Geological Text Summarization Using Generative Large Language Models

Matheus Stein de Aguiar and Rafael Oleques Nunes and Dennis Giovani Balreira
Institute of Informatics, UFRGS
msaguiar@inf.ufrgs.br, ronunes@inf.ufrgs.br, dgbalreira@inf.ufrgs.br

Abstract

Large generative language models have demonstrated impressive performance in various Natural Language Processing (NLP) tasks. However, the geological domain presents unique challenges for NLP due to its specialized language, which is full of technical terms. Therefore, pre-trained language models on generic corpora may not be suitable for performing geological domain-specific tasks. This article compares several models to identify those with the best performance in the Portuguese geological domain for a text summarization task. We applied the models to a *Revista Geologia USP* dataset. The dataset consists of abstracts of scientific texts and their respective titles, which we aim for the models to approximate with the summarization task. We tested the models in various scenarios, providing examples or not, and at two temperature levels. We then evaluated the models' performance using quantitative metrics and a brief qualitative analysis comparing the titles proposed by the models with the original title. The results show that the Gemma3:27b model was better in some scenarios, while the Llama3:8b model performed best in others.

1 Introduction

It is widely recognized that we are living in an unprecedented era of data generation, and the geological domain is no exception (Landim, 2003). This scenario creates the need for efficient methods to organize, interpret, and summarize complex technical information. Geological reports, scientific articles, and technical descriptions in the field often contain dense and detailed language, making it difficult for specialists and decision-makers to quickly access essential information (Rowe and Frewer, 2000). In this context, the task of automatic text summarization emerges as a promising solution to facilitate the analysis and interpretation of such documents.

In recent years, Large Language Models (LLMs) have demonstrated exceptional effectiveness in Natural Language Processing (NLP) tasks, including translation, classification, and text summarization. Unlike traditional approaches based on sentence extraction, LLMs are also capable of performing abstractive summarization, producing more cohesive, concise, and semantically enriched summaries (Liu and Lapata, 2019).

Despite significant advances in the field, the publication of articles in the geological domain in Portuguese using NLP techniques remains relatively scarce, whether due to the lack of open data or the challenges presented by the domain's specialized vocabulary. Therefore, this paper aims to explore the applicability and limitations of LLMs in the task of summarizing geological texts. We focus on summarizing the abstracts of articles into titles, evaluating the quality of the generated summaries, and discussing the opportunities and challenges associated with the use of this technology in such a specialized domain. We used the following models: deepseek-r1 (version 8B) (Guo et al., 2025), gemma3 (versions 4B, 12B, and 27B) (Team et al., 2025), qwen3 (version 8B) (Yang et al., 2025), llama3 (version 8B) (Grattafiori et al., 2024), GAIA-PT-BR (version 4B) (Camilo-Junior et al., 2025), more information on how we chose these models can be found in the Subsection 4.3.

Below, we summarize the contributions of this work:

- Availability of a new corpus composed of publications from the *Revista Geologia USP* of the University of São Paulo (USP) (Instituto de Geociências, 2001), a Brazilian geoscientific journal. We collected all the issues published in the journal into a single dataset;
- Comparison of LLMs for the task of geological text summarization;

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 111–119
April 13-16, 2026

- Comparative analysis of the human-written title and the LLM-generated summary in representing the content of the text.

Further sections of this paper are organized as follows. Section 2 introduces the key concepts essential for understanding this study, particularly the metrics employed, and Section 3 presents and discusses the main approaches identified in the literature, outlining where this work aims to position itself. Section 4 provides information about the corpus, models, and pipeline. Our experimental evaluation and results in Section 5. Section 6 presents conclusions and perspectives for future research, followed by Section 6 with the limitations, and acknowledgments in Section 6.

2 Background

This section defines key concepts essential for a deeper understanding of this work, ranging from the task being performed to the metrics used for evaluating results. The definitions presented here are based on the concepts introduced in Nenkova et al. (Nenkova et al., 2011).

2.1 Summarization

Summarization is an NLP task that aims to synthesize the knowledge contained in one or more documents while preserving the most important information (Allahyari et al., 2017). The challenge in this task lies in crafting a concise and fluent summary. Another important challenge is evaluating the resulting summary, as commonly used metrics such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE) focus on n-gram overlap. For instance, ROUGE may assign a high score to the sentences "João is 80 years old and weighs 30 kilograms" and "João is 30 years old and weighs 80 kilograms," even though, upon qualitative inspection, they convey entirely different meanings.

