Revue-IRS

Revue Internationale de la Recherche Scientifique (Revue-IRS)

ISSN: 2958-8413

Vol. 4, No. 1, Janvier 2026

This is an open access article under the CC BY-NC-ND license.

Evaluation des LLM quantizés dans un pipeline RAG à faibles ressources pour la protection des données

Dieudonné MPUTU MWAN

Université de Kinshasa, RD Congo

Abstract: Large Language Models (LLMs) have undergone significant advancements in recent years. However, they exhibit certain limitations, notably the constraint that their knowledge remains confined to the data on which they were trained. To overcome this limitation, specialized enrichment techniques have emerged, such as Retrieval-Augmented Generation (RAG). RAG introduces a novel paradigm aimed at enhancing the performance of LLMs by grounding response generation in an external knowledge base. This study contributes to the democratization of LLM evaluation by identifying the key components of a RAG pipeline within a context characterized by limited computational resources. It enables users seeking to assess LLMs to select the model most appropriate for their specific needs by developing their own evaluation benchmarks. Finally, the study applies a micro-benchmark focusing on data protection, revealing that larger models are not necessarily the most suitable ones.

Keywords: LLM, AI, RAG, benchmark, evaluation.

Résumé : Les grands modèles de langage (LLM) ont connu une évolution remarquable ces dernières années. Cependant, ils présentent certaines limites telles que le fait que leurs connaissances restent figées aux données de leur entrainement. Face à cette contrainte, des techniques d'enrichissement spécialisé sont développées, c'est le cas de la génération augmentée par récupération (RAG). Le RAG introduit un nouveau paradigme pour améliorer les performances des LLM en ancrant la génération des réponses dans une base de connaissance externes à ces derniers. Cette étude se veut une contribution dans la démocratisation de l'évaluation des LLM en identifiant les éléments constitutifs d'un pipeline RAG dans un contexte de faibles ressources computationnelles. Elle permet aux usagers désireux d'évaluer des LLM de pourvoir choisir le mieux adapté à leur besoin en établissant leur propre benchmark d'évaluation. Elle évalue enfin un micro-benchmark sur la protection des données dont il ressort que les LLM les plus grands ne sont pas forcément les mieux adaptés.

Mots clés : LLM, IA, RAG, benchmark, évaluation.

Digital Object Identifier (DOI): https://doi.org/10.5281/zenodo.**18346326**

Introduction

Ces dernières années, les grands modèles de langage (LLM) ont connu du grand succès. En dépit de leur évolution remarquable, ces derniers ont, dans certains contextes, fait preuve des limites dont les plus criantes sont l'hallucination – qui consiste en ce que le LLM génère des textes ou des réponses qui ne sont pas basées sur la réalité (Schaeffer, 2025) – et le fait que les connaissances qu'ils possèdent restent figées aux données de leur entrainement. Mettre à jour les connaissances d'un LLM suppose un nouvel entrainement, chose qui n'est pas simple considérant le coût financier y relatif. Suite à cette contrainte et pour adapter les LLM à des contextes spécifiques, des techniques d'enrichissement spécialisé ont été développées, entre autres le fine-tuning et la génération augmentée par récupération (RAG), cette dernière offrant plus de pertinence (Bouvard, Ciancone, Gourru, & Schaeffer, 2024). Le RAG intervient donc avec un nouveau paradigme pour améliorer les performances des LLM, celui d'ancrer la génération des réponses dans une base de connaissance externes aux LLM (Khang, Park, Hong, & Jung, 2025) et facilitant ainsi la communication, la centralisation des ressources et les réponses aux questions fréquentes (Steffenel & Lucas, 2025). Cependant, le RAG n'est pas une solution magique clé à la main, il est constitué de plusieurs briques d'outils à implémenter en vue de cette amélioration spécifique de LLM. Il en ressort la question de recherche de cette étude : Comment choisir le LLM pour un pipeline RAG ?

