From Bones to Rocks: A Systematic Evaluation of Specialized Definition Generation for Portuguese

Rafael Oleques Nunes, Dennis Giovani Balreira, Joel Luís Carbonera
Federal University of Rio Grande do Sul
Porto Alegre, Brazil
{ronunes, dgbalreira, jlcarbonera}@inf.ufrgs.br

Abstract

This work presents a systematic evaluation of Large Language Models (LLMs) for generating specialized definitions in Portuguese, focusing on the medical and geological domains. We introduce a robust benchmark and employ a rigorous, statistically grounded evaluation framework, including 5-fold cross-validation and significance testing, to ensure the reliability and generalizability of our findings. Our comprehensive experiments with various open-source, decoder-only LLMs explore in-context learning (ICL) with diverse prompting strategies, ranging from zero-shot to few-shot and contextual information. The evaluated models include multilingual architectures and one model that underwent continued pretraining specifically for Portuguese, allowing us to assess the impact of language adaptation on definition generation quality. The results indicate that most evaluated models perform effectively in this task, with relatively small performance differences among the top models. Statistical analyses confirmed that these differences are not consistently significant, suggesting that several open LLMs, regardless of their size, multilingual capacity, or language specialization, offer comparable effectiveness for Portuguese definition generation. These findings provide valuable insights for selecting and adapting models for specialized NLP tasks in low-resource languages like Portuguese.

Keywords: definition generation, specialized domains, Large Language Models, low-resource, Portuguese language, geological terminology, medical terminology

1. Introduction

Glossaries, thesauri, and ontologies are essential resources for obtaining definitions of general and domain-specific terms (Lopes et al., 2024). These definitions can be applied in various natural language processing (NLP) tasks, such as Word Sense Disambiguation (Navigli and Velardi, 2005; AlMousa et al., 2022), ontology alignment (Lopes et al., 2024), and improving the comprehensibility of domain-specific documents, such as legal, historical and medical texts (Nunes et al., 2018; Giulianelli et al., 2023; Chouhan and Gertz, 2024).

Despite the existence of curated lexical resources and structured knowledge bases (Gadetsky et al., 2018; Riaño et al., 2019; Navigli et al., 2021; Furtado et al., 2024), achieving comprehensive coverage of specialized terminology across languages and domains remains a fundamental challenge — particularly in low-resource settings and technical fields. To address this limitation, recent work has focused on the task of Definition Generation (DG), a relatively novel NLP task (Gadetsky et al., 2018), which aims to generate a definition given a specific term.

Although efforts have been made to generate definitions for both general and specialized domains (Giulianelli et al., 2023; Chouhan and Gertz, 2024; Furtado et al., 2024), there is a notable lack of research targeting languages other than English, such as Portuguese (Furtado et al., 2024). To the best of our knowledge, the only study that evaluates DG in Portuguese is DORE (Furtado et al., 2024). However, that work focuses solely on general knowledge definitions obtained from dictionaries such as *Dicio* and *Portuguese Wiktionary*. While these are reliable sources, they are not designed to represent specialized domain knowledge, such as in Geology or Medicine.

In this work, we present a systematic evaluation of definitions generated by decoder-only Large Language Models (LLMs) in specialized domains for the Portuguese language. Our key contributions are threefold:

- **Pioneering evaluation in Portuguese**: We conduct, to the best of our knowledge, the first systematic assessment of definition generation in specialized domains, specifically medicine and geology, for the Portuguese language;
- **Comprehensive benchmarking**: We establish a robust benchmark that compares multiple LLMs under diverse prompting strategies, including zero-shot and in-context learning (ICL), with and without the inclusion of domain-specific contextual information;
- **Rigorous evaluation methodology**: We adopt a statistically grounded evaluation frame-

Source code available at https://github.com/RafaelOleques/definition_generation_lrec2026.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3794–3804
11–16 May 2026.

work, employing 5-fold cross-validation and significance testing to ensure the robustness, reliability, and generalizability of our findings.

2. Related Work

Definition Generation (DG) was first proposed by Noraset et al. (2017), who framed it as a word-to-sequence task using word embeddings to generate natural-language definitions. However, polysemy poses a significant challenge: a term like "bank" may require different definitions depending on context. To address this, early work incorporated local context and global context (Ni and Wang, 2017; Gadetsky et al., 2018; Ishiwatari et al., 2019).

Subsequent advances shifted from RNN-based architectures (Ni and Wang, 2017; Ishiwatari et al., 2019; Li et al., 2020) to transformer-based models. Encoder–decoder frameworks demonstrated improved fluency and accuracy (Giulianelli et al., 2023; Furtado et al., 2024), while encoder-only and decoder-only approaches explored prompt-based and few-shot generation strategies (Chouhan and Gertz, 2024; Furtado et al., 2024).

In Portuguese, DG remains underexplored. To the best of our knowledge, the only dedicated evaluation is Furtado et al. (2024)'s DORE benchmark, which relies on general-domain dictionaries (Dicio, Portuguese Wiktionary) and does not cover specialized vocabularies. Our work extends this line by targeting medical and geological domains and introducing rigorous cross-validation and statistical testing.

3. Definition Generation in Specialized Domains

This section describes the corpora and resources used to support definition generation in specialized Portuguese domains. We detail the construction and use of domain-specific glossaries and contextual corpora of Named Entity Recognition (NER). These complementary resources form the basis for training and evaluating our definition generation models.

3.1. Domain-Specific Glossaries

For the medical domain, we used the Medical Subject Headings (MeSH) thesaurus[1], incorporating its multilingual definitions provided through the DeCS/MeSH initiative (dec, 2018). MeSH was selected for its reliable, expert-curated definitions maintained by PubMed. We adopted the 2018 version, which contains 28,939 definitions and 121,318 terms.

For the geological domain, we employed the glossary provided by the Brazilian Commission of Geological and Paleobiological Sites[2] (SIGEP), using the version published on May 11, 2025[3]. This glossary compiles domain-specific terminology related to Brazilian geological heritage and was chosen for its authoritative content and focus on national geoscientific resources. The glossary contains 1,514 definitions and 2,026 terms.

