Determinants of Geographic Information Quality in Large Language Models: Effects of Model Family, Scale, Language, Quantization, and 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, France
[2]Université Lumière Lyon 2, ERIC UR 3083, 5 avenue Pierre Mendès France, 69500 Bron, France
[3]INRAE, France

Abstract

We investigate the impact of model and prompt configurations on the quality of geographical information in large language models, considering factors such as model family, scale, quantization, instruction fine-tuning, prompt design, and language. In addition, we assess the quality of internal representations, with a particular focus on small generative models that struggle to follow instructions. Our results show 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.

Keywords

LLM, Geographic evaluation, Probing

1. Introduction

Large Language Models (LLMs), particularly through various conversational or personal assistant web services, are progressively becoming competitors to traditional search engines for information retrieval and verification[1]. Notably, the third most common use case for LLMs is answering geography-related questions [1], despite such information being readily accessible through online encyclopedias or mapping services, such as Wikipedia or OpenStreetMap.

Today, the range of LLMs available to users is rapidly expanding, with models released in various sizes and variants (notably through instruction fine-tuning and different levels of quantization). For users seeking to host models locally or avoid reliance on privately hosted LLMs, model selection and configuration involve a trade-off between hardware resource costs and the relevance of the generated responses.

The objective of this work is to evaluate the criteria that have the greatest impact on the quality of geographic information produced by LLMs. We are particularly interested in geographic information and use cases related to France; therefore, we center our experiments on its metropolitan territory, including French as a prompting language. We focus exclusively on the intrinsic geographic knowledge of models (without the use of agents or external data sources) and compare the impact of the following criteria:

- Model family
- Model size
- Level of quantization
- Training type (*base* or *instruct*)
- Prompt type
- Language (*English* or *French*)

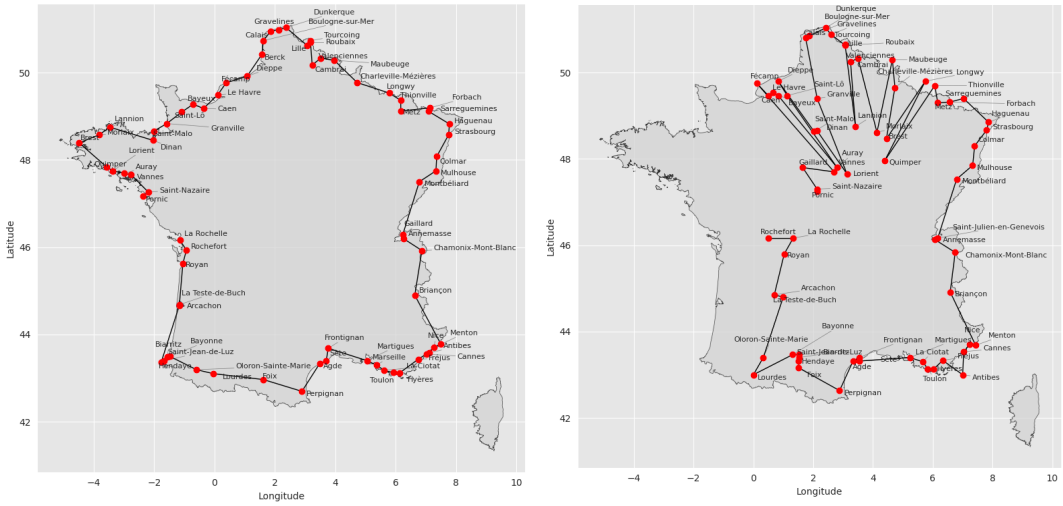
GeoExT 2026: Fourth International Workshop on Geographic Information Extraction from Texts at ECIR 2026, April 2, 2026, Delft, The Netherlands

*Corresponding author.
†These authors contributed equally.
remy.decoupes@inrae.fr (R. Decoupes); adrien.guille@univ-lyon2.fr (A. Guille)
https://orcid.org/0000-0003-0863-9581 (R. Decoupes); https://adrienguille.github.io/ (A. Guille)

[1]Report "From Googling to Asking ChatGPT" by the NGO AI Forensics: https://aiforensics.org/work/governing-ai-search

Additionally, since smaller models (<7B) sometimes struggle to follow prompt instructions, evaluating their geographical knowledge solely based on answer correctness may be unfair. Based on the hypothesis that language models encode token semantics linearly in their representation space [2], we investigate whether geographic coordinates can be decoded linearly from the representations of the output layer, following the approach of Gurnee et al. (2023) [3].

