sémantiquement identique à la référence, mais utilisant des synonymes, obtiendra un score très bas malgré une réponse correcte.

Pour évaluer nos modèles de manière plus fine, nous avons donc utilisé une métrique basée sur la similarité sémantique, à savoir BERTScore (Zhang *et al.*, 2019) Rappel, Précision et F1, notées respectivement BERT-R, BERT-P et BERT-F1. D'une part, cela nous permet de situer les performances de nos approches avec les métriques R2, et d'autre part, d'obtenir une évaluation plus précise grâce aux similarités sémantiques.

4.3 Module de RI

Nous avons construit un module de RI en utilisant Pyserini, une bibliothèque Python open-source dérivée d'Anserini, intégrant plusieurs techniques de RI.

Dans un premier temps, nous avons indexé l'ensemble des citations de MEDLINE, à l'exception

2. https ://pubmed.ncbi.nlm.nih.gov/

de celles dont le résumé était indisponible ($n \approx 25$ millions de citations). Pour chaque requête, nous avons récupéré $m = 1000$ articles les plus pertinents pour y répondre. Cette approche suit les observations de Almeida *et al.* (2023), qui montrent que cette architecture atteint un R2-R@1000 supérieur à 90 %. Compte tenu des économies en ressources et en temps de calcul, nous estimons que cette solution est suffisante.

4.4 Sélection des Phrases

Une fois les publications susceptibles de contenir le contexte nécessaire pour répondre à la requête identifiées, l'étape suivante consiste à sélectionner les informations pertinentes parmi toutes les publications. Nous avons choisi de travailler au niveau de la phrase afin d'incorporer les connaissances en lien avec la requête.

Nous avons opté pour le calcul d'un plongement de chaque phrase à l'aide d'un modèle basé sur un encodeur, permettant ainsi une comparaison sémantique entre elles. Pour ce faire, nous avons utilisé la bibliothèque SentenceTransformer (Reimers & Gurevych, 2019) ainsi que BioLinkBERT-large (Yasunaga *et al.*, 2022) pour produire ces plongements. BioLinkBERT est une version de LinkBERT pré-entraînée sur des corpus biomédicaux et obtenant les meilleures performances globales sur le benchmark BLURB (Gu *et al.*, 2021).

Nous avons décidé de regrouper les phrases par sujet en appliquant une méthode de groupement (clustering) sur leurs plongements. Étant donné que plusieurs milliers de phrases devaient être comparées, nous avons utilisé l'algorithme *community_detection* implémenté dans SentenceTransformer, car il est conçu pour gérer un grand nombre de phrases. Cet algorithme calcule la similarité cosinus entre les plongements pour déterminer les groupes et intègre plusieurs optimisations pour traiter de vastes collections ($seuil = 92\%$).

Après avoir regroupé les phrases sémantiquement, il est nécessaire d'identifier celles qui composeront le contexte pour chaque thématique. Pour cela, nous avons implémenté une version de l'algorithme de TextRank (Mihalcea & Tarau, 2004) et l'avons appliqué à chaque groupe pour identifier leurs phrases importantes ($l = 4$, $l = 10$, ou $l = 15$ phrases par sujet).

4.5 Ordonnancement

Afin d'obtenir le contexte le plus précis possible, il peut être bénéfique de sélectionner les thématiques pertinentes et, éventuellement, de supprimer celles qui ne le sont pas.

En nous appuyant sur nos travaux précédents sur l'ordonnancement de documents dans une architecture de récupération multi-étapes (Lesavourey & Hubert, 2024), nous établissons une analogie entre le classement de publications scientifiques et le classement des thématiques. La première tâche vise à classer les documents par ordre de pertinence par rapport à une requête. Nous avons démontré que modifier la granularité de ces documents et sélectionner les phrases pertinentes plutôt que de considérer l'intégralité du document est bénéfique. Le classement des thématiques suit une logique similaire, à la différence près qu'il s'agit ici de groupes (clusters) composés de phrases sémantiquement proches.

Nous avons appliqué un encodeur croisé BioLinkBERT, affiné sur le jeu de données BioASQ-TaskB. Ce modèle calcule une probabilité de pertinence qui est utilisée pour classer les thématiques. Une fois la liste des sujets ordonnée, nous en avons choisi un nombre fixe à utiliser comme contexte

pour les requêtes (c'est-à-dire les $c = 5$, $c = 10$ ou $c = 15$ premières thématiques en fonction des expérimentations).

