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

Générer pour mieux tester : vers des datasets diversifiés pour une évaluation fiable des systèmes de QA

Louis Jourdain[1] Skander Hellal[1]
(1) ChapsVision, France
ljourdain@chapsvision.com, shellal@chapsvision.com

RÉSUMÉ

L'évaluation des modèles d'IA générative repose sur des datasets contenant des valeurs de référence attendues pour une entrée donnée. Cependant, la constitution de ces jeux de données est un processus complexe et coûteux. Cet article explore la génération automatique de datasets de questions diversifiées pour tester notamment les systèmes de RAG (*Retrieval Augmented Generation*). Nous proposons un cadre méthodologique combinant modèles de langage à grande échelle (LLMs) et techniques traditionnelles de traitement du langage naturel (NLP) et de data science, incluant les graphes de connaissances, la similarité sémantique voire le *topic modeling*. L'approche proposée repose sur un système modulaire exploitant diverses sources documentaires et intégrant des mécanismes avancés de filtrage afin de garantir la qualité et la diversité des questions produites.

ABSTRACT

Generate to Better Test the Generations : Towards Diverse Datasets for Reliable Evaluation of Question Answering Systems

The evaluation of generative AI models relies on datasets containing expected reference values for given inputs. However, building such datasets is a complex and costly process. This paper explores the automatic generation of diverse question datasets for testing Retrieval-Augmented Generation (RAG) systems. We propose a methodological framework that combines large language models (LLMs) with traditional natural language processing (NLP) and data science techniques, including knowledge graphs, semantic similarity, and even topic modeling. The proposed approach is based on a modular system that leverages various documentary sources and integrates advanced filtering mechanisms to ensure the quality and diversity of the generated questions.

MOTS-CLÉS : génération de données, évaluation, RAG, LLMs, NLP, diversité, dataset synthétique.

KEYWORDS: data generation, evaluation, RAG systems, LLMs, NLP, diversity, synthetic dataset.

Introduction

Si les LLMs ont démontré des performances remarquables sur de nombreuses tâches comme le résumé automatique ou la classification de texte (Brown *et al.*, 2020), leur évaluation pour les tâches génératives complexes demeure un défi majeur. On se concentrera sur la tâche de *Question Answering* (QA) qui a de nombreuses applications pratiques et des implémentations populaires comme l'architecture RAG (Lewis *et al.*, 2021).

Une évaluation rigoureuse de ces tâches suppose des jeux de données de qualité, ce qui est difficile en

Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM), pages 204–227.

raison de :

— leur coût et temps de collecte
— leur obsolescence rapide, car l'évolution rapide des modèles d'IA entraîne la saturation des benchmarks existants (Ott *et al.*, 2022).
— leur manque de diversité qui remet en question la pertinence de l'évaluation réalisée (Lima *et al.*, 2024).

Face à ces limitations, la génération de datasets synthétiques à partir d'une base documentaire s'est imposée comme une alternative prometteuse (Guo *et al.*, 2024). Cependant, l'adoption de ces jeux de données soulève plusieurs questions essentielles autour de leur qualité, de leur diversité et de leur fiabilité pour mener à bien une évaluation. Cet article propose une approche innovante de génération automatique de datasets de questions combinant modèles génératifs récents et techniques traditionnelles de NLP. Nous proposons de travailler sur la diversité des datasets synthétiques en multipliant les sources de données et les modes de génération des questions. Nous présenterons ensuite l'implémentation de notre système dont nous discuterons des perspectives d'amélioration. Les contributions principales de ce travail sont les suivantes :

1. Une synthèse des approches existantes sur l'utilisation de datasets synthétiques pour l'évaluation des tâches de question-réponse et de RAG.
2. Une réflexion critique sur la diversité des datasets synthétiques
3. La conception d'un système combinant LLMs avec des techniques de NLP (recherche vectorielle, graphes de connaissances) pour générer des questions plus diversifiées.

1 Des datasets synthétiques pour évaluer les systèmes de QA

1.1 Évaluer les tâches d'IA générative

Des tâches comme le résumé automatique ou le *Question Answering* qui nécessitaient des modèles spécifiques sont désormais réalisées par des LLMs sans entraînement préalable à partir d'instructions précises et de quelques exemples bien choisis (Brown *et al.*, 2020). L'élargissement des capacités de l'IA à des tâches plus complexes rend leur évaluation plus difficile. L'absence de métriques standardisées pour ces nouvelles tâches est problématique car les métriques classiques sont obsolètes, et si de nouvelles méthodes comme *LLM as a Judge*, sont explorées (Gu *et al.*, 2025), leur fiabilité reste incertaine (Shankar *et al.*, 2024). L'amélioration des performances des modèles conduit à l'obsolescence rapide des benchmarks existants. Des jeux de données comme TriviaQA (Joshi *et al.*, 2017) autrefois standards sont devenus trop simples pour des modèles modernes comme GPT-4, qui atteignent des performances humaines (OpenAI *et al.*, 2024). Cette saturation nuit à l'évaluation fine des progrès des modèles (Ott *et al.*, 2022). La difficulté à évaluer de nouvelles tâches de NLP comme le RAG traduit ce changement de paradigme. L'évaluation des systèmes RAG requiert des métriques précises intégrant pertinence, cohérence et fidélité des réponses(Chen *et al.*, 2023). Il est ainsi indispensable de disposer de datasets de référence contenant des réponses attendues (*gold answers*) et les étapes intermédiaires du système. Or, ces jeux de données doivent être suffisamment diversifiés et représentatifs des requêtes réelles pour garantir une évaluation pertinente (Kim *et al.*, 2024). La constitution manuelle de jeux de données atteint ses limites, ce qui positionne la génération automatique comme une alternative méthodologiquement robuste.

1.2 Des datasets synthétiques pour évaluer

L'annotation manuelle, longue et coûteuse, rend la constitution de datasets difficilement soutenable. Les datasets synthétiques, générés à partir de corpus documentaires, sont devenus incontournables en *machine learning* et l'essor LLMs a encouragé cette approche. Initialement destinés à l'entraînement des modèles (Guo *et al.*, 2024; Josifoski *et al.*, 2023; Mao *et al.*, 2024) ces datasets synthétiques sont désormais exploités pour tester les performances des systèmes de *Question Answering*, où la conservation des sources d'information est essentielle(Braga *et al.*, 2024; Jeronymo *et al.*, 2023). L'usage des LLMs pour la génération de datasets offre a priori une solution rapide, économique et souple, ce qui peut être particulièrement précieux pour les domaines où la collecte de données annotées est difficile en raison de contraintes légales, éthiques ou d'un manque de ressources (Mishra *et al.*, 2025; Pranida *et al.*, 2025).

1.3 Limitations des datasets synthétiques

L'utilisation de datasets synthétiques pour l'évaluation soulève toutefois plusieurs problèmes. Si certains tels que les biais et les hallucinations (Huang *et al.*, 2025) sont connus de la discipline, la pertinence d'une évaluation basée sur des données artificielles reste débattue. Certains travaux explorent les limites des datasets synthétiques pour l'entraînement (Liu *et al.*, 2024) ou leur valeur pour l'évaluation (Long *et al.*, 2024). Un enjeu clé est la qualité et la représentativité des jeux de données qui doivent refléter la diversité des productions humaines (Maheshwari *et al.*, 2024). Bien que certaines études montrent une corrélation entre les scores obtenus sur des datasets traditionnels et synthétiques, cette validation reste reste précaire et sujette à des biais de confirmation.

Les datasets synthétiques pèchent également par leur manque de diversité, ce qui peut dégrader les performances des modèles d'IA (Zhu *et al.*, 2024). Des travaux récents, comme MIMDE (Francis *et al.*, 2024), soulignent que les datasets synthétiques sont souvent plus homogènes et prévisibles. Les LLMs ont également tendance à produire des questions plus longues et factuelles que les annotateurs humains (Zhang *et al.*, 2025). En outre, la diversité des intentions de recherche complique le transfert de performances d'une tâche à l'autre (Dai *et al.*, 2022). Enfin, les benchmarks existants étant eux-mêmes des artefacts, leur alignement avec les requêtes réelles des utilisateurs reste incertain.

2 Diversifier la façon de poser les questions

La diversité des questions est un enjeu central pour la qualité des jeux de données en QA, mais elle est souvent sous-estimée. Assurer une réelle diversité ne se limite pas à une simple génération aléatoire de questions ; elle nécessite une approche théorique rigoureuse et une implémentation réfléchie à chaque étape du processus, depuis la sélection des sources documentaires à la génération des questions, en passant par le filtrage des sorties produites.

2.1 Une diversité limitée dans les systèmes existants

Les travaux existants montrent un décalage entre l'évolution des systèmes de QA et la nature des questions utilisées pour les évaluer. Alors que les modèles récents sont capables de raisonnements

complexes, les benchmarks actuels restent largement dominés par des questions factuelles simples (Lima *et al.*, 2024). Par exemple, une analyse menée sur HotpotQA (Yang *et al.*, 2018) révèle que plus de la moitié des questions posées sont purement factuelles. Si les travaux scientifiques reconnaissent les défis posés par les données synthétiques (hallucinations, biais, perte de qualité) (Liu *et al.*, 2024), rares sont ceux qui proposent une méthodologie rigoureuse pour garantir une diversité réelle des questions générées.

