Analyse de la littérature sur les stratégies d'augmentation de données dans des contextes à faible ressources

Benedictus Kent Rachmat[1,2]
(1) LISN, CNRS, Université Paris-Saclay, F-91405 Orsay, France
(2) AI-LAB, SLB, F-92140 Clamart, France
rachmat@lisn.fr

RÉSUMÉ

Les grands modèles de langage (LLMs) ont révolutionné le traitement automatique des langues (TAL), mais leur succès demeure largement limité aux domaines généralistes disposant de ressources abondantes. En revanche, l'application des LLMs à des domaines spécialisés à faibles ressources soulève des défis majeurs liés à la rareté des données d'entraînement, à la dérive de domaine et aux contraintes terminologiques strictes. Cette revue propose un état de l'art des approches actuelles pour le question-réponse (QA) en contexte spécialisé et à faibles ressources avec les LLMs. Nous commençons par analyser la couverture et la représentativité des jeux de données de QA spécialisés en les comparant à de grands ensembles de référence, que nous appelons *ParentQA*. Sur la base de cette analyse, nous passons en revue les stratégies centrées sur les données visant à accroître la diversité des entrées, notamment à travers des techniques d'augmentation. Nous abordons également les métriques d'évaluation adaptées aux tâches spécialisées et les considérations éthiques associées. En cartographiant les méthodologies existantes et en identifiant les questions de recherche ouvertes, cette étude vise à orienter les futurs travaux sur l'adaptation des LLMs pour une utilisation robuste et responsable dans des environnements contraints en ressources et spécifiques à un domaine.

ABSTRACT

Literature review on data augmentation strategies in low-resource settings

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP), but their success remains largely confined to high-resource, general-purpose domains. In contrast, applying LLMs to low-resource domains poses significant challenges due to limited training data, domain drift, and strict terminology constraints. This survey provides an overview of the current landscape in domain-specific, low-resource QA with LLMs. We begin by analyzing the coverage and representativeness of specialized-domain QA datasets against large-scale reference datasets what we refer to as *ParentQA*. Building on this analysis, we survey data-centric strategies to enhance input diversity, including data augmentation techniques. We further discuss evaluation metrics for specialized tasks and consider ethical concerns. By mapping current methodologies and outlining open research questions, this survey aims to guide future efforts in adapting LLMs for robust and responsible use in resource-constrained, domain-specific environments.

MOTS-CLÉS : Grands modèles de langue, Faible ressources, Augmentation de Données, Adaptation au domaine, Méthodes d'évaluation.

KEYWORDS: LLMs, Low-Resource, Data Augmentation, Domain Adaptation, Evaluation Metrics.

1 Introduction

Au fil des années, les grands modèles de langue (LLM) (OpenAI *et al.*, 2023; Gemini *et al.*, 2024; DeepSeek-AI *et al.*, 2025) ont démontré des performances remarquables dans diverses tâches de traitement du langage naturel (TAL). Cependant, ces avancées restent essentiellement limitées aux domaines pour lesquels les ensembles de données sont disponibles, proposant des corpus d'entraînement massifs (Kaplan *et al.*, 2020). En revanche, les ensembles de données à faible ressource (Ravichander *et al.*, 2019; Möller *et al.*, 2020) présentent des défis majeurs pour les LLM en raison de la rareté des données et de leur sous-représentation. Le manque de données en quantité et en qualité suffisantes engendre des lacunes en matière de couverture lexicale (Hangya *et al.*, 2022), de connaissances culturelles (Li *et al.*, 2024) et de nuances syntaxiques (Lucas *et al.*, 2024). Par conséquent, les performances des LLM dans des contextes à faible ressource sont nettement inférieures à celles obtenues avec des ensembles de données bien dotés. Cette disparité limite fortement les avancées en IA dans les domaines concernés.

Cet article de synthèse met en lumière les méthodes et les évaluations utilisées dans les contextes à faible ressource et dans les domaines spécialisés. Nous soutenons que la diversité et la qualité des ensembles de données sont plus importantes que l'accumulation de grandes quantités de données médiocres. Cette perspective est corroborée par des études démontrant que la qualité des données d'entraînement a un impact significatif sur la performance des modèles de langage, en particulier dans les environnements à faible ressource (Micallef *et al.*, 2022; Sajith & Kathala, 2024). Pour pallier la rareté des données, l'augmentation des données s'est imposée comme une solution efficace (Seo *et al.*, 2024), permettant de générer des exemples supplémentaires afin d'améliorer la robustesse des modèles.

Le traitement automatique du langage naturel (TALN) englobe un large éventail de tâches, telles que la synthèse de texte, la modélisation de sujets et la génération de texte (Wikipedia LLMs, 2025). Dans cette étude, nous nous concentrons explicitement sur la tâche de question-réponse *(Question Answering, QA)*, car elle s'impose comme un domaine de recherche particulièrement dynamique, notamment dans les contextes à faible ressource. Dans les applications spécifiques à un domaine, notamment dans le secteur privé et les milieux de recherche indépendants, les systèmes de QA et de chatbots (Afzal *et al.*, 2024; Megahed *et al.*, 2024) sont couramment utilisés pour faciliter l'interaction des utilisateurs avec les ensembles de données et évaluer les capacités des modèles. De plus, avec l'avènement des grands modèles de langue, les systèmes de QA peuvent être adaptés pour exécuter d'autres tâches de TALN via la restructuration des données et l'affinage des modèles. Toutefois, malgré ces avancées, les applications spécifiques à un domaine continuent de rencontrer des défis majeurs dans les environnements à faible ressource.

2 Problématique

Présentation Les environnements à faibles ressources pour les grands modèles de langage (LLMs) sont des contextes où des ressources essentielles comme des corpus larges et variés, des jeux de données annotés, une expertise métier ou la disponibilité des données sont fortement limitées ou totalement absentes. Ces contraintes dépassent largement les défis habituellement associés aux langues peu dotées. Même dans des langues riches en ressources comme l'anglais, de nombreux domaines spécialisés, tels que certaines branches de la médecine ou de la recherche scientifique, souffrent d'un

manque chronique de données (Seo *et al.*, 2024). Comme les LLMs sont en grande partie préentraînés sur des corpus génériques de grande taille, ils échouent souvent à se généraliser à des tâches qui exigent une connaissance fine et spécialisée du domaine. Par exemple, dans le domaine biomédical, bien qu'il existe un grand volume de textes médicaux généraux, les jeux de données portant sur des maladies rares ou sur des essais cliniques spécifiques restent rares voire inexistants, ce qui provoque des décalages de distribution et une baisse de performance des modèles (Chen *et al.*, 2024b).

