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

Étude des déterminants impactant la qualité de l'information géographique chez les LLMs : famille, taille, langue, quantization et fine-tuning

Rémy Decoupes[1, 3] Adrien Guille[2]
(1) TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE. Maison de la Télédétection 500, rue J.F. Breton 34090 Montpellier
(2) Université Lumière Lyon 2, ERIC UR 3083, 5 av Pierre Mendes France, 69500 Bron, France
(3) INRAE remy.decoupes@inrae.fr, adrien.guille@univ-lyon2.fr

RÉSUMÉ

Nous analysons l'impact de plusieurs facteurs d'optimisation sur la qualité des informations géographiques contenues dans des grands modèles de langue (LLMs) : famille, taille, «quantization», «instruction fine-tuning», prompt et langue. Nous évaluons également la qualité des représentations internes, en particulier pour les modèles génératifs ayant des difficultés à suivre les instructions. Nos résultats montrent que la quantization dégrade nettement les performances, tandis que les versions conversationnelles («Instruct») perdent généralement en qualité d'informations par rapport à leur version «base», à l'exception des modèles de petite taille. L'ensemble de notre protocole d'évaluation est entièrement reproductible et disponible en accès libre.

ABSTRACT

Study of the determinants impacting the quality of geographic information in LLMs : family, size, language, quantization, and fine-tuning

We investigate the impact of optimization factors on the geographical information quality of large language models, including model family, size, quantization, instruction fine-tuning, prompt and language. Additionally, we assess the quality of internal representations, particularly for generative models that struggle with instruction following. Our findings indicate that quantization substantially degrades performance, while Instruction fine-tuning generally harms intrinsic knowledge — except in smaller models. The full evaluation protocol is reproducible and publicly available.

MOTS-CLÉS : LLM, information géographique, evaluation.

KEYWORDS: LLM, geographic information, evaluation.

1 Introduction

Les grands modèles de langue (*Large Language Models*, LLMs), notamment via les différents services web d'agent conversationnel ou assistant personnel, deviennent progressivement des concurrents aux moteurs de recherche pour la vérification et la recherche d'informations. En particulier, le troisième usage le plus fréquent des LLMs est de répondre à des questions de géographie (Zheng *et al.*, 2024), bien que ces informations soient aisément accessibles via des encyclopédies en ligne ou des services de cartographie comme Wikipédia ou OpenStreetMap.

Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM), pages 108–119.

Aujourd'hui, l'offre de LLMs disponible pour les utilisateurs est en pleine expansion, avec des modèles déclinés en différentes tailles et variantes (notamment par *instruction fine-tuning* et par niveau de *quantization*). Pour les utilisateurs souhaitant héberger localement un modèle, ou du moins éviter d'utiliser des LLMs hébergés sur des plateformes privées, le choix du modèle repose sur un compromis entre le coût des ressources matérielles et la pertinence des réponses fournies.

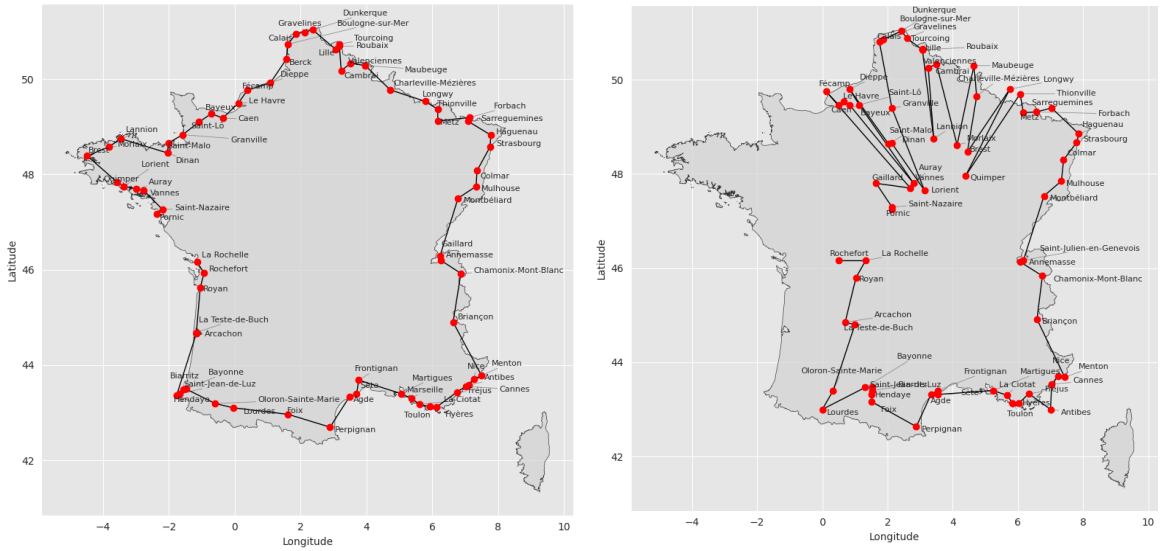
L'objectif de ce travail est d'évaluer les critères ayant le plus d'impact sur la qualité des informations géographiques des modèles. La géographie étant un domaine où la vérification des réponses est aisée, cette évaluation sert également de proxy pour identifier les facteurs influençant la qualité des informations détenues par les modèles.