La majorité des approches récentes, qu'elles soient issues de l'industrie (Amazon Bedrock, Microsoft Azure, Databricks, Hugging Face) ou du monde académique, adoptent un schéma classique de type *generate then filter*, où un LLM produit des questions à partir d'un document source avant filtrage des sorties inadaptées. En l'absence de consignes ciblées, les modèles privilégient toutefois des questions simples et factuelles. Les tentatives actuelles pour assurer une certaine diversité reposent principalement sur la variation de prompts pour induire des catégories de questions différentes, une stratégie souvent trop limitée pour refléter toute la richesse des formulations attendues. Plusieurs classifications des types de questions ont été proposées dans la littérature et les benchmarks, mais elles restent généralement peu opérationnalisées dans les pipelines de génération. La première annexe récapitule les typologies proposées dans divers travaux. La majorité distingue les questions factuelles de celles nécessitant un raisonnement. D'autres critères fréquemment mobilisés incluent la longueur de la réponse attendue, le type de réponse ou encore le style de formulation de la question. Cette hétérogénéité des critères et l'absence de consensus soulignent les limites des taxonomies actuellement utilisées.

2.2 Construire une taxonomie

Une taxonomie pertinente doit refléter de manière fidèle la diversité des questions possibles. Or, de nombreux critères différents ont été pris en compte dans les travaux que l'on a mentionnés. Parmi eux :

- Un critère stylistique qui paraît faible pour classer l'ensemble des questions possibles.
- Le degré d'intelligibilité et interprétabilité de la question.
- Le type d'opération qu'un système de QA doit effectuer pour y réponde (simple recherche d'information, raisonnement et déduction à partir de plusieurs données...). C'est le critère le plus fréquemment utilisé dans la littérature mais il présente un risque de circularité (on ne génère que des questions que le système peut techniquement traiter).
- Le contenu de la réponse attendue (HotpotQA), reste un critère descriptif, fortement tributaire du document source et de la question.
- Notion d'informativité (informativeness) d'une question (Mazzaccara *et al.*, 2024), soit la quantité d'information qu'elle véhicule, qui est très complexe à calculer.
- Des critères linguistiques (structure syntaxique, questions WH/oui-non) qui varient selon les langues et ne prennent pas en compte la dimension sémantique.
- Des critères sémantiques pensés pour capturer l'intention de l'utilisateur (Nielsen *et al.*, 2008).
- Enfin, une piste encore peu explorée consiste à inverser la démarche : générer une question à partir des informations requises pour y répondre.

Ces critères peinent à couvrir l'ensemble de l'espace des questions possibles car ils induisent des découpages souvent incompatibles ou redondants. Une alternative consisterait à les traiter non comme des catégories fixes, mais comme des paramètres à faire varier, au prix toutefois d'une explosion combinatoire. Plutôt que d'imposer une typologie arbitraire à la génération, il serait plus efficace de structurer la diversité en amont en diversifiant les sources de connaissance utilisées. Ce changement

de perspective pourrait s'avérer plus efficace pour refléter la richesse des datasets humains.

2.3 Mesurer la diversité

Classer les questions en catégories peut aider à évaluer la diversité d'un dataset mais reste insuffisant car toute typologie est critiquable. De plus, lorsqu'un système de génération s'appuie sur une typologie pour structurer ses sorties, puis l'utilise pour prouver leur diversité, il court le risque d'une prophétie auto-réalisatrice. Évaluer la diversité exige une définition claire du concept, appuyée sur des métriques réellement informatives. Diverses approches quantitatives ont été proposées pour évaluer la diversité (Zhu *et al.*, 2025), notamment des mesures formelles basées sur les n-grammes, ou des comparaisons avec des données de référence. Les méthodes par projection dans un espace d'*embeddings*, comparant les textes sur la base de leur contenu sémantique, stylistique ou intentionnel (Tevet & Berant, 2021) apparaissent plus prometteuses à condition que des distances vectorielles soient vraiment pertinentes pour capturer la diversité perçue par des humains.

Pour l'instant, aucune de ces techniques ne s'impose comme solution évidente, et plusieurs points restent à clarifier :

— Doit-on uniquement évaluer la diversité des questions, ou aussi inclure les réponses pour le calcul d'une métrique ? Les documents sélectionnés ?
— Ces mesures fonctionnent le mieux sur des textes longs alors que les questions sont des textes très brefs.
— La diversité linguistique des questions est directement liée à la diversité thématique et linguistique des documents source.

Il n'y a à ce jour pas de protocole solide et validé par l'ensemble de la communauté ni pour mesurer la diversité d'un dataset de questions, ni même pour comparer la diversité de deux datasets.

2.4 Diversifier les données utilisées pour générer les questions

Plutôt que de partir d'un modèle théorique des types de questions, il est prometteur d'explorer des moyens variés de générer des questions à partir de différentes sources de connaissance car le manque de diversité des jeux de données synthétiques en QA découle souvent d'une dépendance trop forte aux documents sources. En mobilisant les outils de data science et de TAL pour diversifier ces sources (documents multiples, structures sémantiques, inférences, etc.), on favorise *de facto* la production de questions inédites et variées.

Du document isolé au groupe : Au lieu de n'utiliser qu'un document pour générer une question comme c'est couramment pratiqué dans la littérature, on peut utiliser plusieurs documents pour générer des questions portant sur la synthèse ou la confrontation des informations qu'ils contiennent, ce qui est pertinent pour évaluer des solutions RAG qui récupèrent plusieurs documents. Pour ce faire, il est nécessaire de sélectionner des documents au hasard mais des documents proches sémantiquement.

Les graphes de connaissances : Les systèmes de QA qui utilisent une base documentaire ont une vue limitée au contenu des documents qui leur sont fournis et peinent à amalgamer des informations disséminées dans la base contrairement aux graphes de connaissances qui stockent les relations longue distance, d'où leur utilisation pour le QA (Feng *et al.*, 2023). On peut à partir de graphes créer des questions testant l'exhaustivité ou portant sur le lien entre deux éléments qui n'apparaissent pas dans un même document. Cette piste a été développée par le framework open source RAGAS (Es *et al.*,

2023) mais ce système reste difficilement interprétable sur la génération des graphes, difficilement contrôlable, coûteux en appels LLM, et son impact réel sur la diversité des questions générées n'est pas clairement établi.

Les techniques de clustering et de topic modeling : Les utilisateurs posent parfois des questions très génériques ou souhaitent explorer rapidement l'ensemble de leur base documentaire. Or les systèmes de QA se focalisent sur des informations précises. Des techniques comme le clustering et le *topic modeling* permettent à moindre coût computationnel de regrouper les documents similaires (librairies telles que BERTTopic (Grootendorst, 2022)).

3 Construire un générateur de dataset

La génération de questions à partir d'une base documentaire est la tâche complémentaire à celle de *Question Answering*. Il est donc naturel que les techniques utilisées pour les deux tâches évoluent conjointement (Goyal & Mahmoud, 2024). On doit concevoir la génération de datasets comme un véritable problème d'ingénierie logicielle, et non uniquement comme une tâche relevant de l'IA ou du TAL. En l'envisageant comme un pipeline modulaire, composé d'étapes claires (extraction, transformation, génération, filtrage, validation), on peut mieux en contrôler la robustesse, l'évolutivité et la traçabilité, tout en assurant la reproductibilité et la diversité des questions produites.

3.1 Historique des techniques de génération de datasets

Les premières méthodes de génération de datasets de QA reposaient sur des règles simples, comme la suppression d'un élément dans une phrase ou un triplet de graphe, une approche peu coûteuse mais générant des questions monotones. L'apparition des modèles de langue pré-entraînés, spécialisés pour la génération de questions (Zhou *et al.*, 2017), a permis d'obtenir des questions plus variées mais au prix d'un entraînement spécifique, ce qui n'est plus le cas des LLMs.

L'utilisation des LLMs est récemment plébiscitée pour générer des datasets de QA employés plus souvent pour spécialiser des *embeddings* que pour évaluer. Comme le notent Zhu et al. (Zhu *et al.*, 2025), l'emploi de LLMs pour créer des générateurs de datasets peut être développé dans plusieurs directions, soit par *fine-tuning* des LLMs pour la génération de questions (Xu *et al.*, 2025) soit par utilisation de prompts spécifiques pour guider le LLM à générer certains types de questions. L'usage des LLMs a facilité la génération de datasets et amélioré la qualité des questions, mais s'y limiter réduit cette tâche à du *prompt engineering* et cache la complexité inhérente à la tâche.