Ces limitations posent des défis majeurs pour les systèmes de question-réponse dans les domaines à faibles ressources. Les systèmes de QA nécessitent non seulement une couverture lexicale étendue, mais aussi des connaissances factuelles précises, des capacités de raisonnement spécifiques au domaine, et la faculté d'extraire ou d'inférer des informations à partir du contexte. Lorsque les corpus spécialisés sont peu nombreux, les modèles QA ont du mal à apprendre la terminologie, les connaissances de fond et les schémas d'inférence nécessaires à la production de réponses précises et pertinentes. De plus, en l'absence d'annotations conçues par des experts, il est difficile d'ajuster les modèles pour traiter des types de questions spécialisés, ce qui augmente le taux d'hallucinations et réduit la fiabilité des réponses. Bien qu'il n'existe pas de seuil universellement reconnu pour définir un environnement à faibles ressources, nous considérons qu'un jeu de données relève de cette catégorie lorsqu'il n'est pas couramment utilisé pour l'entraînement préliminaire des grands modèles de langage. C'est notamment le cas des ensembles absents des référentiels standards.

Questions de recherche Nous proposons aussi d'explorer plusieurs questions de recherche. Premièrement, il est essentiel d'identifier des stratégies efficaces pour accroître la quantité et la qualité des données spécifiques à un domaine en utilisant les LLMs, en particulier dans les domaines où ces données sont rares. Deuxièmement, nous cherchons à comprendre quelles approches permettent d'améliorer l'adaptation des LLMs aux tâches spécifiques à un domaine. Troisièmement, il est nécessaire d'établir des cadres et des métriques d'évaluation robustes afin de mesurer avec précision la performance des modèles dans ces contextes. Enfin, il est crucial de considérer les implications éthiques, de confidentialité et d'équité lors du déploiement des LLMs dans des domaines spécialisés. En conséquence, nous formulons les questions de recherche suivantes :

— **Q1** : Comment accroître efficacement les données spécifiques à un domaine en utilisant les LLMs ?
— **Q2** : Quelles sont les approches permettent d'améliorer l'adaptation des LLMs aux tâches spécifiques à un domaine ?
— **Q3** : Comment évaluer la performance des LLMs dans des contextes à faible ressource, et quelles sont les métriques les plus appropriées ?
— **Q4** : Quelles sont les implications éthiques, de confidentialité et d'équité à prendre en compte ?

3 Travaux connexes

Ding *et al.* (2024a) proposent une analyse du domaine selon deux axes, centrée à la fois sur les données et sur l'apprentissage. Ils définissent quatre « perspectives sur les données » (création, annotation, reformulation, co-annotation) et présentent différents paradigmes d'apprentissage allant de l'ajustement supervisé à l'apprentissage par alignement. Ils illustrent aussi des applications concrètes, comme Dr. LLaMA pour la question-réponse médicale (où ChatGPT ou GPT-4 réécrivent ou génèrent de nouvelles paires question–réponse) et la stratégie de masquage sélectif de DALE. Chai

et al. (2025) complètent cette approche par une taxonomie technique claire, incluant des méthodes simples, basées sur des prompts, sur la récupération d'information, ou hybrides. Cependant, aucune de ces deux études ne propose une comparaison systématique des différents paradigmes appliqués aux contraintes spécifiques des domaines biomédical ou juridique à faibles ressources, telles que les contraintes de confidentialité ou les variations de distribution.

Notre revue s'appuie sur ces travaux en se concentrant spécifiquement sur l'augmentation de données pour la question-réponse dans les contextes biomédicaux et juridiques à faibles ressources. À l'aide de jeux de données ciblés, nous analysons dans quelle mesure les différentes méthodes d'augmentation répondent aux contraintes propres à ces domaines. Plutôt que de proposer un nouveau cadre théorique, notre contribution consiste en une comparaison détaillée, fondée sur les données, qui met en lumière la pertinence pratique des approches dans des contextes sensibles.

4 Revue de la littérature et analyse

4.1 Méthodologie d'identification des articles et d'analyse de la diversité

Identification des articles Pour mener notre analyse, nous visons à identifier les sous-ensembles de données sous-représentés au sein de leurs domaines respectifs. Notre attention se porte spécifiquement sur les ensembles de données des domaines du biomédical et du juridique, car ces deux domaines ont été largement étudiés dans la communauté de recherche sur les grands modèles de langage (LLMs). Bien qu'une littérature importante existe pour ces domaines, il reste difficile d'identifier des jeux de données à faibles ressources disponibles publiquement, souvent en raison de préoccupations liées à la confidentialité, de restrictions d'accès ou de l'absence de référentiels standardisés. Par conséquent, pour chaque domaine, nous limitons notre analyse à 3 ou 4 types de jeux de données accessibles et suffisamment documentés pour être analysés.

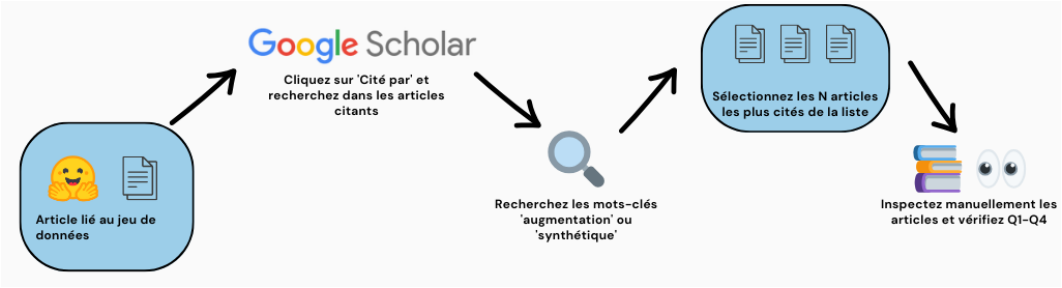


FIGURE 1 – Flux de travail pour l'identification des articles pertinents sur l'augmentation des ensembles de données

Comme illustré sur la figure 1, nous avons mis en place un flux de travail structuré afin d'identifier les travaux de recherche pertinents portant sur l'augmentation des ensembles de données et la génération de données synthétiques. Pour explorer cette problématique de manière systématique, nous avons réalisé une revue de la littérature en nous concentrant sur les techniques d'augmentation et de génération de données synthétiques appliquées aux ensembles de données étudiés.

À l'aide de Google Scholar, nous avons recherché des articles contenant soit le mot-clé *augmentation*,

soit le mot-clé *synthetic*, rédigés en anglais, puis nous les avons filtrés pour ne conserver que ceux en lien avec le traitement automatique des langues (TAL). Ces deux mots-clés ont été choisis afin de couvrir largement la littérature pertinente sur l'augmentation de données, et l'indexation en texte intégral de Google Scholar nous a permis d'identifier des travaux dans lesquels ces termes apparaissent au-delà du titre ou du résumé. Cette approche a facilité l'identification des contributions potentiellement pertinentes. Nous avons ensuite sélectionné jusqu'à N articles de recherche primaire les plus cités pour chaque jeu de données, avec $N \leq 3$ [1], en excluant les articles de revue ainsi que ceux qui ne mentionnent les techniques d'augmentation que dans leur section de travaux connexes. Les articles de revue ont été exclus car, bien qu'ils fournissent des synthèses utiles, ils ne présentent généralement pas d'analyses méthodologiques détaillées ni de résultats empiriques spécifiques aux jeux de données étudiés. Ce filtrage, fondé à la fois sur le nombre de citations et le type de publication, nous a permis de nous concentrer sur les contributions les plus influentes et techniquement substantielles aux méthodologies d'augmentation de données.