There are two main types of summarization:

- **Extractive:** Summaries are created by selecting and concatenating exact fragments from the original text.
- **Abstractive:** Summaries are written to convey the main ideas using rephrased or newly generated sentences. In general, an abstractive summary should reflect the summary author's interpretation of the original content.

In this work, we evaluate abstractive summarization through generative LLMs.

2.2 Metrics

There is no consensus in the literature on the most appropriate metric for studies of this type. This section presents the metrics we have chosen to use, along with their respective limitations.

2.2.1 ROUGE

The most widely used metric for text summarization is ROUGE (Barbella and Tortora, 2022). It compares the summary generated by the model to a reference summary, which is the one we would expect the model to produce, with a score closer to 1.0 indicating better performance. In this study, we use three versions of the ROUGE metric:

- **ROUGE-1:** Measures the number of overlapping unigrams between the generated and reference summaries.
- **ROUGE-2:** Measures the overlap of bigrams.
- **ROUGE-L:** Measures the length of the longest common subsequence between the generated and reference summaries.

As previously mentioned, this metric, especially ROUGE-1, has the limitation of not assessing whether two summaries are semantically equivalent, as it relies solely on surface-level n-gram overlap.

2.2.2 BLEU

The next metric used is BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002). Initially designed for evaluating machine translation, BLEU can also be applied to summarization tasks, although it is not ideal, and low scores are expected. Summarization is generally more open-ended, while BLEU emphasizes exact n-gram matches. In general, BLEU computes the precision of n-gram matches between the generated and reference summaries, applies a brevity penalty when the output is too short, and then takes the geometric mean across 1-gram to 4-gram precision, with a score closer to 1.0 being better. As mentioned, BLEU tends to underperform in summarization tasks due to its reliance on literal overlap, without accounting for synonyms or changes in word order.

2.2.3 METEOR

The Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Banerjee and Lavie, 2005) aims to address some of BLEU's limitations. It considers synonyms, word stems (stemming),

and reordering (with a penalty). Unlike BLEU, which focuses solely on precision, METEOR combines both precision and recall using a harmonic mean, while also penalizing excessive word reordering; a score closer to 1.0 is considered better.

2.2.4 BERTScore

Unlike the metrics above, BERTScore (Zhang et al., 2019) is based on language embeddings. It leverages the Bidirectional Encoder Representations from Transformers (BERT) model to compute semantic similarity. Instead of comparing words directly, it compares their meanings. Each word in both the generated and reference summaries is transformed into a contextualized vector using BERT. Cosine similarity is then calculated between the vectors, performing a soft alignment that selects the best-matching words in the reference. BERTScore outputs precision, recall, and F1-score based on these semantic alignments.

3 Related work

The task of summarization has evolved over the years and has been the subject of study for a long time, employing a variety of techniques such as concept-based methods (Miller, 1995), extractive statistical techniques (Gupta and Lehal, 2010), more recent approaches using embeddings (Kobayashi et al., 2015), and Recurrent Neural Networks (RNNs) (Chen and Le Nguyen, 2019). However, as the scope narrows, the number of studies becomes more limited.

In the context of summarization for texts in Portuguese, existing works have mainly addressed general domains (e.g., articles, news, etc.) (Paiola et al., 2022; Antiqueira et al., 2009). As we move toward more technical areas, we observe a greater number of applications in the legal domain (Dias, 2024) and in the medical field (Fraile Navarro et al., 2025).

On the other hand, for geological applications, most summarization studies are focused on texts in Chinese (Wang and Wang, 2017; Ma et al., 2022). In other languages, such as Hindi (Sobhana et al., 2010) and English (Enkhsaikhan et al., 2021), research tends to be more concentrated on information retrieval rather than summarization. Studies involving Portuguese texts in the geological domain are extremely scarce (Nunes et al., 2025), with most related works focusing on Named Entity Recognition (NER) (Nunes et al., 2025). The lack of summarization studies in this domain is likely due to the absence of publicly available corpora tailored to this task. In Portuguese, there is the REGIS corpus (Reference Corpus for Geoscientific Information Systems), which focuses on information retrieval (Lima de Oliveira et al., 2021), and for NER tasks, there are corpora such as PetroNER (Freitas et al., 2023), PetroKGraph (Freitas et al., 2023), and the GeoCorpus (Amaral et al., 2017; Consoli et al., 2020; Gomes et al., 2021; Nunes et al., 2024).

