Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Génération de Questions-réponses expertes : YourExpertBench+

Markarit Vartampetian[1] Diandra Fabre[1] Philippe Mulhem[1] Fouad Hassani[2] Didier Schwab[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP[*], LIG, 38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr
(2) ARTELIA, 4 Rue Germaine Veyret-Verner, 38130 Échirolles, FRANCE

RÉSUMÉ

Nous présentons dans cet article YourExpertBench+, une extension de YourBench qui permet de générer des questions-réponses (QR) expertes, sourcées, à partir de documents écrits par des experts. Appliqués à des publications académiques en Ingénierie Hydraulique, notre approche est capable de générer plus de 13 000 couples questions-réponses de haute qualité, à un coût relativement faible. Nous utilisons les QR générées par YourExpertBench+ dans un cadre d'affinage supervisé, et nous montrons expérimentalement qu'un modèle général s'améliore nettement grâce à notre ensemble de données sur 2 collections de test.

ABSTRACT

Expert Question–Answer Generation : YourExpertBench+

In this article, we present YourExpertBench+, an extension of YourBench framework that enables the generation of expert, source-grounded question-answer (QA) pairs from expert-written documents. Applied to academic publications in Hydraulic Engineering, our approach is able to generate more than 13,000 high-quality question–answer pairs at a relatively low cost. We use the QA pairs generated by YourExpertBench+ in a supervised fine-tuning setting, and we experimentally show that a general-purpose LLM is very positively impacted by our dataset on two test collections.

MOTS-CLÉS : LLM, Questions-Réponses Synthétiques, Affinage, Évaluation, Ingénierie hydraulique.

KEYWORDS: LLM, Synthetic Question–Answering, Fine-tuning, Evaluation, Hydraulic Engineering.

1 Introduction

Les grands modèles de langue (LLM) ont démontré de fortes capacités de généralisation sur un large éventail de tâches (Li *et al.*, 2024b). Toutefois, leurs performances restent limitées dans les domaines spécialisés (Ji *et al.*, 2023; Lu *et al.*, 2025), à cause de la nature statique ou trop généraliste de leurs données d'entraînement qui tendent à sous-représenter les connaissances spécialisées. L'adaptation et l'évaluation des LLM dans ces contextes nécessitent de grands jeux de données annotés, spécifiques au domaine. Or, ces ressources reposent majoritairement sur des annotations manuelles coûteuses à acquérir auprès d'experts, ce qui limite leur disponibilité à grande échelle. Face à ces contraintes, la

∗. Institute of Engineering Univ. Grenoble Alpes

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 273–286.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

génération de données synthétiques constitue une alternative prometteuse. Elle permet de produire automatiquement, et à grande échelle, des ensembles de données ciblés, en incluant aussi des scénarios rares, critiques ou propres à un domaine, sous-représentés dans les jeux de données existants. Nous proposons ici une extension du cadre YourBench (Shashidhar *et al.*, 2025) pour la génération de Questions-Réponses (QR) ouvertes expertes dans un domaine spécialisé comme l'Ingénierie Hydraulique. Nous adaptons YourBench à la génération bout en bout de questions ouvertes à partir de documents académiques longs, en conservant en parallèle sa modularité et son adaptabilité à d'autres domaines ou tâches. Nous évaluons également l'apport des données synthétiques générées dans un contexte d'évaluation et d'adaptation des LLM au domaine considéré. Nos principales contributions sont les suivantes : i) Une extension méthodologique de YourBench pour la génération de QR ouvertes expertes en Hydraulique, ii) La construction et la mise à disposition d'un jeu de données spécialisé en ingénierie hydraulique en français, ancré dans des documents techniques, et son utilisation pour l'adaptation de LLM en Ingénierie Hydraulique et pour la tâche de questions-réponses.

2 État de l'art

Jeux de données QR Les protocoles d'évaluation (benchmarks) des connaissances spécialisées des LLM reposent majoritairement sur le format QR, privilégié pour sa reproductibilité et son évaluation automatique à grande échelle. En Ingénierie Hydraulique, les benchmarks évaluant des tâches en langage naturel demeurent limités et présentent des contraintes marquées selon la langue, le format (QCM vs. questions ouvertes), la taille et la localisation. Par exemple, HydroSE-Bench (Hu *et al.*, 2025) couvre l'Hydrologie et les Structures hydrauliques en chinois. Il inclut environ 4 000 QR générées semi-automatiquement à partir d'examens universitaires, de publications spécialisées et de normes industrielles chinoises. De même, QualBench (Hong *et al.*, 2025) regroupe près de 17 000 QR issues d'examens chinois de qualification professionnelle en Ingénierie des ressources en eau et en Hydroélectricité, avec une forte dépendance au contexte local chinois. Cependant, la traduction automatique de ces ressources peut introduire des erreurs de terminologie métier ou produire des paires QR peu pertinentes lorsque les questions reposent sur des éléments sans équivalent direct dans le contexte cible, notamment des contextes éducatifs spécifiques ou des cadres normatifs, réglementaires et institutionnels propres aux pays où elles ont été conçues (Rajaee *et al.*, 2025; Niklaus *et al.*, 2025; Singh *et al.*, 2025; Wu *et al.*, 2025). En anglais, SuperGPQA (Team *et al.*, 2025b), et sa version en français (Vartampetian *et al.*, 2025) contiennent des QR sur de nombreux domaines, parmi lesquelles 218 QR portent sur l'Hydraulique (Hydrologie, Aménagement hydraulique, Hydroélectricité). Ces benchmarks sont construits à partir de manuels, de notes de cours et de programmes universitaires. Malgré l'existence de nombreux jeux de données QR en français, les ressources demeurent rares dans des domaines techniques comme l'Hydraulique.