3.2. Named Entity Recognition Corpora

To provide contextual information for the specialized terms, we utilized NER corpora from two Portuguese-language domains: medicine and geology. NER corpora are especially useful in this task because they contain annotated entities representing domain-specific concepts, ensuring that terms appear in realistic and meaningful textual contexts. Including such context can help models better understand how a term is used in practice and may assist in disambiguating its meaning in possible cases of polysemy.

For the medical domain, we employed the **BETE** corpus (Pavanelli et al., 2023), which focuses on diabetes-related texts. BETE offers expert annotations on relevant medical entities, providing a concentrated source of terminology crucial for evaluating term definitions in healthcare contexts. Although smaller in size, BETE's high-quality annotations make it valuable for our task.

In the geological domain, we used **GeoCorpus-3** (Gomes et al., 2021; Nunes et al., 2024), which consists of texts about Brazilian sedimentary basins. This corpus contains a wider range of geological terms and diverse contexts, reflecting the complexity and richness of geological language. Its larger scale complements the medical corpus by covering a different specialized domain with distinct terminology.

The two corpora differ in size and lexical variety. GeoCorpus-3 includes 5,272 sentences, with a total of 166,462 tokens and 16,694 unique tokens, making it a rich resource for geological terminology in context. BETE contains 439 sentences, 43,198 tokens, and 2,844 unique tokens, offering a more focused but domain-critical set of medical terms.

By integrating these NER corpora, we ensure that the context used for definition generation is both relevant and representative of real-world usage in specialized domains. This approach strengthens the evaluation by grounding model outputs in authentic textual evidence.

[1] https://meshb.nlm.nih.gov/

[2] Original in Portuguese: *Comissão Brasileira de Sítios Geológicos e Paleobiológicos*.

[3] Available at https://sigep.eco.br/glossario/index.html

3.3. NER and Glossary Alignment

To connect formal domain knowledge with real-world language use, we aligned glossary entries with contextual instances drawn from specialized NER corpora. This alignment process was designed to identify where glossary terms, including their lexical variants, appear in natural text. The outcome is a term-context-definition triad that supports more precise and context-aware definition generation.

We started by constructing a reverse index that maps each term and its synonyms to a canonical concept identifier. This indexing step is crucial, as many glossary entries include not only a preferred label but also a set of alternative lexical forms used in practice. By flattening these variants into a unified term-to-ID mapping, our system could detect both base terms and synonyms in the corpus and correctly associate them with their glossary definitions.

For term matching, all sentences and glossary terms were lowercased to ensure consistent comparisons. We adopted a strict word-boundary regular expression strategy, ensuring that only exact, standalone terms were matched—this avoided spurious hits such as matching "renal" inside "adrenal". Importantly, we deliberately avoided the use of stemming or lemmatization, since these techniques, while useful for general IR or classification tasks, risk merging semantically distinct terms in specialized contexts. In high-precision tasks such as definition generation, such conflation introduces unacceptable noise, reducing the reliability of alignments.

We also chose not to incorporate embedding-based similarity measures (e.g., cosine distance between contextualized or static word vectors) to expand term coverage. Although embedding distances have been shown to increase recall in term matching tasks, the literature reports that their effectiveness is highly sensitive to threshold tuning, with accuracy varying substantially depending on the semantic density of the domain (Nunes et al., 2018). Given our focus on precision and reproducibility across domains, we prioritized deterministic and interpretable matching techniques.

Once terms were identified in the corpus, we retrieved their corresponding scope notes from the glossary and assembled a structured dataset containing: the matched term or synonym, the set of sentences in which it appeared, and its definition. This resource is the basis for the subsequent stages of our pipeline, enabling targeted prompt construction and supporting evaluations of term coverage and generation quality.

By grounding lexical knowledge in high-confidence contextual instances, our alignment strategy not only reinforces the semantic fidelity of the glossary but also enhances its operational value in language generation tasks. This dual benefit ensures that definitions are both semantically accurate and pragmatically useful, particularly when applied to specialized domains where precision and contextual appropriateness are critical.

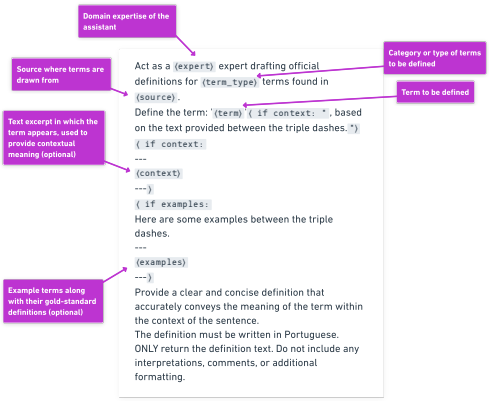

Figure 1: Components of the prompt used in this work.

3.4. Models and Prompting

We evaluate DG using both multilingual and monolingual decoder-only LLMs. We chose to exclusively use decoder-only models due to their strong performance in prior benchmarks for the Portuguese language, as reported by Furtado et al. (2024).

In line with previous work, we selected models with approximately 7 billion parameters or more. The smallest model included in our evaluation is GAIA (CAMILO-JUNIOR et al., 2025), which has 4 billion parameters. We opted to include it because, to the best of our knowledge, it is the only open-source Portuguese LLM that has been instruction-tuned, rather than relying on parameter-efficient fine-tuning methods such as LoRA (Garcia et al., 2024), or being distributed as a closed model (Pires et al., 2023; Almeida et al., 2024; Abonizio et al., 2024).

All models were run locally using the OLlama framework for unified execution and consistent resource usage. The models evaluated are as follows:

- **GAIA (Gemma-based)** (CAMILO-JUNIOR et al., 2025) – 4B parameters, monolingual (Portuguese);
- **Gemma-3** (Team et al., 2025) – 12B parameters, multilingual;
- **Llama-3.1** (Grattafiori et al., 2024) – 8B parameters, multilingual;

- **Mistral** (Jiang et al., 2023) – 7B parameters, multilingual;
- **Phi-3** (Abdin et al., 2024) – 14B parameters, multilingual;
- **Qwen-3** (Yang et al., 2025) – 14B parameters, multilingual.