4.6 Génération de Réponse

Nous avons vu dans les sections précédentes comment établir un contexte pour répondre aux questions biomédicales. Pour générer une réponse à partir d'une question et de son contexte, nous avons développé un outil de génération de réponses basé sur LLaMA. Nous avons utilisé la troisième version de ce modèle dans sa version à 8 milliards de paramètres [3], désignée sous le nom de *llama3.1-8B* dans la suite de cet article. L'architecture de ce modèle, dont les poids sont ouverts, est optimisée pour offrir des performances élevées sur les tâches de question-réponse. De plus, son tokenizer performant permet de traiter un grand nombre de tokens en entrée, ce qui est particulièrement important pour le ICL.

Nous avons également choisi ce modèle afin de réduire les coûts de calcul et la consommation énergétique par rapport aux LLM dotés d'un plus grand nombre de paramètres. Pour réduire davantage ces coûts, nous avons appliqué une quantification du modèle et utilisé une précision de 4-bits pour la représentation des nombres flottants, au lieu de 32-bits.

Les prompts que nous avons utilisés pour paramétrer le modèle sont présentés dans les Figures 1, 2 et 3. Nous avons également mené une expérience sans contexte afin d'évaluer l'apport de l'ajout du contexte.

Prompt Système

You are a biomedical expert providing answers. I will give a question and several context texts about the question. Based on the context, give a short answer to the question.

Prompt Utilisateur

QUESTION : **Une question biomédicale**
CONTEXTS : **Phrases extraites de la littérature**
ANSWER :

FIGURE 1 – Spécifications du Prompt avec Incorporation de Contexte

Prompt Système

You are a biomedical expert providing answers. I will ask a question and your role is to give a short answer to the question.

Prompt Utilisateur

QUESTION : **Une question biomédicale**
ANSWER :

FIGURE 2 – Spécifications du Prompt sans Incorporation de Contexte

3. https ://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Prompt Système

You are a biomedical expert providing answers. I will give a question and several context texts about the question. Based on the context, give a short answer to the question. Moreover I will give you 3 questions and their corresponding answers as examples.

Prompt Utilisateur

EXAMPLES : *Un ensemble de 3 questions/réponses dépendant du type de question**
QUESTION : **Une question biomédicale**
CONTEXTS : **Phrases extraites de la littérature**
ANSWER :

FIGURE 3 – Spécifications du Prompt avec ICL et Incorporation de Contexte

5 Résultats

Dans cette section, nous présentons les résultats expérimentaux obtenus en appliquant notre approche sur les deux jeux de données décrits dans la Section 4.1. Nous avons évalué ses performances en analysant l'impact de l'intégration et de la restructuration du contexte. Ensuite, nous avons testé l'effet du ICL en ajoutant des exemples de paires (requête, réponse).

5.1 Influence du Contexte

L'objectif de ces premières expériences est de montrer l'effet des différents types de restructuration du contexte. Nous évaluons si un contexte textuel réduit est suffisant pour que le LLM obtienne de bonnes performances ou si chaque information doit être répétée afin d'être prise en compte.

Nous avons développé trois variantes du modèle afin d'établir des bases de comparaison. Tout d'abord, nous avons généré des réponses en utilisant *llama3.1-8B* sans intégrer aucun contexte. Ensuite, nous avons utilisé *llama3.1-8B* sur le même jeu de données, mais en incorporant un contexte en sélectionnant $l = 4$ phrases par groupe, sans appliquer de classement des sujets. Enfin, nous avons extrait ce que nous appelons le « Contexte Exact », qui correspond aux extraits pertinents fournis dans le jeu de données BioASQ. Dans un scénario réel, une telle information n'est pas disponible, et cette variante nous permet uniquement d'estimer les scores maximaux atteignables avec cette configuration de modèle.