This work addresses a gap not previously covered by studies by introducing a new corpus specifically designed for abstractive summarization in the geological domain in Portuguese. In addition to making this dataset available, we conducted a comparative evaluation of several open-source LLMs under various prompting conditions and analyzed the models' ability to generate meaningful titles based on article abstracts. These contributions enable further research in this underexplored domain and offer a baseline for future evaluations of generative summarization models in technical contexts.

4 Methodology

This section details the procedures adopted to collect, process, and evaluate data in our study. We describe the construction of the dataset, the selected language models, and the evaluation strategy employed to compare generated outputs.

4.1 Corpus

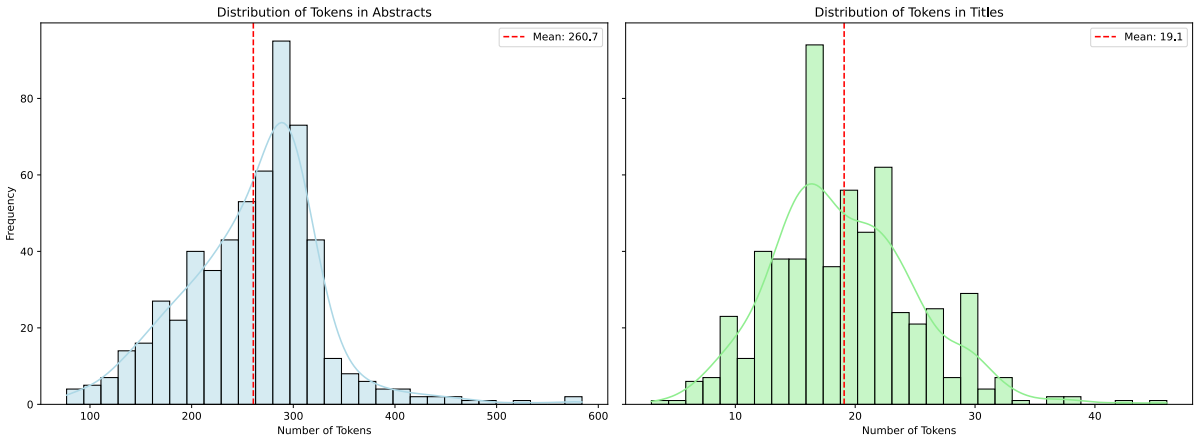
The corpus used in this study was extracted from the journal *Geologia USP* (Instituto de Geociências, 2001) from the Institute of Geosciences at the University of São Paulo. It contains 583 instances, each corresponding to an article's title, abstract, and a link to the full article. We selected all articles that have been published in the journal since 2001. Figure 1 shows the distribution of tokens per abstract and per title. Abstracts range from 77 to 584 tokens, with an average of 260 tokens, while titles range from 3 to 46 words, with an average of 19. The full dataset is publicly available on GitHub[1]. No data cleaning was necessary during the preprocessing phase.

4.2 Overview

The first step of this work was web scraping to build the dataset used. With the collected data, we tested several variations of four of the most frequently accessed models on the Ollama platform at the time

[1] https://github.com/MatheusSteinDeAguiar/GeoSummarization

Figure 1: Distribution of the number of tokens by frequency, on the left analysis of summaries and on the right analysis of titles



of writing, as well as an additional model specifically trained for Portuguese. Further details on each model are provided in Section 4.3. For each selected model, we experimented with variations in the number of parameters, number of examples (zero-shot and one-shot), and temperature settings.

After running the models, the generated results were evaluated using the metrics explained previously in Subsection 2.2. We compared model performance across various metrics by evaluating the generated titles against the original author-provided titles. Additionally, BERTScore was employed to assess the semantic similarity between the generated titles and the abstract text.