For the prompting strategies, we followed approaches proposed in prior work on definition generation for specialized terms (Chouhan and Gertz, 2024; Zhang and Soh, 2024). The main structure of our prompt is based on the format introduced by Chouhan and Gertz (2024).

We introduce two key adaptations to tailor the prompting strategy to our multi-domain scenario. First, we modified the domain-specific sections of the prompt to ensure applicability across different specialized areas. For example, in the geological domain, we instructed the model to define *geological* terms as if it were an expert in *geoscience*, drawing on *scientific reports and bulletins in geoscience* as contextual sources. In the medical domain, the model was prompted to define *medical* terms from the perspective of a *medical* expert, leveraging knowledge akin to that found in *PubMed citations*. This contrasts with the original prompt, which was restricted to the legal domain.

Second, we added support for ICL by incorporating example terms along with their definitions, inspired by prior work on example-based prompting (Zhang and Soh, 2024). Example selection for ICL was guided by semantic similarity: we computed the cosine distance between the target term and all candidate terms in the training set, selecting the top-k most similar examples (e.g., one for 1-shot, two for 2-shot, and so on).

To isolate the individual contributions of examples and contextual information, we designed two distinct prompt progressions, each anchored at a different baseline:

- **Example settings:** starts from the pure baseline (no examples, no context) and then adds ICL examples (examples only).
- **Context settings:** starts from the context-only baseline and then adds the same ICL examples (examples + context).

These separate baselines allow us to clearly compare how much examples improve performance on their own versus the additional gains they provide when combined with contextual information. Figure 1 illustrates the prompt format employed.

4. Experimental Evaluation

Our experiments were conducted using an RTX 4090 GPU. To run the models, we used the OLlama framework in combination with LangChain. For retrieving semantically similar terms, we employed the Sentence-Transformers library with the *all-MiniLM-L6-v2* model. All hyperparameters were kept at their default values, and a temperature of 0 was set to ensure reproducibility.

Regarding evaluation metrics, we used BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) for syntactic similarity analysis, and BERTScore (Zhang* et al., 2020) for semantic similarity. All the metrics are used to compare the original definition with the one generated by the LLMs. To increase the reliability of the results, we performed 5-fold cross-validation with a random split. We ensured that terms identified as synonyms in the thesaurus were grouped within the same fold using the *GroupKFold* method from the `scikit-learn` library.

For our statistical analysis, we adopted non-parametric tests, as the assumption of normality was violated according to the Shapiro–Wilk test. To compare model performance with and without contextual information, we applied the Wilcoxon signed-rank test. To evaluate the impact of the number of examples in the ICL setup, we used the Friedman test, followed by Holm's post-hoc correction for multiple comparisons.

These tests were chosen based on established best practices for robust model comparison in Machine Learning (Demšar, 2006; Garcia and Herrera, 2008) and Natural Language Processing (Dror et al., 2018; Schuff et al., 2023), particularly under non-normal data distributions. This approach allows us to draw more reliable conclusions regarding model behavior across different prompting strategies.

5. Results and Discussion

This section presents the main findings of our evaluation on definition generation in specialized domains. We focus primarily on the F1 score of BERTScore, as it better captures semantic similarity than exact surface overlap. For comparison, additional metrics are reported in Tables 1 and 2. In these tables, **bolded values** indicate the best zero-shot and few-shot scores for each metric, while underlined values highlight the best result per model.

We structure our discussion around three main research questions, each addressing a different aspect of model behavior. **RQ1** considers which models perform best overall and what limitations they exhibit across prompting conditions. **RQ2** asks whether model size reliably predicts performance. Finally, **RQ3** explores broader trends that emerge from statistical analyses of the results.