Domaine	Articles citant des ensembles de données	Q1	Q2	Q3	Q4
Biomédical	Möller *et al.* (2020), corpus COVID-QA	–	✓	✓	–
	↪ Reddy *et al.* (2020)	✓	✓	✓	–
	↪ Siriwardhana *et al.* (2023)	✓	✓	✓	–
	↪ Samuel *et al.* (2024)	✓	✓	✓	–
	Wang *et al.* (2024), corpus ReDis-QA	✓	✓	✓	–
	↪ Wang *et al.* (2025a)	✓	✓	✓	✓
	↪ Li *et al.* (2025)	✓	✓	–	✓
	Arias-Duart *et al.* (2025), corpus CareQA	✓	✓	✓	–
	↪ Wang *et al.* (2025b)	✓	✓	✓	✓
	Chen *et al.* (2024a), corpus Medbullets	–	–	✓	✓
	↪ Kim *et al.* (2025)	✓	✓	✓	✓
	↪ Wang *et al.* (2025b)	✓	✓	✓	✓
	↪ Wang *et al.* (2025a)	✓	✓	✓	✓
Juridique	Ravichander *et al.* (2019), corpus PrivacyQA	–	–	✓	–
	↪ Vold & Conrad (2021)	–	✓	✓	–
	↪ Parvez *et al.* (2023)	✓	✓	✓	✓
	↪ Nayak *et al.* (2024)	✓	✓	✓	–
	Ahmad *et al.* (2020), corpus PolicyQA	–	–	✓	–
	Lin *et al.* (2022), corpus TruthfulQA	–	–	✓	✓
	↪ Wang *et al.* (2023)	–	✓	✓	✓
	↪ Kim *et al.* (2023)	✓	✓	✓	✓
	↪ Ding *et al.* (2024b)	✓	✓	✓	✓

TABLE 1 – Aperçu de l'intersection entre chaque question de recherche (Q1 à Q4) et les articles décrivant des corpus dans les deux domaines étudiés. La coche ✓ indique que la question est traitée, le tiret qu'elle ne l'est pas, et les flèches ↪ indiquent la réutilisation de ces jeux de données pour diverses méthodes d'augmentation de données.

Dans le tableau 1, nous avons adopté une approche structurée pour analyser chacune des quatre questions de recherche dans les domaines du biomédical et du juridique. Ce cadre méthodologique permet un examen systématique des techniques d'augmentation appliquées à différents ensembles de données à faible ressource. Nous avons sélectionné trois à quatre ensembles de données par domaine.

1. Certains jeux de données sont récents et disposent encore de peu de méthodes spécifiques.

En associant les approches d'augmentation aux différents types de données, notre étude offre des perspectives aux chercheurs souhaitant améliorer les performances des grands modèles de langue (LLMs) dans des environnements à faible ressource.

4.2 Domaine biomédical

4.2.1 Présentation des articles collectés

Le domaine biomédical reste l'un des plus critiques pour l'application de l'IA, en raison de son potentiel à révolutionner le diagnostic, la planification des traitements et la gestion des patients. Malgré ces promesses, ce domaine est confronté aux défis des ressources limitées ou inaccessibles en dehors des hôpitaux. Bien que les données médicales puissent être capturées sous diverses formes, y compris des images, des vidéos et d'autres supports, nous limitons cette étude aux données textuelles afin de maintenir un cadre cohérent.

En appliquant notre méthodologie, nous avons sélectionné quatre jeux de données de questions-réponses médicales à faibles ressources pour une analyse approfondie. Afin d'établir une définition de référence du concept de « faibles ressources » dans ce domaine, nous avons comparé chaque jeu de données à des ensembles de référence largement adoptés, tels que MedQA (Jin *et al.*, 2021) et MedMCQA (Pal *et al.*, 2022), qui regroupent plus de 182 000 instances de questions-réponses couvrant divers sous-domaines médicaux. Nous désignons ce corpus de référence sous le nom de `ParentQA`. Les quatre jeux de données spécialisés retenus sont les suivants :

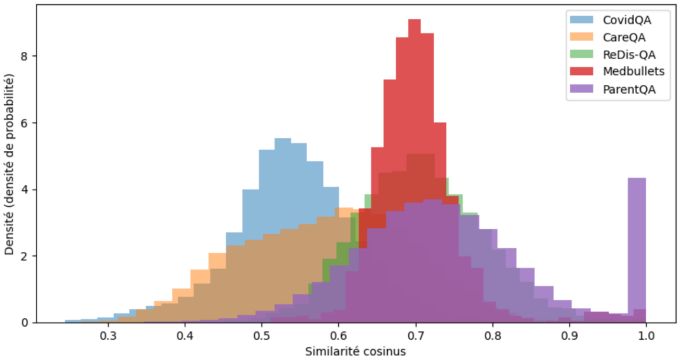
— **COVID-QA** (Möller *et al.*, 2020) : 2 019 paires de questions-réponses annotées par des experts sur la COVID-19, suivant une approche d'annotation inspirée de SQuAD
— **ReDis-QA** (Wang *et al.*, 2024) : 205 maladies rares à travers 1 360 paires de questions-réponses de haute qualité
— **CareQA** (Arias-Duart *et al.*, 2025) : 8 390 instances annotées couvrant des questions à réponse ouverte et fermée dans des domaines tels que la médecine, les soins infirmiers, la biologie, la chimie, la psychologie et la pharmacologie
— **Medbullets** (Chen *et al.*, 2024a) : 616 cas cliniques réels permettant d'évaluer le raisonnement et la prise de décision dans des scénarios cliniques complexes

4.2.2 Analyse de la diversité

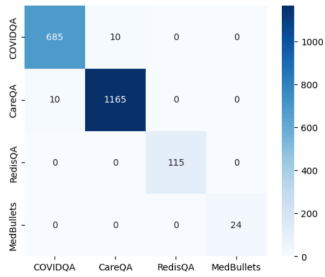
Pour mesurer la diversité de chaque corpus par rapport au domaine concerné, nous utilisons le modèle de plongements `all-MiniLM-L6-v2` [2]. Ce modèle a été choisi en raison de son bon compromis entre performance et efficacité. Il permet un encodage rapide de grandes quantités de données, ce qui est particulièrement adapté à notre contrainte de ressources computationnelles limitées. Avec environ 22 millions de paramètres, une dimension d'embedding de 384, et une empreinte mémoire modérée (environ 86 Mo), il projette chaque phrase dans un espace vectoriel dense où la proximité entre vecteurs reflète leur similarité de sens. Il constitue une solution efficace pour établir une première analyse comparative, en permettant notamment de calculer la similarité sémantique entre phrases.