Nous nous concentrons uniquement sur les informations géographiques intrinsèques aux modèles (sans agent et sans accès externe) et comparons l'impact des critères suivants :

- La famille du modèle
- La taille du modèle
- Le niveau de *quantization*
- L'entraînement (*base* ou *instruct*)
- Le type du *prompt*
- La langue choisie

Par ailleurs, faisant l'hypothèse que les modèles de langue encodent linéairement la sémantique des tokens dans leur espace de représentation (Jiang *et al.*, 2024), nous étudions si les coordonnées géographiques peuvent être décodées linéairement à partir des représentations de la couche de sortie, à l'image de ce que font (Gurnee & Tegmark, 2024).

Notre évaluation, reproductible et accessible sur `https://github.com/AdrienGuille/geo-llm`, aboutit à quelques recommandations : la *quantization* a un impact important, il est préférable d'utiliser un modèle avec moins de paramètres mais sans *quantization*. La phase d'ajustement aux instructions (*i.e. instruction fine-tuning*) appliquée aux modèles de base pour obtenir les variantes "Instruct" ou "Chat" détériore les informations géographiques intrinsèques aux modèles (hormis pour les modèles de petite taille).



(a) Mistral-Small-24B

(b) Mistral-7B_v0.1

FIGURE 1 – Anamorphose de la carte de France avec les coordonnées prédites par les LLMs

La figure 1 illustre les différences de qualité de l'information géographique restituée par les LLMs.

Cette carte affiche les coordonnées prédites des villes frontalières françaises montrant ainsi l'impact sur les frontières en chaînant par sens trigonométrique les coordonnées prédites.

2 État de l'art

Bien avant l'émergence des modèles de langue, des travaux ont mis en évidence que l'information géographique pouvait être implicitement encodée dans le langage naturel. En analysant les cooccurrences de toponymes dans des corpus, (Louwerse & Zwaan, 2009) ont montré qu'il était possible d'inférer, avec une certaine fiabilité, la proximité géographique entre les lieux mentionnés.

Cette information géographique, présente dans les textes, est naturellement captée par les modèles de langue. En particulier, les représentations vectorielles (embeddings) issues de ces modèles intègrent en partie des dimensions géographiques et culturelles (Decoupes *et al.*, 2025; Gurnee & Tegmark, 2024).

Cependant, les modèles de langue, souvent considérés comme des boîtes noires, sont connus pour reproduire ou amplifier divers biais, qu'ils soient politiques, raciaux, sexistes, etc (Kotek *et al.*, 2023). De plus, plusieurs études récentes (Decoupes *et al.*, 2025; Kruspe & Stillman, 2024; Liu *et al.*, 2024) ont mis en évidence l'existence de biais géographiques, définis comme une représentation inégale ou un traitement différencié selon les régions, les pays ou les cultures. Cette problématique est d'autant plus préoccupante que les jeux de données utilisés pour l'entraînement et l'alignement de ces modèles sont rarement accessibles, du fait d'une concurrence accrue entre les acteurs industriels. Or, une part importante des biais réside précisément dans ces jeux de données.

Face à cela, la communauté en traitement automatique des langues (TAL) a développé plusieurs benchmarks pour évaluer les performances des modèles. Néanmoins, l'information géographique y reste largement sous-représentée. Par exemple, le corpus Common Crawl ne contient que 18,7 % de documents géolocalisables (Ilyankou *et al.*, 2024), avec une couverture très inégale du globe. Les zones urbaines, économiquement prospères et anglophones y sont nettement surreprésentées.

Plusieurs travaux se sont intéressés à l'étude des biais géographiques, en analysant par exemple la présence de stéréotypes ou les variations de performance selon les régions. Toutefois, à notre connaissance, aucune étude ne s'est encore penchée sur l'impact des différents paramètres de configuration des modèles sur ces biais.

3 Méthodologie générale

Le processus d'évaluation de l'impact des différents critères sur la qualité des informations géographiques se compose de plusieurs étapes. La première consiste à charger les différentes variantes de modèles, puis à leur soumettre une série de prompts afin de récupérer deux types de sorties : d'une part, la réponse textuelle du modèle, et d'autre part, l'*embedding* associé au lieu recherché.

Deux traitements distincts sont ensuite réalisés. Le premier, en langage naturel, consiste à calculer la distance géographique entre les coordonnées GPS inférées par le modèle et les coordonnées exactes. Le second vise à entraîner un modèle de régression linéaire afin de prédire les coordonnées GPS à partir des *embeddings* des lieux. L'évaluation des LLMs repose également sur le calcul de la

distance géographique entre les coordonnées calculées par la régression à partir des *embeddings* et les emplacements réels.