Model	Context	Shots	Precision (%)	Recall (%)	F1 (%)	BLEU (%)	METEOR (%)
gaia:4b	with context	0	67.07 ± 0.40	67.30 ± 0.32	67.07 ± 0.32	0.58 ± 0.55	11.66 ± 0.49
		1	70.10 ± 0.56	70.15 ± 0.64	70.01 ± 0.54	7.14 ± 1.20	18.50 ± 1.03
		2	70.19 ± 0.67	70.63 ± 0.57	70.29 ± 0.53	8.59 ± 2.71	18.99 ± 1.47
		4	70.87 ± 0.58	71.50 ± 0.86	71.08 ± 0.71	11.55 ± 4.43	20.92 ± 2.40
		8	70.92 ± 0.84	71.42 ± 1.00	71.07 ± 0.89	**12.38 ± 2.26**	21.82 ± 3.03
	without context	0	69.14 ± 0.82	69.17 ± 0.56	69.05 ± 0.50	**1.28 ± 0.48**	14.20 ± 0.56
		1	71.01 ± 0.63	71.12 ± 0.42	70.97 ± 0.42	5.96 ± 2.40	17.53 ± 1.27
		2	70.84 ± 0.53	71.39 ± 0.67	71.02 ± 0.48	4.98 ± 2.41	18.13 ± 1.82
		4	71.22 ± 0.58	71.67 ± 0.91	71.34 ± 0.69	5.45 ± 3.70	19.22 ± 1.85
		8	71.45 ± 0.84	71.93 ± 1.11	71.58 ± 0.92	7.58 ± 1.83	20.13 ± 1.56
gemma3:12b	with context	0	67.94 ± 0.36	67.11 ± 0.57	67.41 ± 0.30	0.25 ± 0.36	10.71 ± 0.56
		1	70.39 ± 0.53	68.60 ± 0.53	69.35 ± 0.44	3.11 ± 1.03	12.87 ± 1.10
		2	70.64 ± 0.51	69.18 ± 0.65	69.77 ± 0.56	4.15 ± 1.23	14.57 ± 1.29
		4	71.00 ± 0.82	69.35 ± 0.95	70.04 ± 0.82	4.48 ± 2.97	14.36 ± 1.82
		8	72.55 ± 0.67	69.85 ± 0.91	71.03 ± 0.77	7.26 ± 2.23	16.63 ± 2.40
	without context	0	68.61 ± 0.86	69.09 ± 0.52	68.76 ± 0.64	1.19 ± 0.30	13.59 ± 0.58
		1	69.59 ± 0.69	70.09 ± 0.65	69.73 ± 0.50	1.88 ± 0.25	14.85 ± 0.42
		2	69.21 ± 0.70	70.50 ± 0.71	69.76 ± 0.58	1.83 ± 0.37	15.69 ± 1.12
		4	69.29 ± 0.54	70.92 ± 0.70	70.00 ± 0.54	2.23 ± 0.27	16.47 ± 0.68
		8	69.61 ± 0.50	71.37 ± 1.00	70.39 ± 0.70	2.40 ± 0.25	17.11 ± 0.87
llama3.1:8b	with context	0	69.20 ± 0.49	68.62 ± 0.62	68.82 ± 0.51	0.74 ± 0.51	12.66 ± 0.58
		1	71.70 ± 0.67	70.47 ± 0.95	70.98 ± 0.76	4.81 ± 0.94	16.87 ± 2.36
		2	71.94 ± 0.57	70.85 ± 1.02	71.28 ± 0.73	5.92 ± 1.37	18.18 ± 2.78
		4	72.69 ± 0.90	71.53 ± 1.36	71.99 ± 1.11	8.19 ± 1.89	20.32 ± 3.88
		8	73.27 ± 0.74	72.23 ± 1.22	**72.64 ± 0.97**	9.82 ± 2.30	21.69 ± 3.32
	without context	0	**69.23 ± 0.50**	**69.68 ± 0.51**	**69.37 ± 0.33**	0.97 ± 0.61	14.85 ± 0.80
		1	72.24 ± 0.41	70.70 ± 0.96	71.34 ± 0.62	4.95 ± 0.84	17.75 ± 2.00
		2	72.64 ± 0.96	71.53 ± 1.12	71.96 ± 0.98	5.97 ± 1.35	19.62 ± 2.78
		4	72.88 ± 0.81	71.87 ± 0.95	72.24 ± 0.79	6.95 ± 0.52	20.31 ± 1.71
		8	**73.30 ± 0.75**	72.13 ± 1.19	72.58 ± 0.87	7.35 ± 1.51	20.46 ± 2.81
mistral:7b	with context	0	66.06 ± 0.48	67.77 ± 0.47	66.82 ± 0.35	0.69 ± 0.24	13.18 ± 0.65
		1	69.52 ± 0.20	70.75 ± 0.77	70.02 ± 0.45	7.84 ± 1.52	19.91 ± 1.72
		2	69.63 ± 0.26	70.95 ± 0.73	70.16 ± 0.42	8.68 ± 1.45	20.00 ± 1.73
		4	69.48 ± 0.27	70.93 ± 0.59	70.08 ± 0.30	8.24 ± 1.14	20.05 ± 1.44
		8	68.83 ± 0.45	70.83 ± 0.69	69.68 ± 0.52	7.90 ± 1.42	19.46 ± 1.62
	without context	0	65.72 ± 0.48	68.52 ± 0.53	67.01 ± 0.34	0.92 ± 0.38	**16.06 ± 1.17**
		1	70.83 ± 0.97	72.05 ± 0.79	71.33 ± 0.78	11.58 ± 2.89	22.78 ± 1.71
		2	71.08 ± 0.70	72.52 ± 0.96	71.69 ± 0.76	10.11 ± 2.74	22.39 ± 2.10
		4	71.47 ± 0.62	72.91 ± 0.84	72.06 ± 0.60	11.28 ± 2.57	**22.98 ± 2.59**
		8	71.36 ± 0.57	**73.15 ± 0.46**	72.10 ± 0.36	9.22 ± 1.25	22.81 ± 0.84
phi3:14b	with context	0	62.99 ± 5.00	66.24 ± 5.09	64.51 ± 5.01	0.59 ± 0.33	12.63 ± 4.08
		1	58.56 ± 11.49	61.61 ± 12.14	59.95 ± 11.78	3.35 ± 3.32	13.23 ± 8.24
		2	57.21 ± 13.18	60.22 ± 14.31	58.59 ± 13.69	3.64 ± 3.58	13.70 ± 8.70
		4	46.23 ± 28.62	48.64 ± 30.23	47.33 ± 29.35	4.10 ± 3.77	12.50 ± 9.95
		8	49.53 ± 24.77	51.82 ± 26.29	50.57 ± 25.47	4.44 ± 4.10	13.43 ± 10.36
	without context	0	65.02 ± 4.28	68.23 ± 4.06	66.50 ± 4.14	0.89 ± 0.59	14.48 ± 4.87
		1	66.03 ± 9.66	68.33 ± 9.30	67.05 ± 9.45	5.80 ± 3.54	18.16 ± 7.57
		2	66.59 ± 8.35	69.02 ± 8.16	67.68 ± 8.21	6.66 ± 4.06	18.86 ± 7.66
		4	66.49 ± 8.46	69.17 ± 8.14	67.69 ± 8.25	7.03 ± 4.63	19.40 ± 7.95
		8	63.41 ± 13.21	65.58 ± 13.13	64.37 ± 13.12	6.85 ± 5.46	18.53 ± 8.37
qwen3:14b	with context	0	52.91 ± 0.50	62.07 ± 0.18	57.06 ± 0.33	0.07 ± 0.04	8.32 ± 0.50
		1	52.68 ± 0.51	63.00 ± 0.25	57.30 ± 0.36	0.33 ± 0.17	9.49 ± 0.44
		2	52.65 ± 0.57	62.82 ± 0.23	57.20 ± 0.32	0.30 ± 0.28	9.38 ± 0.57
		4	52.77 ± 0.53	63.22 ± 0.35	57.44 ± 0.28	0.32 ± 0.28	10.07 ± 0.62
		8	52.79 ± 0.54	63.16 ± 0.32	57.42 ± 0.35	0.47 ± 0.25	9.79 ± 0.57
	without context	0	52.79 ± 0.43	63.00 ± 0.32	57.37 ± 0.25	0.19 ± 0.04	9.36 ± 0.55
		1	52.79 ± 0.46	63.69 ± 0.47	57.65 ± 0.24	0.77 ± 0.24	10.93 ± 0.62
		2	52.99 ± 0.55	64.18 ± 0.36	57.97 ± 0.32	0.63 ± 0.27	11.37 ± 0.37
		4	53.01 ± 0.51	64.23 ± 0.31	58.00 ± 0.23	0.49 ± 0.24	11.45 ± 0.89
		8	53.09 ± 0.46	64.41 ± 0.41	58.11 ± 0.17	0.85 ± 0.31	11.70 ± 0.65