Outre le domaine, la nature des documents sources influence également la conception des jeux de données QR. Les benchmarks fondés sur des articles académiques reposent le plus souvent sur une curation humaine, incluant la sélection des documents, la formulation experte des questions et la validation des réponses. Certaines ressources se limitent à des formats courts ou fermés (oui/non, réponses extractives), dérivés uniquement des titres et résumés d'articles (Dasigi *et al.*, 2021; Jin *et al.*, 2019). D'autres proposent des questions expertes ouvertes, restant centrés sur des domaines de l'IA/ML (Singh *et al.*, 2024; Baumgärtner *et al.*, 2025; Huang *et al.*, 2025; Lee *et al.*, 2023). Enfin, SciQAG (Wan *et al.*, 2024) propose une approche proche de la nôtre pour la génération synthétique de paires QR à partir d'articles scientifiques, mais se limite aux sciences fondamentales (chimie, science des matériaux, physique) et repose sur des modèles propriétaires.

Génération de QR Synthétiques La rareté des ressources dans les domaines spécialisés pousse à s'intéresser à la génération automatique de questions-réponses synthétiques pour le développement et l'amélioration des modèles (préentraînement, post-entraînement, évaluation). Certaines approches reposent sur l'agrégation d'exemples existants conçus par des humains afin de produire de nouveaux jeux de données (Li *et al.*, 2025a; Chen *et al.*, 2024; Gandhi *et al.*, 2024; Ziegler *et al.*, 2025; Shahgir *et al.*, 2025). Bien que ces méthodes permettent une génération à grande échelle, la diversité des données produites dépend largement de la qualité et de la variété des exemples sources ou de la disponibilité d'experts du domaine (Li *et al.*, 2024a). D'autres travaux s'appuient sur des ressources en accès ouvert, telles que Wikipédia (Krishna *et al.*, 2025; Li *et al.*, 2025b; Lupidi *et al.*, 2025), pour générer automatiquement des QR. La génération à partir de documents techniques présente cependant des contraintes distinctes : la longueur des documents complique leur traitement intégral dans la fenêtre contextuelle des LLM, et leur organisation thématique est moins localisée. Un paragraphe peut couvrir plusieurs thématiques, tandis qu'une même thématique peut apparaître de manière fragmentée dans le document. Les publications scientifiques s'adressent en outre à un public spécialisé, emploient un lexique propre au domaine et suivent une structure différente du format encyclopédique. La génération synthétique de QR est également utilisée pour l'entraînement et l'évaluation de systèmes RAG, Filice *et al.* (2025) proposant une approche propriétaire, tandis que (Es *et al.*, 2024) repose sur des étapes de prétraitement coûteuses, telles que la construction de graphes de connaissances. Enfin, d'autres travaux ciblent la génération de jeux de données QR au format QCM, qui diffère de notre approche centrée sur des questions ouvertes, bien que celle-ci puisse être adaptée à un format QCM.

3 Méthodologie adoptée

Dans cette section, nous présentons le cadre YourBench, puis YourExpertBench+, qui étend le cadre original afin de générer des QR synthétiques expertes, mono-passage [1] (réponse dérivée d'un passage unique) et multi-passage (réponse nécessitant la combinaison de plusieurs passages issus d'un même document) à partir de publications académiques en Ingénierie Hydraulique.

3.1 YourBench : cadre original

Shashidhar *et al.* (2025) proposent un cadre automatique pour la génération de jeux de données à partir de collections documentaires publiques. Il mobilise un ensemble de LLM (modèles propriétaires et *open-weight*) pour générer des paires QR à choix multiples (QCM) ancrées dans le contexte source. (1) **Ingestion et prétraitement :** les documents d'entrée sont normalisés puis segmentés en passages cohérents selon la similarité sémantique entre phrases consécutives et sous contraintes de longueur en tokens ; plusieurs de ces passages sont ensuite combinés aléatoirement pour former des contextes multi-passages. Un résumé global est généré pour chaque document, qui, combiné aux passages formés précédemment, constitue le contexte d'entrée pour la génération des QR. (2) **Génération de QR :** Les LLM génèrent, par deux prompts spécifiques, des questions mono-passage (MonoP) et multi-passages (MultiP) dans un format structuré (json). Ils extraient également des citations *verbatim* issues du passage fourni afin d'ancrer la réponse dans le texte source. La génération suit une typologie de dix catégories de questions prédéfinies (ex. analytique, conceptuelle, application). Une auto-estimation de la difficulté est effectuée pour chaque QR ($\in \{1, ..., 10\}$). (3) **Filtrage et évaluation finale :** Les QR candidates sont ensuite filtrées en deux étapes : par validation des citations et par déduplication sémantique. Ces étapes génèrent un dataset de plus de 150 000 QR à choix multiples. En outre, une validation manuelle complémentaire est réalisée par 20 annotateurs sur 2 000

1. Nous appelons *passage* un extrait court d'un document.

QR échantillonnées. Enfin, une évaluation par paires selon le principe d'évaluation par un LLM juge est effectuée.

3.2 YourExpertBench+

Notre proposition étend les étapes (1) et (3) de YourBench et nous adaptons l'étape de génération de QR (2) afin de cibler des QR expertes en Ingénierie Hydraulique.

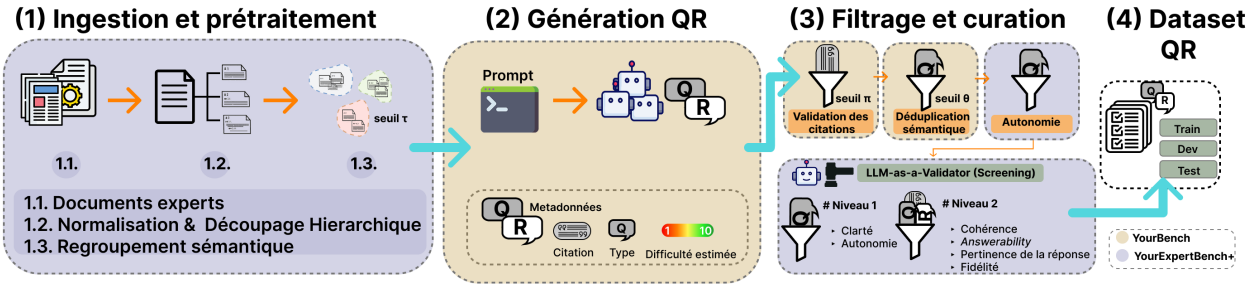

FIGURE 1 – Les composants de YourExpertBench+.