La figure 2 illustre les distributions de similarité cosinus obtenues pour chaque jeu de données à faibles ressources par rapport au jeu de données parent, mettant en évidence les principaux schémas

2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2



(a) Similarité cosinus entre chaque corpus spécialisé et *ParentQA*



(b) Chevauchement du vocabulaire

FIGURE 2 – Analyses des données médicales par rapport aux données `ParentQA` en termes de similarité cosinus (gauche) et de chevauchement du vocabulaire (droite)

de recouvrement de domaine et de diversité linguistique. Le pic prononcé à 1.0 pour `ParentQA` reflète sa forte auto-similarité, puisque chaque plongement comparé à lui-même donne naturellement une similarité cosinus de 1.0. Nous observons que les jeux de données spécialisés (ReDis-QA et Medbullets) ont tendance à se regrouper autour de 0.7, suggérant un alignement partiel avec `ParentQA` tout en mettant en évidence des terminologies spécifiques au domaine et des structures de questions distinctes. Notamment, certains jeux de données présentent une répartition plus large des valeurs (CovidQA et CareQA), ce qui implique une plus grande hétérogénéité du contenu et un chevauchement sémantique plus faible avec le corpus parent. Lorsqu'une distribution est davantage décalée vers la gauche (c'est-à-dire avec une similarité plus faible), cela indique une divergence encore plus marquée en termes de style des questions et de vocabulaire spécialisé. Par exemple, la distribution de COVID-QA est située plus à gauche, reflétant son vocabulaire spécifique à la pandémie. La matrice de chevauchement lexical (Figure 2b) complète ces observations en mettant en lumière les caractéristiques linguistiques uniques qui distinguent chaque jeu de données à faibles ressources du corpus parent.

4.2.3 Positionnement face aux questions de recherche

Parmi les méthodes examinées, la Q1 (*comment accroître les données*) se décline en deux paradigmes. D'une part, la génération *few-shot* suivie de filtrage (*round-trip consistency*) exemplifiée par (Samuel *et al.*, 2024) sur CovidQA permet d'obtenir rapidement des gains notables sans corpus massif préalable . D'autre part, les pipelines *chain-of-thought* à grande échelle combinent extraction de raisonnement, synthèse et révision document-référence pour produire des centaines de milliers voire des centaines de milliards de tokens médicaux, mais requièrent un accès étendu à des manuels, graphes de connaissances ou bases cliniques (Kim *et al.*, 2025; Wang *et al.*, 2025b).

Concernant la Q2 (*quelles approches pour l'adaptation des LLM*), trois axes se dégagent. Le fine-tuning sur corpus annoté (RoBERTa + COVID-QA) apporte des améliorations stables à partir de quelques milliers d'exemples experts (Möller *et al.*, 2020). L'instruction-tuning *CoT* accroît la précision sur divers benchmarks médicaux en intégrant explicitement la chaîne de pensée lors de l'apprentissage (Kim *et al.*, 2025). Enfin, les architectures RAG end-to-end ou multi-phase combinent

retrieval adapté et phases de RL pour un alignement plus fin sur les critères cliniques, mais sont fortement liés aux connaissances externes et aux métriques spécifiques au domaine (Siriwardhana *et al.*, 2023; Wang *et al.*, 2025b).

En ce qui concerne la Q3 (*évaluation et métriques*), les indicateurs *close-ended* (Exact Match, F1) et *perplexité* restent le socle générique applicable à tous les domaines (Möller *et al.*, 2020; Samuel *et al.*, 2024). Les mesures sémantiques (BERTScore, BLEURT) et les juges automatiques (G-Eval) (Chen *et al.*, 2024a; Arias-Duart *et al.*, 2025) enrichissent l'analyse qualitative des réponses générées, tandis que l'évaluation humaine demeure nécessaire pour valider la cohérence et la factualité dans les contextes cliniques (Wang *et al.*, 2025a).

Enfin, concernant la Q4 *(principes éthiques)*, la plupart des articles omettent ces questions ou ne les mentionnent que brièvement, soulignant une lacune critique dans les applications en santé, où la sécurité des patients, la confidentialité des données et l'accès équitable sont primordiaux (Wang *et al.*, 2025b,a). Étant donné les risques potentiels liés à des conseils médicaux incorrects ou biaisés (Li *et al.*, 2025), il est essentiel que les recherches futures intègrent des *analyses de biais, des protocoles de préservation de la vie privée et des cadres réglementaires* dans les stratégies d'augmentation de données en contexte de ressources limitées pour le domaine biomédical.

Dans l'ensemble, on distingue deux familles de méthodes : d'une part, les **méthodes génériques** génération few-shot, l'instruction-tuning CoT, et fine-tuning léger sur petits corpus annotés couplés aux métriques Exact Match, F1 et perplexité qui offrent une mise en œuvre rapide et des gains de 5–10 % avec seulement quelques dizaines d'exemples (Möller *et al.*, 2020; Samuel *et al.*, 2024; Chen *et al.*, 2024a). D'autre part, les **méthodes spécifiques** nécessitent un accès à des ressources spécialisées (manuels, graphes de connaissances, annotations expertes) ainsi qu'un réglage fin des prompts, des composants architecturaux et une intégration dans des pipelines de fine-tuning complexes. Ces méthodes sont souvent mobilisées après l'application des techniques génériques afin d'établir une ligne de base, puis d'optimiser davantage les performances en ciblant les spécificités du domaine. Leur efficacité accrue s'accompagne toutefois d'une moindre transférabilité, car elles requièrent une adaptation préalable.

4.2.4 Autres approches

Nous observons que plusieurs méthodes d'augmentation de données ont été proposées en complément de l'approche décrite ci-dessus. Le cadre XAIQA génère des paires question-réponse synthétiques à partir de dossiers médicaux électroniques en s'appuyant sur un classificateur de documents et un modèle explicatif pour extraire des concepts médicaux et produire des questions, les réponses étant issues de passages textuels fortement informatifs (Stremmel *et al.*, 2023). De même, CLINGEN propose une génération de données synthétiques pour le TAL clinique par infusion de connaissances, extrayant des sujets cliniques de graphes de connaissances et de LLMs, tout en intégrant des styles d'écriture propres au domaine pour garantir diversité et pertinence des textes générés (Xu *et al.*, 2024).