Our evaluation, reproducible and available at https://github.com/adrienguille/geo-llm, leads to several recommendations. Quantization has a strong negative impact; it is preferable to use a smaller model without quantization. Moreover, instruction fine-tuning applied to base models to obtain "Instruct" or "Chat" variants generally degrades the intrinsic geographic information encoded in the models, with the exception of small-scale models.



(a) Mistral-Small-24B

(b) Mistral-7B_v0.1

Figure 1: Anamorphosis of the map of France using coordinates predicted by LLMs.

One of the main results of this study is illustrated by Figure 1, which shows the differences in the quality of geographic information retrieved by LLMs (*Mistral-Small-24B* and *Mistral-7B_v0.1*). The map displays the predicted coordinates of French border cities, highlighting the distortions of national boundaries when the predicted coordinates are connected in radial order. The greatest distortions observed for *Mistral-7B_v0.1* result from difficulties in predicting negative values (see Figure 5).

2. Related works

Long before the emergence of language models, several studies had shown that geographic information could be implicitly encoded in natural language. By analyzing the co-occurrences of toponyms in textual corpora, Louwerse et al. (2009) [4] demonstrated that it was possible to infer, with reasonable reliability, the geographic proximity between the mentioned places. This geographic information, present in texts, is naturally captured by language models. In particular, the vector representations (embeddings) produced by these models partly integrate geographic and cultural dimensions [5, 3].

Building on this implicit geographic knowledge, several studies have sought to systematically evaluate the geographic reasoning capabilities of language models [6, 7, 5, 8]. One common method is to prompt the model with geographic questions and measure the errors between its outputs and factual information [8, 5, 6]. Called probing, this approach has also been applied to GPS coordinates [6], similar to our study but focused on the impact of few-shot learning rather than on the effect of different model configurations. Other studies have explored whether embeddings used by language models encode geographic distance information [5] highlighting the existence of geographic biases [9, 10, 11], defined as unequal representation or differentiated treatment depending on regions. Previous research has

additionally examined how different languages influence the geographic capabilities of LLMs [12, 13], again the impact of varying model configurations has not been evaluated

Unfortunately, geographic biases remain prevalent in both LLMs and their training data. For instance, the Common Crawl corpus contains only 18.7% of geolocatable documents [14], with highly uneven global coverage. Urban, economically prosperous, and English-speaking areas are clearly overrepresented.

While prior studies have evaluated the impact of prompts, model size, and the choice of natural language on geographic capabilities [8, 15], they have not investigated the effects of model quantization, the choice between base and instruction fine-tuned versions, or the geographic knowledge of smaller generative models that are incapable of following instructions necessary for standard geographic benchmarking. Our study addresses these gaps, providing a more comprehensive evaluation of LLMs' geographic knowledge across different configurations and model scales.

3. General Methodology

The process used to evaluate the impact of different criteria on the quality of geographic information consists of several steps. The first step involves submitting a series of three prompts (both in English and French) in order to retrieve two types of outputs: the model's textual response, and, the embedding associated with the queried location. Here are the prompts:

- **P1** **City name only**: The prompt provides only the name of the city, *e.g.*, "La Rochelle".
- **P2** **Simple question**: The prompt asks for the location of the city, *e.g.*, "Où se trouve la ville de La Rochelle?" / "Where is the city of La Rochelle?"
- **P3** **Request for geographic coordinates**: The prompt explicitly asks for GPS coordinates, *e.g.*, "Quelles sont les coordonnées géographiques de la ville de La Rochelle?" / "What are the geographical coordinates of the city of La Rochelle?"

Two distinct analyses are then performed. The first, in natural language and similar to [6], consists of computing the geographic distance between the GPS coordinates inferred by the model and the ground-truth coordinates. To do so, with a regular expression, we attempt to extract coordinates directly from the text generated by the models in response to prompt **P3**. Note that we generate model outputs by sampling the most probable token at each decoding step, ensuring result reproducibility. This approach is justified since the expected answer is unique, namely a single pair of latitude/longitude coordinates.

The second aims to train a linear regression model to predict GPS coordinates directly from the embeddings [3]. Across all three prompts, we train two linear transformations between the model's final-layer city representations (embeddings) and, respectively, latitude and longitude.

Then, the computed error is the geographic distance between the coordinates predicted by the LLM or by the regression and the actual locations.