4.3 Models

All selected models are freely available for use[2]. Model selection was based first on free availability, followed by popularity (measured by number of downloads/accesses). The models tested were: deepseek-r1 (version 8B) (Guo et al., 2025), gemma3 (versions 4B, 12B, and 27B) (Team et al., 2025), qwen3 (version 8B) (Yang et al., 2025), llama3 (version 8B) (Grattafiori et al., 2024), GAIA-PT-BR (version 4B) (Camilo-Junior et al., 2025) — although not among the most downloaded models, GAIA is a Portuguese-focused model derived from continued training of gemma3, and we found it relevant to include in the comparison.

5 Experiments and Results

For the experiments, we used the infrastructure of the High-Performance Computing Park (PCAD[3]) at INF/UFRGS. The machine configuration included an NVIDIA GeForce RTX 4090 GPU, 224 GB of RAM, and an Intel(R) Xeon(R) E5-2620 v4 processor with 2.10 GHz, eight cores, and 16 threads.

The experiments conducted for each model described in Section 4.3 were designed to test different settings, starting with the number of examples shots passed in the prompt. We tested both zero-shot and one-shot settings. Additionally, two different temperature values were evaluated: 0.0, which ensures reproducibility, and 0.7, which introduces more creativity into the model outputs at the expense of determinism. The prompt used is shown in Citation 1. It includes the instruction text and the abstract to be summarized, represented by the placeholder input. For one-shot variations, an example is prepended to the prompt. The prompt was developed empirically and tailored specifically for this project.

Citation 1:

> You need to clearly and precisely summarize the idea of the text in a single sentence. I will provide you with the abstract of a scientific article, and I need you to return a title for that article. Respond only with the title.
>
> Abstract: input 1

For this study, we conducted three experiments to test different scenarios.

- Test 1: zero-shot and temperature set to 0, ensuring the reproducibility of the experiment;

[2] https://ollama.com/

[3] http://gppd-hpc.inf.ufrgs.br

- Test 2: one-shot and temperature set to 0, ensuring the reproducibility of the experiment and providing an example for the model to be based on;

- Test 3: zero-shot and temperature set to 0.7, losing reproducibility but gaining creativity in the creation of titles;

We chose these temperatures because the first, 0, ensures the reproducibility of the test, and 0.7 is approximately the limit at which the performance of the models for summarization in Li et al. (2025) began to decline.

We present the results of the Test 1 — zero-shot with temperature set to 0 — in the Table 1. This test utilizes BERTscore to compare the titles generated by the models to the original titles of the articles. Noteworthy observations include the significantly longer runtime of the DeepSeek model compared to the others, the expected performance gains of Gemma3 as model size increases, and the underperformance of the Gaia model despite being a Portuguese-specific continuation of Gemma3:4b.

Figure 2 presents the Nemenyi post-hoc statistical test results using the BERTScore F1 as the evaluation metric. The analysis reveals two models with statistically superior performance (Gemma3:27b and Gemma3:12b), while others, such as Llama, DeepSeek, and Gaia, form a lower-performing group. Some models, such as Qwen and Gaia, exhibit statistical similarity to each other. However, overall, several model pairs show no statistically significant difference at the 0.05 level.

Figure 2: Statistical significance test among models (BERTScore F1)

We also evaluated the quality of generated titles based on their alignment with the original abstract — not just the reference title — using BERTScore. Table 2 shows the results for both approaches, and Figure 3 shows a heatmap of statistical significance for the zero-shot approach.

Interestingly, four of the seven tested models produced titles with semantic quality comparable to the original title. Only Llama, Qwen, and Gaia had statistically inferior performance. This finding suggests that even when metric scores are not high compared to the reference title, some models are just as effective as the human author in summarizing the abstract into a title.

Figure 3: Statistical significance between models and original title (abstract alignment)

The Table 1 shows results for the one-shot scenario (Test 2). Due to computational limitations, only four models were tested. For the one-shot scenarios, we use the first row of the data set as an example, so the tests in this scenario have one less sample. Most models showed improvements in all metrics (1–2 percentage points), especially Llama:8b. Notably, Gemma3:27b underperformed compared to its zero-shot performance. Surprisingly, Gemma3:12b outperformed the 27b model in BERTScore, which is also reflected in abstract alignment results in Table 2.