Ingestion et prétraitement : Contrairement à YourBench, nous adoptons Docling (Team, 2024) comme outil principal de conversion afin de préserver la structure des publications académiques (titres, sections, tableaux, formules), complétée par des vérifications heuristiques. Notre approche se base sur des publications académiques, à la différence de YourBench. De tels documents possèdent une organisation hiérarchique explicite et intentionnelle, nous choisissons alors de réaliser un découpage basé sur cette structure, qui respecte l'organisation conceptuelle du contenu (sections, sous-sections). Pour former des questions multi-passages (MultiP), nous ne pouvons pas présupposer l'existence de modèles d'embeddings spécifiquement adaptés au domaine ou la pertinence d'un regroupement aléatoire de passages issus d'une similarité sémantique locale, nous adoptons donc une approche hybride combinant trois signaux complémentaires : (1) **Références croisées explicites :** les références internes détectées par expressions régulières (ex. « voir Section 5.11 ») sont privilégiées, car elles correspondent à des connexions conceptuelles fournies par les auteurs. (2) **Proximité structurelle :** les segments partageant une même section parente, ou issus du découpage d'une section volumineuse, sont regroupés, pour capturer ainsi la continuité et la cohérence. (3) **Similarité sémantique :** calculée via la similarité cosinus (seuil $\tau \geq 0{,}8$)[2] entre segments appartenant à des sections parentes différentes, afin d'identifier des connexions sémantiques distantes. Pour intégrer le document dans sa globalité lors de la génération de QR, YourBench repose sur un résumé généré par LLM. Les publications scientifiques utilisées comportant déjà un résumé et des passages introductifs, nous remplaçons cette approche par un résumé extractif léger.

Génération de QR : Nous conservons la stratégie à deux niveaux (mono-passage et multi-passage par document) de YourBench, en utilisant un ensemble de LLM. Néanmoins, nous nous adaptons à des paires de QR expertes et à la langue française. Dans un environnement professionnel d'ingénierie, les questions pertinentes ne relèvent généralement pas d'un format QCM ; nous privilégions donc des questions ouvertes (QRO) dont la résolution requiert un raisonnement professionnel et une expertise du domaine. Les prompts demandent au modèle de se considérer comme un expert du domaine

2. Les représentations vectorielles denses sont obtenues à partir du modèle BGE-M3, retenu pour ses performances élevées en français (Ciancone *et al.*, 2024; Muennighoff *et al.*, 2023).

(ex. ingénieur en hydraulique) et exigent une couverture de différentes perspectives professionnelles (chercheur, ingénieur, consultant, auditeur). Ils précisent également que les QR générées ne doivent pas pouvoir être résolues par une simple recherche documentaire ou par la reprise directe d'un fragment du texte, mais nécessiter une interprétation experte et une mise en relation complexe des informations. Les prompts de génération demandent, en outre, une mise en correspondance explicite entre segments ; le modèle doit préciser la nature de la relation (par ex. chaîne causale) avant la formulation de la QR, comme étape interne de génération, visant à favoriser une synthèse non superficielle. La génération est conditionnée par les passages sources (mono-passage ou multi-passages) et le résumé global extrait, fournissant à la fois le détail local et le contexte global. On demande aux LLM de générer des questions expertes spécifiques au domaine selon une structure de prompt en plusieurs volets : (1) définition du rôle expert et des objectifs ; (2) phase d'analyse du contenu documentaire ; (3) planification des questions à générer ; (4) consignes détaillées pour la formulation et critères de qualité ; (5) types de questions autorisés ; (6) échelle de difficulté (7) format de sortie structuré. Le modèle adapte la quantité de questions à la richesse du contenu et à sa pertinence selon différents profils métiers (chercheur, ingénieur, consultant, auditeur). Les catégories de questions et l'échelle de difficulté reprennent celles de YourBench. Les catégories proposées ne sont pas intrinsèquement liées au format QCM et peuvent être appliquées à des QRO. Le passage au format QRO introduit cependant des exigences supplémentaires : contrairement aux QCM, les QRO ne limitent pas la réponse à un ensemble fermé d'options et exigent la production d'une réponse de référence complète, claire et cohérente. Dans notre contexte, cela implique de limiter les questions dont la réponse se réduit à la reprise directe d'un fragment du texte et de veiller à ce que les réponses générées restent pertinentes, fidèles et ancrées dans les documents sources, aspects que nous traitons à l'étape de filtrage décrite ci-après.

Filtrage et curation : Étant donné que les LLM génèrent des erreurs, des hallucinations ou des redondances, des traitements ultérieurs sont appliqués sur les QR générées afin d'assurer l'ancrage au texte source et la non-redondance des QRO finales. Ces traitements successifs sont :

(1) **Validation des citations :** Comme dans YourBench, l'ancrage des réponses au texte source est vérifié par correspondance floue (*fuzzy string matching*) à l'aide de PartialRatio, une mesure dérivée de la distance de Levenshtein, qui estime le chevauchement lexical entre les citations extraites et le passage source. Le score d'ancrage consiste en la moyenne des scores sur l'ensemble des citations. Les paires dépassant un seuil $\pi = 0.85$ sont conservées, ce dernier permettant de tenir compte de légères variations de surface telles que des différences de formatage LaTeX des équations, l'omission de marques bibliographiques ou des variations dans les éléments de mise en forme markdown (emphases, listes).