3.2 L'importance de l'amont : Diversifier la source d'informations

Si la phase de génération de questions est essentielle, les étapes antérieures de sélection des données fournies dans le prompt sont cruciales et souvent négligées. Long et al. (Long *et al.*, 2024) préconisent soit l'entraînement de modèles spécialisés capables de maîtriser certains paramètres de génération, soit de s'inspirer d'approches de type RAG pour intégrer des connaissances spécifiques. (Ziegler *et al.*, 2024). Ces réflexions soulignent l'importance de mécanismes de récupération pour la génération de questions. La phase de sélection des documents (sampling) pour générer les questions est également importante. Là où les documents sont le plus souvent sélectionnés aléatoirement, il pourrait

être intéressant d'étudier des méthodes de *sampling* plus raffinées, qui favorisent par exemple des documents considérés comme plus représentatifs de la base de données (plus centraux dans l'espace vectoriel latent). Cela permettrait de limiter les risques liés aux aléas de la constitution de la base documentaire.

3.3 Les différents modes de création des questions

Les approches employant des LLMs pour la génération de datasets reposent sur la rédaction d'un ou de plusieurs prompts spécifiques prenant en argument un ou plusieurs extraits de documents. Il est important de raffiner cette étape de génération de questions sur plusieurs points. L'emploi de techniques plus avancées de *prompt engineering* telles que le *few-shot prompting* ou l'utilisation de *persona* est susceptible d'influencer la qualité des datasets générés. Aucun travail n'a mené d'études poussées pour évaluer leur influence. Des approches par chaînage de prompt, demandant par exemple d'extraire d'un document des prédicats logiques et interroger sur les conséquences de ces prédicats, permettent de mieux tester les capacités de raisonnement des systèmes de QA génératifs (Liu *et al.*, 2025). Alors que la plupart des systèmes ne génèrent qu'une seule version par question, reformuler les questions initiales constitue une stratégie pertinente pour enrichir la diversité. Certaines approches, comme celles d'Amazon ou de Deepeval, introduisent du bruit ou modifient le style afin de tester la robustesse des systèmes à la variabilité linguistique. Ces transformations peuvent toutefois générer des questions non ancrées dans le document source, voire issues d'hallucinations.

3.4 L'importance de l'aval : tri des questions

Si les questions générées par des LLMs à partir de documents sont en général bien formulées et fidèles aux informations fournies en contexte, les LLM génèrent parfois des contenus fictifs même dans des tâches contraintes. À titre d'exemple, notre système a produit la question suivante : "Quels sont les thèmes principaux abordés dans le livre de Jean Michelin, Jonquille ?". Or, le document utilisé pour générer cette question était un extrait d'interview dans lequel le nom de l'auteur n'apparaissait pas. Le LLM produit également des questions qui font directement référence au contexte donné et auxquelles on ne peut pas interpréter sans accès au contexte ("Quel est le produit présenté dans cet article ?")

Une question générée peut être évaluée selon plusieurs critères : le fait que l'on puisse répondre à la question (answerability), le fait que la question soit bien basée sur le document fourni en contexte (**groundedness**), le fait que la question soit pertinente pour l'utilisateur final (**relevance**), le fait que l'on puisse répondre à la question à partir du document uniquement (**stand-alone**), ou encore le fait que la question challenge les connaissances du LLM et sa bonne utilisation des données fournies en contexte, et ne porte pas sur une information générale (**uncommoness**)

C'est pourquoi la plupart des systèmes existants reposent sur le principe de *generate then filter* avec des heuristiques pour éliminer les questions déficientes. Le filtrage sur les datasets peut améliorer de plusieurs points les performances du système final (Dai *et al.*, 2022). L'approche la plus utilisée pour ce tri est d'évaluer les questions générées avec un LLM (Schimanski *et al.*, 2024), ce qui reste coûteux et peu fiable. D'autres méthodes (Alberti *et al.*, 2019) proposent de mesurer la pertinence de la question (*round-trip consistency*) en utilisant un système de QA pour montrer que l'on peut bien répondre à la question. L'inconvénient est que l'on dépend de la performance du système de QA

pour juger de la pertinence de la question et qu'on éliminera des questions pertinentes auxquelles le système de QA échoue. Une approche complètement différente consisterait à utiliser un *reranker*, un modèle entraîné à produire un score numérique jugeant la pertinence d'un document pour répondre à une question donnée.

4 Proposition de pipeline de génération de question

La génération de dataset de questions ne saurait donc se réduire à une simple tâche de *prompt engineering* mais nécessite le développement d'un pipeline complet prenant en compte la sélection d'information, des mécanismes complexes de génération de la question et une phase de tri du produit final. Nous avons conçu un système permettant de produire un nombre défini de questions à partir d'une base documentaire. Nous assurons la diversité en variant les méthodes de génération selon la structure des documents, et l'architecture, indépendante du type de base de données, repose sur des connecteurs.

4.1 Description technique du système

Le pipeline permet de construire un dataset à partir d'une base documentaire indexée dans une base de données dotée de la recherche vectorielle. Les vecteurs ont été calculés avec le modèle BAAI/bge-m3 et les textes stockés dans le moteur de recherche Elastic. Les textes ont été enrichis à l'aide d'un pipeline de NLP hybride interne à l'entreprise, combinant approches par règles (ontologie construite sur le corpus), modèles légers et LLM pour détecter les entités, organisations et concepts ainsi que les relations les connectant (triplets). Ces relations forment un graphe de connaissance qui est stocké dans un second index. Le système permet la sélection du modèle de génération.

La seconde annexe présente un schéma de l'architecture. Il est possible de construire des générateurs de questions qui cherchent dans les index des données spécifiques (un ou plusieurs documents, une sous-section du graphe), intègrent ces entrées dans des prompts adaptés et génèrent une question. La troisième annexe présente les différents types de questions actuellement implémentés dans le système. Ces questions satisfont plusieurs critères de diversité : différence de formulation, de typologie linguistique, de type de réponse attendue, d'étapes nécessaires pour répondre à la question... L'ajout d'un nouveau type de question est possible en codant la récupération de certaines données depuis les bases (k documents sémantiquement proches, une marche aléatoire dans le graphe...) puis l'intégration de ces données dans un schéma de prompt. Certains des prompts utilisés sont disponibles en quatrième annexe. Cette conception modulaire permet d'ajouter facilement de nouveaux types de questions. Par exemple, les questions *multi-hop* sont souvent difficiles pour les systèmes de QA. L'annexe d'un article (Gandhi *et al.*, 2024) suggère rapidement une idée pour générer des questions*multi-hop* par composition de sous-questions portant sur une entité commune. Cette approche s'intègre aisément à notre architecture. Le générateur agit comme un orchestrateur, articulant des modules de récupération d'information et des étapes de génération pilotées par des LLMs. Cette approche rejoint les tendances identifiées par Guo et al. (Guo *et al.*, 2024), qui suggèrent que la conception d'un générateur de données automatisé présente des similitudes avec le développement d'agents IA industriels ("*We believe it would also be quite valuable to develop a data generation agent for industrial applications.*").

4.2 Évaluation de la diversité des datasets générés

Pour évaluer la diversité des questions générées, nous invitons le lecteur à consulter un ensemble de 600 questions produites à partir d'articles de presse sur le conflit en Ukraine, disponible sur Hugging Face. Les documents sources ne sont pas publiés pour des raisons de droits d'auteur. Afin de démontrer que notre système produit des questions plus variées qu'une approche naïve par simple prompt, nous avons sélectionné 100 documents sur le conflit au Mali. À partir de ceux-ci, nous avons produit 100 questions avec notre système (moitié simples, 1/6 multi-documents, 1/6 à base de graphes, 1/6 volontairement dégradées) et 100 questions avec un prompt naïf. Certaine sont disponibles en annexe. Le modèle utilisé est Llama 3.1 7B (Grattafiori *et al.*, 2024), avec une température basse (0.1). Les documents ont été sélectionnés aléatoirement et un même document peut être utilisé dans plusieurs instances de génération.Une seconde expérience envisagée consisterait à générer 25 questions à partir d'un même document. Aucun protocole robuste et consensuel n'existe à ce jour pour mesurer, ni même comparer, la diversité de jeux de données de questions. Toutefois, plusieurs indicateurs peuvent être mobilisés. Sur le plan lexical, on peut calculer le Type-Token Ratio (TTR), l'*overlap* de n-grammes, ou encore l'entropie de la distribution des mots. Sur le plan sémantique, des représentations vectorielles des questions permettent d'estimer la diversité via des mesures de centralité, comme la similarité moyenne entre paires de vecteurs.

Métrique	TTR	Distinct-1	Distinct-2	Distinct-3	Entropy	Embedding Diversity
Vanilla	0.27	0.27	0.50	0.58	5.04	0.60
Notre système	0.27	0.27	**0.59**	**0.73**	**5.16**	**0.62**

TABLE 1 – Tentative d'analyse quantitative de la diversité des datasets générés

La projection des vecteurs de questions (après réduction de dimensionnalité), présentée en dernière annexe, ne révèle pas non plus de différence manifeste, les variations sémantiques pouvant par ailleurs être expliquées par la nature différente des sources utilisées.