3.1. Models

We consider language models of varying sizes, ranging from 500 million to 72 billion parameters, in both their pre-trained (i.e. the "base" version) and instruction-tuned versions (identified by the suffix "-Instruct" for all models, except for Llama 2 where the suffix is "-chat"). These models belong to three different families, including Llama [16, 17], Mistral [18, 19], and Qwen [20]. More specifically, our experiments are conducted with the models presented in Table 1, all downloadable from the HuggingFace Model Hub. Furthermore, for models with at least 70B parameters, only 4-bit quantization was feasible, as these models could not fit into the memory of a single GPU. This setting reflects typical small or medium enterprise deployment constraints, where inference is often limited to one GPU per server.

Parameter Encoding. For each model listed above, we consider the HuggingFace-released parameters in 16-bit bfloat format, as well as compressed variants encoded in 4-bit and 8-bit precision using

Family	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
List of evaluated models, by family, version, and type (Base / Instruct or Chat)

the bitsandbytes library [21]. Other quantization optimization techniques, such as Quantization-Aware Training, have not been compared.

3.2. Geographic Data

The geographic data used in this study come from GeoNames and were downloaded from the Open-DataSoft platform (see Section 6 for reproducibility details). The dataset contains French municipalities with more than 1,000 inhabitants (a total of 8,853 municipalities). To limit LLM inference time, we selected the 1,000 most populated municipalities in France.

4. Natural Language Querying

In this section, we present the methodology and corresponding results for the natural language approach.

4.1. Methodology

Only prompt **P3** is submitted to each model variant. For example, the output of the model Mistral-7B-v0.1 includes the sentence: *"La Rochelle is located at 46° 10′ 00″ N, 1° 20′ 00″ W."*

The responses are then processed using regular expressions to extract and convert the geographic coordinates into decimal format. Then, we compute the geographic distances (which we report in km) using the Haversine formula between the predicted points and the true coordinates.

When more than one quarter of a model variant's responses (i.e., 250 out of 1,000 cities) cannot be processed, that variant is excluded from the analysis to ensure a fair comparison between models. Indeed, a variant might display seemingly good overall results but only for a limited subset of cities, which would bias the evaluation.

Limitations We observed that models with 7B parameters or fewer, particularly under French prompts, sometimes include non-standard Unicode characters in their raw outputs that prevent reliable regex parsing, truncating coordinates at the unit level. These characters seem to disappear when the text is exported to a file, making debugging complex. Thus, the issue affects the evaluation.

4.2. Results

The set of models (29) with their quantization variants (3) and the P3 prompts in English and French generated approximately 130 experiments. In this section, we propose to analyze a behavior that we deem interesting. All visualizations can be consulted at https://adrienguille.github.io/geo-llm/.

Figure 2 provides an overview of the results by displaying the mean distance (in km) between the predicted and true coordinates on a logarithmic scale for all model variants. The **Mistral-Small-24B-Base-2501** model achieves the best performance, with a mean error of 87 km. The 70B Llama family models, unfortunately, produced too many unusable outputs due to their level of *quantization*. Overall, the *base* versions offer the best performance for larger models (except for Qwen 32B), while for smaller models (< 7B), only the *instruct* or *chat* variants yielded usable data. For this reason, in Section 5, we propose analyzing the models not through their textual outputs but through their internal representations (*embeddings*).