In Table 3, we evaluated models under zero-shot and temperature set to 0.7 (Test 3) to observe how increased creativity affects output. As expected, reproducibility was lost, but performance remained similar to that at the 0 °C setting. Gemma3:27b remained the best-performing model in this scenario.

Table 1: Results for zero and one-shot (temperature = 0.0)

Approach	Model	Rouge1	Rouge2	RougeL	Meteor	BLEU	BERTscore
zero-shot	Gemma3:27b	**0.4308**	**0.2650**	**0.3729**	**0.3943**	**0.1102**	**0.7637**
zero-shot	Gemma3:12b	0.4000	0.2237	0.3370	0.3393	0.0829	0.7483
zero-shot	deepseek-r1:8b	0.3695	0.2011	0.3067	0.2930	0.0844	0.7440
zero-shot	Llama:8b	0.3881	0.2247	0.3269	0.3091	0.0852	0.7417
zero-shot	Qwen3:8b	0.3793	0.2051	0.3117	0.2831	0.0840	0.7416
zero-shot	Gemma3:4b	0.3865	0.2077	0.3217	0.3196	0.0746	0.7404
zero-shot	Gaia-PT-BR:4b	0.3676	0.1930	0.3030	0.3188	0.0735	0.7407
one-shot	Llama:8b	**0.4480**	**0.2834**	**0.3931**	**0.4088**	**0.1303**	**0.7733**
one-shot	Gemma3:12b	0.4158	0.2407	0.3534	0.3650	0.1191	0.7727
one-shot	Gemma3:27b	0.4250	0.2454	0.3592	0.3711	0.1014	0.7611
one-shot	Gemma3:4b	0.3926	0.2203	0.3339	0.3561	0.0899	0.7567

Table 2: BERTScore between abstract and generated title (zero-shot and one-shot)

Approach	Model	Precision	Recall	F1-score
	Original Title	**0.5935**	**0.7560**	**0.6646**
zero-shot	deepseek-r1:8b	0.5921	0.7406	0.6576
zero-shot	Qwen3:8b	0.5906	0.7276	0.6517
zero-shot	Gemma3:27b	0.5744	0.7375	0.6456
zero-shot	Gemma3:12b	0.5718	0.7210	0.6375
zero-shot	Gaia:4b	0.5728	0.7195	0.6375
zero-shot	Gemma3:4b	0.5707	0.7217	0.6371
zero-shot	Llama:8b	0.5714	0.7375	0.6354
one-shot	Gemma3:12b	0.5851	0.7577	0.6599
one-shot	Llama:8b	0.5796	0.7592	0.6569
one-shot	Gemma3:4b	0.5752	0.7411	0.6473
one-shot	Gemma3:27b	0.5774	0.7357	0.6467

5.1 Qualitative Analysis

To better understand the qualitative aspects of the generated titles, we selected examples from each model under the zero-shot setting, with the temperature set to 0, as shown in Table 4. This analysis aims to complement the automatic metrics by illustrating how different models interpret the summarization task and whether the generated titles align with standard scientific writing conventions.

Many of the titles generated appear plausible and relevant, often capturing the central theme of the abstract. However, stylistic and structural differences between models were noticeable. For instance, Gemma:12b, Gemma:27b, and Gaia tended to produce two-part titles separated by colons, resembling the structure often found in academic articles. In contrast, Qwen and Deepseek sometimes included unnecessary formatting such as asterisks or the prefix "Title:".

In particular, Gemma:27b often generated longer and more descriptive titles, whereas Llama:8b favored brevity and simplicity, sometimes omitting key concepts from the abstract. These qualitative differences underscore the importance of combining automatic metrics with manual inspection, particularly when evaluating outputs in specialized domains such as geology.

Table 5 presents the titles generated by Gemma:27b across all three scenarios, along with the original title. Despite variations, the generated titles consistently capture the essential elements of the abstract and remain relevant.

6 Conclusion

In this study, we explored and analyzed the summarization of texts taken from the *Revista Geologia USP* from the University of São Paulo. The task consisted of generating a suitable title for an article based on its abstract. We evaluated seven models via the Ollama framework, testing them under three different conditions: zero-shot (no examples in the prompt), one-shot (one example in the prompt), and zero-shot with temperature 0.7, aimed at encouraging greater variability in the generated titles.