(2) **Déduplication sémantique :** La génération peut introduire une redondance sémantique. À l'instar de Shashidhar *et al.* (2025), nous projetons les questions dans un espace vectoriel puis les regroupons par DBSCAN selon leur similarité cosinus ($> \theta = 0.9$), afin de ne regrouper que des questions quasi-identiques, généralement ancrées dans les mêmes citations et conduisant ainsi à des réponses proches. Le médoïde de chaque cluster est retenu comme représentant, avec un poids proportionnel à la taille du cluster afin de préserver la saillance conceptuelle.

(3) ***Screening*** Cette étape comporte deux sous-étapes successives : un préfiltrage par expressions régulières et un *screening* par LLM (*LLM-as-a-Validator*). Le préfiltrage élimine les paires QR contenant des méta-références explicites au document (ex. « selon le texte », « dans ce chapitre »), assurant ainsi l'autonomie des questions (*self-contained*). Celui-ci permet d'écarter des cas explicites, mais reste limité à des indices de surface. Nous introduisons ensuite, comme étape de filtrage automatique à grande échelle, le *screening* par LLM à deux niveaux : le premier

évalue les questions seules selon des critères de clarté et d'autonomie, tandis que le second correspond à un filtrage contextuel qui évalue conjointement la paire QR et les passages sources associés, selon des critères de cohérence, de répondabilité (*answerability*), de pertinence et de fidélité. Le screening repose sur des décisions binaires (accepter / rejeter). À l'issue de cette chaîne de filtrage et de curation, les QR retenues constituent le corpus synthétique exploité pour les expérimentations décrites dans la section suivante.

4 Génération de QR synthétiques pour l'Hydraulique

4.1 Documents source et sélection des modèles LLM

La génération des QR synthétiques nécessite un corpus documentaire spécialisé en ingénierie hydraulique. Nous avons collecté 350 documents scientifiques (articles, communications et thèses) en accès ouvert issus du portail HAL via son API, des actes du Comité Français des Barrages et Réservoirs et de La Houille Blanche par *scrapping* [3]. Ce corpus source contient environ 2,5 millions de mots, les documents individuels variant entre 2 000 et 90 000 mots (moyenne 7 300 mots/document).

Nous utilisons des modèles *Instruct* capables de suivre des instructions et de raisonner à partir de prompts. Nous retenons des modèles de taille comparable et d'architectures différentes afin d'augmenter la diversité des QR générées, suivant Shashidhar *et al.* (2025). Nous sélectionnons les modèles : Mistral-Small-3.2-24B-Instruct-2506, Qwen3-32B, DeepSeek-R1-Distill-Qwen-32B et Llama-3.3-70B-Instruct, notés dans la suite Mistral, Qwen, DeepSeek et Llama. Tous reçoivent les mêmes prompts et documents sources, avec des paramètres de génération fixés selon les recommandations pour chaque modèle. Nous avons exploré des modèles de 8B et 7B paramètres dans des expérimentations non rapportées ici, mais avons constaté une dégradation de la qualité des sorties (ex. sorties json non conformes, QR hors domaine ou surreprésentant les questions factuelles, ainsi que « pseudo-QR multi-passages » dont la réponse peut en réalité être déduite d'un seul passage). Nous ne les utilisons donc pas ici.

4.2 Corpus de question-réponse synthétique HydrauSynthQR-FR

Le processus décrit en Section 3.2, appliqué aux 350 documents présentés en 4.1, génère initialement 102 867 QR « brutes » en ingénierie hydraulique. Au terme du processus de filtrage, 13 041 QR sont conservées, soit un taux de rétention global de 12,7%, correspondant à des QR ouvertes susceptibles d'être formulées par des experts du domaine. Cette réduction substantielle résulte de l'application successive de critères de filtrage stricts, définis afin de constituer un corpus synthétique spécialisé plus contrôlé, dont l'utilité est évaluée en aval (cf. 6), avec un volume final situé dans les ordres de grandeur couramment mobilisés pour la spécialisation de modèles.

De ces 102 867 QR initiales, le filtrage linguistique est d'abord appliqué afin de conserver uniquement les QR rédigées en français, éliminant 1 088 QR (1,1%) [4]. Les étapes de filtrage décrites en partie 3.2 réduisent progressivement le volume restant : (1) la validation des citations a filtré 13,1% des QR ; (2) la déduplication sémantique 17,6% ; (3) et le *screening* par LLM 55,6%. Ces valeurs montrent

3. https://github.com/scrapy/scrapy

4. Notons que cette étape s'est révélée nécessaire, DeepSeek, produisant occasionnellement des sorties en anglais malgré des données sources et des prompts exclusivement en français.

que chaque étape de filtrage a un effet significatif, chacune éliminant plus de 10% des QR générées. Nous constatons cependant des disparités marquées entre modèles. Llama est le modèle le plus volumineux parmi ceux évalués (70B paramètres), mais présente le taux de rejet par citation le plus élevé (49%), tandis que Qwen3-32B apparaît comme le plus fidèle aux passages sources (11%). La longueur des citations n'est que faiblement corrélée au score d'ancrage (Spearman $r = 0,19$), avec des comportements hétérogènes selon les modèles. En termes de diversité, Qwen génère les questions les plus variées (94% de questions uniques), tandis que Mistral est davantage redondant. L'analyse inter-modèles montre que DeepSeek et Mistral partagent le plus grand nombre de groupes de questions similaires, alors que Llama et Qwen sont les plus différents. Durant le *screening*, étape la plus sélective, le critère d'autonomie étant responsable de la majorité des rejets. Llama est le moins filtré (68% de rejet), tandis que Qwen est le plus filtré ($\approx 79\%$).

Configuration et coût Pour la génération de QR, nous avons fixé la taille maximale des passages mono-passage à 3 000 tokens, un compromis entre les contraintes liées aux longs contextes (limite de fenêtre contextuelle, charge computationnelle, stabilité de la génération) et la nécessité de disposer d'un contenu suffisamment informatif. Pour les questions multi-passage, 2 à 5 segments sont combinés dans une fenêtre maximale de 9 000 tokens. L'ensemble du processus de génération a été exécuté en environ 12 heures sur deux GPU MI250X pour tous les modèles (x4 pour Llama-3.3-70B), en précision BF16, avec le moteur d'inférence vLLM (Kwon *et al.*, 2023). Ceci montre que la production d'un volume substantiel de QR est relativement peu coûteuse au regard des ressources mobilisées.