Table 1: Summary of all metrics by Model, Context, and Number of Shots (BETE using MeSH definitions).

RQ1: Which models demonstrate the best overall performance, and what are the main limitations observed across the evaluated settings?

Tables 1 and 2 summarize the performance of all evaluated models under prompting strategies that vary in the presence of examples and contextual information. Overall, the models that achieved the highest and most consistent performance were **GAIA**, **Gemma-3**, **Llama-3.1**, and **Mistral**. These models stood out across both domains, with BERTScores approaching 70% in the medical domain and 68% in the geological domain.

These same models also proved robust to prompting variations, maintaining strong performance across both "with context" and "without context" settings. Notably, even in the most minimal zero-shot configurations (i.e., no examples, no context), GAIA, Gemma-3, and Mistral still reached close to 66% BERTScore in geology and around 68% in medicine. Their relative insensitivity to the addition or absence of context and examples suggests a strong internal representation of definitional patterns in Portuguese, independent of elaborate prompt engineering.

In contrast, **Phi-3** consistently yielded the lowest performance. In the geological domain, it under-

Model	Context	Shots	Precision (%)	Recall (%)	F1 (%)	BLEU (%)	METEOR (%)
gaia:4b	with context	0	67.51 ± 0.98	65.57 ± 0.69	66.40 ± 0.76	0.71 ± 0.35	10.14 ± 0.92
		1	68.39 ± 0.81	66.55 ± 0.75	67.32 ± 0.67	1.42 ± 0.76	11.90 ± 1.37
		2	68.61 ± 0.73	66.95 ± 0.64	67.64 ± 0.58	1.79 ± 1.09	12.77 ± 1.47
		4	68.74 ± 0.80	67.39 ± 0.59	67.92 ± 0.66	2.23 ± 1.01	13.46 ± 1.41
		8	68.50 ± 0.58	67.67 ± 0.74	67.94 ± 0.59	3.01 ± 1.38	14.50 ± 1.39
	without context	0	68.75 ± 0.81	67.01 ± 0.35	67.73 ± 0.51	0.91 ± 0.31	11.58 ± 0.86
		1	68.77 ± 0.79	67.73 ± 0.69	68.13 ± 0.62	1.90 ± 1.05	13.39 ± 1.28
		2	68.76 ± 0.92	67.86 ± 0.71	68.17 ± 0.72	2.24 ± 1.70	13.82 ± 1.93
		4	68.91 ± 1.00	67.97 ± 0.66	68.30 ± 0.71	2.10 ± 1.81	14.32 ± 1.37
		8	68.55 ± 0.73	68.03 ± 0.80	68.15 ± 0.64	2.91 ± 2.12	14.58 ± 1.90
gemma3:12b	with context	0	69.10 ± 0.45	64.67 ± 0.27	66.67 ± 0.34	0.25 ± 0.11	8.23 ± 0.72
		1	69.83 ± 0.57	66.02 ± 0.41	67.75 ± 0.47	0.65 ± 0.27	10.37 ± 0.77
		2	69.69 ± 0.69	66.23 ± 0.47	67.79 ± 0.51	0.88 ± 0.65	10.61 ± 1.01
		4	69.89 ± 0.72	66.51 ± 0.62	68.04 ± 0.63	0.95 ± 0.57	10.95 ± 0.86
		8	70.28 ± 0.49	66.24 ± 0.57	68.07 ± 0.50	1.10 ± 0.65	10.88 ± 0.97
	without context	0	**69.95 ± 0.80**	**67.94 ± 0.55**	**68.78 ± 0.54**	0.91 ± 0.34	**12.28 ± 0.80**
		1	69.52 ± 0.59	68.84 ± 0.60	69.03 ± 0.50	1.62 ± 0.47	14.69 ± 0.72
		2	68.99 ± 0.84	68.92 ± 0.58	68.80 ± 0.57	1.75 ± 0.52	15.13 ± 1.21
		4	69.03 ± 0.67	69.09 ± 0.65	68.90 ± 0.51	1.70 ± 0.43	15.24 ± 0.86
		8	69.21 ± 0.67	**69.11 ± 0.76**	69.01 ± 0.58	1.98 ± 0.86	15.53 ± 1.16
llama3.1:8b	with context	0	69.89 ± 1.03	66.66 ± 0.65	68.11 ± 0.79	0.57 ± 0.18	10.75 ± 0.92
		1	70.80 ± 0.99	67.45 ± 0.75	68.96 ± 0.79	1.23 ± 0.66	12.67 ± 1.78
		2	70.82 ± 1.04	67.58 ± 0.93	69.04 ± 0.94	1.57 ± 1.08	13.09 ± 2.24
		4	70.82 ± 1.16	67.69 ± 1.11	69.09 ± 1.09	1.82 ± 1.17	13.59 ± 2.38
		8	**71.21 ± 1.05**	68.17 ± 1.12	**69.52 ± 1.04**	2.36 ± 1.27	15.01 ± 1.98
	without context	0	69.36 ± 0.88	67.12 ± 0.73	68.09 ± 0.76	0.61 ± 0.25	11.89 ± 1.51
		1	70.05 ± 1.08	67.35 ± 0.78	68.54 ± 0.88	1.26 ± 0.80	12.51 ± 1.63