Au-delà des approches présentées, deux cadres méthodologiques se sont imposés pour renforcer la robustesse et la qualité des représentations dans les contextes à faibles ressources : l'entraînement adversarial, qui génère des exemples perturbés afin de sanctionner et de consolider les frontières de décision du modèle, et l'apprentissage contrastif, qui organise l'espace d'embeddings en rapprochant des paires sémantiquement proches et en écartant des exemples hétérogènes. Ces paradigmes se sont

récemment révélés prometteurs en QA biomédical. Par exemple, Zhao *et al.* (2023) exploitent l'algorithme ILAG pour générer des exemples négatifs « hard » et intègrent un entraînement adversarial en mode low-resource, améliorant la robustesse d'un système de recherche de réponses. De même, Mahbub *et al.* (2022) proposent un cadre d'adaptation de domaine par apprentissage adversarial, permettant de transférer efficacement des modèles préentraînés vers des données cliniques sans annotations supplémentaires. Enfin, Zhang *et al.* (2022) introduisent QFCL (Focus-Driven Contrastive Learning for Medical Question Summarization), un protocole contrastif qui génère des exemples négatifs difficiles en fonction du point focal des questions, améliorant ainsi la reformulation et la compréhension des requêtes.

4.3 Domaine juridique

4.3.1 Présentation des articles collectés

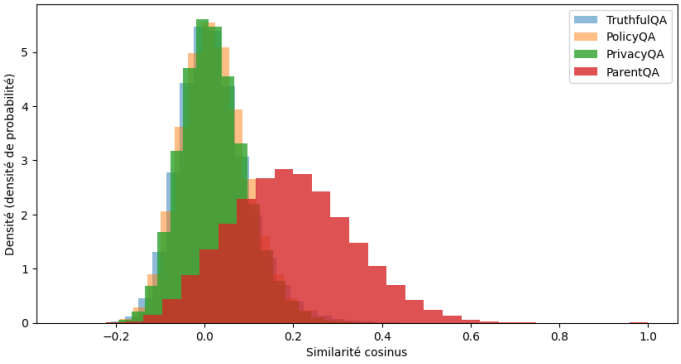
À mesure que le volume des affaires juridiques augmente, l'intelligence artificielle joue un rôle crucial dans la réduction de la charge de travail, la minimisation des erreurs humaines et l'accélération des décisions judiciaires tout en garantissant leur cohérence. En automatisant les tâches répétitives et chronophages, telles que l'analyse de documents et la recherche juridique, l'IA permet aux professionnels du droit de se concentrer davantage sur la prise de décision stratégique et l'évaluation nuancée des dossiers. De plus, l'analyse prédictive aide à anticiper les résultats, favorisant ainsi la transparence et la cohérence des décisions judiciaires (Lai *et al.*, 2024).

En appliquant notre méthodologie à ce domaine, nous avons identifié trois jeux de données QA juridiques pertinents pour une analyse approfondie. Nous avons retenu un seul corpus comme jeu de données `ParentQA`, le sous-ensemble juridique de MMLU (Hendrycks *et al.*, 2021), incluant les catégories *international law*, *professional law* et *US foreign policy*. Ces sous-ensembles, largement utilisés pour le pré-entraînement des grands modèles de langage, contiennent environ 2 000 exemples. Les trois jeux de données spécialisés retenus pour cette étude sont les suivants :

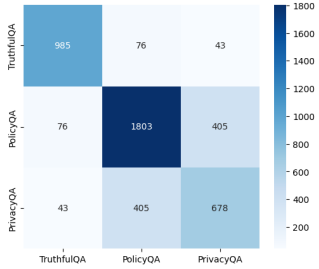
— **PolicyQA** (Ahmad *et al.*, 2020) : un jeu de données de lecture de compréhension portant sur les politiques de confidentialité des sites web, comprenant plus de 25 000 triplets question-passage-réponse visant des réponses concises.
— **PrivacyQA** (Ravichander *et al.*, 2019) : un jeu de données QA sur les politiques de confidentialité, contenant 1 750 questions et plus de 3 500 annotations réalisées par des experts, alliant perspectives juridiques et informatiques.
— **TruthfulQA** (Lin *et al.*, 2022) : un benchmark composé de 817 questions adversariales réparties en 38 catégories, dont un sous-ensemble dédié aux questions juridiques, conçu pour évaluer la véracité des réponses générées par les modèles de langage.

4.3.2 Analyse de la diversité

En utilisant le même modèle de plongements, `all-MiniLM-L6-v2`, La figure 3 illustre les distributions de similarité cosinus pour chaque jeu de données juridique à faibles ressources par rapport au jeu de données parent. La distribution de `ParentQA` présente une hauteur plus faible, ce qui est attendu étant donné que ce corpus contient moins de données. En revanche, les distributions de TruthfulQA, PolicyQA et PrivacyQA affichent des pics plus proches de zéro, suggérant un alignement



(a) Similarité cosinus entre chaque corpus spécialisé et *ParentQA*



(b) Chevauchement du vocabulaire

FIGURE 3 – Analyses des données juridique par rapport aux données *ParentQA*

significativement plus faible avec le jeu de données parent. Cela indique que le vocabulaire et les structures de questions dans ces jeux de données juridiques sont plus diversifiés et distincts de ceux de `ParentQA`. La matrice de chevauchement lexical (Figure 3b) corrobore ces observations, montrant un recouvrement lexical limité entre les différents jeux de données. Cela s'explique par le fait que MMLU contient principalement des questions liées au droit, tandis que PolicyQA et PrivacyQA sont axés sur les politiques de confidentialité, ce qui se traduit par une faible similarité lexicale.

4.3.3 Positionnement face aux questions de recherche

Parmi les méthodes examinées, la Q1 (comment accroître les données spécifiques au domaine) met en œuvre des stratégies de génération et de récupération : génération de perturbations sémantiquement équivalentes via paraphrase LLM (Ding *et al.*, 2024b), synthèse de corpus par comparaison d'outputs (Kim *et al.*, 2023), extraction d'exemples par multi-retrieveurs (Parvez *et al.*, 2023) et génération massive d'instructions à partir de méta-templates (Nayak *et al.*, 2024).

Concernant la Q2 (quelles approches pour l'adaptation des LLM), les travaux combinent pré-entraînement continu, fine-tuning et renforcement : PolicyQA affine un modèle BERT pré-entraîné sur un corpus de politiques de confidentialité, afin de l'adapter spécifiquement à la tâche de question-réponse extractive dans ce domaine sensible (Ahmad *et al.*, 2020). Rowen active un mécanisme de « retrieve-only-when-needed » générique (Ding *et al.*, 2024b), ALMoST enchaîne reward modelling, démonstrations synthétiques et RL (Kim *et al.*, 2023), Citrus combine CPT, SFT et RL réflexif pour tâches cliniques (Wang *et al.*, 2025b), et (Vold & Conrad, 2021) démontrent le gain de F1 (+31 %) et MRR (+41 %) d'un fine-tuning de RoBERTa sur PrivacyQA .

En ce qui concerne la Q3 (évaluation et métriques), les études mobilisent des mesures standard adaptées à chaque tâche : EM et F1 pour la QA extractive (Ahmad *et al.*, 2020), précision, rappel, F1 et MRR pour la classification et le ranking (Ravichander *et al.*, 2019). Ces métriques sont largement reconnues pour leur robustesse et leur capacité à refléter la performance dans des contextes à faibles ressources.

