Benchmarking LLMs in Geoscience: A Serverless Approach using GeoBench and AWS

Otávio Parraga[1], Arthur Fachel[1], Rodolfo S. Antunes[2], Luiz Gonzaga Jr[2], Maurício Roberto Veronez[2], Rodrigo C. Barros[1], Lucas S. Kupssinskü[1]

[1]Pontifical Catholic University of Rio Grande do Sul (PUCRS)
Porto Alegre – RS – Brazil

[2]University of the Sinos River Valley (UNISINOS)
Porto Alegre – RS – Brazil

{otavio.parraga,arthur.fachel}@edu.pucrs.br,

{rodrigo.barros,lucas.kupssinsku}@pucrs.br

{rsantunes,lgonzaga,veronez}@unisinos.br

Abstract. *This paper presents an evaluation of open-weight Large Language Models (LLMs) for geoscientific tasks using the GeoBench benchmark. To overcome local hardware limitations when assessing massive models, we implemented a serverless cloud infrastructure on AWS, utilizing API Gateway, Lambda, and Amazon Bedrock. This architecture enabled high-throughput inference and automated data augmentation.*

Resumo. *Este artigo apresenta uma avaliação sistemática de Large Language Models (LLMs) de pesos abertos para tarefas geocientíficas, utilizando o benchmark GeoBench. Para superar restrições de hardware local ao avaliar modelos massivos, implementamos uma infraestrutura em nuvem serverless na AWS, utilizando API Gateway, Lambda e Amazon Bedrock. Essa arquitetura permitiu inferência em larga escala e o aumento automatizado de dados.*

1. Introduction

The incorporation of modern Artificial Intelligence (AI) into data-driven processes has completely reshaped geoscientific field research [Whitmeyer et al. 2010]. This evolution has triggered a massive influx of digital and multimodal data, making manual synthesis so complex that it is virtually impossible for an individual expert to manage without computational support [Marques Jr et al. 2020]. To alleviate these analytical bottlenecks, Machine Learning (ML) has become an essential asset [Dramsch 2020]. Most recently, the discipline has increasingly adopted Large Language Models (LLMs), leveraging their natural language capabilities to interrogate semantic databases and extract intricate insights from unstructured text [Deng et al. 2024, Zhao et al. 2023].

Although versatile, general-purpose LLMs often fail to capture the technical intricacies of geoscientific literature, leading to inaccuracies and "hallucinations" [Parraga et al. 2023]. To address these limitations, researchers typically adapt models to the domain via fine-tuning or integration with external knowledge bases [Deng et al. 2024, Lin et al. 2023]. Nevertheless, providing practitioners with reliable

recommendations for model selection necessitates evaluation against standardized benchmarks, such as GeoBench [Deng et al. 2024][1]. In this context, the choice of evaluation metrics is just as critical as the data itself for accurately assessing model quality.

To address this problem, this paper extends previous work by evaluating a diverse suite of open-source language models using the GeoBench benchmark [Deng et al. 2024], and by assessing compact models in both their base and fine-tuned configurations alongside massive, state-of-the-art architectures [Garcez et al. 2025]. Because local hardware constraints prevent deploying models with hundreds of billions of parameters, we designed a scalable cloud infrastructure on Amazon Web Services (AWS) to support this research. Specifically, we implemented a serverless pipeline utilizing AWS API Gateway and Lambda to interface directly with Amazon Bedrock, allowing us to evaluate these large-scale models without the overhead of maintaining dedicated high-tier GPUs. To ensure the assessment of open-ended geoscientific tasks, our evaluation framework employed a custom-trained RM to serve as a fine-grained, domain-specific metric. Ultimately, training this RM was only made possible by leveraging our AWS infrastructure to automatically generating the plausible-yet-incorrect responses needed to augment the GeoSignal dataset at scale [Deng et al. 2024].

2. Methodology

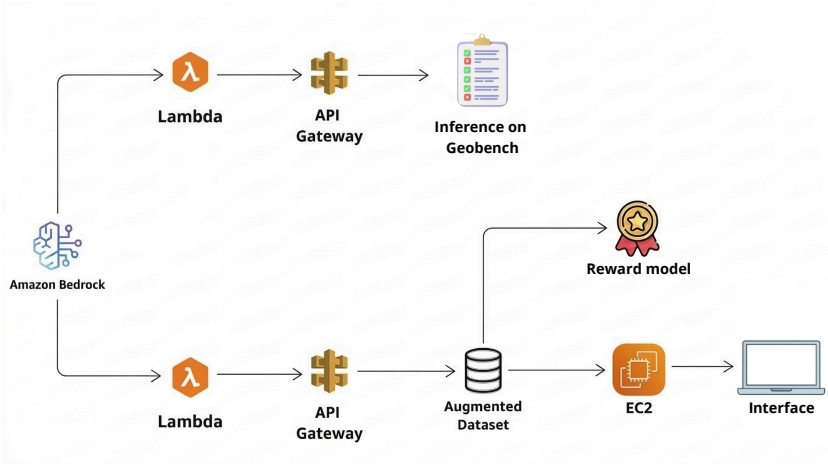

Figura 1. Developed pipeline for evaluating the models and augmenting reward dataset.