		2	70.18 ± 0.97	67.72 ± 0.89	68.79 ± 0.87	1.69 ± 1.19	13.60 ± 2.31
		4	70.35 ± 0.96	68.03 ± 0.90	69.03 ± 0.88	2.12 ± 1.52	14.14 ± 2.15
		8	70.59 ± 1.14	68.09 ± 1.08	69.17 ± 1.05	2.49 ± 1.65	14.68 ± 2.37
mistral:7b	with context	0	67.74 ± 0.99	66.51 ± 0.52	66.98 ± 0.70	**1.02 ± 0.45**	11.27 ± 0.74
		1	68.38 ± 0.82	67.07 ± 0.69	67.57 ± 0.63	1.51 ± 0.63	12.66 ± 1.54
		2	68.38 ± 0.90	67.43 ± 0.91	67.76 ± 0.82	2.02 ± 0.86	13.26 ± 2.00
		4	68.54 ± 0.93	67.78 ± 0.91	68.01 ± 0.85	2.45 ± 0.93	14.00 ± 2.10
		8	68.36 ± 0.73	67.83 ± 0.83	67.95 ± 0.67	2.88 ± 1.17	14.26 ± 2.16
	without context	0	65.87 ± 0.45	65.99 ± 0.43	65.79 ± 0.37	0.70 ± 0.49	12.17 ± 0.90
		1	68.65 ± 0.69	67.76 ± 0.72	68.06 ± 0.64	2.10 ± 1.07	13.99 ± 1.53
		2	68.81 ± 0.65	68.33 ± 0.83	68.42 ± 0.67	2.80 ± 1.61	15.18 ± 1.70
		4	68.97 ± 0.85	68.54 ± 0.72	68.60 ± 0.70	3.19 ± 1.23	15.39 ± 1.62
		8	68.79 ± 0.82	68.57 ± 0.94	68.52 ± 0.80	**4.30 ± 1.73**	**16.27 ± 2.14**
phi3:14b	with context	0	44.73 ± 1.76	43.02 ± 2.23	43.75 ± 2.01	0.00 ± 0.00	2.46 ± 0.30
		1	48.05 ± 2.77	45.87 ± 3.04	46.83 ± 2.89	0.00 ± 0.00	2.54 ± 0.40
		2	47.65 ± 1.37	45.59 ± 1.09	46.48 ± 1.21	0.00 ± 0.00	2.49 ± 0.23
		4	39.47 ± 2.94	37.55 ± 2.65	38.40 ± 2.78	0.00 ± 0.00	2.00 ± 0.20
		8	21.28 ± 1.52	20.00 ± 1.63	20.54 ± 1.57	0.00 ± 0.00	0.93 ± 0.10
	without context	0	54.32 ± 0.77	53.11 ± 0.96	53.62 ± 0.83	0.00 ± 0.00	3.00 ± 0.27
		1	45.96 ± 2.22	43.79 ± 2.02	44.72 ± 2.10	0.00 ± 0.00	2.30 ± 0.24
		2	46.28 ± 1.47	44.25 ± 1.47	45.14 ± 1.47	0.00 ± 0.00	2.36 ± 0.25
		4	43.20 ± 3.63	40.60 ± 3.94	41.75 ± 3.79	0.00 ± 0.00	1.98 ± 0.22
		8	34.03 ± 1.74	32.50 ± 1.69	33.16 ± 1.70	0.00 ± 0.00	1.73 ± 0.13
qwen3:14b	with context	0	54.57 ± 0.67	61.97 ± 0.13	57.92 ± 0.43	0.21 ± 0.08	9.40 ± 0.47
		1	54.81 ± 0.73	62.86 ± 0.27	58.42 ± 0.43	0.16 ± 0.04	9.97 ± 0.40
		2	54.91 ± 0.69	63.01 ± 0.34	58.54 ± 0.41	0.41 ± 0.13	10.48 ± 0.30
		4	54.92 ± 0.59	63.06 ± 0.47	58.57 ± 0.29	0.30 ± 0.14	10.64 ± 0.64
		8	54.91 ± 0.69	63.14 ± 0.29	58.60 ± 0.40	0.41 ± 0.14	10.77 ± 0.49
	without context	0	55.24 ± 0.70	63.44 ± 0.25	58.93 ± 0.49	0.29 ± 0.06	10.72 ± 0.34
		1	55.61 ± 0.70	64.10 ± 0.23	59.40 ± 0.39	0.45 ± 0.10	10.98 ± 0.30
		2	55.78 ± 0.67	64.44 ± 0.42	59.64 ± 0.44	0.47 ± 0.22	11.35 ± 0.63
		4	55.75 ± 0.81	64.46 ± 0.41	59.63 ± 0.56	0.57 ± 0.26	11.27 ± 0.74
		8	55.82 ± 0.73	64.52 ± 0.11	59.71 ± 0.43	0.67 ± 0.27	11.70 ± 0.74

Table 2: Summary of all metrics by Model, Context, and Number of Shots (GeoCorpus-3 using SIGEP definitions).

performed in all prompting conditions and was particularly affected by the presence of context: while its zero-shot scores (without context) ranged from 62% to 65%, these dropped to as low as 54% in the setting using context. Phi-3 also showed the highest standard deviation across runs, ranging from 0.83 to 3.79 in the no-context setup and 1.21 to 2.89 with context, indicating pronounced instability.

Although Phi-3 appeared slightly more stable in the medical domain, with average scores between 64% and 66% (no context) and 47% to 64% (with context), its variance was substantially higher: standard deviations ranged from 4 to 13 without context and from 5 to 29 with context. This inconsistency underscores its unreliability and supports prior claims about the need for repeated trials in model evaluation (Gorman and Bedrick, 2019).