La performance du pipeline RAG dépend essentiellement de la performance, de l'adaptation contextuelle du LLM dont il est constitué d'une part, et de comment la base de connaissances est construite d'autre part. Il sied de noter, cependant, que tout ce que sait faire un LLM c'est prédire le mot suivant en se basant sur le prompt qui lui est soumis (Boulle, 2024). La littérature sur le choix contextuel de LLM est très vaste et prend une immense diversité d'orientations. Certaines recherches ont théorisé sur le choix du LLM. En matière d'applications des LLM, la demande de solutions de services qui offrent un équilibre optimal entre les performances et les coûts reste cruciale. L'évaluation des LLM ne peut se faire que dans une logique du rapport qualité-prix. Malgré la diversité des LLM, aucun modèle ne peut à lui seul être performant dans toutes les tâches ou applications, surtout lorsqu'il faut concilier performance et coût (Hu & et al., 2024). Par ailleurs, un LLM est censé être une « IA digne de confiance[1] », c'est-à-dire, il doit être licite, éthique et robuste (Duprieu & Berkouk, 2024). Toutefois, une évaluation des LLM peut se faire aussi sur base des limites que présente chacune, il ne suffit donc pas d'évaluer leurs capacités, mais aussi leurs limites (Kooli, Flament, Dutrey, Diniz, & Claveau, 2024). Outre ces approches théoriques, nombreuses d'autres études prennent une approche du benchmarking. Chaque benchmark suit une approche particulière en prenant en compte l'un et/ou l'autre critère qu'il évalue : Chart-to-Experience (Kim, 2025), FullFront (Sun, Will Wang, Gu, Li, & Cheng, 2025), Execute, (Edman, Fraser, & Schmid, 2025), CUTE (Edman, Schmid, & Fraser, CUTE: Measuring LLMs' Understanding of Their Tokens, 2024), etc. En revanche, la validité de ces benchmarks est à mettre en cause si l'on considère le fait que dans la plupart des cas, ces benchmarks visent des tâches trop générales et sont statiques, basés sur des jeux des données prédéfinis qui, au bout d'un moment, deviennent obsolètes (Babonnaud, 2024).

Considérant le fait qu'il y n'a pas de benchmark qui tienne dans toutes les situations, notre étude propose une démarche permettant de mettre place les outils d'évaluation des LLM en sciences humaines, en l'occurrence sur la protection des données. Ceci permet à quiconque voudrait évaluer des LLM en vue d'en choisir la mieux adaptée à son besoin de pouvoir établir son propre benchmark d'évaluation. Pour ce faire, nous avons identifié quelques bonnes pratiques

[1] Selon le Groupe d'experts de haut niveau de la Commission Européenne (HLEG-IA), cf. Duprieu H.

pour mettre en place un pipeline RAG et nous avons évalué un micro-benchmark sur la protection des données. Il en ressort que les plus grands LLM ne sont pas forcément les mieux adaptés.

1. Constitution du pipeline d'évaluation

Dans ce premier point, nous présentons une fiche technique de notre pipeline RAG et ensuite, en présentant notre benchmark, nous focaliserons sur le choix du LLM qui constitue l'une des pièces centrales du pipeline à côté de la base des connaissances.

1.1. Fiche technique du pipeline

Le pipeline RAG comprend essentiellement trois étapes, à savoir : la récupération (retrieval), l'augmentation (augmentation) et la génération (generation). L'illustration ci-dessous donne un aperçu intégral d'un pipeline RAG :