Our objective was to experiment with and compare a diverse set of language models with distinct parameter counts across the GeoBench benchmark [Deng et al. 2024], which combines both open- and closed-ended geoscientific tasks. To capture performance across these varying formats, we employed a multifaceted evaluation framework. Closed-ended questions were evaluated using standard exact-match accuracy scoring. For open-ended tasks, we utilized an LLM-as-a-judge methodology to assess prompt alignment, correctness, and answer relevance, supplemented by scores from a custom-trained, geoscience-specific RM [Garcez et al. 2025].

[1] `https://huggingface.co/datasets/daven3/geobench`

For smaller models with 15 billion parameters or fewer, we conducted on-premises evaluations and fine-tuning experiments. However, scaling this evaluation to massive, state-of-the-art architectures—such as Llama 3.1 (405B) [Dubey et al. 2024], DeepSeek R1 [DeepSeek-AI et al. 2025], and Mistral Large 3 [Mistral AI 2025]—becomes computationally prohibitive, as the compute (specially vram) scales with the model size.

To overcome the local hardware constraints restricting our experiments, we designed a scalable cloud infrastructure on AWS based on a serverless architecture to interact with these massive open-weight models (Figure 1). Within this environment, an AWS API Gateway manages incoming evaluation requests, routing them seamlessly to an AWS Lambda function. This function then executes the complex geoscientific prompts directly against foundation models hosted on Amazon Bedrock. This API Gateway-Lambda-Bedrock pipeline facilitates a high-throughput evaluation of LLMs, eliminating the overhead associated with maintaining dedicated high-tier GPUs.

2.1. Reward Modeling

Evaluating open-ended geoscientific tasks requires a nuanced assessment of conceptual understanding, which prompted the development of a custom-trained RM calibrated specifically for the geosciences [Zhong et al. 2025]. The RM acts as a regression-based discriminator trained with a preference-based objective, teaching the model to distinguish between a correct "chosen" response and a suboptimal "rejected" response [Ouyang et al. 2022].

Because the foundational GeoSignal instruction dataset [Deng et al. 2024][2] provides only gold-standard instructions and answers, we repurposed our identical AWS architectural pipeline to augment the dataset with the required negative samples. By systematically querying a high-capacity teacher model—specifically Llama 4 Maverick [Meta AI 2025]—through the Bedrock API, we automatically generated geoscientific responses that were explicitly prompted to be "plausible but incorrect". Generating these authoritative-sounding negative samples at scale using the serverless infrastructure was essential for the preference-based training phase. It provided the comprehensive data needed for the RM to effectively learn domain boundaries, penalize geoscientific hallucinations, and accurately score the open-ended evaluations.

3. Results

The implementation of our cloud infrastructure was fundamental in enabling the comprehensive evaluation of language models across different scales, overcoming local hardware restrictions. We used the cloud architecture described in Section 2 to run inference tests on large models for the GeoBench tasks. In this scenario, DeepSeek R1 stood out as the best-performing model, achieving 72.52% accuracy on multiple-choice questions and 75.37% on true/false questions.

In the open-ended geoscientific tasks, we observed that while larger models achieve markedly higher prompt alignment scores, their gains in factual correctness remain comparatively modest. This discrepancy reveals that scaling alone is insufficient to guarantee reliable geoscientific reasoning, highlighting persistent challenges in factual groun-

[2] `https://huggingface.co/datasets/daven3/geosignal`

Tabela 1. Closed-ended tasks results for base and instruct models.

Model	MULTIPLE CHOICE	TRUE/FALSE
Llama 3.3 (70B)	64.28	65.67
Llama 3.1 (405B)	67.58	68.65
Mistral Large 3	60.98	57.46
DeepSeek R1	**72.52**	**75.37**
Llama 4 Maverick	68.68	66.41
Llama 4 Scout	65.38	69.40
Mixtral 8x7B	48.90	71.64

Tabela 2. Results of the larger models for the open-ended tasks.