3.1 Modèles

Nous considérons des modèles de langue de tailles variées, de 500 millions à 72 milliards de paramètres, dans leur version pré-entraînée et leur version spécialisée (repérées par le suffixe "-Instruct" pour tous les modèles, à l'exception de Llama 2 pour lequel le suffixe est "-chat") pour répondre aux instructions, tirées de différentes familles : Llama, Mistral et Qwen. Plus spécifiquement, nous menons nos expériences avec les modèles présentés dans le tableau 1, téléchargeables depuis le "Model Hub" HuggingFace. Par ailleurs, seul le niveau de *quantization int4* n'a pu être utilisé pour les modèles $\geq 70B$ car leur chargement était impossible sur une seule carte graphique (Nvidia A100 80GB).

Famille	Version	Base	Instruct / Chat
LLaMA	2	7B-hf, 13B	7B-chat-hf, 13B-chat
	3.1	8B, 70B	8B-Instruct, 70B-Instruct
	3.2	1B, 3B	1B-Instruct, 3B-Instruct
Mistral	v0.1	7B	Instruct-v0.1
	v0.2	N/A	Instruct-v0.2
	v0.3	7B	Instruct-v0.3
	Small-2501	24B-Base	24B-Instruct
Qwen	2.5	0.5B, 7B, 14B, 32B, 72B	0.5B-Instruct, 7B-Instruct, 14B-Instruct, 32B-Instruct, 72B-Instruct

TABLE 1 – Liste des modèles évalués, par famille, version et type (Base / Instruct ou Chat)

Codage des paramètres. À noter que, pour chacun des modèles listés ci-dessus, nous considérons les paramètres tels que partagés sur HuggingFace, codés sur 16 bits au format "BrainFloat", mais également les paramètres compressés et codés sur 4 bits et 8 bits à l'aide de la bibliothèque "bitsandbytes".

3.2 Données géographiques

Les données géographiques utilisées dans cette étude proviennent de GeoNames et ont été téléchargées à partir du site OpenDataSoft, voir section 7 pour reproduire notre étude. Le fichier contient les communes françaises de plus de 1000 habitants (soit 8853 communes). Afin de limiter le temps d'inférence des LLMs, nous avons sélectionné les 1000 communes les plus peuplées de France.

3.3 Protocole

Prompts. Nous considérons trois types de prompts, plus ou moins spécifiques, en français et en anglais :

P1 **Nom de la ville uniquement** : Le prompt fournit simplement le nom de la ville, *e.g.* « La Rochelle ».

P2 **Question simple** : Le prompt interroge la localisation de la ville, *e.g.* « Où se trouve la ville de La Rochelle ? » / « Where is the city of La Rochelle ? ».

P3 **Demande de coordonnées géographiques** : Le prompt demande explicitement les coordonnées GPS, *e.g.* « Quelles sont les coordonnées géographiques de la ville de La Rochelle ? » / « What are the geographical coordinates of the city of La Rochelle ? ».

Nous évaluons la capacité des modèles à approcher les coordonnées géographiques des villes à partir de ces prompts de deux façons, décrites brièvement ci-après.

Interrogation en langue naturelle. Dans un premier temps, nous tentons d'extraire les coordonnées à partir du texte généré par ces modèles en réponse au prompt **P3**. Il est à noter que nous générons les réponses en piochant le token le plus probable à chaque étape du processus, de sorte à assurer la reproductibilité des résultats – ce qui paraît raisonnable, dans la mesure où la réponse attendue est unique, à savoir une paire de coordonnées latitude/longitude.

Régression linéaire sur les représentations. Dans un second temps et sur l'ensemble des trois prompts, nous apprenons deux transformations linéaires entre la représentation des villes calculées par le modèle dans sa dernière couche et, d'une part, la latitude et, d'autre part, la longitude.

Évaluation. Les coordonnées prédites, qu'elles soient générées textuellement par un modèle ou linéairement estimées d'après les représentations calculées par un modèle, sont comparées avec les coordonnées réelles selon la distance Haversine, qui donne la distance géodésique entre ces points.

4 Interrogation en langue naturelle

4.1 Méthodologie

Seul le prompt **P3** « Quelles sont les coordonnées géographiques de la ville de `La Rochelle` » est soumis à chacune des variantes des modèles. Par exemple, la sortie du modèle Mistral-7B-v0.1 contient la phrase suivante *« La Rochelle est située à 46° 10′ 00″ N, 1° 20′ 00″ O »*.

Les réponses sont ensuite traitées à l'aide d'expressions régulières afin d'extraire et de convertir les coordonnées géographiques dans le format décimal. Cela permet de calculer la distance géographique, en utilisant la formule de Haversine, entre les points prédits et les coordonnées réelles.

Lorsque plus d'un quart des réponses d'une variante (soit 250 réponses sur 1000 villes) ne peuvent être exploitées, nous excluons cette variante de l'analyse afin de garantir une comparaison équitable entre modèles. En effet, une variante pourrait présenter de bons résultats globaux, mais uniquement sur un sous-ensemble limité de villes, ce qui biaiserait l'évaluation.