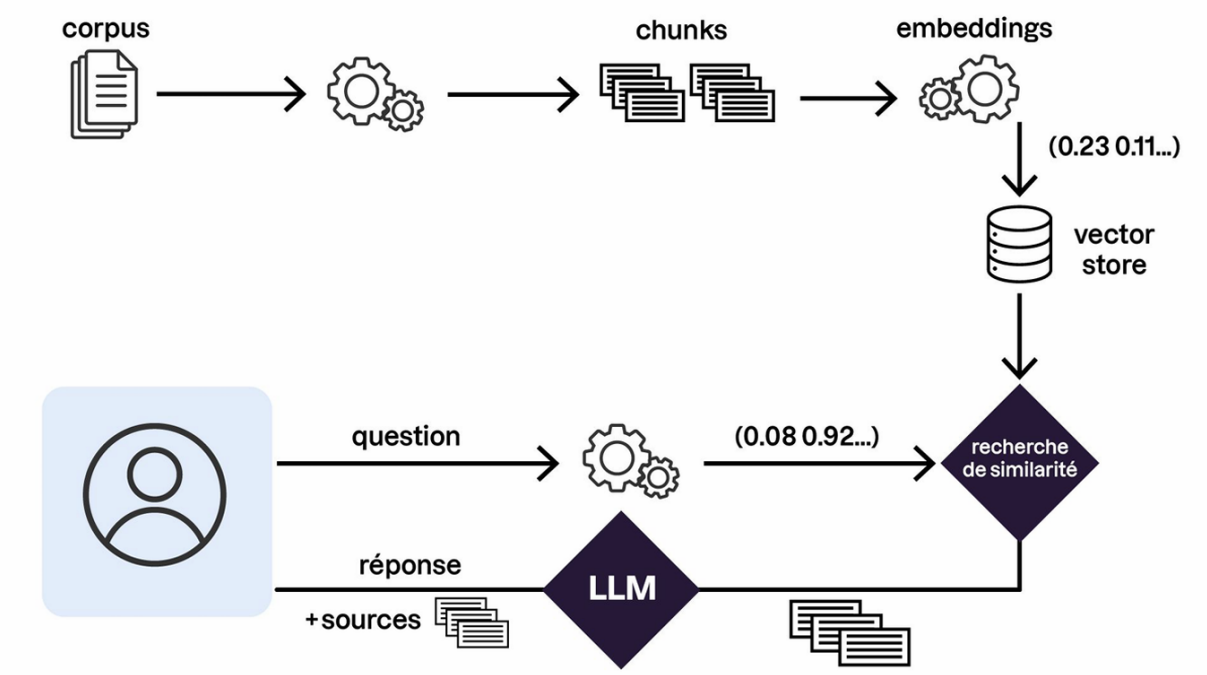


Figure 1 : illustration du pipeline RAG. Source : OpenClassroom

Se référant à cette illustration, la composition de notre pipeline mise en place se présente comme suit :

1. Corpus : 213 textes dont des textes légaux, des textes de la CNIL, des textes sur la cybersécurité, des textes de l'Edpb, des modèles des documents, etc. ;
2. Chunking : text_splitter de langchain,
   - Nombre des caractères par chunk : 1500
   - Nombre des caractères de chevauchement : 150
   - Résultats : 223576 chunks obtenus
3. Embedding : SentenceTransformers avec le modèle all-MiniLM-L6-v2
4. Vector store : Faiss.
5. Paramètres LLM :
   - temperature : 0,2
   - top_k : 20

- top_p : 0.8
- repetition_penalty : 1.1

6. Prompt système :

```
Vous êtes un assistant virtuel à {ENTITE_NAME}.
Répondez à la question de l'utilisateur en vous basant UNIQUEMENT sur le
contexte fourni ci-dessous.
Si l'information n'est pas dans le contexte, dites que vous ne savez pas ou
que l'information n'est pas disponible dans les documents fournis.
Soyez concis et précis. Citez vos sources si possible (par exemple, en
mentionnant le nom du fichier ou la catégorie trouvée dans les
métadonnées).

Contexte fourni:
---
{context_str}
---
Réponds uniquement à partir du contexte fourni. Ne réponds pas si
l'information est absente.

Si l'utilisateur a posé une question qui semble concerner des informations
spécifiques à {ENTITE_NAME}, mais aucune information pertinente n'a été
trouvée dans notre base de connaissances, indiquez poliment que vous n'avez
pas cette information spécifique et suggérez à l'utilisateur de reformuler
sa question ou de contacter directement {ENTITE_NAME}.
N'inventez pas d'informations sur {ENTITE_NAME}.
Réponds uniquement à partir du contexte fourni. Ne réponds pas si
l'information est absente.
```

Figure 2 : illustration du prompt système. Source : Notre code

1.2. Présentation du benchmark et choix du LLM

Hugging Face établit un tableau de bord d'évaluation sur base de 7 critères, à savoir : IFEval, BBH, MATH, GPQA, MUSR, MMLU et CO_2 Cost[2]. L'avantage qu'offre un tel outil est que l'évaluation est toujours actualisée en suivant l'évolution des modèles.

- IFEval (Instruction-Following Evaluation : Évaluation du suivi des instructions) : test de la capacité du modèle à suivre des instructions de formatage explicites (suivi des instructions, formatage et génération).
- BBH (Big Bench Hard) : l'ensemble de tâches stimulantes pour le LLM dans différents domaines, par exemple : compréhension du langage, raisonnement mathématique et sens commun et connaissance du monde.
- MATH (Mathematics Aptitude Test of Heuristics : Test d'aptitude aux mathématiques, niveau 5) : problèmes mathématiques de niveau lycée (Algèbre complexe, Problèmes de géométrie, Calcul différentiel et intégral).
- GPQA (Graduate-Level Google-Proof Q&A): questions à choix multiples de niveau doctorat en sciences chimie, biologie et physique.
- MUSR (Raisonnement souple multi-étapes) : raisonnement et compréhension de textes longs, compréhension du langage, capacités de raisonnement et raisonnement contextuel long.