Les scores obtenus par ces trois variantes sont présentés dans le Tableau 1. Nous observons que le système de base (sans contexte) obtient des performances relativement bonnes en termes de Rappel, mais très faibles en termes de Précision, qu'elle soit sémantique (BERT-P) ou lexicale (R2-P). L'intégration d'un contexte sans classement est indéniablement bénéfique, car elle améliore les performances du système de base. Cependant, les améliorations les plus importantes sont observées pour les métriques sémantiques, donc, bien que le LLM puisse exploiter le contexte, il ne s'aligne pas complètement avec le vocabulaire utilisé par les annotateurs.

TABLE 1 – Baselines

Context	R2-R	R2-P	R2-F1	BERT-R	BERT-P	BERT-F1
None	28.86	14.50	16.50	58.52	31.50	35.50
Unranked Topics	30.60	14.78	16.73	69.90	60.77	64.40
Contexte Exact	37.02	26.76	26.70	73.43	67.68	69.70

5.2 Influence de la Structuration du Contexte

Nous avons étudié l'impact de la structuration du contexte. Pour ce faire, nous avons ajouté le module de classement des groupes à nos expériences précédentes et généré des réponses en faisant varier les paramètres définissant le format du contexte, c'est-à-dire le nombre de groupes sélectionnés et le nombre de phrases par groupe. Étant donné que chaque groupe est associé à un sujet présent dans le corpus, l'objectif ici est de déterminer la taille de contexte nécessaire pour générer des réponses précises et d'évaluer si le LLM a besoin d'informations répétées pour les traiter efficacement.

Le Tableau 2 présente les scores obtenus pour ces expériences. Tout d'abord, nous observons qu'avec $l = 4$ phrases par sujet, comme dans les expériences précédentes, la sélection des clusters tend à diminuer les scores de Rappel, tant lexical que sémantique. Ce résultat était attendu, car nous avons intentionnellement limité la quantité d'information récupérée. Cependant, cette perte est compensée par un gain en Précision lorsque $c = 10$ clusters sont sélectionnés, comme l'indiquent les améliorations des scores F1. Il semble que choisir trop ou trop peu de sujets diminue les performances : un nombre insuffisant de groupes entraîne une perte d'information, tandis qu'un trop grand nombre introduit du bruit. Cette observation est cohérente avec le fait que le modèle de classement (encodeur croisé BioLinkBERT) est optimisé pour retourner une liste de 10 documents pertinents.

Ensuite, nous avons étudié l'effet de l'augmentation du nombre de phrases par sujet en fixant le nombre de groupes. Nous avons généré des réponses avec $c = 5$ et $c = 10$ groupes, et pour chaque configuration, nous avons réalisé l'expérience avec $l = 4$, $l = 10$ et $l = 15$ phrases. Nous observons que chaque configuration obtient de meilleurs scores à mesure que le nombre de phrases augmente. Dans ce cas, les sujets moins pertinents sont éloignés de la requête (en termes de distance de tokens dans la séquence) sans être supprimés. Cela permet de donner plus de poids aux sujets les plus pertinents. Il semble donc judicieux d'aider le LLM à focaliser son attention sur les informations les plus importantes sans pour autant supprimer celles qui le sont moins.

Les meilleurs scores de cette série d'expériences sont obtenus en utilisant les paramètres ayant conduit aux meilleures performances pour chaque étude, c'est-à-dire en considérant $c = 10$ groupes et $l = 15$ phrases par groupe. De plus, nous avons réalisé un *t-test* entre cette variante et les résultats obtenus par la ligne de base intitulée « Unranked Topics » présentée dans la section précédente. Les *valeurs-p* obtenues sont inférieures à $0,05$ pour l'ensemble des métriques, ce qui indique que toutes les améliorations sont significatives.

5.3 Influence du ICL

Nous avons décidé de compléter l'incorporation du contexte structuré en le combinant avec du ICL. Pour chaque type de question, nous avons extrait aléatoirement 3 exemples de paires (question,