4.2 Résultats

La figure 2 propose une vue globale des résultats en affichant la distance moyenne (en Km) entre les coordonnées prédites et les vrais coordonnées en échelle logarithme pour l'ensemble des variations des modèles. Le modèle **Mistral-Small-24B-Base-2501** est la meilleure variante avec une erreur

moyenne de 87 km. Les versions $70B$ des familles Llama ont malheureusement généré trop de sorties non exploitables du fait de leur niveau de *quantization*. De manière globale, les versions *base* proposent les meilleures performances pour les modèles de grandes tailles (hormis pour Qwen $32B$) alors que pour les modèles de tailles $< 7B$, seules les versions *instruct* ou *chat* ont fourni des données exploitables. C'est pourquoi, nous proposons en section 5 d'analyser les modèles non pas sur leurs réponses mais sur leurs représentations internes (*embedding*).

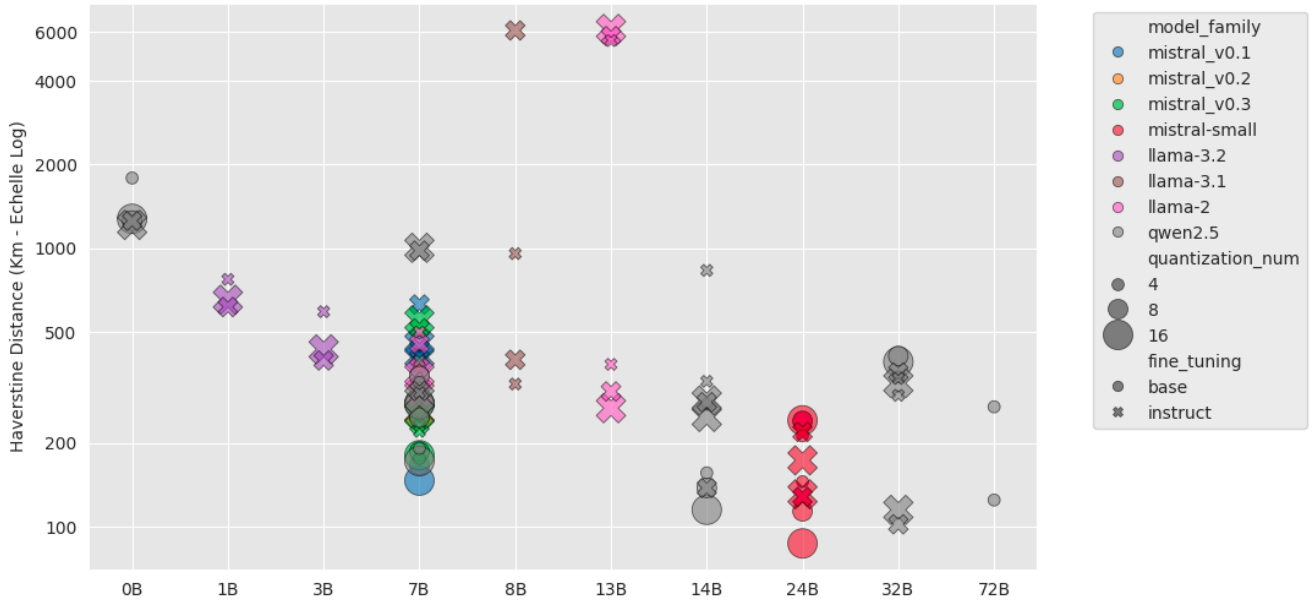


FIGURE 2 – Comparaison des critères pour l'interrogation en langue naturelle

Un autre aspect intéressant à prendre en compte est la variabilité de la qualité de l'information géographique inter-familles de modèles. La figure 3 montre que les différentes versions de Mistral et Qwen2.5 ont très peu de variabilité contrairement aux modèles des différentes familles de Llama.

Pour finir, le figure 4 propose d'évaluer l'impact du *fine-tuning*, de la langue et du niveau de *quantization* sur la variabilité de la qualité des informations géographiques.

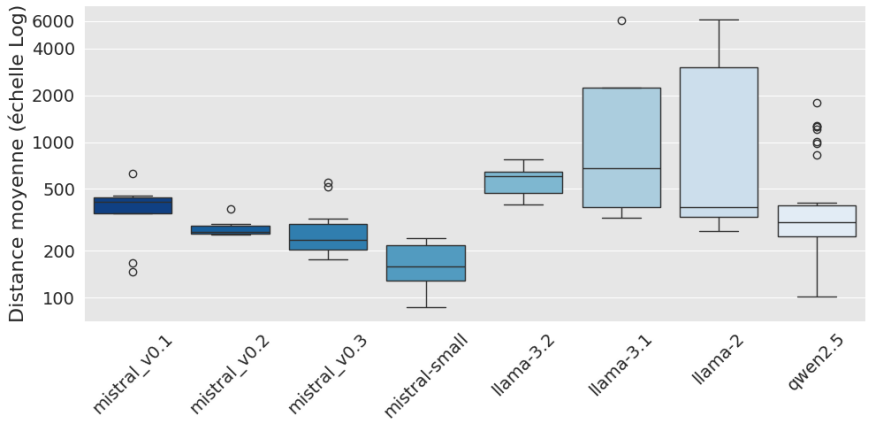
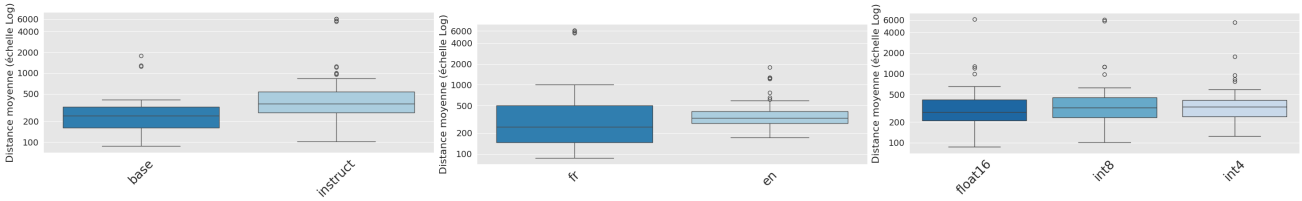