Enfin, concernant la Q4 (principes éthiques), TruthfulQA met en garde contre les risques de désinfor-

mation et la perte de confiance liés aux réponses fallacieuses, appelant à des garde-fous rigoureux (Lin *et al.*, 2022). ALMoST se fonde sur un benchmark HHH (helpful, harmless, honest) pour aligner les modèles sur des valeurs humaines et réduire les sorties nocives (Kim *et al.*, 2023). Cependant, on constate que la plupart des travaux ne traitent pas pleinement les implications éthiques, de confidentialité et d'équité. Or ces dimensions sont essentielles pour garantir la confiance des utilisateurs, prévenir les biais algorithmiques et se conformer aux régulations.

Les approches génériques reposent sur des mécanismes de paraphrase, de récupération et de transfert de connaissances, offrent rapidité de prototypage et extensibilité à tout domaine faible en ressources, mais restent limitées par la cohérence et la richesse du modèle de base (Kim *et al.*, 2023; Ding *et al.*, 2024a; Nayak *et al.*, 2024). À l'inverse, les solutions domain-spécifiques tirent parti de corpus et workflows experts pour atteindre des performances maximales, au prix d'un surcoût en collecte de données spécialisées, expertise métier et charge computationnelle (Vold & Conrad, 2021; Wang *et al.*, 2023). Dès lors, il convient de démarrer par un fine-tuning minimal sur un transformer générique, puis d'ajouter progressivement modules architecturaux et corpus ciblés pour répondre aux exigences métier et garantir une adoption éthique.

4.3.4 Autres approches

Outre l'approche décrite précédemment, des méthodes récentes comme **KGDG** (Zhou *et al.*, 2025) et **DALE** (Ghosh *et al.*, 2023) proposent des stratégies efficaces d'augmentation de données juridiques. KGDG s'appuie sur une base de connaissances juridiques et combine trois modules (génération, correction, vérification) pour produire des exemples de raisonnement juridique de qualité, permettant d'entraîner LAWGPT, un modèle performant surpassant les modèles juridiques existants. DALE utilise un pré-entraînement avec masquage sélectif de spans corrélés pour apprendre le langage juridique, suivi d'une génération conditionnelle d'augmentations cohérentes et diversifiées. Testé sur 13 jeux de données, DALE obtient des gains allant jusqu'à +49.8%. Ces méthodes offrent une opportunité prometteuse pour enrichir les jeux de données introduits précédemment dans un contexte à faibles ressources.

Malgré ces avancées, un défi majeur demeure dans la disponibilité et la structuration des ensembles de données juridiques. De nombreux cas restent non documentés ou inaccessibles, ce qui accentue la complexité inhérente au langage spécifique au domaine, aux fréquentes évolutions réglementaires, ainsi qu'à la nécessité de disposer de données annotées de haute qualité (Abdallah *et al.*, 2023). De plus, plusieurs sous-domaines juridiques restent largement inexplorés dans le monde des LLMs, notamment les accords commerciaux internationaux [3], le droit spatial [4], le droit du Traité sur l'Antarctique [5], le droit des brevets en biotechnologie et génétique [6], parmi tant d'autres. Les ensembles de données disponibles dans ces domaines sont encore à l'état brut et non structurés, nécessitant un prétraitement conséquent avant de pouvoir être exploités efficacement pour la recherche ou l'analyse juridique.

3. https://datatopics.worldbank.org/dta/table.html
4. https://www.unoosa.org/oosa/en/ourwork/spacelaw/index.html
5. https://www.ats.aq
6. https://www.wipo.int/wipolex/en/

5 Limites

Bien que cette étude apporte des éléments de compréhension sur l'augmentation de données et la génération de données synthétiques dans le contexte des ensembles de données à faibles ressources, plusieurs limites doivent être soulignées.

Spécificité par domaine Cette analyse se limite aux domaines du biomédical et du juridique. Bien que ces domaines présentent des défis riches et variés, l'extension de l'étude à d'autres secteurs tels que l'énergie, la finance ou les sciences pourrait révéler des nuances supplémentaires et une applicabilité plus large des techniques d'augmentation.

Contraintes de recherche aux mots-clés La recherche bibliographique s'est appuyée exclusivement sur les mots-clés *augmentation* et *synthétique*. Cette approche ciblée peut exclure des travaux pertinents utilisant des terminologies ou des méthodologies alternatives, limitant ainsi la portée de nos résultats.

Sélection des ensembles de données parents Les ensembles de données « parents » sous-jacents à notre analyse se composent uniquement de deux collections à grande échelle, avec l'hypothèse que leur diversité est suffisante pour établir une base de référence robuste. Cependant, l'augmentation du nombre et de la diversité de ces ensembles de données améliorerait probablement la portée et la généralisation de notre analyse.

Généralisation du modèle de plongements Dans cette étude, nous avons utilisé un modèle de plongements généraliste, `all-MiniLM-L6-v2`, qui pourrait ne pas saisir de manière optimale les spécificités terminologiques et contextuelles propres à chaque domaine. Une piste d'amélioration, envisagée lorsque les ressources computationnelles le permettront, consistera à recourir à des modèles de plongements spécialisés. Par exemple, `PubMedBERT` [7] pour le domaine biomédical afin d'obtenir des représentations plus précises et potentiellement d'améliorer la qualité des résultats.

6 Conclusion

Dans cet article, nous avons présenté une analyse approfondie des stratégies d'augmentation de données dans des contextes à faibles ressources, en nous concentrant sur les domaines biomédical et juridique. Nous avons mené une revue de la littérature en identifiant d'abord les articles décrivant des jeux de données pertinents, puis en analysant les travaux les plus cités sur Google Scholar proposant des méthodes d'augmentation de données en lien avec ces jeux de données. Nous avons évalué leur traitement de quatre questions de recherche clés : comment accroître les données spécifiques à un domaine, quelles approches utiliser pour adapter les LLMs, comment évaluer leurs performances, et quelles implications éthiques doivent être prises en compte. Cette revue s'est appuyée sur des analyses de diversité (similarité cosinus et chevauchement lexical) afin de mettre en évidence les différences

7. https://huggingface.co/NeuML/pubmedbert-base-embeddings

entre les jeux de données spécialisés et leurs corpus parents, révélant ainsi des défis importants liés à la rareté et à la spécificité des données.

Dans la continuité de ce travail, une évaluation empirique comparative des différentes stratégies d'augmentation appliquées à chaque jeu de données constitue une prochaine étape importante. Cette première étude ouvre également la voie à l'identification de méthodes d'augmentation adaptées aux contextes à faibles ressources, en cohérence avec les objectifs de ma thèse. Je prévois également d'étendre cette analyse à d'autres langues et domaines spécialisés, afin d'évaluer plus en profondeur la robustesse, la généralisabilité et les limites des approches étudiées.

Références

ABDALLAH A., PIRYANI B. & JATOWT A. (2023). Exploring the state of the art in legal qa systems. *Journal of Big Data*, **10**(1), 127.