TABLE 2 – Effet de la structuration du contexte sur la génération de réponse

#sentences/cluster	#clusters	R2-R	R2-P	R2-F1	BERT-R	BERT-P	BERT-F1
4	5	29.27	15.61	17.04	68.84	60.97	64.09
4	10	30.36	16.69	18.13	69.48	61.67	64.77
4	15	29.59	15.31	17.02	69.35	60.31	63.95
10	5	29.95	15.88	17.43	69.23	61.42	64.46
15	5	31.5	17.05	18.38	69.56	61.78	64.73
10	10	31.22	**17.56**	**18.77**	70.00	61.87	65.02
15	10	**31.52**	17.34	18.74	**70.43**	**62.43**	**65.54**

réponse) issues du dataset BioASQ10. Le ICL est censé aider le LLM à mieux comprendre comment structurer ses réponses et potentiellement à s'aligner sur le vocabulaire utilisé par les annotateurs. Les Tables 3 et 4 présentent les scores obtenus sur les datasets BioASQ11 et BioASQ12, respectivement. Nous avons mené des expériences en faisant varier les mêmes paramètres que dans la section précédente et en utilisant le prompt illustré dans la Figure 3.

Tout d'abord, nous observons que, pour des paramètres identiques, l'ajout du ICL diminue systématiquement les performances sur les deux mesures de Rappel : en moyenne -5,17% sur R2-R et -0,92% sur BERT-R. Cette légère perte est largement compensée par des gains plus importants en Précision et en F1-score : en moyenne +22,34% sur R2-P et +4,32% sur BERT-P. La Table 5 indique le nombre moyen de tokens dans les réponses de référence et dans celles générées avec $l = 10$ phrases et $c = 10$ clusters. Les réponses générées sont beaucoup plus longues que la vérité terrain, et le ICL tend à réduire leur longueur. Cela conduit à récupérer légèrement moins d'informations, mais celles retournées sont bien plus précises. Ce phénomène est observable sur les deux jeux de données. Les scores obtenus sur les métriques lexicales sont plus faibles sur le dataset BioASQ12, ce qui peut s'expliquer par le fait qu'un nouvel annotateur a été impliqué dans sa création. Par conséquent, le modèle n'a pas d'éléments préalables sur le vocabulaire utilisé par cet annotateur. Nous avons effectué un *t-test* entre la meilleure variante du Tableau 3 et notre référence « Unranked Topics ». Nous avons constaté que la *valeur-p* associée à BERT-R était supérieure à 0,05, ce qui indique que la perte n'est pas significative. Toutes les autres *valeurs-p* étaient inférieures à 0,05. Les gains observés sur les métriques de précision sont significatifs, mais la perte sur R2-R l'est également.

TABLE 3 – Effet du ICL sur BioASQ11

#sentences/cluster	#clusters	R2-R	R2-P	R2-F1	BERT-R	BERT-P	BERT-F1
4	5	28.31	18.68	18.63	68.28	63.16	65.00
4	10	27.73	19.58	19.05	68.48	63.85	65.35
10	5	28.74	19.23	19.13	69.00	63.94	65.63
10	10	**29.75**	**23.05**	**21.28**	**69.22**	**65.61**	**66.58**

Nous avons comparé nos résultats avec d'autres systèmes soumis dans la Phase A+ du challenge BioASQ12[4]. Les meilleures soumissions en termes de R2-R (de 32,01 à 38,68 selon le batch) ont

4. https://participants-area.bioasq.org/results/12b/phaseAplus/

TABLE 4 – Effet du ICL sur BioASQ12

#sentences/cluster	#clusters	R2-R	R2-P	R2-F1	BERT-R	BERT-P	BERT-F1
4	5	26.91	17.70	18.10	68.44	62.33	64.69
4	10	27.03	18.87	18.77	68.66	63.53	65.37
10	5	27.64	**19.75**	18.38	69.56	61.78	64.73
10	10	**28.60**	19.67	**19.61**	**70.10**	**64.41**	**66.50**

TABLE 5 – Nombre moyen de tokens

Origine de la réponse	BioASQ11	BioASQ12
Sans ICL	97.0	110.5
Avec ICL	74.6	80.2
Vérité Terrain	42.1	50.3

une Précision significativement plus faible (R2-F1 variant de 12,44 à 19,23) que notre meilleur système. Cela indique que notre compromis entre Rappel et Précision est plus efficace. Par ailleurs, les systèmes obtenant les meilleurs R2-F1 (de 25,03 à 28,62) présentent un meilleur équilibre mais affichent des scores de Rappel inférieurs à ceux obtenus par notre meilleur modèle (R2-R variant de 22,62 à 27,23).