FIGURE 3 – Distribution des erreurs de distance par famille de modèles



(a) Fine-tuning (b) Langue (c) Quantization

FIGURE 4 – Analyse de l'impact des différents critères sur la variabilité de la qualité des informations geographiques

4.3 Analyse qualitative

L'ensemble des modèles (29) avec leurs variantes de *quantization* (3) et les prompts **P3** en anglais et français ont généré approximativement 130 expérimentations. Nous proposons, dans cette section, d'analyser certains comportements que nous avons jugés intéressants. L'ensemble des visualisations peut être consulté sur https://adrienguille.github.io/geo-llm/.

Discrétisation de l'espace géographique Certains modèles *instruct* de Mistral ont tendance à attribuer les mêmes coordonnées géographiques à des villes proches géographiquement, donnant l'impression d'une discrétisation de l'espace, comme illustré par la figure 5 où apparaît un quadrillage des points inférés (illustrés en rouge).

Positivité de la longitude Par ailleurs, on note chez certains modèles Mistral et Llama l'incapacité à formuler des longitudes de signe négatif. On observe une certaine symétrie axiale autour du méridien de Greenwich, particulièrement visible pour les villes de Bretagne, les plus proches de la pointe bretonne se trouvant d'autant plus décalées à l'est dans les réponses des modèles, comme illustrée par la figure 5.

5 Régression linéaire sur les représentations

5.1 Méthodologie

Nous échantillonnons 100 villes parmi les 1000 dont nous disposons, selon lesquelles nous ajustons, pour chaque modèle, deux transformations linéaires pour approcher respectivement la latitude et la longitude. Plus spécifiquement, nous procédons à une régression Ridge, le poids optimal du terme de pénalité L2 étant choisi par validation croisée sur ces 100 villes. Pour les villes dont le nom est découpé en plusieurs tokens – les modèles considérés étant tous des décodeurs avec un mécanisme d'attention uniquement vers la gauche, nous ne conservons que la représentation du dernier token. Nous mesurons les distances d'après les coordonnées estimées pour les 900 villes non utilisées pour le calcul de l'ajustement linéaire, que nous analysons dans la sous-section suivante.

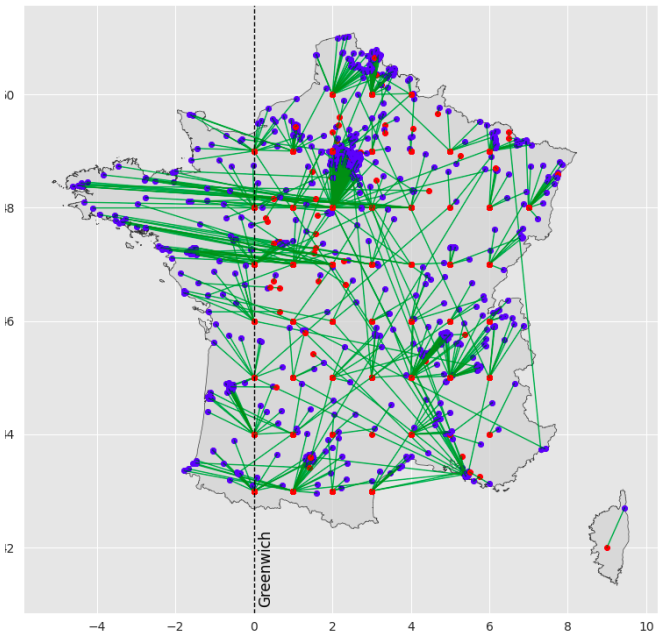


FIGURE 5 – Discrétisation de l'espace pour le modèle Mistral-7B-Instruct-v0.3_float16. Les points bleus correspondent aux vraies coordonnnées alors que les points rouges correspondent aux coordonnées inférées par le modèle.