The results were satisfactory, with the generated titles generally containing the key information from the original titles. The Gemma3:27b model performed best in the zero-shot scenarios with both temperature settings (0 and 0.7), while Llama:8b achieved the best results in the one-shot setting. We also evaluated the semantic similarity between the abstract and the generated title using BERTScore, and observed that most models performed as well as, statistically, the original author-created titles.

Future work could address these limitations, explore a broader range of models—including commercial LLMs such as GPT, Gemini, or Grok, incorporate human evaluation of the generated titles or use LLM as a judge.

Limitations

This study has some limitations. First, the available dataset was relatively small. Second, computational constraints limited the scope of experiments, since all models were run locally. Another concern is potential data leakage: although all the data used is publicly available, it is not centralized in any

Table 3: Metrics for zero-shot models (temperature = 0.7)

Model	Rouge1	Rouge2	RougeL	Meteor	BLEU	BERTscore
Gemma3:27b	**0.4269**	**0.2433**	**0.3575**	**0.3581**	**0.0974**	**0.7516**
Gemma3:12b	0.4007	0.2250	0.3371	0.3440	0.0802	0.7454
deepseek-r1:8b	0.3664	0.1910	0.2986	0.2825	0.0803	0.7423
Gaia-PT-BR:4b	0.3675	0.1931	0.3042	0.3179	0.0734	0.7412
Qwen3:8b	0.3705	0.1990	0.3017	0.2748	0.0819	0.7404
Gemma3:4b	0.3883	0.2049	0.3250	0.3216	0.0729	0.7402
Llama:8b	0.3687	0.2050	0.3006	0.2830	0.0777	0.7385

Table 4: Titles found by each model (zero-shot, temperature = 0)

Model	Example 1	Example 2	Example 3
Original Title	Respostas do microplâncton marinho durante os Eventos Anóxicos Oceânicos (OAEs) do Cretáceo	Ocorrências de fosfato sedimentar em coberturas fanerozoicas na região de Arraias - TO, borda oeste da Bacia do São Francisco, Brasil	Gênese dos beachrocks da Ilha oceânica da Trindade, Brasil
Gemma3:27b	Respostas do Microplâncton Marinho aos Eventos Anóxicos Oceânicos (OAEs) no Cretáceo: Uma Revisão Sistemática	Ocurrencência de Estratos Isolados de Rochas Sedimentares Fanerozoicas na Bacia do São Francisco	Diagenese e Formação de Beachrocks na Ilha da Trindade: Um Estudo de Caso
Gemma3:12b	Microfósseis e Eventos Anóxicos Oceânicos no Cretáceo: Uma Revisão das Respostas do Microplâncton	Campos Belos e Arraias: Ocorrência de Conglomerados Fanerozoicos com Potencial Fosfatífero.	Beachrocks da Ilha da Trindade: Registro Diagenético e Microbiológico em Ambiente de Intermare
Llama:8b	"Microfósseis Marinhos como Indicadores de Eventos Anóxicos Oceânicos no Cretáceo"	"Estratos Fanerozoicos Fosfatados na Borda Oeste da Bacia do São Francisco: Novas Contribuições para o Conhecimento da Formação Sete Lagoas"	"Formação e Diagenese de Beachrocks na Ilha da Trindade: Um Estudo sobre Processos Sedimentares e Paleoníveis Marinhos"
deepseek-r1:8b	**Resposta dos Microfósseis ao Eventos Anóxicos Oceânicos do Cretáceo**	Novas informações sobre a ocorrência de litotipos fosfatados fanerozoicos na borda oeste da Bacia do São Francisco.	**Título:** Beachrocks da Praia Leste da Ilha da Trindade: Formação e Composição de Cimentos Carbonáticos
Gemma3:4b	Microfósseis como Indicadores dos Eventos Anóxicos Oceânicos do Cretáceo.	Fosforitos e conglomerados polimíticos na Bacia do São Francisco: Implicações para a exploração de recursos fosfáticos.	Beachrocks da Ilha da Trindade: Processos Diagenéticos e Registro Paleomarino.
Qwen3:8b	Respostas dos Microfósseis Marinhos aos Eventos Anóxicos Oceânicos no Cretáceo e sua Relação com Ciclos Biogeoquímicos	Fosforitos e conglomerados polimíticos na Formação Sete Lagoas: implicações para exploração de fosfato na Bacia do São Francisco.	Diagenese e Formação de Beachrocks na Ilha da Trindade: Processos Químicos, Microbiológicos e a Presença de Clorita Singenética
Gaia:4b	Microfósseis marinhos e a resposta aos eventos anóxicos oceânicos do Cretáceo: uma análise biogeoquímica e evolutiva.	Fosforitos e Fosfatos Fanerozoicos na Bacia do São Francisco: Implicações Sedimentológicas e Exploratórias	Beachrocks da Ilha da Trindade: Processos Físico-Químicos e Microbiológicos na Formação de Sedimentos Costeiros