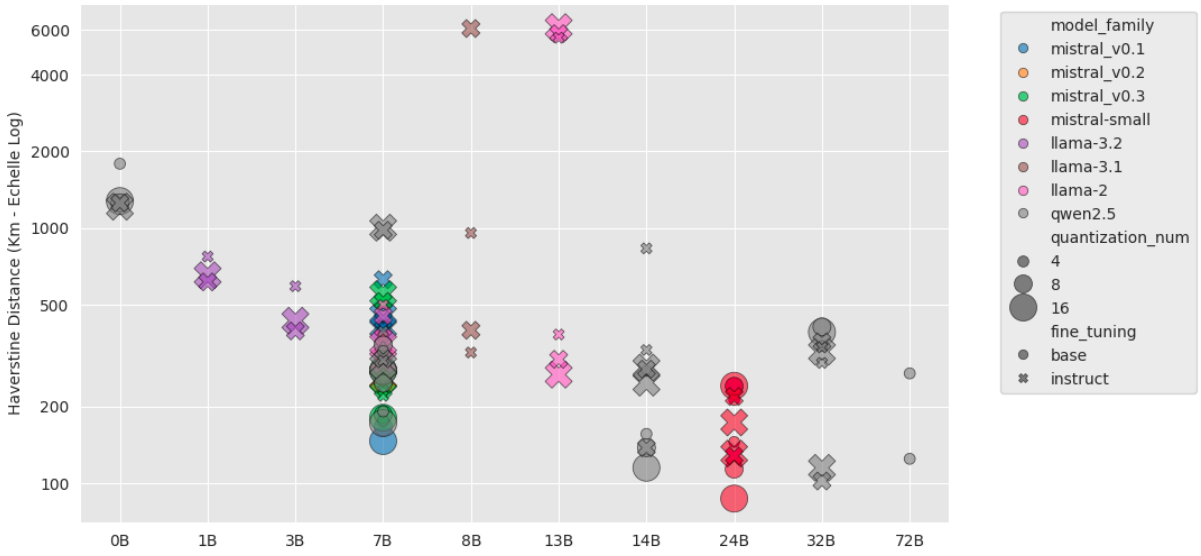


Figure 2: Comparison of criteria for natural language querying

Another interesting aspect to consider is the variability in the quality of geographic information across model families. Figure 3 shows that the different versions of Mistral and Qwen2.5 have very little variability, unlike the models from the different Llama families.

Figure 4 proposes evaluating the impact of fine-tuning, language, and quantization level on the variability of geographic information quality.

Finally, we observe in certain Mistral and Llama models an inability to formulate longitudes with a negative sign. We observe a certain axial symmetry around the Greenwich meridian, particularly visible for cities in Brittany. The cities closest to the tip of Brittany are shifted further to the east in the models' responses, as illustrated by Figure 5.

5. Linear Regression on Representations

In this section, we present the methodology and results for the regression approach based on embeddings to predict GPS coordinates. This approach is motivated by the fact that some smaller LLM configurations were unable to follow the instructions required by prompt **P3**; however, we aim to assess whether their embedding representations nonetheless capture geographic information.

5.1. Methodology

Leveraging the observation made by Gurnee and Tegmark (2024), according to which LLMs learn a linear representation of time and space in their last layer, we fit two linear models to predict, respectively, the latitude and longitude, from the last layer's embeddings. More specifically, we randomly sample 100 cities from the 1,000 most populous cities in France and, for each model configuration, perform a Ridge regression, with the L2 regularization strength selected by cross-validation on the sampled cities.

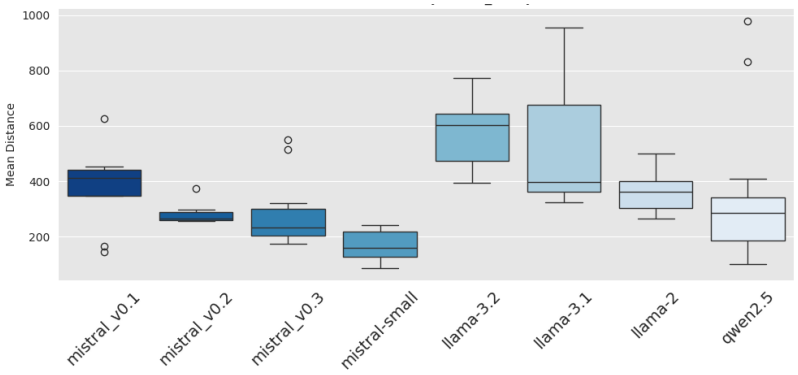
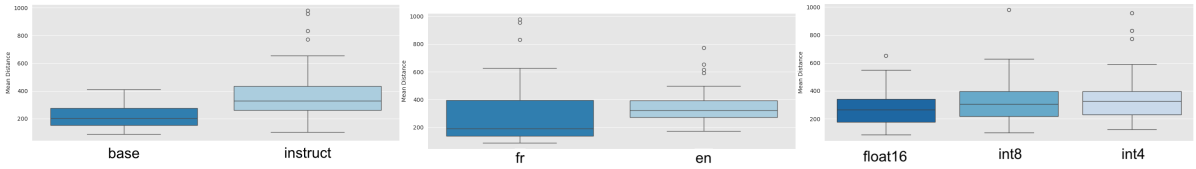


Figure 3: Distribution of distance errors by model family



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

Figure 4: Analysis of the impact of different criteria on the variability of geographic information quality