[2] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/

- MMLU-Pro (Compréhension linguistique multitâche massive - Professionnel) : questions à choix multiples révisées par des experts dans différents domaines, par exemple : médecine et santé, droit et éthique, ingénierie et mathématiques.
- CO_2 Cost (Émissions de dioxyde de carbone) : impact environnemental de l'entraînement des modèles. Les modèles volumineux peuvent avoir une empreinte carbone importante.

Cette évaluation est complétée par une autre évaluation faite en collaboration avec le Gouvernement Français, il s'agit d'un autre tableau de bord de comparaison de LLM génératifs sur des jeux de données adaptés à la langue française[3]. Ce tableau de bord est établi sur base d'un jeu de données construit à partir des sujets du Baccalauréat (examen de fin d'études secondaires en France), d'une évaluation des connaissances au niveau doctorat (GPQA) et de test de capacité des modèles à suivre des instructions, traduit et adapté culturellement (IFEval). Le jeu de données est fourni par le Ministère de l'Éducation nationale et extrait par le Laboratoire National de Métrologie et d'Essais (LNE) et l'Institut National de Recherche en Informatique et en Automatique (Inria).

Les top 3 modèles de cette comparaison de modèles d'IA génératifs sur des jeux de données adaptés à la langue française sont les suivants :

1. deepseek-ai/DeepSeek-R1-Distill-Llama-70B, avec une note de 55,93%
2. mistralai/Mistral-Large-Instruct-2411, avec une note de 49,41%
3. meta-llama/Llama-3.3-70B-Instruct, avec une note de 48,54%

Par ailleurs, la question de l'équilibre entre performance, qualité des réponses et efficacité computationnelle est centrale dans le contexte actuel d'optimisation des LLM pour des usages spécifiques. Cette étude s'inscrit dans une démarche d'évaluation comparative de différents niveaux de quantization d'un modèle. Nous choisissons le modèle Meta-Llama-3.1-Instruct. Un modèle dit « Instruct » est optimisé pour les usages chat : les conversations, les résumés et les réponses aux questions. Considérant le fait que notre pipeline est implémenté pour des raisons d'évaluation, l'objectif n'a pas été de choisir le modèle le mieux coté à l'instar des tops 3 ci-dessus évoqués, mais par contre de prendre un modèle qui offre plusieurs variants à comparer. Il est aussi question pour nous de prendre un modèle susceptible de fonctionner correctement avec nos ressources computationnelles disponibles avec pour objectif de mesurer leur impact sur la qualité des réponses générées dans un pipeline RAG. Nous avons travaillé avec les versions suivantes, toutes disponibles sur Hugging Face :

1. Meta-Llama-3.1-8B-Instruct-GGUF:F32 : version ayant 8 milliards de paramètres et conservant une précision maximale en float 32 bits, elle pèse 32 GB ;
2. Meta-Llama-3.1-8B-Instruct-GGUF:Q8_0 : version ayant 8 milliards de paramètres et quantizée en 8 bits, elle pèse 8.5 GB ;
3. Meta-Llama-3.1-8B-Instruct-GGUF:Q2_K : version ayant 8 milliards de paramètres et quantizée en 2 bits, elle pèse 3.2 GB ;
4. Meta-Llama-3.1-70B-Instruct-GGUF:Q2_K : version avec 70 milliards de paramètres et quantizée en 2 bits, elle pèse 26 GB.

La quantization est une technique d'optimisation visant à réduire la taille mémoire et les besoins en calcul des modèles, en représentant les poids à des niveaux de précision inférieurs (Egashira, Vero, Staab, He, & Vec, 2024). Cette compression permet de faciliter le déploiement des LLM

[3] https://huggingface.co/spaces/fr-gouv-coordination-ia/llm_leaderboard_fr#/

sur des infrastructures moins coûteuses (edge computing, CPU, GPU modestes) ou dans des environnements à contraintes de performance. Cependant, ce gain en efficacité s'accompagne parfois d'une dégradation de la qualité des réponses, ce qui rend nécessaire une évaluation rigoureuse et contextualisée. En effet, pour la plupart de méthodes de quantization post-apprentissage la performance reste au rendez-vous jusqu'à 8 bits. Mais qu'avec des précisions binaires inférieures, ces méthodes s'avèrent inefficace (Liu & et al., 2023).