4.3 Perspectives et critiques

Notre méthode d'évaluation quantitative ne garantit pas une description satisfaisante de la diversité attendue, la mesure de diversité à partir des *embeddings* présente des limites. Par exemple, les questions de type oui/non et celles à choix multiples diffèrent structurellement, ce qui fausse leur comparaison avec les questions factuelles. De plus, la différence de sens entre deux questions n'est pas toujours un indicateur pertinent de leur diversité réelle. Une alternative plus rigoureuse consisterait à exploiter un dataset annoté manuellement et à comparer la diversité des questions générées avec celles issues de ce corpus. Toutefois, les jeux de données disponibles datent pour la plupart de l'ère du QA extractif et sont principalement conçus pour tester les capacités d'extraction, ce qui en fait des artefacts non représentatifs des usages des utilisateurs. Une analyse approfondie devrait intégrer une comparaison qualitative des productions issues de différents LLMs, ainsi qu'une étude d'ablation sur les composants du pipeline. Un axe d'amélioration consisterait à permettre la création de datasets hybrides en combinant les résultats de plusieurs modèles génératifs. Enfin, l'utilisation de jeux de données réels issus d'applications RAG en production pourrait permettre d'aligner les questions générées sur les attentes des utilisateurs. Une approche complémentaire serait d'intégrer

des méthodes de personnalisation des datasets en fonction des profils d'utilisateurs et des domaines d'application, comme suggéré par Braga et al. (Braga *et al.*, 2024). Cette démarche permettrait d'aligner les questions générées sur les besoins réels des systèmes et d'en renforcer la pertinence contextuelle.

Certains axes nécessiteraient ainsi une exploration approfondie :

- **Évaluation via des modèles de QA :** Il serait pertinent de tester plusieurs modèles de QA à partir des questions générées, afin de vérifier si celles-ci présentent bien un défi accru par rapport aux benchmarks existants.
- **Validation de la diversité :** Une comparaison des performances des modèles sur des datasets humains et synthétiques, accompagnée d'analyses quantitatives de la diversité et de la pertinence, permettrait de vérifier la corrélation entre scores sur dataset synthétique et dataset manuel.
- **Uné étude d'ablation sur les différents composants de l'architecture :** Le générateur repose largement sur des LLMs, mais l'article ne détaille pas suffisamment l'influence du choix du modèle sur la diversité et la qualité des questions produites ou encore la nécessité d'ajuster les hyperparamètres du LLM pour optimiser la diversité et l'adéquation des questions générées.
- **Une mise à l'épreuve sur une diversité de corpus documentaires :**. L'efficacité du générateur est conditionné par la qualité et de la structure des documents sources, mais cela n'est pas pleinement exploré. Il faudrait étudier comment le système réagit-il à des bases documentaires hétérogènes.
- **Une comparaison plus poussée avec l'état de l'art existant :**. Il serait utile de positionner l'approche par rapport aux autres méthodes de génération, ce qui est complexe car beaucoup sont des solutions industrielles.

Conclusion

Face à la saturation des datasets existants et à la difficulté d'en constituer de nouveaux, la génération automatique de datasets synthétiques apparaît comme une solution prometteuse et adaptable aux besoins spécifiques des systèmes évalués. Toutefois, il faut s'assurer qu'ils soient d'une diversité suffisante pour mettre au défi les systèmes. Si la génération assistée par LLMs permet d'automatiser et d'accélérer la création de jeux de données, elle nécessite une supervision rigoureuse et des stratégies de contrôle qualité. De plus, l'absence de métriques standardisées pour mesurer la diversité des questions limite la comparabilité des approches et soulève des interrogations quant à la fiabilité des évaluations obtenues. Cette étude propose d'ouvrir la voie à des systèmes d'évaluation plus adaptatifs, capables de suivre l'évolution rapide des modèles d'IA et d'accompagner le développement de nouvelles capacités de raisonnement et de génération de contenu. Toutefois, la question centrale demeure : un dataset synthétique peut-il réellement capturer toute la complexité des questions humaines ? C'est en intégrant une réflexion sur la qualité des données que l'on pourra garantir des évaluations fiables et exploitables pour le développement de systèmes d'IA toujours plus performants.

Références

ALBERTI C., ANDOR D., PITLER E., DEVLIN J. & COLLINS M. (2019). Synthetic QA Corpora Generation with Roundtrip Consistency. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 6168–6173, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1620.

BOLOTOVA V., BLINOV V., SCHOLER F., CROFT W. B. & SANDERSON M. (2022). A Non-Factoid Question-Answering Taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 1196–1207, Madrid Spain : ACM. DOI : 10.1145/3477495.3531926.

BRAGA M., KASELA P., RAGANATO A. & PASI G. (2024). Synthetic Data Generation with Large Language Models for Personalized Community Question Answering. arXiv :2410.22182 [cs], DOI : 10.48550/arXiv.2410.22182.

BROWN T. B., MANN B., RYDER N., SUBBIAH M., KAPLAN J., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D. M., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language Models are Few-Shot Learners. arXiv :2005.14165 [cs], DOI : 10.48550/arXiv.2005.14165.

CHEN J., LIN H., HAN X. & SUN L. (2023). Benchmarking Large Language Models in Retrieval-Augmented Generation. arXiv :2309.01431 [cs], DOI : 10.48550/arXiv.2309.01431.

DAI Z., ZHAO V. Y., MA J., LUAN Y., NI J., LU J., BAKALOV A., GUU K., HALL K. B. & CHANG M.-W. (2022). Promptagator : Few-shot Dense Retrieval From 8 Examples. arXiv :2209.11755 [cs], DOI : 10.48550/arXiv.2209.11755.

ES S., JAMES J., ESPINOSA-ANKE L. & SCHOCKAERT S. (2023). RAGAS : Automated Evaluation of Retrieval Augmented Generation. arXiv :2309.15217 [cs], DOI : 10.48550/arXiv.2309.15217.

FENG S., BALACHANDRAN V., BAI Y. & TSVETKOV Y. (2023). FactKB : Generalizable Factuality Evaluation using Language Models Enhanced with Factual Knowledge. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 933–952, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.59.

FILICE S., HOROWITZ G., CARMEL D., KARNIN Z., LEWIN-EYTAN L. & MAAREK Y. (2025). Generating Diverse Q&A Benchmarks for RAG Evaluation with DataMorgana. arXiv :2501.12789 [cs] version : 1, DOI : 10.48550/arXiv.2501.12789.

FRANCIS J., ESNAASHARI S., POLETAEV A., CHAKRABORTY S., HASHEM Y. & BRIGHT J. (2024). MIMDE : Exploring the Use of Synthetic vs Human Data for Evaluating Multi-Insight Multi-Document Extraction Tasks. arXiv :2411.19689 [cs], DOI : 10.48550/arXiv.2411.19689.

GANDHI S., GALA R., VISWANATHAN V., WU T. & NEUBIG G. (2024). Better Synthetic Data by Retrieving and Transforming Existing Datasets. arXiv :2404.14361 [cs], DOI : 10.48550/arXiv.2404.14361.

GOYAL M. & MAHMOUD Q. H. (2024). A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics*, **13**(17), 3509. Number : 17 Publisher : Multidisciplinary Digital Publishing Institute, DOI : 10.3390/electronics13173509.

GRATTAFIORI A., DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AL-DAHLE A., LETMAN A., MATHUR A., SCHELTEN A., VAUGHAN A., YANG A., FAN A., GOYAL A., HARTSHORN