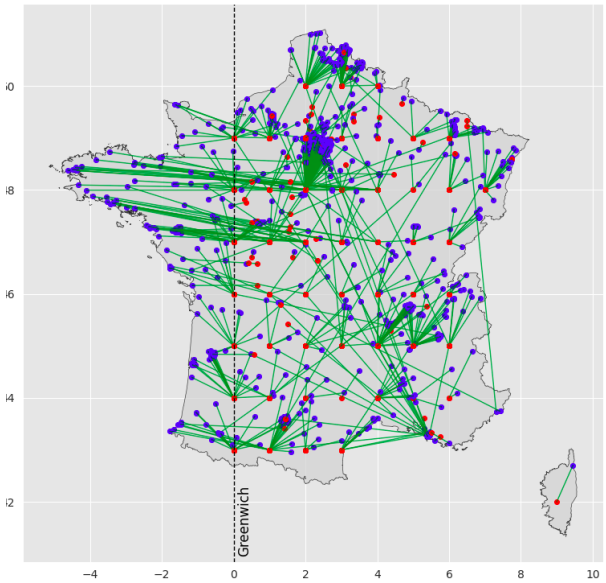


Figure 5: Positivity of Longitude error with the Mistral-7B-Instruct-v0.3_float16 model. The blue dots correspond to the true coordinates, while the red dots correspond to the coordinates inferred by the model, and green lines indicate the connections between the true and the predicted points.

For city names tokenized into multiple tokens, and since all considered models are decoder-only with left-to-right attention, we retain only the representation of the final token. We measure the distances based on the estimated coordinates for the 900 cities not used to calculate the linear adjustment, which we analyze in the following subsection.

5.2. Results

Effect of the Number of Parameters Unsurprisingly, with parameters encoded in 16 bits, we observe that the precision of the adjustment tends to increase with the growth in the number of

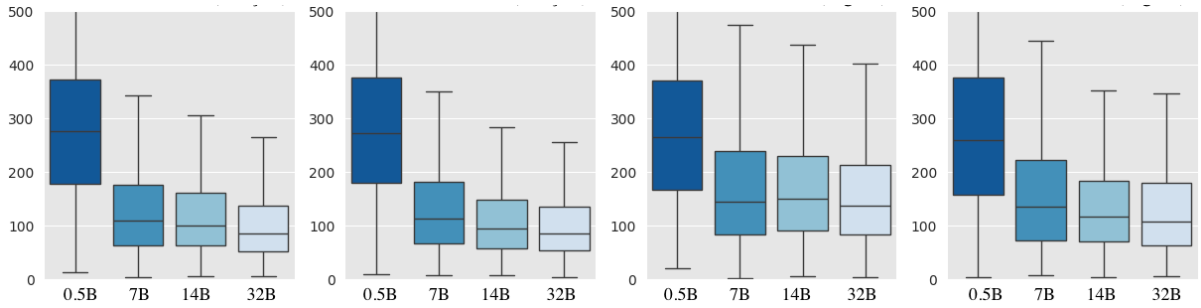


Figure 6: Distributions of distances in km by linear adjustment, Qwen base. Sub-figures from left to right: (a) P2 - French, (b) P3 - French, (c) P2 - English, (d) P3 - English.

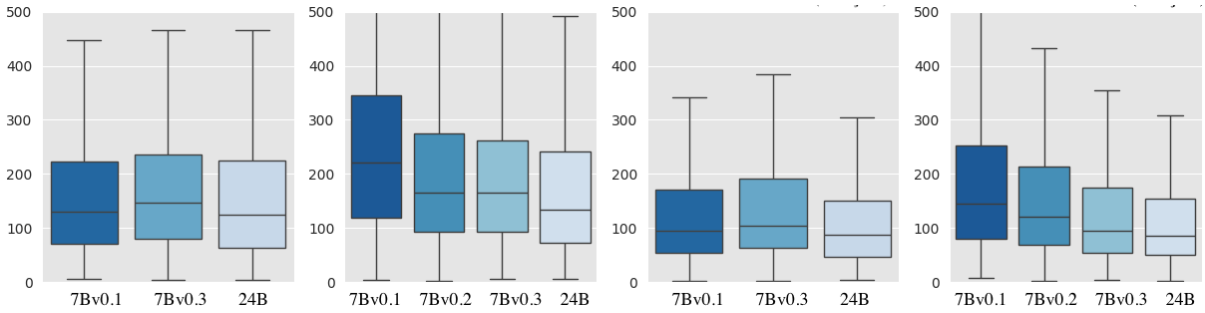


Figure 7: Distributions of distances in km by linear adjustment, for Mistral "base" and "instruct" versions. Sub-figures from left to right: (a) Base P1, (b) Instruct P1, (c) Base P3 French, (d) Instruct P3 French.