En revanche, la méthode de comparaison des LLM que nous proposons n'est pas conçue pour évaluer seulement les différents niveaux de quantization, elle peut servir pour comparer n'importe quels LLM. Néanmoins, nous considérons plus l'avantage que peut offrir les LLM quantizés en ce qui concerne l'économie des ressources computationnelle, minimisant ainsi la consommation énergétique.

La base de connaissances qui alimente le pipeline RAG que nous utilisons pour l'évaluation est constituée de la documentation juridique sur la protection des données. Notre évaluation est donc réalisée dans ce domaine spécifique : la protection des données. Le jeu de données d'évaluation est constitué de 40 questions (prompts) reparties de la manière suivante :

- 25 sur les définitions : en droit, savoir définir les concepts avec précision est indispensable. Ces questions nous permettent de savoir quel LLM donne des définitions précises.
- 10 sur des exemples des cas : les études des cas nous permettent à évaluer la capacité des LLM à répondre à des problématiques plus ou moins complexes.
- 3 sur les procédures : étant donné que le pipeline RAG est conçu en mode chatbot, il peut servir d'un assistant virtuel à la protection de données, il est indispensable de le faire tourner avec un LLM capable d'orienter pertinemment ses utilisateurs sans les induire en erreur.
- 2 sur les modèles de documents : comme pour la précédente série, un assistant virtuel se doit de se montrer à la hauteur d'aider ses utilisateurs à être plus productifs. Générer automatiquement des documents ou des modèles des documents est une qualité indispensable.

L'appréciation des réponses générées est faite en mettant en parallèle les réponses générées et ensuite attribuer une cote à chaque réponse selon les critères d'évaluation suivantes : la qualité de la réponse et la durée d'exécution du LLM. La cote va de 0 à 4, ce qui ne laisse pas la possibilité de donner une cote du milieu de l'échelle. 0 étant la valeur nulle, la cote va de 1 à 4 : 1 pour insuffisant, 2 pour bien, 3 pour très bien et 4 pour excellent.

La qualité de la réponse est déterminée par la pertinence, la concision et le fact-checking (Guo & et al., 2023).

- Pertinence : la réponse est-elle correcte ?
  La pertinence vérifie dans quelle mesure la réponse fournie par le LLM répond avec justesse à la question posée, en apportant une information correcte, cohérente et adaptée au contexte.

- Concision : est-elle directe et sans digression ?
  La concision évalue si la réponse est claire, structurée, et dépourvue de contenu redondant ou hors sujet. Cela favorise l'efficacité cognitive de l'utilisateur.

- Fact-checking : s'appuie-t-elle correctement sur les sources ?
  Le fact-checking mesure si le contenu généré par le LLM est vérifiable, référencé, et basé sur des sources fiables, qu'elles soient explicites ou implicites.

2. Résultats d'évaluation

2.1. Présentation des résultats

Tableau 1 : Temps d'exécution en secondes des LLM

	8B-F32	8B-Q8	8B-Q2	70B-Q2
Total du temps	827,12	358,76	195,97	836,07
Moyenne	20,678	8,969	4,89925	20,90175

Pour calculer la durée d'exécution, on récupère le timestamp Unix (nombre des secondes écoulées depuis le 1er janvier 1970) avant l'envoi du prompt dans une variable *start_time* et le timestamp Unix après la génération de la réponse dans la variable *end_time*.

La durée d'exécution est la différence entre le *start_time* et *le end_time* :

Latance = end_time – start_time

D'après nos données, le LLM avec un plus grand nombre des paramètres, en l'occurrence Meta-Llama-3.1-70B-Instruct-GGUF:Q2_K, met beaucoup plus de temps en dépit de sa forte quantization.

Par contre, pour les LLM ayant le même nombre des paramètres, 8B en l'occurrence, la durée d'exécution baisse avec une forte quantization.

Tableau 2 : Pertinence des réponses

	8B-F32	8B-Q8	8B-Q2	70B-Q2
Total	107	86	84	116
Moyenne	2,675	2,15	2,1	2,9

Il résulte de ces données que :

- Pour les LLM avec le même nombre des paramètres, les modèles perdent la performance, en ce qui est de la pertinence des réponses, avec la forte quantization en petit nombre des bits.
- Le modèle ayant plus des paramètres a obtenu le meilleur score en dépit de sa quantiztion en petit nombre des bits.

Il s'avère donc que les modèles avec plus des paramètres sont plus performants, quant à la pertinence de leurs réponses, que ceux avec moins des paramètres.