A., Yang A., Mitra A., Sravankumar A., Korenev A., Hinsvark A., Rao A., Zhang A., Rodriguez A., Gregerson A., Spataru A., Roziere B., Biron B., Tang B., Chern B., Caucheteux C., Nayak C., Bi C., Marra C., McConnell C., Keller C., Touret C., Wu C., Wong C., Ferrer C. C., Nikolaidis C., Allonsius D., Song D., Pintz D., Livshits D., Wyatt D., Esiobu D., Choudhary D., Mahajan D., Garcia-Olano D., Perino D., Hupkes D., Lakomkin E., AlBadawy E., Lobanova E., Dinan E., Smith E. M., Radenovic F., Guzmán F., Zhang F., Synnaeve G., Lee G., Anderson G. L., Thattai G., Nail G., Mialon G., Pang G., Cucurell G., Nguyen H., Korevaar H., Xu H., Touvron H., Zarov I., Ibarra I. A., Kloumann I., Misra I., Evtimov I., Zhang J., Copet J., Lee J., Geffert J., Vranes J., Park J., Mahadeokar J., Shah J., Linde J. v. d., Billock J., Hong J., Lee J., Fu J., Chi J., Huang J., Liu J., Wang J., Yu J., Bitton J., Spisak J., Park J., Rocca J., Johnstun J., Saxe J., Jia J., Alwala K. V., Prasad K., Upasani K., Plawiak K., Li K., Heafield K., Stone K., El-Arini K., Iyer K., Malik K., Chiu K., Bhalla K., Lakhotia K., Rantala-Yeary L., Maaten L. v. d., Chen L., Tan L., Jenkins L., Martin L., Madaan L., Malo L., Blecher L., Landzaat L., Oliveira L. d., Muzzi M., Pasupuleti M., Singh M., Paluri M., Kardas M., Tsimpoukelli M., Oldham M., Rita M., Pavlova M., Kambadur M., Lewis M., Si M., Singh M. K., Hassan M., Goyal N., Torabi N., Bashlykov N., Bogoychev N., Chatterji N., Zhang N., Duchenne O., Çelebi O., Alrassy P., Zhang P., Li P., Vasic P., Weng P., Bhargava P., Dubal P., Krishnan P., Koura P. S., Xu P., He Q., Dong Q., Srinivasan R., Ganapathy R., Calderer R., Cabral R. S., Stojnic R., Raileanu R., Maheswari R., Girdhar R., Patel R., Sauvestre R., Polidoro R., Sumbaly R., Taylor R., Silva R., Hou R., Wang R., Hosseini S., Chennabasappa S., Singh S., Bell S., Kim S. S., Edunov S., Nie S., Narang S., Raparthy S., Shen S., Wan S., Bhosale S., Zhang S., Vandenhende S., Batra S., Whitman S., Sootla S., Collot S., Gururangan S., Borodinsky S., Herman T., Fowler T., Sheasha T., Georgiou T., Scialom T., Speckbacher T., Mihaylov T., Xiao T., Karn U., Goswami V., Gupta V., Ramanathan V., Kerkez V., Gonguet V., Do V., Vogeti V., Albiero V., Petrovic V., Chu W., Xiong W., Fu W., Meers W., Martinet X., Wang X., Wang X., Tan X. E., Xia X., Xie X., Jia X., Wang X., Goldschlag Y., Gaur Y., Babaei Y., Wen Y., Song Y., Zhang Y., Li Y., Mao Y., Coudert Z. D., Yan Z., Chen Z., Papakipos Z., Singh A., Srivastava A., Jain A., Kelsey A., Shajnfeld A., Gangidi A., Victoria A., Goldstand A., Menon A., Sharma A., Boesenberg A., Baevski A., Feinstein A., Kallet A., Sangani A., Teo A., Yunus A., Lupu A., Alvarado A., Caples A., Gu A., Ho A., Poulton A., Ryan A., Ramchandani A., Dong A., Franco A., Goyal A., Saraf A., Chowdhury A., Gabriel A., Bharambe A., Eisenman A., Yazdan A., James B., Maurer B., Leonhardi B., Huang B., Loyd B., Paola B. D., Paranjape B., Liu B., Wu B., Ni B., Hancock B., Wasti B., Spence B., Stojkovic B., Gamido B., Montalvo B., Parker C., Burton C., Mejia C., Liu C., Wang C., Kim C., Zhou C., Hu C., Chu C.-H., Cai C., Tindal C., Feichtenhofer C., Gao C., Civin D., Beaty D., Kreymer D., Li D., Adkins D., Xu D., Testuggine D., David D., Parikh D., Liskovich D., Foss D., Wang D., Le D., Holland D., Dowling E., Jamil E., Montgomery E., Presani E., Hahn E., Wood E., Le E.-T., Brinkman E., Arcaute E., Dunbar E., Smothers E., Sun F., Kreuk F., Tian F., Kokkinos F., Ozgenel F., Caggioni F., Kanayet F., Seide F., Florez G. M., Schwarz G., Badeer G., Swee G., Halpern G., Herman G., Sizov G., Guangyi, Zhang, Lakshminarayanan G., Inan H., Shojanazeri H., Zou H., Wang H., Zha H., Habeeb H., Rudolph H., Suk H., Aspegren H., Goldman H., Zhan H., Damlaj I., Molybog

I., TUFANOV I., LEONTIADIS I., VELICHE I.-E., GAT I., WEISSMAN J., GEBOSKI J., KOHLI J., LAM J., ASHER J., GAYA J.-B., MARCUS J., TANG J., CHAN J., ZHEN J., REIZENSTEIN J., TEBOUL J., ZHONG J., JIN J., YANG J., CUMMINGS J., CARVILL J., SHEPARD J., MCPHIE J., TORRES J., GINSBURG J., WANG J., WU K., U K. H., SAXENA K., KHANDELWAL K., ZAND K., MATOSICH K., VEERARAGHAVAN K., MICHELENA K., LI K., JAGADEESH K., HUANG K., CHAWLA K., HUANG K., CHEN L., GARG L., A L., SILVA L., BELL L., ZHANG L., GUO L., YU L., MOSHKOVICH L., WEHRSTEDT L., KHABSA M., AVALANI M., BHATT M., MANKUS M., HASSON M., LENNIE M., RESO M., GROSHEV M., NAUMOV M., LATHI M., KENEALLY M., LIU M., SELTZER M. L., VALKO M., RESTREPO M., PATEL M., VYATSKOV M., SAMVELYAN M., CLARK M., MACEY M., WANG M., HERMOSO M. J., METANAT M., RASTEGARI M., BANSAL M., SANTHANAM N., PARKS N., WHITE N., BAWA N., SINGHAL N., EGEBO N., USUNIER N., MEHTA N., LAPTEV N. P., DONG N., CHENG N., CHERNOGUZ O., HART O., SALPEKAR O., KALINLI O., KENT P., PAREKH P., SAAB P., BALAJI P., RITTNER P., BONTRAGER P., ROUX P., DOLLAR P., ZVYAGINA P., RATANCHANDANI P., YUVRAJ P., LIANG Q., ALAO R., RODRIGUEZ R., AYUB R., MURTHY R., NAYANI R., MITRA R., PARTHASARATHY R., LI R., HOGAN R., BATTEY R., WANG R., HOWES R., RINOTT R., MEHTA S., SIBY S., BONDU S. J., DATTA S., CHUGH S., HUNT S., DHILLON S., SIDOROV S., PAN S., MAHAJAN S., VERMA S., YAMAMOTO S., RAMASWAMY S., LINDSAY S., LINDSAY S., FENG S., LIN S., ZHA S. C., PATIL S., SHANKAR S., ZHANG S., ZHANG S., WANG S., AGARWAL S., SAJUYIGBE S., CHINTALA S., MAX S., CHEN S., KEHOE S., SATTERFIELD S., GOVINDAPRASAD S., GUPTA S., DENG S., CHO S., VIRK S., SUBRAMANIAN S., CHOUDHURY S., GOLDMAN S., REMEZ T., GLASER T., BEST T., KOEHLER T., ROBINSON T., LI T., ZHANG T., MATTHEWS T., CHOU T., SHAKED T., VONTIMITTA V., AJAYI V., MONTANEZ V., MOHAN V., KUMAR V. S., MANGLA V., IONESCU V., POENARU V., MIHAILESCU V. T., IVANOV V., LI W., WANG W., JIANG W., BOUAZIZ W., CONSTABLE W., TANG X., WU X., WANG X., WU X., GAO X., KLEINMAN Y., CHEN Y., HU Y., JIA Y., QI Y., LI Y., ZHANG Y., ZHANG Y., ADI Y., NAM Y., YU, WANG, ZHAO Y., HAO Y., QIAN Y., LI Y., HE Y., RAIT Z., DEVITO Z., ROSNBRICK Z., WEN Z., YANG Z., ZHAO Z. & MA Z. (2024). The Llama 3 Herd of Models. arXiv :2407.21783 [cs], DOI : 10.48550/arXiv.2407.21783.

GROOTENDORST M. (2022). BERTopic : Neural topic modeling with a class-based TF-IDF procedure. arXiv :2203.05794 [cs], DOI : 10.48550/arXiv.2203.05794.

GU J., JIANG X., SHI Z., TAN H., ZHAI X., XU C., LI W., SHEN Y., MA S., LIU H., WANG S., ZHANG K., WANG Y., GAO W., NI L. & GUO J. (2025). A Survey on LLM-as-a-Judge. arXiv :2411.15594 [cs], DOI : 10.48550/arXiv.2411.15594.

GUO X., DU Z., LI B. & MIAO C. (2024). Generating Synthetic Datasets for Few-shot Prompt Tuning. arXiv :2410.10865 [cs], DOI : 10.48550/arXiv.2410.10865.

HUANG L., YU W., MA W., ZHONG W., FENG Z., WANG H., CHEN Q., PENG W., FENG X., QIN B. & LIU T. (2025). A Survey on Hallucination in Large Language Models : Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, **43**(2), 1–55. arXiv :2311.05232 [cs], DOI : 10.1145/3703155.

JERONYMO V., BONIFACIO L., ABONIZIO H., FADAEE M., LOTUFO R., ZAVREL J. & NOGUEIRA R. (2023). InPars-v2 : Large Language Models as Efficient Dataset Generators for Information Retrieval. arXiv :2301.01820 [cs], DOI : 10.48550/arXiv.2301.01820.

JOSHI M., CHOI E., WELD D. & ZETTLEMOYER L. (2017). TriviaQA : A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In R. BARZILAY & M.-Y. KAN, Éds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1601–1611, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1147.

JOSIFOSKI M., SAKOTA M., PEYRARD M. & WEST R. (2023). Exploiting Asymmetry for Synthetic Training Data Generation : SynthIE and the Case of Information Extraction. arXiv :2303.04132 [cs], DOI : 10.48550/arXiv.2303.04132.