Étant donné que les meilleures soumissions ont utilisé des modèles avec un nombre de paramètres bien plus élevé (ex. GPT-3.5, GPT-3.5 Turbo, GPT-4), ont employé des techniques de fine-tuning et ont potentiellement optimisé leurs résultats en fonction des métriques (ex. génération de réponses plus longues pour améliorer le Rappel, usage d'un module de traduction pour maximiser le chevauchement des bi-grammes), nous pouvons conclure que notre approche est à la fois pertinente et efficace.

6 Conclusion

Dans cet article, nous avons présenté plusieurs approches visant à intégrer des connaissances biomédicales dans un LLM pour améliorer les réponses générées dans le cadre d'une tâche de BQA. Nous avons montré que la génération de réponses contextualisées influence davantage la Précision que le Rappel. De plus, les améliorations sur les métriques sémantiques sont plus importantes que sur les métriques lexicales, ce qui signifie que les réponses générées ne s'alignent pas facilement avec un vocabulaire donné.

Nous avons proposé de structurer les phrases des articles pertinents en groupes sémantiques. Le classement de ces groupes améliore les scores sous certaines conditions. Il est crucial d'en sélectionner un nombre suffisant pour capturer les informations pertinentes, mais en récupérer trop peut introduire du bruit et dégrader les performances. De plus, il semble bénéfique d'augmenter le nombre de phrases par groupe. Cela permet au LLM de concentrer son attention sur les informations pertinentes en éloignant les moins pertinentes, sans pour autant les supprimer.

Nous avons ensuite étudié si l'incorporation d'exemples de paires (question, réponse) était bénéfique. Nous montrons que l'intégration du ICL dans une architecture RAG, malgré une légère perte en Rappel, permet d'obtenir des améliorations majeures en termes de Précision et de F1-score. La comparaison avec certains des meilleurs modèles du challenge BioASQ montre que notre approche atteint des résultats compétitifs.

Les travaux futurs seront consacrés à l'exploration d'autres méthodes d'intégration du contexte pour la génération de réponses, par exemple en utilisant des bases de connaissances biomédicales pour structurer l'information sous forme de triplets sémantiques (sujet-prédicat-objet) (Agrawal *et al.*, 2024; Kilicoglu *et al.*, 2012). De plus, des recherches approfondies sur l'optimisation de la sélection du contexte pourraient améliorer à la fois la qualité et la lisibilité des réponses dans des applications biomédicales réelles. Enfin, il serait pertinent de citer les articles constituant le contexte dans les réponses générées afin que les lecteurs puissent facilement valider les informations qui lui sont retournées.

Références

AGRAWAL G., KUMARAGE T., ALGHAMDI Z. & LIU H. (2024). Can knowledge graphs reduce hallucinations in LLMs ? : A survey. In K. DUH, H. GOMEZ & S. BETHARD, Éds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 3947–3960, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.naacl-long.219.

ALMEIDA T., JONKER R. A., REIS J., ALMEIDA J. R. & MATOS S. (2024). Bit. ua at bioasq 12 : From retrieval to answer generation. *CLEF Working Notes*.

ALMEIDA T., JONKER R. A. A., POUDEL R., SILVA J. M. & MATOS S. (2023). Bit. ua at bioasq 11b : Two-stage ir with synthetic training and zero-shot answer generation. In *CLEF (Working Notes)*, p. 37–59.

ATEIA S. & KRUSCHWITZ U. (2024). Can open-source llms compete with commercial models ? exploring the few-shot performance of current GPT models in biomedical tasks. In G. FAGGIOLI, N. FERRO, P. GALUSCÁKOVÁ & A. G. S. DE HERRERA, Éds., *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 de *CEUR Workshop Proceedings*, p. 78–98 : CEUR-WS.org.

AYALA O. & BECHARD P. (2024). Reducing hallucination in structured outputs via retrieval-augmented generation. In Y. YANG, A. DAVANI, A. SIL & A. KUMAR, Éds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 6 : Industry Track)*, p. 228–238, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.naacl-industry.19.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

CHALKIDIS I., FERGADIOTIS M., MALAKASIOTIS P., ALETRAS N. & ANDROUTSOPOULOS I. (2020). LEGAL-BERT : The muppets straight out of law school. In T. COHN, Y. HE & Y. LIU, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 2898–2904, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.261.