5 Affinage supervisé et protocole d'évaluation

Dans la suite, nous étudions l'apport des QR synthétiques pour la spécialisation de modèles par affinage supervisé (*Supervised Fine-Tuning*, SFT). Cette section décrit d'abord la partition du corpus, puis la procédure d'entraînement, et enfin le protocole d'évaluation, qui repose sur deux jeux de données (1) HydrauSynthQR-FR pour la génération de réponses ouvertes ; et (2) SuperGPQA-HCE pour la réponse à des questions à choix multiples.

Partition du corpus HydrauSynthQR-FR Les QR générées sont réparties en un sous-ensemble d'entraînement (*train*) de 10 744 QRO, de 655 QRO pour la validation (*dev*) et 1 642 QRO pour le *test*, sans intersection. Parmi les 13 041 QR validées, celles issues de documents publiés en 2025 sont conservées pour l'ensemble *dev* et *test*, afin de limiter tout risque de contamination. La partition en *dev*/*test* est ensuite effectuée par document selon un ratio $1 : 2$, assurant qu'aucun document n'apparaît dans les deux ensembles.

Configuration expérimentale Pour nos expérimentations, nous adaptons les modèles *Qwen3-8B-Base* et *Qwen3-8B-Instruct* par SFT avec LoRA (Hu *et al.*, 2021). L'entraînement est effectué sur deux GPU AMD MI250X. Nous utilisons un rang LoRA $r = 16$ et $\alpha = 32$ avec un *dropout* de 0.05. L'optimisation est effectuée avec AdamW, un taux d'apprentissage de 1×10^{-5} et un ordonnanceur *cosinus*. L'entraînement est réalisé pendant 5 époques avec une taille du batch de 128 et 4 étapes d'accumulation du gradient.

Protocole d'évaluation et métriques L'évaluation s'appuie sur deux jeux de données en hydraulique et deux configurations. L'évaluation sur **HydrauSynthQR-FR** mesure l'effet direct de l'affinage sur la génération de réponses ouvertes spécialisées, tandis que celle sur **SuperGPQA-HCE** vise à en évaluer la généralisation, à la fois à des connaissances en hydraulique externes aux données d'affinage et à un format différent (QCM).

En particulier, l'évaluation sur **HydrauSynthQR-FR** est conduite en zéro-shot avec décodage *greedy* en configuration dite *closed-book*, où le modèle répond uniquement à la question, sans accès au document source. Nous utilisons des métriques d'évaluation complémentaires pour mesurer les performances des modèles. Toutes les métriques comparent la réponse générée $RCand$ avec la réponse de référence $RRef$ afin d'estimer leur similarité. ROUGE-L mesure la similarité lexicale selon la plus longue sous-séquence commune et F1 calcule la moyenne harmonique entre précision et rappel des tokens communs (*overlapping*). Pour évaluer l'ancrage factuel par rapport au texte source K, nous utilisons K-Precision (Adlakha *et al.*, 2024), qui calcule la proportion de tokens de $RCand$ apparaissant dans K, et K-Precision++, qui exclut les tokens présents dans la question afin de réduire les recouvrements artificiels. BERTScore (Schmidtova *et al.*, 2024) évalue la similarité sémantique, par la similarité cosinus entre les représentations vectorielles contextualisées des $RCand$ et $RRef$. L3Score (Pramanick *et al.*, 2024) repose sur les probabilités de log-vraisemblance attribuées par un LLM, ici Gemma-3-27B-it (Team *et al.*, 2025a), aux jugements (yes/no) d'équivalence entre $RCand$ et $RRef$, sans reposer sur des seuils arbitraires. L'évaluation en QRO est intrinsèquement complexe ; les LLM produisent des réponses lexicalement variées qui peuvent paraître différentes en étant sémantiquement équivalentes et factuellement correctes. Nous mobilisons ainsi des métriques complémentaires afin de couvrir différentes dimensions d'évaluation, en considérant leurs limites respectives. Les métriques lexicales, comme ROUGE et F1, mesurent une similarité de surface, mais sont insensibles aux paraphrases et aux synonymes et peuvent échouer à capturer la qualité du raisonnement. K-Precision et K-Precision++ offrent une approximation de l'ancrage des réponses ; reposant elles aussi sur une similarité de surface, elles peuvent toutefois pénaliser des paraphrases valides et, à l'inverse, surestimer des réponses qui reprennent des termes du domaine ou des fragments de la source sans en préserver correctement le sens. BERTScore permet de réduire cette dépendance à la similarité de surface en comparant des représentations vectorielles contextualisées, mais peut rester insuffisant pour détecter des inexactitudes factuelles. Enfin, L3Score apporte un signal complémentaire, susceptible de mieux capter des correspondances sémantiques fines entre la réponse générée et la réponse de référence, notamment lorsque celles-ci dépendent d'indices contextuels subtils. En tant que métrique fondée sur un LLM juge, il soulève toutefois des questions de biais, tels que les préférences stylistiques, les biais de famille de modèles ou les formes d'auto-préférence (Xu *et al.*, 2024; Gu *et al.*, 2025). Chaque métrique comportant ses propres limites, leur usage conjoint permet d'examiner plusieurs signaux d'évaluation, dont les résultats sont interprétés comme des indicateurs complémentaires.