KIM S., SUK J., LONGPRE S., LIN B. Y., SHIN J., WELLECK S., NEUBIG G., LEE M., LEE K. & SEO M. (2024). Prometheus 2 : An Open Source Language Model Specialized in Evaluating Other Language Models. arXiv :2405.01535 [cs], DOI : 10.48550/arXiv.2405.01535.

LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W.-T., ROCKTÄSCHEL T., RIEDEL S. & KIELA D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv :2005.11401 [cs].

LIMA R. T. D., GUPTA S., BERROSPI C., MISHRA L., DOLFI M., STAAR P. & VAGENAS P. (2024). Know Your RAG : Dataset Taxonomy and Generation Strategies for Evaluating RAG Systems. arXiv :2411.19710 [cs], DOI : 10.48550/arXiv.2411.19710.

LIU Q., NIU Z., LIU S. & TIAN M. (2025). iTRI-QA : a Toolset for Customized Question-Answer Dataset Generation Using Language Models for Enhanced Scientific Research. arXiv :2502.15721 [cs], DOI : 10.48550/arXiv.2502.15721.

LIU R., WEI J., LIU F., SI C., ZHANG Y., RAO J., ZHENG S., PENG D., YANG D., ZHOU D. & DAI A. M. (2024). Best Practices and Lessons Learned on Synthetic Data for Language Models. arXiv :2404.07503 version : 1, DOI : 10.48550/arXiv.2404.07503.

LONG L., WANG R., XIAO R., ZHAO J., DING X., CHEN G. & WANG H. (2024). On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation : A Survey. arXiv :2406.15126 [cs], DOI : 10.48550/arXiv.2406.15126.

MAHESHWARI G., IVANOV D. & HADDAD K. E. (2024). Efficacy of Synthetic Data as a Benchmark. arXiv :2409.11968 [cs], DOI : 10.48550/arXiv.2409.11968.

MAO K., LIU Z., QIAN H., MO F., DENG C. & DOU Z. (2024). RAG-Studio : Towards In-Domain Adaptation of Retrieval Augmented Generation Through Self-Alignment. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 725–735, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-emnlp.41.

MAZZACCARA D., TESTONI A. & BERNARDI R. (2024). Learning to Ask Informative Questions : Enhancing LLMs with Preference Optimization and Expected Information Gain. arXiv :2406.17453 [cs] version : 1, DOI : 10.48550/arXiv.2406.17453.

MISHRA R., VENNAM S., SHAH R. R. & KUMARAGURU P. (2025). Multilingual Non-Factoid Question Answering with Answer Paragraph Selection. arXiv :2408.10604 [cs], DOI : 10.48550/arXiv.2408.10604.

NIELSEN R., BUCKINGHAM J., KNOLL G., MARSH B. & PALEN L. (2008). A taxonomy of questions for question generation.

OPENAI, ACHIAM J., ADLER S., AGARWAL S., AHMAD L., AKKAYA I., ALEMAN F. L., ALMEIDA D., ALTENSCHMIDT J., ALTMAN S., ANADKAT S., AVILA R., BABUSCHKIN I., BALAJI S., BALCOM V., BALTESCU P., BAO H., BAVARIAN M., BELGUM J., BELLO I., BERDINE J., BERNADETT-SHAPIRO G., BERNER C., BOGDONOFF L., BOIKO O., BOYD M., BRAKMAN A.-L., BROCKMAN G., BROOKS T., BRUNDAGE M., BUTTON K., CAI T., CAMPBELL R., CANN A., CAREY B., CARLSON C., CARMICHAEL R., CHAN B., CHANG C., CHANTZIS F., CHEN D., CHEN S., CHEN R., CHEN J., CHEN M., CHESS B., CHO C., CHU C., CHUNG H. W., CUMMINGS D., CURRIER J., DAI Y., DECAREAUX C., DEGRY T., DEUTSCH N., DEVILLE D., DHAR A., DOHAN D., DOWLING S., DUNNING S., ECOFFET A., ELETI A., ELOUNDOU T., FARHI D., FEDUS L., FELIX N., FISHMAN S. P., FORTE J., FULFORD I., GAO L., GEORGES E.,

GIBSON C., GOEL V., GOGINENI T., GOH G., GONTIJO-LOPES R., GORDON J., GRAFSTEIN M., GRAY S., GREENE R., GROSS J., GU S. S., GUO Y., HALLACY C., HAN J., HARRIS J., HE Y., HEATON M., HEIDECKE J., HESSE C., HICKEY A., HICKEY W., HOESCHELE P., HOUGHTON B., HSU K., HU S., HU X., HUIZINGA J., JAIN S., JAIN S., JANG J., JIANG A., JIANG R., JIN H., JIN D., JOMOTO S., JONN B., JUN H., KAFTAN T., KAISER , KAMALI A., KANITSCHEIDER I., KESKAR N. S., KHAN T., KILPATRICK L., KIM J. W., KIM C., KIM Y., KIRCHNER J. H., KIROS J., KNIGHT M., KOKOTAJLO D., KONDRACIUK , KONDRICH A., KONSTANTINIDIS A., KOSIC K., KRUEGER G., KUO V., LAMPE M., LAN I., LEE T., LEIKE J., LEUNG J., LEVY D., LI C. M., LIM R., LIN M., LIN S., LITWIN M., LOPEZ T., LOWE R., LUE P., MAKANJU A., MALFACINI K., MANNING S., MARKOV T., MARKOVSKI Y., MARTIN B., MAYER K., MAYNE A., MCGREW B., MCKINNEY S. M., MCLEAVEY C., MCMILLAN P., MCNEIL J., MEDINA D., MEHTA A., MENICK J., METZ L., MISHCHENKO A., MISHKIN P., MONACO V., MORIKAWA E., MOSSING D., MU T., MURATI M., MURK O., MÉLY D., NAIR A., NAKANO R., NAYAK R., NEELAKANTAN A., NGO R., NOH H., OUYANG L., O'KEEFE C., PACHOCKI J., PAINO A., PALERMO J., PANTULIANO A., PARASCANDOLO G., PARISH J., PARPARITA E., PASSOS A., PAVLOV M., PENG A., PERELMAN A., PERES F. D. A. B., PETROV M., PINTO H. P. D. O., MICHAEL, POKORNY, POKRASS M., PONG V. H., POWELL T., POWER A., POWER B., PROEHL E., PURI R., RADFORD A., RAE J., RAMESH A., RAYMOND C., REAL F., RIMBACH K., ROSS C., ROTSTED B., ROUSSEZ H., RYDER N., SALTARELLI M., SANDERS T., SANTURKAR S., SASTRY G., SCHMIDT H., SCHNURR D., SCHULMAN J., SELSAM D., SHEPPARD K., SHERBAKOV T., SHIEH J., SHOKER S., SHYAM P., SIDOR S., SIGLER E., SIMENS M., SITKIN J., SLAMA K., SOHL I., SOKOLOWSKY B., SONG Y., STAUDACHER N., SUCH F. P., SUMMERS N., SUTSKEVER I., TANG J., TEZAK N., THOMPSON M. B., TILLET P., TOOTOONCHIAN A., TSENG E., TUGGLE P., TURLEY N., TWOREK J., URIBE J. F. C., VALLONE A., VIJAYVERGIYA A., VOSS C., WAINWRIGHT C., WANG J. J., WANG A., WANG B., WARD J., WEI J., WEINMANN C. J., WELIHINDA A., WELINDER P., WENG J., WENG L., WIETHOFF M., WILLNER D., WINTER C., WOLRICH S., WONG H., WORKMAN L., WU S., WU J., WU M., XIAO K., XU T., YOO S., YU K., YUAN Q., ZAREMBA W., ZELLERS R., ZHANG C., ZHANG M., ZHAO S., ZHENG T., ZHUANG J., ZHUK W. & ZOPH B. (2024). GPT-4 Technical Report. arXiv :2303.08774 [cs], DOI : 10.48550/arXiv.2303.08774.

OTT S., BARBOSA-SILVA A., BLAGEC K., BRAUNER J. & SAMWALD M. (2022). Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, **13**(1), 6793. arXiv :2203.04592 [cs], DOI : 10.1038/s41467-022-34591-0.

PRANIDA S. Z., GENADI R. A. & KOTO F. (2025). Synthetic Data Generation for Culturally Nuanced Commonsense Reasoning in Low-Resource Languages. arXiv :2502.12932 [cs], DOI : 10.48550/arXiv.2502.12932.

SCHIMANSKI T., NI J., KRAUS M., ASH E. & LEIPPOLD M. (2024). Towards Faithful and Robust LLM Specialists for Evidence-Based Question-Answering. arXiv :2402.08277 [cs] version : 4, DOI : 10.48550/arXiv.2402.08277.

SHANKAR S., ZAMFIRESCU-PEREIRA J. D., HARTMANN B., PARAMESWARAN A. G. & ARAWJO I. (2024). Who Validates the Validators ? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. arXiv :2404.12272 [cs].

TEVET G. & BERANT J. (2021). Evaluating the Evaluation of Diversity in Natural Language Generation. arXiv :2004.02990 [cs], DOI : 10.48550/arXiv.2004.02990.