CHIH B., HAN J. & TSAI R. T. (2024). NCU-IISR : enhancing biomedical question answering with GPT-4 and retrieval augmented generation in bioasq 12b phase B. In G. FAGGIOLI, N. FERRO, P. GALUSCÁKOVÁ & A. G. S. DE HERRERA, Éds., *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 de *CEUR Workshop Proceedings*, p. 99–105 : CEUR-WS.org.

CUCONASU F., TRAPPOLINI G., SICILIANO F., FILICE S., CAMPAGNANO C., MAAREK Y., TONELLOTTO N. & SILVESTRI F. (2024). The power of noise : Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, p. 719–729, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3626772.3657834.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

DONG Q., LI L., DAI D., ZHENG C., MA J., LI R., XIA H., XU J., WU Z., CHANG B., SUN X., LI L. & SUI Z. (2024). A survey on in-context learning. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 1107–1128, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.64.

DONG Q., LIU Y., CHENG S., WANG S., CHENG Z., NIU S. & YIN D. (2022). Incorporating explicit knowledge in pre-trained language models for passage re-ranking. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, p. 1490–1501, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3477495.3531997.

DOUZE M., GUZHVA A., DENG C., JOHNSON J., SZILVASY G., MAZARÉ P.-E., LOMELI M., HOSSEINI L. & JÉGOU H. (2024). The faiss library. *arXiv preprint arXiv :2401.08281*.

DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AL-DAHLE A., LETMAN A., MATHUR A., SCHELTEN A., YANG A., FAN A. *et al.* (2024). The llama 3 herd of models. *arXiv preprint arXiv :2407.21783*.

ERKAN G. & RADEV D. R. (2004). Lexrank : graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, **22**(1), 457–479.

GAO Y., ZONG L. & LI Y. (2024). Enhancing biomedical question answering with parameter-efficient fine-tuning and hierarchical retrieval augmented generation. In *CLEF (Working Notes)*, p. 117–129.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, **3**(1). DOI : 10.1145/3458754.

GUO J., CAI Y., FAN Y., SUN F., ZHANG R. & CHENG X. (2022). Semantic models for the first-stage retrieval : A comprehensive review. *ACM Trans. Inf. Syst.*, **40**(4). DOI : 10.1145/3486250.

JI Z., LEE N., FRIESKE R., YU T., SU D., XU Y., ISHII E., BANG Y. J., MADOTTO A. & FUNG P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, **55**(12), 1–38.

JIANG A. Q., SABLAYROLLES A., ROUX A., MENSCH A., SAVARY B., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., HANNA E. B., BRESSAND F. *et al.* (2024). Mixtral of experts. *arXiv preprint arXiv :2401.04088*.

JIN Q., YUAN Z., XIONG G., YU Q., YING H., TAN C., CHEN M., HUANG S., LIU X. & YU S. (2022). Biomedical question answering : a survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, **55**(2), 1–36.

KANAKARAJAN K. R., KUNDUMANI B. & SANKARASUBBU M. (2021). BioELECTRA :pretrained biomedical text encoder using discriminators. In D. DEMNER-FUSHMAN, K. B. COHEN, S. ANANIADOU & J. TSUJII, Éds., *Proceedings of the 20th Workshop on Biomedical Language Processing*, p. 143–154, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.bionlp-1.16.

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 B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *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.

KILICOGLU H., SHIN D., FISZMAN M., ROSEMBLAT G. & RINDFLESCH T. C. (2012). Semmeddb : a pubmed-scale repository of biomedical semantic predications. *Bioinformatics*, **28**(23), 3158–3160.

KRITHARA A., NENTIDIS A., BOUGIATIOTIS K. & PALIOURAS G. (2023). BioASQ-QA : A manually curated corpus for Biomedical Question Answering. *Scientific Data*, **10**(1), 170.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2019). Biobert : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240. DOI : 10.1093/bioinformatics/btz682.

LESAVOUREY M. & HUBERT G. (2024). Enhancing Biomedical Document Ranking with Domain Knowledge Incorporation in a Multi-Stage Retrieval Approach. In *12th BioASQ Workshop at CLEF 2024*, volume 3740, Grenoble, France. HAL : hal-04744454.