AFZAL A., KOWSIK A., FANI R. & MATTHES F. (2024). Towards optimizing and evaluating a retrieval augmented qa chatbot using LLMs with human in the loop. *arXiv preprint arXiv :2407.05925*.

AHMAD W., CHI J., TIAN Y. & CHANG K.-W. (2020). PolicyQA : A reading comprehension dataset for privacy policies. In T. COHN, Y. HE & Y. LIU, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 743–749, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.66.

ARIAS-DUART A., MARTIN-TORRES P. A., HINJOS D., BERNABEU-PEREZ P., GANZABAL L. U., MALLO M. G., GURURAJAN A. K., LOPEZ-CUENA E., ALVAREZ-NAPAGAO S. & GARCIA-GASULLA D. (2025). Automatic evaluation of healthcare LLMs beyond question-answering. *arXiv preprint arXiv :2502.06666*.

CHAI Y., XIE H. & QIN J. S. (2025). Text data augmentation for large language models : A comprehensive survey of methods, challenges, and opportunities. *arXiv preprint arXiv :2501.18845*.

CHEN H., FANG Z., SINGLA Y. & DREDZE M. (2024a). Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv :2402.18060*.

CHEN X., MAO X., GUO Q., WANG L., ZHANG S. & CHEN T. (2024b). Rarebench : Can LLMs serve as rare diseases specialists ? In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, p. 4850–4861.

DEEPSEEK-AI, GUO D., YANG D., ZHANG H., SONG J., ZHANG R., XU R., ZHU Q., MA S., WANG P., BI X., ZHANG X., YU X., WU Y., WU Z. F., GOU Z., SHAO Z., LI Z., GAO Z., LIU A., XUE B., WANG B., WU B., FENG B., LU C., ZHAO C., DENG C., ZHANG C., RUAN C., DAI D., CHEN D., JI D., LI E., LIN F., DAI F., LUO F., HAO G., CHEN G., LI G., ZHANG H., BAO H., XU H., WANG H., DING H., XIN H., GAO H., QU H., LI H., GUO J., LI J., WANG J., CHEN J., YUAN J., QIU J., LI J., CAI J. L., NI J., LIANG J., CHEN J., DONG K., HU K., GAO K., GUAN K., HUANG K., YU K., WANG L., ZHANG L., ZHAO L., WANG L., ZHANG L., XU L., XIA L., ZHANG M., ZHANG M., TANG M., LI M., WANG M., LI M., TIAN N., HUANG P., ZHANG P., WANG Q., CHEN Q., DU Q., GE R., ZHANG R., PAN R., WANG R., CHEN R. J., JIN R. L., CHEN R., LU S., ZHOU S., CHEN S., YE S., WANG S., YU S., ZHOU S., PAN S., LI S. S., ZHOU S., WU S., YE S., YUN T., PEI T., SUN T., WANG T., ZENG W., ZHAO W., LIU W., LIANG W., GAO W., YU W., ZHANG W., XIAO W. L., AN W., LIU X., WANG X., CHEN X., NIE X., CHENG X., LIU X., XIE X., LIU X., YANG X., LI X., SU X., LIN X., LI X. Q., JIN X., SHEN X., CHEN X., SUN X., WANG X., SONG X., ZHOU X., WANG X., SHAN X., LI

Y. K., WANG Y. Q., WEI Y. X., ZHANG Y., XU Y., LI Y., ZHAO Y., SUN Y., WANG Y., YU Y., ZHANG Y., SHI Y., XIONG Y., HE Y., PIAO Y., WANG Y., TAN Y., MA Y., LIU Y., GUO Y., OU Y., WANG Y., GONG Y., ZOU Y., HE Y., XIONG Y., LUO Y., YOU Y., LIU Y., ZHOU Y., ZHU Y. X., XU Y., HUANG Y., LI Y., ZHENG Y., ZHU Y., MA Y., TANG Y., ZHA Y., YAN Y., REN Z. Z., REN Z., SHA Z., FU Z., XU Z., XIE Z., ZHANG Z., HAO Z., MA Z., YAN Z., WU Z., GU Z., ZHU Z., LIU Z., LI Z., XIE Z., SONG Z., PAN Z., HUANG Z., XU Z., ZHANG Z. & ZHANG Z. (2025). Deepseek-r1 : Incentivizing reasoning capability in LLMs via reinforcement learning.

DING B., QIN C., ZHAO R., LUO T., LI X., CHEN G., XIA W., HU J., LUU A. T. & JOTY S. (2024a). Data augmentation using LLMs : Data perspectives, learning paradigms and challenges. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2024*, p. 1679–1705, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-acl.97.

DING H., PANG L., WEI Z., SHEN H. & CHENG X. (2024b). Retrieve only when it needs : Adaptive retrieval augmentation for hallucination mitigation in large language models. *arXiv preprint arXiv :2402.10612*.

GEMINI, GEORGIEV P., LEI V. I., BURNELL R., BAI L., GULATI A., TANZER G., VINCENT D., PAN Z., WANG S. *et al.* (2024). Gemini 1.5 : Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv :2403.05530*.

GHOSH S., EVURU C. K. R., KUMAR S., RAMANESWARAN S., SAKSHI S., TYAGI U. & MANOCHA D. (2023). DALE : Generative data augmentation for low-resource legal NLP. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 8511–8565, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.528.

HANGYA V., SAADI H. S. & FRASER A. (2022). Improving low-resource languages in pre-trained multilingual language models. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 11993–12006, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.822.

HENDRYCKS D., BURNS C., BASART S., ZOU A., MAZEIKA M., SONG D. & STEINHARDT J. (2021). Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

JIN D., PAN E., OUFATTOLE N., WENG W.-H., FANG H. & SZOLOVITS P. (2021). What disease does this patient have ? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, **11**(14), 6421.

KAPLAN J., MCCANDLISH S., HENIGHAN T., BROWN T. B., CHESS B., CHILD R., GRAY S., RADFORD A., WU J. & AMODEI D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv :2001.08361*.

KIM H., HWANG H., LEE J., PARK S., KIM D., LEE T., YOON C., SOHN J., PARK J., REYKHART O. *et al.* (2025). Small language models learn enhanced reasoning skills from medical textbooks. *npj Digital Medicine*, **8**(1), 240.

KIM S., BAE S., SHIN J., KANG S., KWAK D., YOO K. & SEO M. (2023). Aligning large language models through synthetic feedback. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 13677–13700, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.844.

LAI J., GAN W., WU J., QI Z. & PHILIP S. Y. (2024). Large language models in law : A survey. *AI Open*.

LI C., CHEN M., WANG J., SITARAM S. & XIE X. (2024). Culturellm : Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, **37**, 84799–84838.

LI J., WANG Y., ZHANG K., CAI Y., HOOI B., PENG N., CHANG K.-W. & LU J. (2025). Fact or guesswork ? evaluating large language model's medical knowledge with structured one-hop judgment. *arXiv preprint arXiv :2502.14275*.