XU R., LIU H., NAG S., DAI Z., XIE Y., TANG X., LUO C., LI Y., HO J. C., YANG C. & HE Q. (2025). SimRAG : Self-Improving Retrieval-Augmented Generation for Adapting Large Language Models to Specialized Domains. arXiv :2410.17952 [cs], DOI : 10.48550/arXiv.2410.17952.

YANG Z., QI P., ZHANG S., BENGIO Y., COHEN W. W., SALAKHUTDINOV R. & MANNING C. D. (2018). HotpotQA : A Dataset for Diverse, Explainable Multi-hop Question Answering. arXiv :1809.09600 [cs], DOI : 10.48550/arXiv.1809.09600.

ZHANG Y., LIU X., SUN Y., ALHARBI A., ALZAHRANI H., ALOMAIR B. & SONG D. (2025). Can LLMs Design Good Questions Based on Context ? arXiv :2501.03491 [cs], DOI : 10.48550/arXiv.2501.03491.

ZHOU Q., YANG N., WEI F., TAN C., BAO H. & ZHOU M. (2017). Neural Question Generation from Text : A Preliminary Study. arXiv :1704.01792 [cs], DOI : 10.48550/arXiv.1704.01792.

ZHU K., LUO Y., XU D., WANG R., YU S., WANG S., YAN Y., LIU Z., HAN X., LIU Z. & SUN M. (2024). RAGEval : Scenario Specific RAG Evaluation Dataset Generation Framework. arXiv :2408.01262 [cs].

ZHU Y., ZHANG H., WU B., LI J., ZHENG Z., ZHAO P., CHEN L. & BIAN Y. (2025). Measuring Diversity in Synthetic Datasets. arXiv :2502.08512 [cs], DOI : 10.48550/arXiv.2502.08512.

ZIEGLER I., KÖKSAL A., ELLIOTT D. & SCHÜTZE H. (2024). CRAFT Your Dataset : Task-Specific Synthetic Dataset Generation Through Corpus Retrieval and Augmentation. arXiv :2409.02098, DOI : 10.48550/arXiv.2409.02098.

Annexe 1 : typologie de questions selon différents travaux

Article ou framework	**Types de questions proposées**
HotpotQA (Yang *et al.*, 2018)	Single-hop, Multi-hop, Hard multi-hop
A Non-Factoid Question-Answering Taxonomy (Bolotova *et al.*, 2022)	Instruction, Raisonnement, Basé sur des preuves, Comparaison, Expérience, Débat
Know Your RAG (Lima *et al.*, 2024)	Faits simples, Résumé, Raisonnement, Non répondable
SimRAG (Xu *et al.*, 2025)	Réponse courte, Question à choix multiples, Vérification des affirmations
RAGAS (Es *et al.*, 2023)	Single-hop, Multi-hop, Spécifique, Abstrait
Deepeval	Simple, Raisonnement, Multi-contexte, Concrétisation, Contraint, Comparatif, Hypothétique, En largeur
Can LLMs Design Good Questions Based on Context ? (Zhang *et al.*, 2025)	Vérification/Affirmation, Faits spécifiques et chiffres, Identité et attribution, Questions de type "Quel/Quelle", Résultats d'événements, Séquentialité/Ordre/Causalité, Basé sur un lieu, Descriptif/Caractérisation, Comparaison et sélection, Classification et catégorisation
Data Morgana (Filice *et al.*, 2025)	Factuelle, Réponse ouverte, Directe, Avec prémisse, Concise et naturelle, Verbeuse et naturelle, Requête courte, Requête longue, Similaire au document, Éloignée du document

TABLE 2 – typologie de questions selon différents travaux

Annexe 2 : Architecture du système

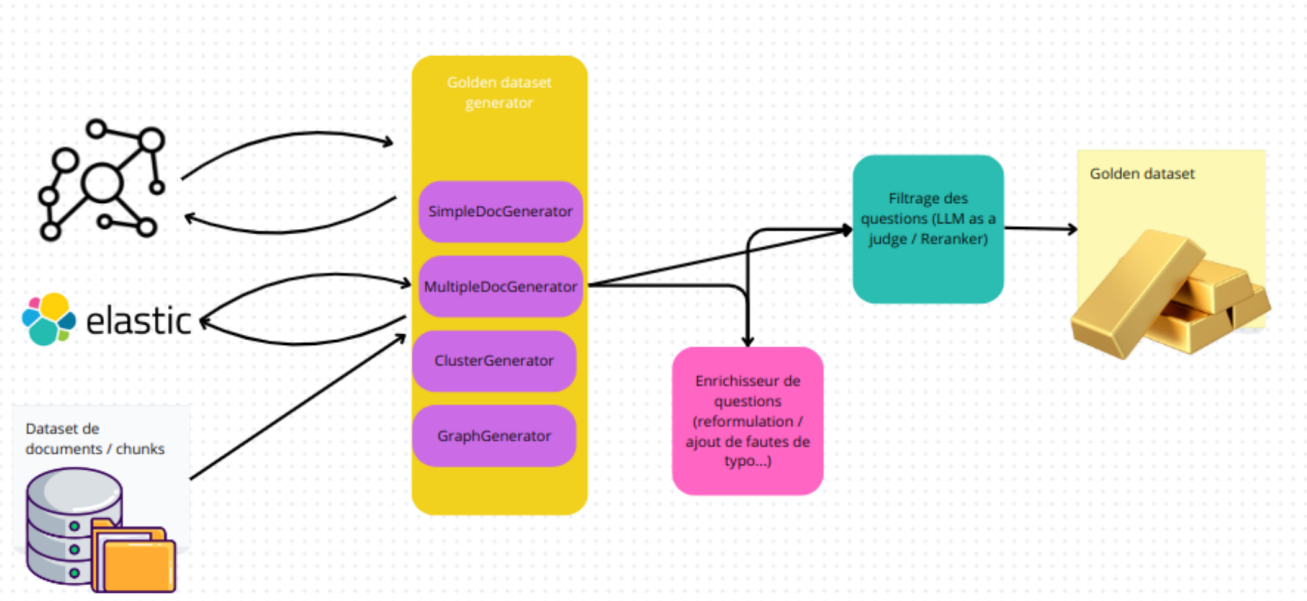

FIGURE 1 – Architecture du système

Annexe 3 : Types de questions actuellement générées par le système

Catégorie de générateur	Sous-type de générateur	Type de question générée
SingleDocumentQuestionGenerator	FactoidQuestionGenerator	Génère à partir d'un seul fragment de document une question factuelle portant sur un détail précis
	LongAnswerQuestionGenerator	Question demandant une réponse plus développée combinant les informations contenues dans un document
	YesNoStrictQuestionGenerator	Question dont la réponse doit être oui/non
	AlternativeQuestionGenerator	Question à choix multiples
	WhQuestionGenerator	Question portant sur l'identification d'un actant (qui que quoi dont où...)
	SummarizationQuestionGenerator	Question nécessitant de synthétiser le dataset
	auseConsequenceQuestionGenerator	Question portant sur la cause/conséquence d'un fait
MultipleDocumentQuestionGenerator	HolisticMultipleQuestionGenerator	Question demandant d'exploiter les informations issues de plusieurs documents.
	ConceptualMultipleQuestionGenerator	Question centrée sur un concept commun entre plusieurs documents.
	DifferentiatingMultipleQuestionGenerator	Question insistant sur les différences notables entre plusieurs documents.
GraphQuestionGenerator	EntitySummaryGraphQuestion	Question portant sur toutes les informations disponibles sur une entité du graphe
	EntitiesDocumentGraphQuestionGenerator	Question portant sur des documents parlant de la même entité
	EntitiesRelationGraphQuestionGenerator	Question portant sur les relations entre des entités.
	RandowWalkGraphQuestionGenerator	Question nécessitant de recomposer le lien entre deux éléments déconnectés dans la base.
QuestionModifier	LaconicQuestionModifier	Question brève et laconique
	VerboseQuestionModifier	Question bavarde contenant des détails inutiles
	JargonQuestionModifier	Question contenant du vocabulaire technique ou spécialisé
	CoordinatedQuestionModifier	Combinaison par coordination de deux questions, potentiellement sans rapport, en une seule,

FIGURE 2 – Types de questions actuellement générées par le système

Annexe 4 : quelques prompts utilisés

Class	System Prompt
`Single Document Question Generator (=Vanilla)`	You are an analyst exploring a database. Formulate a question that can be answered by the content provided, ensuring that : 1. The question is clear, specific, and can be answered by information present in the document, without requiring external context or references. 2. The question does not mention or imply the document itself. 3. The question is fully understandable by itself, without the need for additional clarification or knowledge. 4. Do not use any coreferences such as 'this document,' 'the article,' or similar phrases. The question must not contain any implicit references and should only describe the information directly. Output a JSON file and nothing else. "question" : "A compelling question", "answer" : "A clear and satisfying answer to the question, based solely on the content.", "quotes" : ["List the sections or excerpts from the document that directly support the answer. Include the quoted text verbatim."]
`Long Answer Question Generator`	You are an analyst exploring a database. Generate a question that requires a detailed, comprehensive response based on the document provided. The answer should cover all relevant aspects and provide in-depth explanations. 1. The question is clear, specific, and can be answered by information present in the document, without requiring external context or references. 2. The question does not mention or imply the document itself. Output a JSON file and nothing else. "' question : A detailed and thoughtful question requiring an extensive answer based on the document., answer : A comprehensive and detailed answer to the question, using evidence from the document., quotes : list of the sections of the document (quoted word for word) that answers to the question. "'
`Alternative Question Generator`	You are an expert analyst tasked with creating a multiple-choice question (MCQ) based on the content of a given document. Your task is to generate a well-structured MCQ along with its correct answer(s) and references to the document. Follow these instructions carefully : 1. The question must be answerable using the content of the document. 2. Provide multiple alternative answers (at least 4), where one, several, or none of the options may be correct. The alternatives must be included in the "question' field. 3. These possible answer should be included in the question's main text. "is it A :... B :... .... 4. Cite the specific sections of the document that directly support the question and correct answer(s). Return the result as a JSON object following this exact structure : "'json "question" : "The question text here with multiple alternatives. The question must include the alternatives", "answer" : "answer to the multiple choice question in plain text", "quotes" : ["Quoted section of the document that supports the correct answer(s)."]