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.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

MERKER J. H., BONDARENKO A., HAGEN M. & VIEHWEGER A. (2024). Mibi at bioasq 2024 : retrieval-augmented generation for answering biomedical questions. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France*, volume 3740, p. 176–187.

MIHALCEA R. & TARAU P. (2004). TextRank : Bringing order into text. In D. LIN & D. WU, Éds., *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, p. 404–411, Barcelona, Spain : Association for Computational Linguistics.

NENTIDIS A., KATSIMPRAS G., KRITHARA A., LIMA-LÓPEZ S., FARRÉ-MADUELL E., KRALLINGER M., LOUKACHEVITCH N., DAVYDOVA V., TUTUBALINA E. & PALIOURAS G. (2024a). Overview of BioASQ 2024 : The twelfth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In L. GOEURIOT, P. MULHEM, G. QUÉNOT, D. SCHWAB, L. SOULIER, G. MARIA DI NUNZIO, P. GALUŠČÁKOVÁ, A. GARCÍA SECO DE HERRERA, G. FAGGIOLI & N. FERRO, Éds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*.

NENTIDIS A., KATSIMPRAS G., KRITHARA A. & PALIOURAS G. (2024b). Overview of BioASQ Tasks 12b and Synergy12 in CLEF2024. In G. FAGGIOLI, N. FERRO, P. GALUŠČÁKOVÁ & A. GARCÍA SECO DE HERRERA, Éds., *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*.

PANOU D. N., DIMOPOULOS A. C. & RECZKO M. (2024). Farming open llms for biomedical question answering. In G. FAGGIOLI, N. FERRO, P. GALUSCÁKOVÁ & A. G. S. DE HERRERA, Éds., *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 de *CEUR Workshop Proceedings*, p. 188–196 : CEUR-WS.org.

RADFORD A. & NARASIMHAN K. (2018). Improving language understanding by generative pre-training.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D., SUTSKEVER I. *et al.* (2019). Language models are unsupervised multitask learners. *OpenAI blog*, **1**(8), 9.

RAM O., LEVINE Y., DALMEDIGOS I., MUHLGAY D., SHASHUA A., LEYTON-BROWN K. & SHOHAM Y. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, **11**, 1316–1331.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence embeddings using Siamese BERT-networks. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3982–3992, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

ROBERTSON S., ZARAGOZA H. *et al.* (2009). The probabilistic relevance framework : Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, **3**(4), 333–389.

TAN J., HU J. & DONG S. (2023). Incorporating entity-level knowledge in pretrained language model for biomedical dense retrieval. *Computers in Biology and Medicine*, **166**, 107535.

TINN R., CHENG H., GU Y., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2021). Fine-tuning large neural language models for biomedical natural language processing. *CoRR*, **abs/2112.07869**.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F. *et al.* (2023a). Llama : Open and efficient foundation language models. *arXiv preprint arXiv :2302.13971*.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S. *et al.* (2023b). Llama 2 : Open foundation and fine-tuned chat models. *arXiv preprint arXiv :2307.09288*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

XIE Q., TIWARI P. & ANANIADOU S. (2024). Knowledge-enhanced graph topic transformer for explainable biomedical text summarization. *IEEE Journal of Biomedical and Health Informatics*, **28**(4), 1836–1847. DOI : 10.1109/JBHI.2023.3308064.

YASUNAGA M., LESKOVEC J. & LIANG P. (2022). Linkbert : Pretraining language models with document links.

YATES A., NOGUEIRA R. & LIN J. (2021). Pretrained transformers for text ranking : BERT and beyond. In G. KONDRAK, K. BONTCHEVA & D. GILLICK, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies : Tutorials*, p. 1–4, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-tutorials.1.

ZHANG Q., DING K., LV T., WANG X., YIN Q., ZHANG Y., YU J., WANG Y., LI X., XIANG Z. *et al.* (2024). Scientific large language models : A survey on biological & chemical domains. *ACM Computing Surveys*.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2019). Bertscore : Evaluating text generation with bert. *arXiv preprint arXiv :1904.09675*.

ZHU Y., YUAN H., WANG S., LIU J., LIU W., DENG C., DOU Z. & RONG WEN J. (2023). Large language models for information retrieval : A survey. *ArXiv*, **abs/2308.07107**.