	QA				NOUN				COMPLETION			
Model	Alignment	Correctness	Relevancy	BERT	Alignment	Correctness	Relevancy	BERT	Alignment	Correctness	Relevancy	BERT
Llama 3.3 (70B)	0.93	0.26	0.93	**0.12**	0.97	**0.28**	0.95	**0.16**	**0.60**	0.14	0.81	0.10
Llama 3.1 (405B)	0.70	0.26	0.95	0.11	0.74	0.27	**0.96**	0.13	0.33	0.18	0.79	0.20
Mistral Large 3	0.84	0.24	0.90	0.02	–	–	–	–	0.30	0.12	0.76	0.12
DeepSeek R1	**0.99**	0.25	0.95	0.06	**0.99**	0.25	**0.96**	0.08	**0.60**	**0.20**	**0.87**	**0.27**
Llama 4 Maverick	0.86	**0.29**	0.94	**0.12**	0.89	**0.28**	**0.96**	0.14	0.59	0.18	0.77	0.17
Llama 4 Scout	0.76	0.27	0.94	0.11	–	–	–	–	0.50	0.15	0.80	0.18
Mixtral 8x7B	0.52	0.27	**0.96**	**0.12**	–	–	–	–	0.10	0.15	0.85	-0.01

ding for domain-specific applications. The results indicate that current alignment optimization may preferentially improve a model's structural compliance and form over its substantive scientific accuracy

Conducting this comprehensive evaluation across massive, state-of-the-art architectures was made possible entirely through the implementation of a highly scalable cloud infrastructure. By utilizing a serverless pipeline to interface directly with foundation models via API, we successfully bypassed the local hardware limitations that traditionally restrict the deployment of models with hundreds of billions of parameters. This architectural approach enabled high-throughput inference for extensive benchmarking, providing the computational flexibility needed to rigorously test frontier models under standardized conditions.

4. Conclusion

This study contributes to the ongoing transformation of the geosciences by establishing a comprehensive evaluation of open-weight LLMs using the GeoBench benchmark [Deng et al. 2024]. Our findings demonstrate that performance on geoscientific tasks is driven by a complex interplay of model scale, architectural design, and training strategy, rather than parameter count alone. Crucially, the deployment of a serverless AWS cloud infrastructure—utilizing Bedrock, API Gateway, and Lambda—overcame local hardware limitations, enabling both the scalable evaluation of massive models and the automated dataset augmentation required for our custom RM. Furthermore, by bridging automated processes with expert human validation via an EC2-hosted web interface, our evaluation framework remained strictly aligned with professional geoscientific standards. Ultimately, by combining standardized evaluation metrics with a scalable cloud architecture, this work provides a transparent and reproducible foundation for researchers and practitioners to assess, adapt, and deploy open LLMs within complex geoscientific workflows.

Declaration of AI Technologies in the Writing Process

During the preparation of this work, the authors used Gemini (Google) [3] in order to review and rewrite sections of the manuscript to improve readability and linguistic clarity. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Referências

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J.-M., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Deng, C., Zhang, T., He, Z., Chen, Q., Shi, Y., Xu, Y., Fu, L., Zhang, W., Wang, X., Zhou, C., et al. (2024). K2: A foundation language model for geoscience knowledge understanding and utilization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 161–170.

Dramsch, J. S. (2020). 70 years of machine learning in geoscience in review. *Advances in geophysics*, 61:1–55.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Garcez, V. H., Parraga, O., Marques, A., Spigolon, A. L. D., De Barros, G., Gonzaga, L., Veronez, M. R., Barros, R. C., and Kupssinskü, L. S. (2025). Which is the best llm for geosciences? In *IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium*, pages 6374–6378. IEEE.

Lin, Z., Deng, C., Zhou, L., Zhang, T., Xu, Y., Xu, Y., He, Z., Shi, Y., Dai, B., Song, Y., et al. (2023). Geogalactica: A scientific large language model in geoscience. *arXiv preprint arXiv:2401.00434*.

Marques Jr, A., Horota, R. K., De Souza, E. M., Kupssinskü, L., Rossa, P., Aires, A. S., Bachi, L., Veronez, M. R., Gonzaga Jr, L., and Cazarin, C. L. (2020). Virtual and digital outcrops in the petroleum industry: A systematic review. *Earth-Science Reviews*, 208:103260.

Meta AI (2025). Llama 4: Multimodal intelligence. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`. Accessed: 2026-01-08.

Mistral AI (2025). Introducing mistral 3. `https://mistral.ai/news/mistral-3`. Accessed: 2026-01-29.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Parraga, O., More, M. D., Oliveira, C. M., Gavenski, N. S., Kupssinskü, L. S., Medronha, A., Moura, L. V., Simões, G. S., and Barros, R. C. (2023). Fairness in deep learning: A survey on vision and language research. *ACM Computing Surveys*.

[3] gemini.google.com

Whitmeyer, S. J., Nicoletti, J., and De Paor, D. G. (2010). The digital revolution in geologic mapping. *Gsa Today*, 20(4/5):4–10.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhong, J., Shen, W., Li, Y., Gao, S., Lu, H., Chen, Y., Zhang, Y., Zhou, W., Gu, J., and Zou, L. (2025). A comprehensive survey of reward models: Taxonomy, applications, challenges, and future. *arXiv preprint arXiv:2504.12328*.