Quant à SuperGPQA-HCE, il adopte un format QCM et couvre deux domaines : l'ingénierie hydraulique (106 questions) et le génie civil (189 questions). Ce benchmark est traduit en français à partir du corpus SuperGPQA (Team *et al.*, 2025b) et ne comprend pas de questions nécessitant des calculs. Dans cette étude, nous retenons uniquement les questions portant sur l'hydraulique. Conformément au benchmark original, l'évaluation suit une approche entièrement générative : les modèles testés génèrent une réponse en langue naturelle, puis indiquent la lettre correspondant à l'option choisie. Cette lettre est ensuite extraite de la sortie générée, et la performance est mesurée par correspondance exacte *exact match* entre la lettre prédite et la lettre de référence, en considérant d'éventuelles variations dans le format de la réponse (ex. la bonne réponse est, la réponse est, réponse : $lettre$) ou de la génération directe du contenu textuel de l'option.

6 Résultats

6.1 Résultats sur HydrauSynthQR-FR

Le tableau 1 présente les évaluations sur les données de test de HydrauSynthQR-FR, avec les modèles initiaux et ceux entraînés sur l'ensemble d'entraînement (*train*) d'HydrauSynthQR-FR.

Du tableau 1, nous constatons que le modèle Instruct (avec ou sans fine-tuning) obtient des résultats absolus supérieurs à ceux du modèle Base correspondant, ce qui s'explique par le fait que les modèles Instruct sont intrinsèquement mieux adaptés à la tâche de question-réponse. Sur les métriques lexicales (ROUGE-L, F1, K-Precision), le modèle Base-SFT obtient toutefois des gains relatifs nettement supérieurs (+57% à +80%) par rapport à Instruct-SFT (+46% à +51%). Cette différence peut s'expliquer par le fait qu'un modèle Base avec SFT s'adapte simultanément à la tâche et au domaine, alors que les modèles Instruct s'adaptent principalement au domaine. On observe par ailleurs que les gains relatifs diminuent progressivement lorsque les métriques capturent davantage d'aspects sémantiques, ce qui suggère que le SFT améliore surtout certaines propriétés formelles des réponses (format, structure, concision et utilisation de mots-clés), sans que ces améliorations permettent à elles seules d'isoler une acquisition effective de connaissances de domaine. L'écart entre K-Precision (+80% pour Base, +46% pour Instruct) et K-Precision+ (+31% et +13%) suggère que le SFT favorise d'abord l'apprentissage du vocabulaire du domaine (présence de mots-clés, mesurée par K-Precision) plutôt que leur utilisation dans un contexte pertinent (K-Precision+). En revanche, sur le L3Score, qui mesure la similarité sémantique globale de la réponse, la tendance s'inverse : Instruct-SFT progresse davantage (+5.2%) que Base-SFT (+3.1%). Cette dissymétrie suggère que les fortes améliorations observées sur les métriques lexicales reflètent principalement une adaptation au format de réponse, alors que le L3Score constitue un indice davantage orienté vers l'adéquation sémantique des réponses et l'intégration des connaissances de domaine que vers la simple reprise lexicale. Base-SFT dépasse Instruct sans adaptation sur ROUGE-L (0.228 vs 0.203) et F1 (0.293 vs 0.270), en restant relativement inférieur sur le L3Score (0.492 vs 0.632). Ces résultats indiquent que le SFT est bénéfique pour les modèles Base et Instruct avec des dynamiques complémentaires : le premier bénéficie d'une forte adaptation au format et au vocabulaire du domaine, tandis que le second semble s'orienter vers une meilleure adéquation sémantique des réponses et la mobilisation des connaissances de domaine, selon le gain sur le L3Score (+5.2% contre +3.1%).

TABLE 1 – Résultats de l'évaluation sur HydrauSynthQR-FR. Les **meilleures performances** sont en gras. Les meilleures améliorations après SFT, par mesure, sont soulignées. Les variations (Δ) sont indiquées en pourcentage. Toutes les valeurs sont entre 0 et 1 (↑).

Modèle	ROUGE-L	F1	K-Precision	K-Precision+	BERTScore F1	L3Score
qwen3-8b-base	0.138	0.187	0.161	0.324	0.513	0.477
qwen3-8b-base-sft	0.228 (+65%)	0.293 (+57%)	0.290 (+80%)	0.425 (+31%)	0.598 (+17%)	0.492 (+3.1%)
qwen3-8b-instruct	0.203	0.270	0.262	0.382	0.583	0.632
qwen3-8b-instruct-sft	**0.306** (+51%)	**0.405** (+50%)	**0.383** (+46%)	**0.432** (+13%)	**0.660** (+13%)	**0.665** (+5.2%)

6.2 Résultats sur SuperGPQA-HCE

Le tableau 2 présente les évaluations sur les données de test de SuperGPQA-HCE, avec les modèles initiaux et ceux entraînés sur l'ensemble d'entraînement (train) d'HydrauSynthQR-FR, en zero-shot

et 5-shot. Pour cette dernière configuration, les 5 exemples *few-shot* incluent un raisonnement dans leur réponse, et le prompt final incite le LLM à « réfléchir étape par étape ».

TABLE 2 – Résultats sur SuperGPQA-HCE en 0/5-shot (mode génératif). Les **meilleures performances** sont en gras. Toutes les valeurs sont entre 0 et 1 (↑).

Modèle	**0-shot**	**5-shot (CoT)**
qwen3-8b-base	0.255	0.255
qwen3-8b-base-sft	0.258 (+1.3%)	0.264 (+3.5%)
qwen3-8b-instruct	0.368	0.396
qwen3-8b-instruct-sft	**0.425** (+15.5%)	**0.422** (+6.6%)

Les résultats montrent que Instruct-SFT présente une amélioration en 0-shot (+15.5%), indiquant que le SFT sur des QR synthétiques permet un transfert de connaissances du domaine malgré un format de tâche différent de celui de l'entraînement. En 5-shot, le gain est plus modéré (+6.6%), tandis que cette configuration améliore légèrement le modèle Instruct sans adaptation (0.368 $\rightarrow$ 0.396). Pour le modèle Base, les gains restent plus limités (+1.3% en 0-shot, +3.5% en 5-shot), suggérant que l'adaptation peut aussi apporter une amélioration, même modérée, sur un format de réponse différent.