Tableau 3 : Concision des réponses

Question	8B-F32	8B-Q8	8B-Q2	70B-Q2
Total	102	97	87	123
Moyenne	2,55	2,425	2,175	3,075

D'après les données, le LLM avec le plus grand nombre des paramètres offre plus de concision que ceux avec moins de paramètres. Et pour ceux ayant le même nombre des paramètres, la concision baisse avec la quantisation à de petits nombres des bits.

Tableau 4 : Factualité des réponses

stion	8B-F32	8B-Q8	8B-Q2	70B-Q2
Total	110	99	98	127
Moyenne	2,75	2,475	2,45	3,175

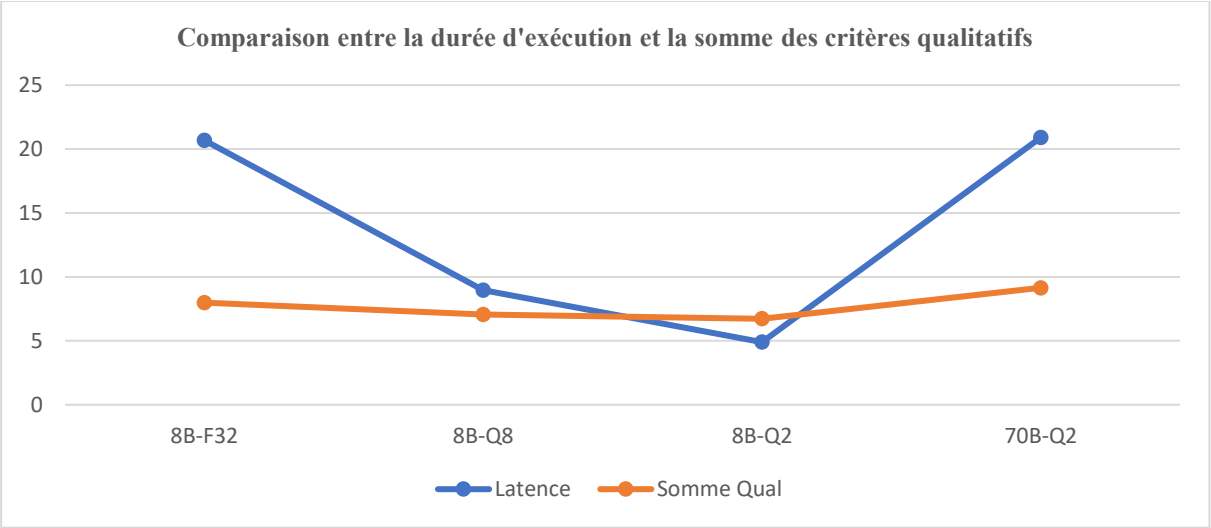
Comme pour les précédents critères, par rapport à la factualité des réponses se référant aux documents fournis, le LLM avec le plus grand nombre des paramètres est mieux coté que les autres, cela en dépit de sa quantization en petit nombre des bits. Par contre, pour les LLM ayant le même nombre des paramètres, le niveau de quantization n'a pas un grand effet, les écarts des cotes sont serrés.

2.2. Interprétation des résultats

Les résultats de l'évaluation ont montré que les LLM ayant plus de paramètres mettent beaucoup plus de temps à répondre, mais ils sont plus pertinents, plus concis et plus factuels.

Quant au niveau de quantization, les LLM fortement quantizés, à de petits nombres de bits, offrent plus de gain de temps en réduisant le temps d'exécution. Toutefois, plus fort le LLM est quantizé, la pertinence et la concision baissent. La quantization n'a pas trop d'influence sur la factualité.

Pour établir une appréciation contextuel basé sur les LLM évalué, la compilation des critères qualitatifs en comparaison avec la durée d'exécution des LLM donne le graphique ci-dessous :



Ce graphique de comparaison permet une prise de position sur le choix du modèle et du niveau de quantization. Les résultats démontrent à travers le graphique que les modèles 8B-Q8 et 8B-Q2 sont plus rapides dans l'exécution, et parfois voix de conséquence, ils consomment moins des ressources computationnelles. Toutefois la qualité des réponses laisse à désirer. En revanche, en dépit de l'écart entre la durée et la qualité du modèle 70B-Q2, celui-ci présente plus de qualité en termes de pertinence, concision et factualité que les deux autres sus-évoqués.