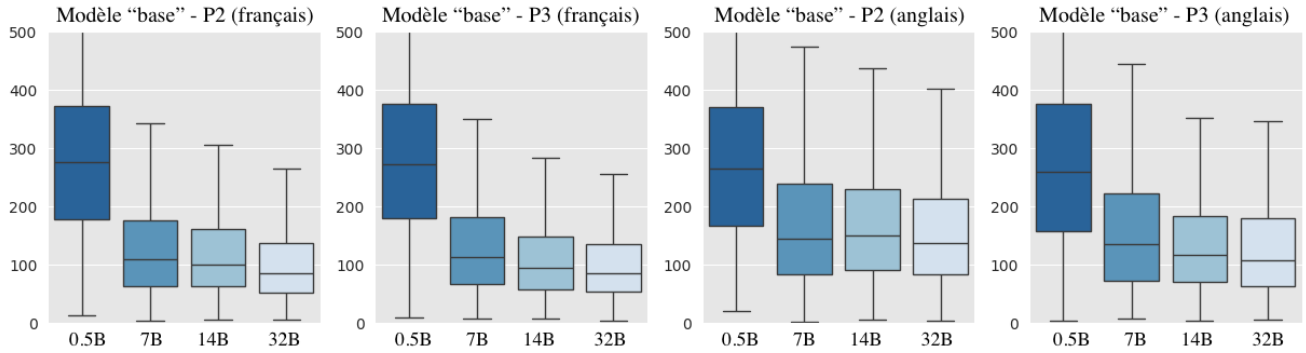


FIGURE 6 – Distributions des distances en km par ajustement linéaire, Qwen base.

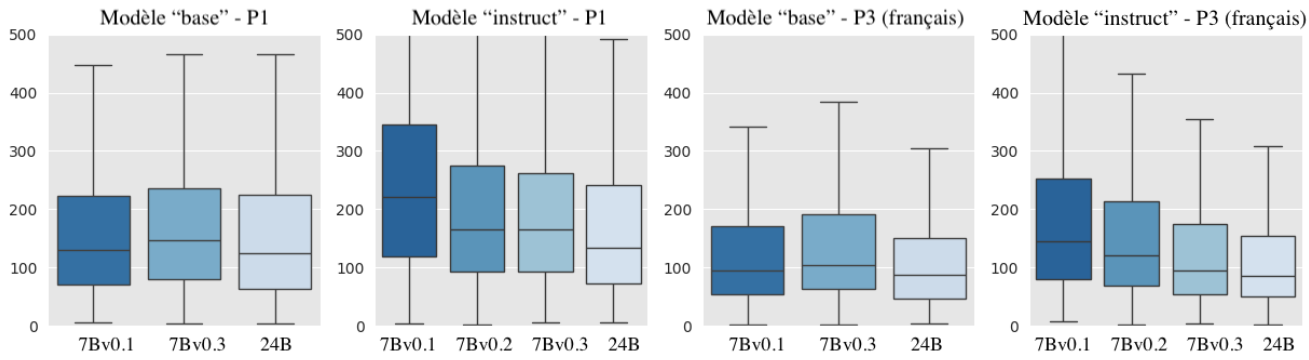


FIGURE 7 – Distributions des distances en km par ajustement linéaire, Mistral.

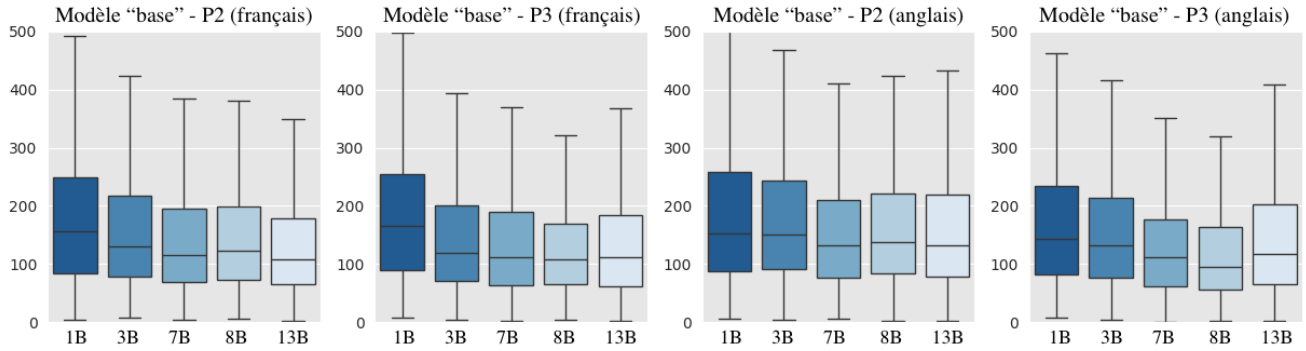


FIGURE 8 – Distributions des distances en km par ajustement linéaire, Llama base.