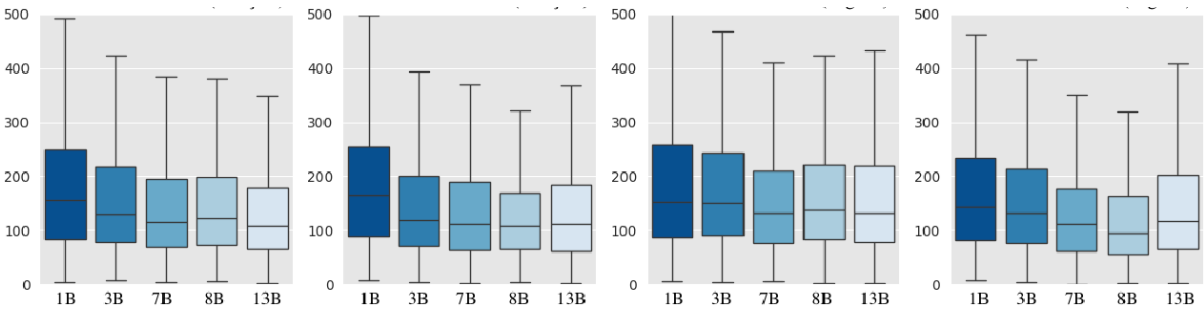


Figure 8: Distributions of distances in km by linear adjustment, LLaMA base. Sub-figures from left to right: (a) P2 - French, (b) P3 - French, (c) P2 - English, (d) P3 - English.

parameters for models from 500 million to 32 billion parameters (the 70 and 72 billion parameter models were only executed in 4-bit precision), as illustrated with the base Qwen models in Figure 6.

Effect of the Prompt and Language The prompt appears to be key to making the geographic information appear in the city's representation. Indeed, a significant gain is systematically observed between the results obtained with prompts **P2** and **P3** compared to those obtained with prompt **P1**, as seen by comparing the left and right parts of Figure 7. Globally, the prompt requesting GPS coordinates (**P3**) leads to the best performance, particularly when written in French. There are exceptions, however, for example concerning the Llama 3.1 model with 8 billion parameters. Refer to Figure 8 for an illustration.

Effect of Model Adjustment to Instructions Depending on the model family considered, different effects are observed. The base models of the Llama family seem to calculate representations that are more favorable to the linear modeling of geographic coordinates than the models specialized for following instructions, which emerges from the comparison of Figures 8 and 9. This aligns with prior work where this linearity property was observed in a Llama base model [3]. In contrast, we observe a

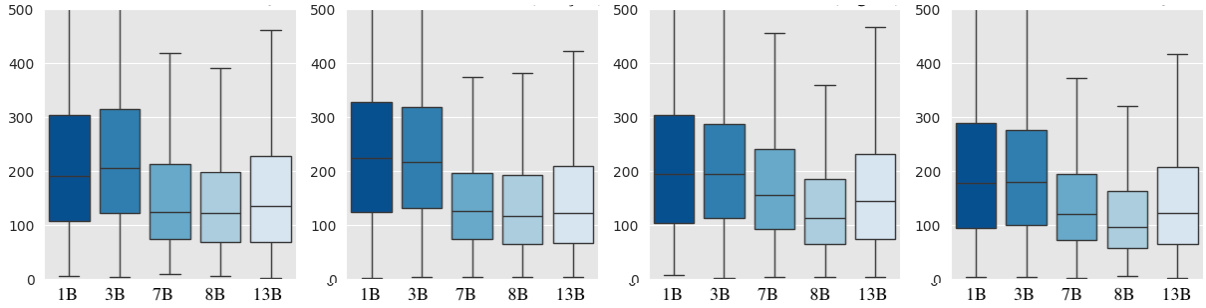


Figure 9: Distributions of distances in km by linear adjustment, LLaMa Instruct. Sub-figures from left to right: (a) P2 - French, (b) P3 - French, (c) P2 - English, (d) P3 - English.