Peu importe sa plus grande durée d'exécution que les autres, le modèle 70B-Q2 offre l'avantage de bonnes qualités des réponses. Et puisque fortement quantizé en petit nombre des bits, il ne pèse que 26 giga octets sur le disque de stockage.

Conclusion

Face au nombre croissant des LLM sur le marché, la question *lequel choisir* était au centre notre étude. Cependant, l'évolution effrénée des données au fil du temps fait que les LLM deviennent obsolètes et qu'il faille les réentraîner. Le coût exorbitant du réentrainement ou d'adaptation aux situations spécifiques a conduit à l'utilisation des techniques plus flexible à l'adaptation contextuelle telle que le RAG.

Au terme de cette étude, il est à noter que le choix du LLM doit prendre plusieurs facteurs en compte, entre autres le coût financier que peut représenter l'acquisition du LLM ou son utilisation. Dans certain contexte, des LLM open sources font bel et bien preuve de robustesse et de performance. Un autre facteur majeur c'est la confidentialité pour laquelle il est généralement nécessaire d'héberger son LLM dans sa propre infrastructure. Dans un tel contexte, l'adaptation à la situation matérielle, aux ressources computationnelles devient primordiale et on peut se voir contraint de sacrifier un peu de qualité au profit d'économiser les ressources computationnelles et énergétiques. Dans ce cas, on a le choix entre des LLM volumineux et moins volumineux (selon le nombre des paramètres disponibles) d'une part et entre la version originale d'un LLM et ses différents niveaux de quantization d'autre part.

Par ailleurs, notre analyse des données a démontré que sur un environnement d'exécution à faibles ressources computationnelles, les écarts des durées d'exécution entre LLM sont très remarquables et les versions ayant moins de paramètres et quantizées à de petit nombre des bits s'exécutent beaucoup plus rapidement que celles ayant plusieurs paramètres ou une quantization moins significative. En revanche, le nombre des paramètres a un impact sur la qualité des réponses, un LLM disposant de plusieurs paramètres génère des réponses de bonne qualité quand bien-même il est quantizé à un petit nombre des bits.

Cette étude constitue une contribution dans la démocratisation de l'évaluation des LLM en présentant les éléments pour mettre en place un pipeline RAG avec un ordinateur classique. L'évaluation des LLM est effectuée dans le contexte bien précis de la protection des données, il n'est pas indéniable que les résultats obtenus changent avec une autre base de connaissance. Toutefois, cette étude peut ouvrir le champ de recherche à d'autres thématique comme l'évaluation de différentes techniques d'embedding, de différentes techniques de chunking ou simplement l'évaluation d'un autre corpus documentaire en adaptant le jeu des données à cet autre domaine.

Bibliographie

Babonnaud, W. (2024). Vers une conceptualisation du micro-benchmarking pour l'évaluation des LLM dans un cadre opérationnel. *EvalLLM2024 - Atelier sur l'évaluation des modèles génératifs (LLM) et challence d'extraction d'information few-shot.* Toulouse : AMIAD, Ministères des Armées.

Boulle, A. (2024). *Introduction pratique aux IA génératives de texte: les grands modèles de langage.*

Bouvard, C., Ciancone, M., Gourru, A., & Schaeffer, M. (2024). Évaluation comparative des approches RAG et fine-tuning. *10 ème Conférence Nationale sur les Applications*

Pratiques de l'Intelligence Artificielle (pp. 38-47). La Rochelle: AFIA-Association Française pour l'Intelligence Artificielle.

Duprieu, H., & Berkouk, N. (2024). *Techniques d'audit des grands modèles de langage.* Commission.

Edman, L., Fraser, A., & Schmid, H. (2025, 05 23). EXECUTE: A Multilingual Benchmark for LLM Token Understanding. *Arxiv*.

Edman, L., Schmid, H., & Fraser, A. (2024). CUTE: Measuring LLMs' Understanding of Their Tokens. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing,* (pp. 3017–3026). Miami: Association for Computational Linguistics.

Egashira, K., Vero, M., Staab, R., He, J., & Vec, M. (2024). *Exploiting LLM Quantization.* ArXiv.

Ganascia, J.-G. (2017). *Intelligence artificielle : Vers une domination programmée ?* Paris: Le Cavalier Bleu.

Guo, Z., & et al. (2023). *Evaluating Large Language Models: A Comprehensive Survey.* Tianjin University.

Hu, Q. J., & et al. (2024). *RouterBench: A Benchmark for Multi-LLM Routing System.* arXiv.