single repository. Additionally, using a higher temperature introduces variability in the results—each run can yield a different output—so a more rigorous evaluation might involve multiple runs and averaging the results.

Acknowledgments

This work was partially funded by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Finance Code 001. We also acknowledge the financial support of the Brazilian funding agency CNPq and Petrobras. Some experiments in this work used the PCAD infrastructure (http://gppdhpc.inf.ufrgs.br) at INF/UFRGS. Parts of this manuscript were written with the support of a generative AI tool (ChatGPT); all content was reviewed and validated by the authors.

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Daniela Amaral, Sandra Collovini, Anny Figueira, Renata Vieira, and Marco Gonzalez. 2017. Processo de construção de um corpus anotado com entidades geológicas visando ren (building an annotated corpus with geological entities for ner)[in portuguese]. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 63–72.

Lucas Antiqueira, Osvaldo N Oliveira Jr, Luciano da Fontoura Costa, and Maria das Graças Volpe Nunes. 2009. A complex network approach to text summarization. *Information Sciences*, 179(5):584–599.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation*

Table 5: Titles found by Gemma:27b in each scenario

	Original	zero-shot, temperature=0	one-shot, temperature=0	zero-shot, temperature=0.7
Ex. 1	Gênese dos beachrocks da Ilha oceânica da Trindade, Brasil	Diagenese e Formação de Beachrocks na Ilha da Trindade: Um Estudo de Caso	Beachrocks da Ilha da Trindade: Registro de Processos Físico-Químicos e Microbiológicos.	Beachrocks da Ilha da Trindade: Gênese, Cimentação e Influência Microbiológica.
Ex. 2	Avaliação geotécnica e análise cinemática da Cachoeira do Mascate como subsídio ao ecoturismo em Coromandel, Minas Gerais	Avaliação Geológica e Análise Cinemática da Cachoeira do Mascate: Implicações para o Ecoturismo Regional	Análise Geológico-Geotécnica e Cinemática da Cachoeira do Mascate: Implicações para o Risco Geológico e Ecoturismo.	Estabilidade Geotécnica e Riscos Geológicos na Cachoeira do Mascate, Triângulo Mineiro.
Ex. 3	Caracterização das variações volumétricas, morfológicas e a influência da antropização em uma praia de enseada, Praia de Guaratuba - PR	Variações do Volume de Sedimentos na Praia de Guaratuba: Dinâmica Natural e Impacto das Obras Urbanas.	Dinâmica Sedimentar e Impactos Costeiros na Praia de Guaratuba, Paraná	Dinâmica Sedimentar e Erosão Costeira na Praia de Guaratuba, Paraná.

measures for machine translation and/or summarization, pages 65–72.

Marcello Barbella and Genoveffa Tortora. 2022. Rouge metric evaluation for text summarization techniques. *Available at SSRN 4120317*.

C. G. Camilo-Junior, S. S. T. Oliveira, L. A. Pereira, M. Amadeus, D. Fazzioni, A. M. A. Novais, and S. A. A. Jordão. 2025. GAIA: An open language model for brazilian portuguese. https://huggingface.co/CEIA-UFG/Gemma-3-Gaia-PT-BR-4b-it. Accessed on July 2025.

Laifu Chen and Minh Le Nguyen. 2019. Sentence selective neural extractive summarization with reinforcement learning. In *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–5. IEEE.

Bernardo Consoli, Joaquim Santos, Diogo Gomes, Fabio Cordeiro, Renata Vieira, and Viviane Moreira. 2020. Embeddings for named entity recognition in geoscience portuguese literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4625–4630.