LIN S., HILTON J. & EVANS O. (2022). TruthfulQA : Measuring how models mimic human falsehoods. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3214–3252, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.229.

LUCAS A., BALADÓN A., PARDIÑAS V., AGÜERO-TORALES M., GÓNGORA S. & CHIRUZZO L. (2024). Grammar-based data augmentation for low-resource languages : The case of Guarani-Spanish neural machine translation. 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. 6385–6397, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.naacl-long.354.

MAHBUB M., SRINIVASAN S., BEGOLI E. & PETERSON G. D. (2022). Bioadapt-mrc : adversarial learning-based domain adaptation improves biomedical machine reading comprehension task. *Bioinformatics*, **38**(18), 4369–4379.

MEGAHED F. M., CHEN Y.-J., ZWETSLOOT I. M., KNOTH S., MONTGOMERY D. C. & JONES-FARMER L. A. (2024). Introducing chatsqc : Enhancing statistical quality control with augmented ai. *Journal of Quality Technology*, **56**(5), 474–497.

MICALLEF K., GATT A., TANTI M., VAN DER PLAS L. & BORG C. (2022). Pre-training data quality and quantity for a low-resource language : New corpus and bert models for maltese. *arXiv preprint arXiv :2205.10517*.

MÖLLER T., REINA A., JAYAKUMAR R. & PIETSCH M. (2020). COVID-QA : A question answering dataset for COVID-19. In K. VERSPOOR, K. B. COHEN, M. DREDZE, E. FERRARA, J. MAY, R. MUNRO, C. PARIS & B. WALLACE, Éds., *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online : Association for Computational Linguistics.

NAYAK N. V., NAN Y., TROST A. & BACH S. H. (2024). Learning to generate instruction tuning datasets for zero-shot task adaptation. *arXiv preprint arXiv :2402.18334*.

OPENAI, ACHIAM J., ADLER S., AGARWAL S., AHMAD L., AKKAYA I., ALEMAN F. L., ALMEIDA D., ALTENSCHMIDT J., ALTMAN S., ANADKAT S. *et al.* (2023). Gpt-4 technical report. *arXiv preprint arXiv :2303.08774*.

PAL A., UMAPATHI L. K. & SANKARASUBBU M. (2022). Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, p. 248–260 : PMLR.

PARVEZ M. R., CHI J., AHMAD W. U., TIAN Y. & CHANG K.-W. (2023). Retrieval enhanced data augmentation for question answering on privacy policies. In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 201–210, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.16.

RAVICHANDER A., BLACK A. W., WILSON S., NORTON T. & SADEH N. (2019). Question answering for privacy policies : Combining computational and legal perspectives. *arXiv preprint arXiv :1911.00841*.

Reddy R. G., Iyer B., Sultan M. A., Zhang R., Sil A., Castelli V., Florian R. & Roukos S. (2020). End-to-end qa on covid-19 : domain adaptation with synthetic training. *arXiv preprint arXiv :2012.01414*.

Sajith A. & Kathala K. C. R. (2024). Is training data quality or quantity more impactful to small language model performance ? *arXiv preprint arXiv :2411.15821*.

Samuel V., Aynaou H., Chowdhury A., Venkat Ramanan K. & Chadha A. (2024). Can LLMs augment low-resource reading comprehension datasets ? opportunities and challenges. In X. Fu & E. Fleisig, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4 : Student Research Workshop)*, p. 307–317, Bangkok, Thailand : Association for Computational Linguistics. doi : 10.18653/v1/2024.acl-srw.36.

Seo M., Baek J., Thorne J. & Hwang S. J. (2024). Retrieval-augmented data augmentation for low-resource domain tasks. *arXiv preprint arXiv :2402.13482*.

Siriwardhana S., Weerasekera R., Wen E., Kaluarachchi T., Rana R. & Nanayakkara S. (2023). Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, **11**, 1–17. doi : 10.1162/tacl_a_00530.

Stremmel J., Saeedi A., Hassanzadeh H., Batra S., Hertzberg J., Murillo J. & Halperin E. (2023). Xaiqa : Explainer-based data augmentation for extractive question answering. *arXiv preprint arXiv :2312.03567*.

Vold A. & Conrad J. G. (2021). Using transformers to improve answer retrieval for legal questions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, p. 245–249, New York, NY, USA : Association for Computing Machinery. doi : 10.1145/3462757.3466102.

Wang B., Zhao H., Zhou H., Song L., Xu M., Cheng W., Zeng X., Zhang Y., Huo Y., Wang Z. *et al.* (2025a). Baichuan-m1 : Pushing the medical capability of large language models. *arXiv preprint arXiv :2502.12671*.

Wang G., Gao M., Yang S., Zhang Y., He L., Huang L., Xiao H., Zhang Y., Li W., Chen L. *et al.* (2025b). Citrus : Leveraging expert cognitive pathways in a medical language model for advanced medical decision support. *arXiv preprint arXiv :2502.18274*.

Wang G., Ran J., Tang R., Chang C.-Y., Chang C.-Y., Chuang Y.-N., Liu Z., Braverman V., Liu Z. & Hu X. (2024). Assessing and enhancing large language models in rare disease question-answering.

Wang Y., Li P., Sun M. & Liu Y. (2023). Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 10303–10315, Singapore : Association for Computational Linguistics.

Wikipedia LLMs (2025). Language model benchmark — Wikipedia, the free encyclopedia. [Online ; accessed 10-March-2025].

Xu R., Cui H., Yu Y., Kan X., Shi W., Zhuang Y., Wang M. D., Jin W., Ho J. & Yang C. (2024). Knowledge-infused prompting : Assessing and advancing clinical text data generation with large language models. In L.-W. Ku, A. Martins & V. Srikumar, Éds., *Findings of the Association for Computational Linguistics : ACL 2024*, p. 15496–15523, Bangkok, Thailand : Association for Computational Linguistics. doi : 10.18653/v1/2024.findings-acl.916.

Zhang M., Dou S., Wang Z. & Wu Y. (2022). Focus-driven contrastive learning for medical question summarization. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond & S.-H. Na, Éds.,

Proceedings of the 29th International Conference on Computational Linguistics, p. 6176–6186, Gyeongju, Republic of Korea : International Committee on Computational Linguistics.

ZHAO B., BAI J., LI C., ZHANG J., RONG W., OUYANG Y. & XIONG Z. (2023). Enhancing biomedical reqa with adversarial hard in-batch negative samples. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **20**(5), 2933–2944. DOI : 10.1109/TCBB.2023.3261315.

ZHOU Z., YU K.-Y., TIAN S.-Y., SHI J.-X., YANG X.-W., SONG P., JIN Y.-X., GUO L.-Z. & LI Y.-F. (2025). LawGPT : Knowledge-guided data generation and its application to legal LLM. *arXiv preprint arXiv :2502.06572*.