Khang, M., Park, S., Hong, T., & Jung, D. (2025). *CReSt: A Comprehensive Benchmark for Retrieval-Augmented Generation with Complex Reasoning over Structured Documents.*

Kim, S. G. (2025, 5 23). Chart-to-Experience: Benchmarking Multimodal LLMs for Predicting Experiential Impact of Charts. *Arxiv*.

Kooli, N., Flament, J., Dutrey, C., Diniz, N., & Claveau, V. (2024). EvalLLM 2024 : présentation de l'atelier Evaluation des LLM et du Challenge en extraction d'information few-shot. *EvalLLM2024 - Atelier sur l'évaluation des modèles génératifs (LLM) et challenge d'extraction d'information few-shot.* Toulouse: AMIAD, Ministère des Armées.

Liu, Z., & et al. (2023). *LLM-QAT: Data-Free Quantization Aware Training for Large Language Models.* ArXiv.

Schaeffer, M. (2025). *Towards efficient Knowledge Graph-based Retrieval. Augmented Generation for conversational agents.* Normandie Université.

Steffenel, L. A., & Lucas, L. (2025). L'Intérêt des RAG dans la Gestion des Connaissances des Processus Administratifs Universitaires à l'ère des LLM. *Atelier KM-IA - Gestion des connaissances tacites en entreprise : réflexions, retours d'expériences, bonnes pratiques et mauvaises surprises de l'intelligence artificielle.* Strasbourg.

Sun, H., Will Wang , H., Gu, J., Li, L., & Cheng, Y. (2025, 05 23). FullFront: Benchmarking MLLMs Across the Full Front-End Engineering Workflow. *Arxiv*.

https://en.wikipedia.org/wiki/ChatGPT?utm_source=chatgpt.com

https://rasa.com

https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/

https://huggingface.co/spaces/fr-gouv-coordination-ia/llm_leaderboard_fr#/

Annexe

Questionnaire d'évaluation

I. Questions sur les définitions : Donne-moi la définition de :

1. Accountability
2. Analyse d'impact
3. Anonymisation
4. Base légale
5. CNIL
6. Consentement
7. Cookies
8. Données personnelles
9. Données sensibles
10. DPO
11. Durée de conservation
12. Finalité du traitement
13. Exercice des droits
14. Minimisation des données
15. Personne concernée
16. Privacy by Default
17. Privacy by Design
18. Profilage
19. Pseudonymisation
20. Registre des traitements
21. Responsable de traitement
22. Sous-traitant
23. Traitement de données
24. Violation de données
25. Conformité

II. Question sur des exemples des cas

26. Mon entreprise veut installer un logiciel qui analyse les mails professionnels des employés pour détecter les fuites d'informations. Avons-nous le droit de le faire ?
27. Une école souhaite publier les photos de ses élèves sur son site web après la fête de fin d'année. Est-ce légal ?
28. Mon médecin m'a proposé de transmettre mes données de santé à un centre de recherche universitaire. Dois-je donner mon consentement ?
29. Une commune a placé une caméra de surveillance sur la place centrale du village, sans panneau d'information. Est-ce conforme à la législation sur la protection des données ?
30. Une entreprise vend sa base de données clients à une autre société après une fusion. Cette transmission est-elle autorisée ?
31. Un employeur exige que tous les employés activent la géolocalisation sur leur téléphone de service. Est-ce légal ?
32. Une société collecte les empreintes digitales de ses employés pour contrôler l'accès aux locaux. Cette méthode est-elle licite ?
33. Un parent installe une application espionne sur le téléphone de son enfant mineur de 16 ans. Est-ce juridiquement acceptable ?
34. Un journaliste publie des documents contenant des données personnelles dans le cadre d'une enquête d'intérêt public. Est-ce conforme au RGPD ?

35. Une mairie veut créer un registre des personnes vulnérables sur son territoire. Quels sont les risques en matière de données ?

III. Question de procédure :

36. Quelles sont les étapes à suivre pour effectuer un audit RGPD ?
37. Je suis l'administrateur du site web de l'université d'Avignon, donne-moi la procédure à suivre pour vérifier la conformité RGPD du site.
38. Mon entreprise souhaite collecter des données personnelles via un formulaire en ligne. Quelles sont les étapes juridiques à respecter avant la mise en ligne du formulaire ?

IV. Question sur les modèles de documents

39. Rédige-moi un modèle de registre de traitement
40. Rédige-moi un modèle de politique de confidentialité