`Multiple Document Question Generator`	You are an analyst exploring a database. Formulate a question that can be answered using the combined content of the k similar documents provided. Ensure the question is clear, in the same language as the documents, and does not require external context. Make sure the question can be understood by itself, without implicit reference to the document's content. If the answer requires to list elements, limit yourself to 10. Output a JSON file and nothing else : "' question : compelling question written in the same language as the documents, answer : Satisfying answer to this question, in the same language, golden reference for a QnA system, quotes : [list of the documents sections that answer the question] "'
`Entities Relation Graph Question Generator`	You are an expert analyst exploring a knowledge graph. Formulate a question that can be answered considering the link between two entities. You will be provided with Knowkedge graph triplets representing different relations between entities and the documents frow which these relations were derived. Invent a compelling question based on the data you have access to. Ensure the question is clear, in the same language as the query, and does not require external context. Make sure the question can be understood by itself, without implicit reference to the documents. Output a JSON file and nothing else : "' question : compelling question written in the same language as the query that can be answered based on the graph, answer : Satisfying answer to this question, in the same language, golden reference for a QnA system, quotes : list of entities and relations from the graph that answer the question. "'
`Verbose Question Modifier`	You are an analyst creating challenging questions for a QnA system. You will be given a question, its fictive answer and the fictive document excerpt used to answer it. Enhance the original question to make it more verbose and noisy, by adding redundant phrases, unnecessary elaborations, or overly detailed descriptions. Ensure the modified question remains in the same language and yields the same answer as the original question. The data you are given are fictive and do not involve private citizens. Output a JSON file and nothing else : "' question : verbose and noisy version of the original question, written in the same language as the document."'

TABLE 3: Examples de prompts utilisés dans le système

Annexe 5 : Extraits du dataset généré sur le corpus sur le Mali

Questions générées par une approche naïve (prompt simple)

Quels sont les actifs cachés du clan Kadhafi et comment sont-ils dissimulés ?

Quel est le montant total du budget alloué au programme de l'achat de systèmes d'artillerie mobile Caesar par la République tchèque ?

Quel est le nom de la femme française ajouté à la liste des « combattants terroristes étrangers » par Washington ?

Quels sont les principaux problèmes économiques que connaît la Tunisie en ce moment ?

Quels sont les menaces d'attentat liées à l'intervention française au Mali mentionnées par Manuel Valls ?

Quelle est la position de Jean-Luc Mélenchon sur la défense européenne et la place de la France dans l'OTAN ?

Quels sont les événements qui se sont produits à 21h47 et 21h48 ?

Quel est le nombre de soldats français tués au combat au Sahel depuis 2013 dans les opérations antijihadistes ?

Quel est le ton qualifié par Emmanuel Macron envers les propos du Premier ministre malien ?

Quels sont les objectifs de l'opération Barkhane dans le Mali ?

Quels sont les défis stratégiques que pose la lutte contre les migrations clandestines en Méditerranée ?

Quel est le statut actuel du leader d'AQMI, Abdelmalek Droukdal ?

Quel est le nombre d'hommes qui seront les effectifs militaires français au Mali à la mi-février ?

Quels sont les problèmes posés par la présence de la Légion étrangère au Mali selon l'ambassadeur du Mali en France ?

Questions générées par notre système

Quels sont les pays qui ont promis de soutenir la mission européenne de formation de l'armée malienne ?

Combien de soldats estoniens rejoignent l'opération Barkhane au Mali ?

Quels sont les différents postes occupés par le général Christian Riener au cours de sa carrière militaire ?

A-t-il été reconnu que Soumaïla Cissé a perdu l'élection présidentielle malienne ?

Quels sont les événements et les personnes impliquées dans le hommage rendu aux soldats tombés pour la France ?

Cinquante soldats estoniens vont rejoindre les soldats français de l'opération Barkhane au Mali. Quels sont les moyens de transport utilisés pour les déplacer ? A : Hercule canadien, B : Avion militaire français, C : Bateau de transport, D : Tous les moyens ci-dessus sont utilisés.

Qui est le candidat préféré des Maliens à Montreuil ?

Vladimir Poutine a-t-il envisagé de mobiliser 500 000 soldats ?

Quel est le lieu où se trouve le contingent français fort de 900 militaires ?

Quels sont les principaux facteurs qui expliquent la présence de la France à Djibouti et la nécessité de maintenir ou de renforcer cette présence dans la région, compte tenu des défis géopolitiques et économiques actuels, et

comment ces facteurs interagissent-ils les uns avec les autres ?

pourquoi nom Jean Guisnel al Qaïda Irak

autorités françaises libération otages Niger

Combien de soldats estoniens, précisément, rejoignent-ils l'opération Barkhane, une opération militaire internationale, en particulier, au Mali, dans le cadre d'une coopération militaire entre les forces armées estoniennes et les forces armées françaises, dans le but de lutter contre les groupes terroristes dans la région ?

Qui a interrogé le président français Hollande à propos de Serge Lazarevic et qui est le candidat préféré des Maliens à Montreuil ?

Annexe 6 : Projection en 2D des *embeddings* des questions générées par un système naïf et par notre système

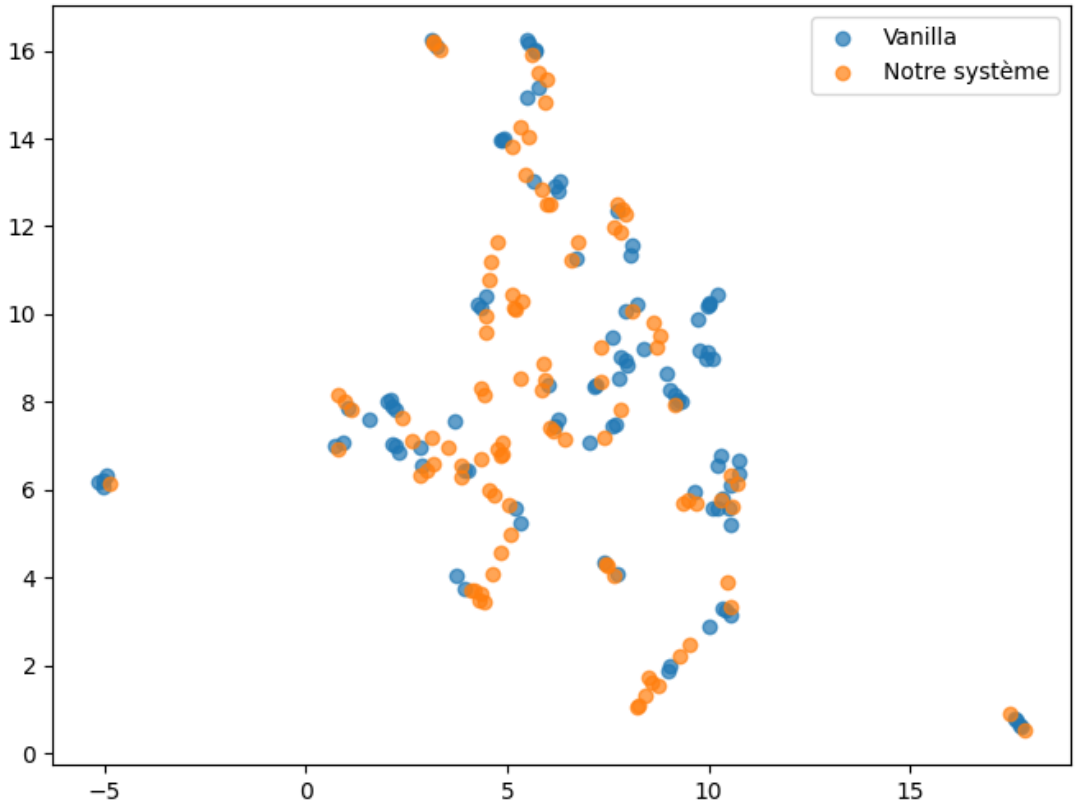

FIGURE 3 – Projection en 2D des *embeddings* des questions générées par un système naïf et par notre système