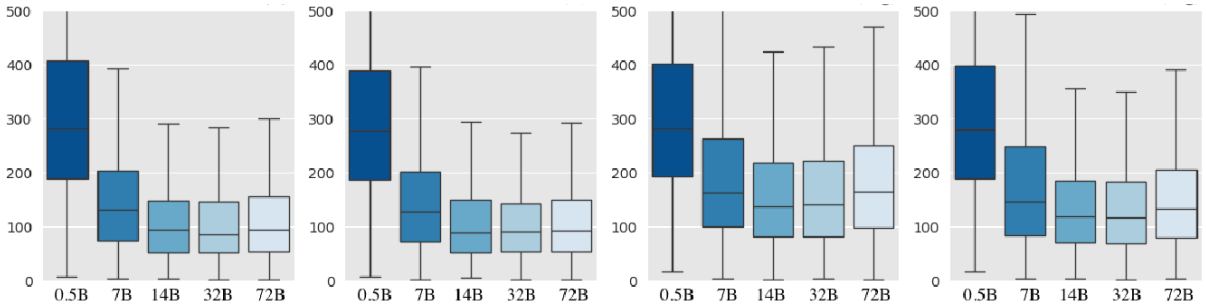


Figure 10: Distributions of distances in km by linear adjustment, Qwen base 4 bits quantization. Sub-figures from left to right: (a) P2 - French, (b) P3 - French, (c) P2 - English, (d) P3 - English

decrease in performance between the fine-tuned models and the base model, which could be interpreted in two ways: (i) while fine-tuning preserves geographical information, it disrupts the linearity of spatial relations, or (ii) that geographical information might be diluted in order to acquire stronger conversational capabilities. Further experiments with non-linear regression would be required in order to assess which interpretation is the most likely. Conversely, no significant difference is noted between the pre-trained models of the Mistral and Qwen families, with the exception of the v0.1 of the $7B$ parameter Mistral model, differences that were progressively corrected by successive adjusted versions (see Figure 7).

Effect of Quantization We observe that the compression of parameters to 4 bits using the "bitsandbytes" library systematically leads to a decline in linear regression performance. More surprisingly, we note that the 70 billion parameter Llama models and the 72 billion parameter Qwen models encoded in 4 bits lead to performances similar to, or even lower than, those obtained with the 7 or 14 billion parameter models encoded in 16 bits (which emerges, for example, by comparing Figures 6 and 10).

6. Reproducibility of the Evaluation

The evaluations, as well as the post-processing and visualizations, are reproducible via the repository https://github.com/adrienguille/geo-llm. For this article, the experiments were performed on a machine with the *Ubuntu 22.04* operating system, equipped with an *NVIDIA A100* graphics card, for an approximate duration of 150 hours (6 days). The disk space consumed to download all model weights is 1.1 TB. The selection and the addition or removal of LLMs, as well as their variants, are configurable.

7. Conclusion and Future Work

This study aims to help select the size of an LLM and its level of "quantization" for an information retrieval task (obtaining geographic coordinates). It allows for the formulation of these recommendations: it

is preferable to choose the largest model that fits GPU memory without quantization and to avoid "Instruction fine-tuned" models as they have lower quality in geographic information, but are capable of following prompt instructions. In future work, we plan to investigate non-linear regression to probe the representation of space in the embeddings of instruct models. Another interesting perspective could be to assess the impact of different quantization techniques, especially calibrated techniques.

Declaration on Generative AI

The authors used Generative Artificial Intelligence (AI) tools to support the preparation of this manuscript. Specifically, AI-based language models were employed to assist in improving the clarity and readability of the text, as well as to help initiate the development of certain parts of the accompanying code. All scientific interpretations, methodological choices, and conclusions remain entirely those of the authors.

References

[1] L. Zheng, W.-L. Chiang, Y. Sheng, T. Li, S. Zhuang, Z. Wu, Y. Zhuang, Z. Li, Z. Lin, E. P. Xing, J. E. Gonzalez, I. Stoica, H. Zhang, LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset, in: ICLR 2024, 2024.

[2] Y. Jiang, G. Rajendran, P. K. Ravikumar, B. Aragam, V. Veitch, On the origins of linear representations in large language models, in: ICML 2024, 2024.

[3] W. Gurnee, M. Tegmark, Language Models Represent Space and Time, in: ICLR 2024, 2024.

[4] M. M. Louwerse, R. A. Zwaan, Language Encodes Geographical Information, Cognitive Science 33 (2009) 51–73. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1551-6709.2008.01003.x. doi:10.1111/j.1551-6709.2008.01003.x.