7 Discussion et conclusion

Dans cet article, nous avons proposé YourExpertBench+, un framework générique permettant de générer automatiquement des paires de question–réponse experts dans des domaines techniques. Notre approche repose sur une génération automatique de QR suivie de plusieurs étapes de filtrage, permettant de contrôler la qualité des données générées. À travers nos expérimentations, nous étudions l'impact respectif de la qualité et de la quantité des données, la première s'avérant cruciale ; notamment dans un contexte de données synthétiques dans des domaines spécialisés. Nous avons démontré l'intérêt de cette approche dans le cas de la génération de QR à partir de documents en ingénierie hydraulique. Les résultats obtenus à travers les différentes évaluations mettent également en évidence des différences d'adaptation selon les types de modèles : les modèles Base avec fine-tuning s'adaptent simultanément à la tâche et au domaine, tandis que les modèles Instruct semblent principalement bénéficier d'un transfert de connaissances du domaine. Par ailleurs, les résultats sur SuperGPQA-HCE montrent que le SFT sur des QR synthétiques peut transférer des connaissances vers un format de tâche différent (QCM vs. question ouverte) dans un cadre de question-réponse.

Notre étude présente plusieurs limites. Les expérimentations portent sur une seule famille de modèles (Qwen3-8B) et un seul domaine (ingénierie hydraulique en français). Le benchmark SuperGPQA-HCE, bien qu'indépendant, reste de taille limitée (106 questions). De plus, le L3Score repose sur un modèle juge, dont les biais potentiels ne sont pas étudiés ici. Bien que la réduction de l'évaluation à un jugement binaire d'équivalence sémantique (oui/non) simplifie le processus d'évaluation, elle ne permet pas de rendre compte d'erreurs plus nuancées, liées par exemple à l'incomplétude des réponses, à des énoncés ambigus, ou à des écarts stylistiques. Un schéma d'évaluation plus détaillé permettrait d'analyser plus précisément le comportement des juges dans des cadres applicatifs réels. Cette évaluation conserve néanmoins un rôle utile pour des évaluations exploratoires ou à grande échelle,

la validation experte humaine restant indispensable. Enfin, l'approche LoRA, bien qu'efficace en termes de coût computationnel, peut restreindre la capacité d'adaptation au domaine. Cette contrainte agit comme une régularisation implicite, en réduisant l'amplitude des mises à jour du modèle par rapport à un affinage supervisé complet. Si cette propriété peut contribuer à préserver la stabilité des capacités acquises lors du préentraînement, elle peut aussi limiter sa capacité à intégrer des adaptations plus profondes aux spécificités du domaine. Des extensions comme DoRA (Liu *et al.*, 2024), qui décompose les poids préentraînés en composantes d'amplitude et de direction, pourraient être explorées afin d'améliorer ce compromis entre adaptation au domaine et stabilité du modèle.

Plusieurs pistes de travaux futurs sont envisagées. Tout d'abord, la qualité des données générées devra être confirmée par une validation humaine réalisée par des experts du domaine. Il sera également nécessaire d'identifier un compromis entre le nombre et la taille des modèles utilisés et la qualité des QR générées afin de limiter l'impact environnemental du processus. Enfin, il est important de valider YourExpertBench+ sur d'autres domaines techniques et d'explorer des configurations complémentaires, notamment l'intégration de méthodes de type RAG et des stratégies d'adaptation par préentraînement continu (CPT) basées sur les QR générées. Des travaux sont actuellement en cours dans cette direction.

8 Remerciements

Ce travail est réalisé dans le cadre de la Chaire AugmentIA portée par la Fondation Grenoble INP grâce au mécénat du Groupe Artelia. Cette chaire bénéficie également d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-23-IACL-0006 (MIAI Cluster). En outre, ce travail a bénéficié d'un accès aux moyens de calcul du CINES au travers de l'allocation de ressources AD011015213 attribuée par GENCI.

Références

ADLAKHA V., BEHNAMGHADER *et al.* (2024). Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, **12**, 681–699. DOI : 10.1162/tacl_a_00667.

BAUMGÄRTNER T., BRISCOE T. & GUREVYCH I. (2025). PeerQA : A scientific question answering dataset from peer reviews. In L. CHIRUZZO, A. RITTER & L. WANG, Éds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 508–544, Albuquerque, New Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2025.naacl-long.22.

CHEN S., GUAN X. *et al.* (2024). REInstruct : Building instruction data from unlabeled corpus. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2024*, p. 6840–6856, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-acl.408.

CIANCONE M., KERBOUA I. *et al.* (2024). Mteb-french : Resources for french sentence embedding evaluation and analysis. *arXiv preprint arXiv :2405.20468*.

DASIGI P., LO K. *et al.* (2021). A dataset of information-seeking questions and answers anchored in research papers. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I.

BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4599–4610, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.365.

ES S., JAMES J. *et al.* (2024). RAGAs : Automated evaluation of retrieval augmented generation. In N. ALETRAS & O. DE CLERCQ, Éds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations*, p. 150–158, St. Julians, Malta : Association for Computational Linguistics. DOI : 10.18653/v1/2024.eacl-demo.16.

FILICE S., HOROWITZ G. *et al.* (2025). Generating Q&A benchmarks for RAG evaluation in enterprise settings. In G. REHM & Y. LI, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6 : Industry Track)*, p. 469–484, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-industry.33.

GANDHI S., GALA R. *et al.* (2024). Better synthetic data by retrieving and transforming existing datasets. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2024*, p. 6453–6466, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-acl.385.

GU J., JIANG X. *et al.* (2025). A survey on llm-as-a-judge.

HONG M., NG W. *et al.* (2025). QualBench : Benchmarking Chinese LLMs with localized professional qualifications for vertical domain evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, p. 5938–5953, Suzhou, China : Association for Computational Linguistics. DOI : 10.18653/v1/2025.emnlp-main.303.