Qwen-3, while also underperforming relative to the top-tier models, demonstrated much greater stability. It maintained flat performance curves around 57% in geology and 58% in medicine across all prompting scenarios, with consistently low standard deviations (around 0.43), as illustrated in Figures 2 and 3. However, its scores remained lower than those of any model except Phi-3, indicating that high stability does not necessarily accompany

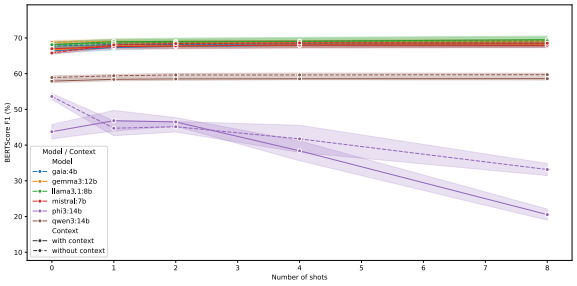

Figure 2: BERTScore (F1) as a function of the number of shots, across prompting scenarios and models, using SIGEP terms from GeoCorpus-3. Shaded areas denote standard deviation.

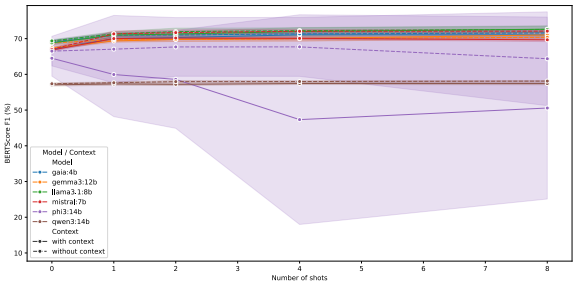

Figure 3: BERTScore (F1) as a function of the number of shots, across prompting scenarios and models, using MeSH/DeCS terms from BETE. Shaded areas denote standard deviation.

strong overall performance in this task.

In summary, the four strongest models, GAIA, Gemma-3, Llama-3.1, and Mistral, delivered high and consistent performance regardless of the prompting configuration, making them reliable choices for definition generation in specialized Portuguese domains. On the other end, Phi-3's poor average results and high variability make it unsuitable for this task, while Qwen-3, despite its stability, remains limited in overall effectiveness.

RQ2: Is model size a reliable indicator of performance in this task?

Although our primary objective was not to conduct a controlled comparison across different model sizes, the range of models evaluated in this work naturally included architectures of varying scales. This diversity allowed us to reflect on the extent to which model size alone explains performance differences in the task of definition generation for specialized Portuguese-language terms.

Surprisingly, model size does not appear to be a reliable predictor of performance in this context. The largest models in our benchmark, **Phi-3** (14B) and **Qwen-3** (14B), consistently delivered the weakest results across both domains. In contrast, smaller models such as **Gemma-3** (12B), **Llama-3.1** (8B), and **Mistral** (7B) achieved strong and stable performance, frequently reaching or surpassing 68% BERTScore in the geological domain and 70% in the medical domain.

A more promising explanatory factor is the nature of each model's pretraining. For example, **Phi-3**, which exhibited both the lowest average performance and the highest variance across few-shot configurations, is essentially a monolingual model trained predominantly on filtered English web data (Abdin et al., 2024). As the authors acknowledge, its base version lacks multilingual capabilities, which likely contributes to its instability and weak results in Portuguese.

Conversely, models explicitly designed for multilingual generalization, such as **Gemma-3**, **Llama-3.1**, and **Mistral**, performed consistently well across all prompting scenarios. Although **Qwen-3** is also advertised as a multilingual model (Yang et al., 2025), it underperformed relative to its peers by roughly 10 percentage points in BERTScore (F1). This is particularly notable given that, in its original evaluation (Yang et al., 2025), Qwen-3 (14B) reported competitive results for Portuguese, even surpassing GPT-4o and Gemini 2.5 Pro in some benchmarks.

A special case in our benchmark is GAIA, the only model explicitly reported to have undergone continued pretraining in Portuguese. Despite being the smallest model (4B parameters), GAIA achieved competitive performance comparable to much larger multilingual models across both domains. Although we cannot directly compare it to other models of similar size, its results suggest that domain and language-specific pretraining can meaningfully boost performance. While it is plausible that the larger models were also exposed to definitions and specialized content during training, GAIA is distinctive for its explicit focus on the Portuguese language.

In summary, while model size alone does not reliably predict performance in this task, our results suggest that architecture, multilingual pretraining objectives, and explicit exposure to the target language may play a more decisive role. These findings may inform future model selection and adaptation strategies for specialized-domain NLP tasks in underrepresented languages such as Portuguese.

RQ3: What trends emerge from statistical analyses?

The statistical analysis revealed a limited number of significant differences between models, particularly in comparisons involving the weakest performers. Among the top-performing models, however, we found little to no statistically significant varia-

tion, suggesting that their performances are largely comparable within the evaluation configurations applied.

In the medical corpus, the Friedman test followed by Holm's post-hoc test detected statistically significant differences between **Llama-3.1** and **Mistral** in the zero-shot setting (without context: $p = 0.0487$; with context: $p = 0.0305$), and between **Gemma-3** and **Llama-3.1** in the 2-shot setting without context ($p = 0.0328$). Still, these are isolated cases and may reflect scenario-specific fluctuations rather than consistent model superiority. In the geological domain, no statistically significant differences were found among the top-tier models, suggesting a high degree of convergence in performance for this task.

As expected, the statistical tests more consistently revealed significant performance gaps between the strongest and weakest models. For instance, **Llama-3.1** exhibited statistically significant differences in 5 settings when compared to **Qwen-3** and in 10 when compared to **Phi-3** in the medical domain. In the geological domain, Llama-3.1 again showed 10 and 8 differences, respectively, with Qwen-3 and Phi-3. Likewise, **Gemma-3** demonstrated six significant differences with Phi-3 and four with Qwen-3 in the geological corpus. Other models showed fewer significant contrasts; for example, **GAIA** and **Mistral** each differed significantly from Qwen-3 in only one setting.