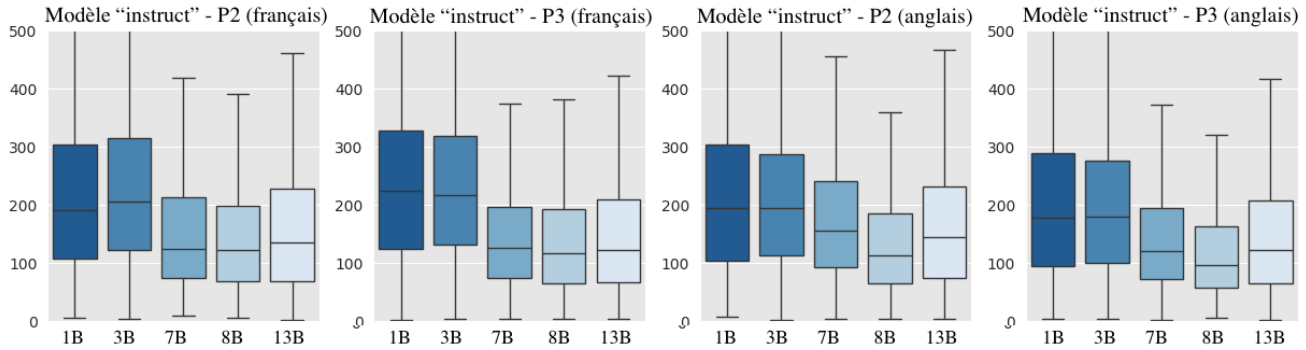


FIGURE 9 – Distributions des distances en km par ajustement linéaire, Llama ajustés aux instructions.

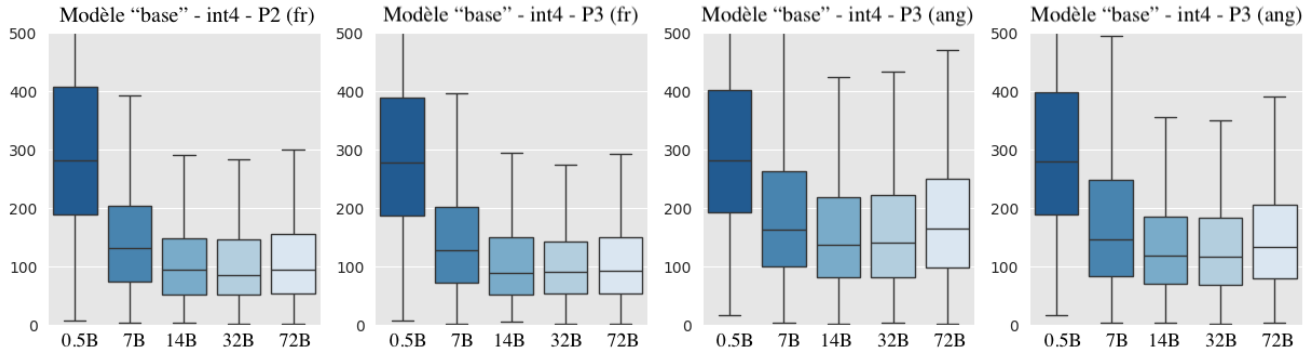


FIGURE 10 – Distributions des distances en km par ajustement linéaire, Qwen base sur 4 bits.

5.2 Résultats

Effet du nombre de paramètres Sans surprise, avec les paramètres codés sur 16 bits, on observe que la précision de l'ajustement tend à croître avec l'augmentation du nombre de paramètres pour les modèles de 500 millions à 32 milliards de paramètres (les modèles à 70 et 72 milliards de paramètres n'ont été exécutés qu'en précision 4 bits), comme illustré avec les modèles Qwen de base dans la figure 6.

Effet du prompt et de la langue Le prompt semble clé pour faire apparaître l'information géographique dans la représentation de la ville. En effet, on observe systématiquement un gain important entre les résultats obtenus avec les prompts **P2** et **P3** par rapport à ceux obtenus avec le prompt **P1**, comme on l'observe en comparant les parties gauche et droite de la figure 7. Globalement, le prompt demandant les coordonnées GPS (**P3**) conduit aux meilleures performances, en particulier lorsqu'il est rédigé en français. Il y a toutefois des exceptions, par exemple s'agissant du modèle Llama 3.1 à 8 milliards de paramètres. Se reporter à la figure 8 pour une illustration.

Effet de l'ajustement des modèles aux instructions Selon la famille de modèles considérée, on observe des effets différents. Les modèles de base de la famille Llama semblent calculer des représentations plus favorables à la modélisation linéaire des coordonnées géographiques que les modèles spécialisés pour répondre aux instructions, ce qui ressort par la comparaison des figures 8 et 9. Autrement dit, sous l'hypothèse de linéarité mentionnée en début d'article, la phase de spécialisation semble avoir engendré un certain oubli en matière de géographie de la France. A contrario, on ne constate pas de différence importante entre les modèles pré-entraînés des familles Mistral et Qwen, à l'exception de la première version spécialisée (*i.e.* v0.1) du modèle Mistral à 7 milliards de paramètres, différences corrigées progressivement par les versions ajustées successives (cf. figure 7).