HU E. J., SHEN Y. *et al.* (2021). Lora : Low-rank adaptation of large language models. https ://arxiv.org/abs/2106.09685.

HU S., SHAN W. *et al.* (2025). Evaluating hydro-science and engineering knowledge of large language models. https ://arxiv.org/abs/2512.03672.

HUANG T., CAO R. *et al.* (2025). Airqa : A comprehensive qa dataset for ai research with instance-level evaluation.

JI Z., LEE N. *et al.* (2023). Survey of hallucination in natural language generation. *ACM computing surveys*, **55**(12), 1–38.

JIN Q., DHINGRA B. *et al.* (2019). PubMedQA : A dataset for biomedical research question answering. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 2567–2577, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1259.

KRISHNA S., KRISHNA K. *et al.* (2025). Fact, fetch, and reason : A unified evaluation of retrieval-augmented generation. In L. CHIRUZZO, A. RITTER & L. WANG, Éds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 4745–4759, Albuquerque, New Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2025.naacl-long.243.

KWON W., LI Z. *et al.* (2023). Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, p. 611–626, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3600006.3613165.

LEE Y., LEE K. *et al.* (2023). QASA : Advanced question answering on scientific articles. In A. KRAUSE, E. BRUNSKILL, K. CHO, B. ENGELHARDT, S. SABATO & J. SCARLETT, Éds.,

Proceedings of the 40th International Conference on Machine Learning, volume 202 de *Proceedings of Machine Learning Research*, p. 19036–19052 : PMLR.

LI H., DONG Q. *et al.* (2024a). Synthetic data (almost) from scratch : Generalized instruction tuning for language models.

LI T., CHIANG W.-L. *et al.* (2025a). From crowdsourced data to high-quality benchmarks : Arena-hard and benchbuilder pipeline. In *International Conference on Machine Learning*, p. 34209–34231 : PMLR.

LI X. L., KAIYOM F. *et al.* (2025b). Autobencher : Towards declarative benchmark construction.

LI Z., XU X. *et al.* (2024b). Leveraging large language models for NLG evaluation : Advances and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 16028–16045, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.896.

LIU S.-Y., WANG C.-Y. *et al.* (2024). Dora : weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24 : JMLR.org.

LU W., LUU R. K. & BUEHLER M. J. (2025). Fine-tuning large language models for domain adaptation : Exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Computational Materials*, **11**(1), 84.

LUPIDI A., GEMMELL C. *et al.* (2025). Source2synth : Synthetic data generation and curation grounded in real data sources.

MUENNIGHOFF N., TAZI N. *et al.* (2023). MTEB : Massive text embedding benchmark. In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2014–2037, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.148.

NIKLAUS J., MERANE J. *et al.* (2025). SwiLTra-bench : The Swiss legal translation benchmark. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 14894–14916, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-long.725.

PRAMANICK S., CHELLAPPA R. & VENUGOPALAN S. (2024). Spiqa : A dataset for multimodal question answering on scientific papers. *NeurIPS*.

RAJAEE S., CHOENNI R. *et al.* (2025). An empirical analysis of machine translation for expanding multilingual benchmarks. In B. HADDOW, T. KOCMI, P. KOEHN & C. MONZ, Éds., *Proceedings of the Tenth Conference on Machine Translation*, p. 1–30, Suzhou, China : Association for Computational Linguistics. DOI : 10.18653/v1/2025.wmt-1.1.

SCHMIDTOVA P., MAHAMOOD S. *et al.* (2024). Automatic metrics in natural language generation : A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, p. 557–583, Tokyo, Japan : Association for Computational Linguistics. DOI : 10.18653/v1/2024.inlg-main.44.

SHAHGIR H. S., LIM C. *et al.* (2025). ExpertGenQA : Open-ended QA generation in specialized domains. In C. CHRISTODOULOPOULOS, T. CHAKRABORTY, C. ROSE & V. PENG, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2025*, p. 2934–2955, Suzhou, China : Association for Computational Linguistics. DOI : 10.18653/v1/2025.findings-emnlp.159.

SHASHIDHAR S., FOURRIER C. *et al.* (2025). Yourbench : Easy custom evaluation sets for everyone. https ://arxiv.org/abs/2504.01833.

SINGH S., ROMANOU A. *et al.* (2025). Global MMLU : Understanding and addressing cultural and linguistic biases in multilingual evaluation. In W. CHE, J. NABENDE, E. SHUTOVA & M. T.

PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 18761–18799, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-long.919.

SINGH S., SARKAR N. *et al.* (2024). SciDQA : A deep reading comprehension dataset over scientific papers. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 20908–20923, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.1163.

TEAM D. S. (2024). *Docling Technical Report*. Rapport interne. DOI : 10.48550/arXiv.2408.09869.

TEAM G., KAMATH A. *et al.* (2025a). Gemma 3 technical report.

TEAM M.-A.-P. *et al.* (2025b). Supergpqa : Scaling llm evaluation across 285 graduate disciplines. https ://arxiv.org/abs/2502.14739.

VARTAMPETIAN M., FABRE D. *et al.* (2025). SuperGPQA-HCE-FR : un corpus spécialisé en français pour le domaine hydraulique et le génie civil. In *Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 253–276, Marseille, France : ATALA & ARIA.

WAN Y., LIU Y. *et al.* (2024). Sciqag : A framework for auto-generated science question answering dataset with fine-grained evaluation.

WU M., WANG W. *et al.* (2025). The Bitter Lesson Learned from 2,000+ Multilingual Benchmarks. DOI : 10.48550/arXiv.2504.15521.

XU W., ZHU G. *et al.* (2024). Pride and prejudice : LLM amplifies self-bias in self-refinement. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15474–15492, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.826.

ZIEGLER I., KÖKSAL A. *et al.* (2025). Craft your dataset : Task-specific synthetic dataset generation through corpus retrieval and augmentation. *Transactions of the Association for Computational Linguistics*, **13**, 1693–1721. DOI : 10.1162/TACL.a.56.