[5] R. Decoupes, R. Interdonato, M. Roche, M. Teisseire, S. Valentin, Evaluation of geographical distortions in language models, Machine Learning 114 (2025) 263. URL: https://link.springer.com/10.1007/s10994-025-06916-9. doi:10.1007/s10994-025-06916-9.

[6] P. Bhandari, A. Anastasopoulos, D. Pfoser, Are Large Language Models Geospatially Knowledgeable?, in: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, ACM, Hamburg Germany, 2023, pp. 1–4. URL: https://dl.acm.org/doi/10.1145/3589132.3625625. doi:10.1145/3589132.3625625.

[7] N. Godey, E. d. l. Clergerie, B. Sagot, On the Scaling Laws of Geographical Representation in Language Models, 2024. URL: http://arxiv.org/abs/2402.19406. doi:10.48550/arXiv.2402.19406, arXiv:2402.19406 [cs].

[8] M. Moayeri, E. Tabassi, S. Feizi, WorldBench: Quantifying Geographic Disparities in LLM Factual Recall, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, ACM, Rio de Janeiro Brazil, 2024, pp. 1211–1228. URL: https://dl.acm.org/doi/10.1145/3630106.3658967. doi:10.1145/3630106.3658967.

[9] A. Kruspe, M. Stillman, Saxony-anhalt is the worst: Bias towards german federal states in large language models, in: A. Hotho, S. Rudolph (Eds.), KI 2024: Advances in Artificial Intelligence, Springer Nature Switzerland, Cham, 2024, pp. 160–174.

[10] Z. Liu, K. Janowicz, K. Currier, M. Shi, Measuring Geographic Diversity of Foundation Models with a Natural Language–based Geo-guessing Experiment on GPT-4, AGILE: GIScience Series 5 (2024) 1–7. URL: https://agile-giss.copernicus.org/articles/5/38/2024/. doi:10.5194/agile-giss-5-38-2024.

[11] M. Stillman, A. Kruspe, Biased Geolocation in LLMs: Experiments on Probing LLMs for Geographic Knowledge and Reasoning, in: Proceedings of The GeoExT 2025: Geographic Information Extraction from Texts Workshop, volume CEUR-WS volume 3969, Lucca, Italy, 2025.

[12] F. Faisal, A. Anastasopoulos, Geographic and Geopolitical Biases of Language Models, in: Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL), Association for Com-

putational Linguistics, Singapore, 2023, pp. 139–163. URL: https://aclanthology.org/2023.mrl-1.12. doi:10.18653/v1/2023.mrl-1.12.

[13] F. Faisal, Y. Wang, A. Anastasopoulos, Dataset Geography: Mapping Language Data to Language Users, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3381–3411. URL: https://aclanthology.org/2022.acl-long.239. doi:10.18653/v1/2022.acl-long.239.

[14] I. Ilyankou, M. Wang, J. Haworth, S. Cavazzi, Quantifying Geospatial in the Common Crawl Corpus, 2024. URL: http://arxiv.org/abs/2406.04952, arXiv:2406.04952 [cs].

[15] R. Manvi, S. Khanna, G. Mai, M. Burke, D. Lobell, S. Ermon, GeoLLM: Extracting Geospatial Knowledge from Large Language Models, 2024. URL: http://arxiv.org/abs/2310.06213. doi:10.48550/arXiv.2310.06213, arXiv:2310.06213 [cs].

[16] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023. URL: http://arxiv.org/abs/2307.09288. doi:10.48550/arXiv.2307.09288, arXiv:2307.09288 [cs].

[17] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, a. et, The Llama 3 Herd of Models, 2024. URL: http://arxiv.org/abs/2407.21783. doi:10.48550/arXiv.2407.21783, arXiv:2407.21783 [cs].

[18] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7B, 2023. URL: http://arxiv.org/abs/2310.06825. doi:10.48550/arXiv.2310.06825, arXiv:2310.06825 [cs].

[19] Mistral Small 3 | Mistral AI, 2025. URL: https://mistral.ai/news/mistral-small-3/.

[20] Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu, Qwen2.5 Technical Report, 2025. URL: http://arxiv.org/abs/2412.15115. doi:10.48550/arXiv.2412.15115, arXiv:2412.15115 [cs].

[21] T. Dettmers, M. Lewis, Y. Belkada, L. Zettlemoyer, Llm.int8(): 8-bit matrix multiplication for transformers at scale, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.