Effet de la *quantization* On observe que la compression des paramètres sur 4 bits avec la bibliothèque "bitsandbytes" entraîne systématiquement un déclin des performances en régression linéaire. Plus surprenant, on note que les modèles Llama à 70 milliards de paramètres et Qwen à 72 milliards de paramètres codés sur 4 bits conduisent à des performances semblables, voire inférieures, à celles obtenues avec les modèles à 7 ou 14 milliards de paramètres codés sur 16 bits (ce qui ressort par exemple en comparant les figures 6 et 10).

5.3 Synthèse des observations

Il ressort de cette expérience que la phase de spécialisation des modèles peut affecter les compétences géographiques du modèle, du moins leur propension à coder linéairement l'information géographique dans l'espace de représentation, ce que l'on a observé de différentes façons pour la famille Llama et pour certains modèles Mistral. Il apparaît aussi que la compression des paramètres sur 4 bits affecte grandement les compétences géographiques, les modèles à 70 ou 72 milliards de paramètres n'étant pas plus performants que des modèles d'une taille inférieure d'un ordre, opérant quant à eux sur 16 bits. Enfin, la langue du prompt a également un effet notable, les prompts rédigés en français conduisant globalement à de meilleurs résultats que ceux rédigés en anglais.

6 Analyse croisée des résultats

La comparaison des résultats obtenus par interrogation directe et par régression linéaire fait ressortir la supériorité des familles Mistral et Qwen pour les deux approches par rapport à la famille Llama. De plus, alors que les performances en régression de Llama 3.1 et Llama 2 sont assez proches de celles des modèles comparables, ceux-ci donnent des résultats nettement plus mauvais que les autres quand on les interroge directement. Ceci suggère une certaine incapacité de ces modèles à formuler dans leurs réponses l'information géographique qu'ils semblent pourtant détenir.

7 Reproductibilité de l'évaluation

Les évaluations ainsi que les post-traitements et les visualisations sont reproductibles à travers le dépôt `https://github.com/AdrienGuille/geo-llm`. Pour cet article, les expérimentations ont été réalisées sur une machine avec le système d'exploitation *Ubuntu 22.04* dotée d'une carte graphique *NVIDIA A100* pour une durée approximative de 150 heures (6 jours). L'espace disque consommé pour télécharger l'ensemble des poids des modèles est de 1,1 To. La sélection et l'ajout ou suppression de LLMs ainsi que leurs variantes sont configurables.

8 Conclusion

Cette étude vise à aider à sélectionner la taille d'un LLM et son niveau de "quantization" pour une tâche de recherche d'information (obtention de coordonnées géographiques). Elle permet de formuler trois recommandations : il est préférable de choisir un modèle avec moins de paramètres mais non "quantizé", d'éviter les modèles "Intruction fine-tuned" car ayant perdu en qualité de l'information géographique et de choisir, malgré tout, un modèle suivant les instructions des prompts. Cette étude intensive est reproductible et adaptable aux futurs nouveaux LLMs.

Références

DECOUPES R., INTERDONATO R., ROCHE M., TEISSEIRE M. & VALENTIN S. (2025). Evaluation of Geographical Distortions in Language Models. In D. PEDRESCHI, A. MONREALE, R. GUIDOTTI, R. PELLUNGRINI & F. NARETTO, Éds., *Discovery Science*, p. 86–100, Cham : Springer Nature Switzerland. DOI : 10.1007/978-3-031-78977-9_6.

GURNEE W. & TEGMARK M. (2024). Language Models Represent Space and Time. In *ICLR 2024*.

ILYANKOU I., WANG M., HAWORTH J. & CAVAZZI S. (2024). Quantifying Geospatial in the Common Crawl Corpus. arXiv :2406.04952 [cs].

JIANG Y., RAJENDRAN G., RAVIKUMAR P. K., ARAGAM B. & VEITCH V. (2024). On the origins of linear representations in large language models. In *ICML 2024*.

KOTEK H., DOCKUM R. & SUN D. Q. (2023). Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, p. 12–24. arXiv :2308.14921 [cs], DOI : 10.1145/3582269.3615599.

KRUSPE A. & STILLMAN M. (2024). Saxony-anhalt is the worst : Bias towards german federal states in large language models. In A. HOTHO & S. RUDOLPH, Éds., *KI 2024 : Advances in Artificial Intelligence*, p. 160–174, Cham : Springer Nature Switzerland.

LIU Z., JANOWICZ K., CURRIER K. & SHI M. (2024). Measuring Geographic Diversity of Foundation Models with a Natural Language–based Geo-guessing Experiment on GPT-4. *AGILE : GIScience Series*, **5**, 1–7. DOI : 10.5194/agile-giss-5-38-2024.

LOUWERSE M. M. & ZWAAN R. A. (2009). Language Encodes Geographical Information. *Cognitive Science*, **33**(1), 51–73. DOI : 10.1111/j.1551-6709.2008.01003.x.

ZHENG L., CHIANG W.-L., SHENG Y., LI T., ZHUANG S., WU Z., ZHUANG Y., LI Z., LIN Z., XING E. P., GONZALEZ J. E., STOICA I. & ZHANG H. (2024). LMSYS-Chat-1M : A Large-Scale Real-World LLM Conversation Dataset. In *ICLR 2024*.