Margarida Rebelo Dias. 2024. Contributions to automatic legal document summarization: Judgements from the portuguese supreme court. Master's thesis, ISCTE-Instituto Universitario de Lisboa (Portugal).

Majigsuren Enkhsaikhan, Wei Liu, Eun-Jung Holden, and Paul Duuring. 2021. Auto-labelling entities in low-resource text: a geological case study. *Knowledge and Information Systems*, 63(3):695–715.

David Fraile Navarro, Enrico Coiera, Thomas W Hambly, Zoe Triplett, Nahyan Asif, Anindya Susanto, Anamika Chowdhury, Amaya Azcoaga Lorenzo, Mark Dras, and Shlomo Berkovsky. 2025. Expert evaluation of large language models for clinical dialogue summarization. *Scientific reports*, 15(1):1195.

Cláudia Freitas, Elvis Souza, Maria Clara Castro, Tatiana Cavalcanti, Patricia Ferreira da Silva, and Fábio Corrêa Cordeiro. 2023. Recursos linguísticos para o pln específico de domínio: o petrolês. *Linguamática*, 15(2):51–68.

Diogo da Silva Magalhães Gomes, Fábio Corrêa Cordeiro, Bernardo Scapini Consoli, Nikolas Laceda Santos, Viviane Pereira Moreira, Renata Vieira, Silvia Moraes, and Alexandre Gonçalves Evsukoff. 2021. Portuguese word embeddings for the oil and gas industry: Development and evaluation. *Computers in Industry*, 124:103347.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Vishal Gupta and Gurpreet Singh Lehal. 2010. A survey of text summarization extractive techniques. *Journal of emerging technologies in web intelligence*, 2(3):258–268.

Universidade de São Paulo Instituto de Geociências. 2001. Geologia usp. série científica. Disponível em: https://www.revistas.usp.br/guspsc. Acesso em: 28 jun. 2025.

Hayato Kobayashi, Masaki Noguchi, and Taichi Yatsuka. 2015. Summarization based on embedding distributions. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1984–1989.

Paulo Milton Barbosa Landim. 2003. *Análise estatística de dados geológicos*. Unesp - Universidade Estadual Paulista.

Lujun Li, Lama Sleem, Niccolo' Gentile, Geoffrey Nichil, and Radu State. 2025. Exploring the impact of temperature on large language models: Hot or cold? *arXiv preprint arXiv:2506.07295*.

Lucas Lima de Oliveira, Regis Kruel Romeu, and Viviane Pereira Moreira. 2021. Regis: A test collection for geoscientific documents in portuguese. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2363–2368.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Kai Ma, Miao Tian, Yongjian Tan, Xuejing Xie, and Qinjun Qiu. 2022. What is this article about? generative summarization with the bert model in the geosciences domain. *Earth Science Informatics*, pages 1–16.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Ani Nenkova, Kathleen McKeown, and 1 others. 2011. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233.

Rafael O Nunes, Andre S Spritzer, Dennis G Balreira, Carla MDS Freitas, and Joel L Carbonera. 2024. An evaluation of large language models for geological named entity recognition. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 494–501. IEEE.

Rafael Oleques Nunes, Andre Suslik Spritzer, Carla Maria Dal Sasso Freitas, Dennis Giovani Balreira, and Joel Luís Carbonera. 2025. Natural language processing in geology: A review of portuguese language resources and techniques. *Procesamiento del Lenguaje Natural*, 74:55–65.

Pedro H. Paiola, Gustavo H. de Rosa, and João P. Papa. 2022. Deep learning-based abstractive summarization for brazilian portuguese texts. In *Brazilian Conference on Intelligent Systems (BRACIS)*, pages 479–493. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Gene Rowe and Lynn J Frewer. 2000. Public participation methods: a framework for evaluation. *Science, technology, & human values*, 25(1):3–29.

NV Sobhana, Alimpan Barua, Monotosh Das, Pabitra Mitra, and SK Ghosh. 2010. Co-occurrence based place name disambiguation and its application to retrieval of geological text. In *International Conference on Web and Semantic Technology*, pages 543–552. Springer.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Boyang Qin Yong Shen Jian Wang and Gang Wang. 2017. Summarization of geological study on low rank coalbed methane in china. *Coal Science and Technology*, 1(1).

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.