These results broadly corroborate earlier analyses: **Phi-3** and **Qwen-3** were consistently outperformed, while the leading models, **GAIA**, **Gemma-3**, **Llama-3.1**, and **Mistral**, exhibited statistically indistinguishable performance in most conditions. However, these findings also suggest a ceiling effect in evaluation: while underperforming models are clearly identified, differences among strong models may be too subtle to be captured reliably under the current experimental setup.

We also investigated whether few-shot prompting configurations led to statistically superior outcomes compared to zero-shot setups. Surprisingly, the Wilcoxon signed-rank test found no significant differences between the best few-shot and zero-shot configurations for the top models. For instance, **Llama-3.1** with 8-shots and context did not significantly outperform its zero-shot counterpart without context in the medical domain. Similarly, no significant difference was observed between this same few-shot setting and the zero-shot no-context configuration of **Gemma-3**. These results reinforce the earlier finding (RQ1) that while prompting strategies can affect performance, top-tier models already perform competitively even in zero-shot conditions.

In summary, the statistical analyses validate the robustness of the evaluation framework in detecting consistent underperformance (e.g., Phi-3 and Qwen-3), but also reveal that performance difference among the strongest models tends to be indistinguishable. Furthermore, the lack of a significant advantage for few-shot prompting underlines the efficiency of the models in this task.

6. Conclusion and Future Work

We presented a comprehensive evaluation of definition generation (DG) in Portuguese for two specialized domains, medicine and geology, by leveraging domain-specific glossaries, contextual corpora, and a structured prompting strategy. Our experiments examined the behavior of several open-source multilingual language models across zero-shot and few-shot scenarios.

The results indicate that most of the multilingual models evaluated perform well in this task, with relatively small performance differences among the top models. Statistical analyses confirmed that these differences are not consistently significant, suggesting that multiple models are comparably effective for Portuguese DG, regardless of their size. Furthermore, we found no consistent advantage of few-shot prompting over zero-shot configurations among the strongest models, highlighting their robustness even with minimal task-specific guidance.

Notably, weaker models such as Phi-3 and Qwen-3 underperformed consistently, which reinforces the importance of multilingual pretraining quality and stability rather than model scale alone. In contrast, GAIA, despite its smaller size, performed competitively, likely benefiting from its continued pretraining in Portuguese.

Future work will focus on expanding the evaluation to additional domains, incorporating more diverse and structured lexical resources, and refining prompt design to further explore generalization capabilities. We also plan to investigate model-specific behaviors through ablation studies and integrate human evaluation to complement automatic metrics and better assess the quality of generated definitions.

7. Bibliographical References

2018. *Health Sciences Descriptors: DeCS*, 2018 edition. BIREME / PAHO / WHO, São Paulo.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2024. Sabi\'a-3 technical report. *arXiv preprint arXiv:2410.12049*.

Thales Sales Almeida, Hugo Abonizio, Rodrigo Nogueira, and Ramon Pires. 2024. Sabi\'a-2: A new generation of portuguese large language models. *arXiv preprint arXiv:2403.09887*.

Mohannad AlMousa, Rachid Benlamri, and Richard Khoury. 2022. A novel word sense disambiguation approach using wordnet knowledge graph. *Computer Speech & Language*, 74:101337.

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

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.

Ashish Chouhan and Michael Gertz. 2024. LexDrafter: Terminology drafting for legislative documents using retrieval augmented generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10448–10458, Torino, Italia. ELRA and ICCL.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Anna Beatriz Dimas Furtado, Tharindu Ranasinghe, Frédéric Blain, and Ruslan Mitkov. 2024. Dore: A dataset for portuguese definition generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5315–5322.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271.

Gabriel Lino Garcia, Pedro Henrique Paiola, Luis Henrique Morelli, Giovani Candido, Arnaldo Cândido Júnior, Danilo Samuel Jodas, Luis CS Afonso, Ivan Rizzo Guilherme, Bruno Elias Penteado, and João Paulo Papa. 2024. Introducing bode: A fine-tuned large language model for portuguese prompt-based task. *CoRR*.

Salvador Garcia and Francisco Herrera. 2008. An extension on" statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of machine learning research*, 9(12).

Mario Giulianelli, Iris Luden, Raquel Fernández, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148.

Diogo da Silva Magalhães Gomes, Fábio Corrêa Cordeiro, Bernardo Scapini Consoli, Nikolas Lacerda 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.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3467–3476.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jiahuan Li, Yu Bao, Shujian Huang, Xinyu Dai, and Jiajun Chen. 2020. Explicit semantic decomposition for definition generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 708–717.

Alcides Lopes, Joel Carbonera, Fabricio Rodrigues, Luan Garcia, and Mara Abel. 2024. How to classify domain entities into top-level ontology concepts using large language models. *Applied Ontology*, (Preprint):1–29.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, Francesco Cecconi, et al. 2021. Ten years of babelnet: A survey. In *IJCAI*, pages 4559–4567. International Joint Conferences on Artificial Intelligence Organization.

Roberto Navigli and Paola Velardi. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE transactions on pattern analysis and machine intelligence*, 27(7):1075–1086.

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard english words and phrases. *IJCNLP 2017*, page 413.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Rafael O Nunes, João E Soares, Henrique DP dos Santos, and Renata Vieira. 2018. Meshxnotes: web-system for clinical notes. In *International Workshop on Artificial Intelligence in Health*, pages 5–12. Springer.

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.

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.

Lucas Pavanelli, Yohan Bonescki Gumiel, Thiago Ferreira, Adriana Pagano, and Eduardo Laber. 2023. Bete: A brazilian portuguese dataset for named entity recognition and relation extraction in the diabetes healthcare domain. In *Brazilian Conference on Intelligent Systems*, pages 256–267. Springer.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*, pages 226–240. Springer.

David Riaño, Mor Peleg, and Annette Ten Teije. 2019. Ten years of knowledge representation for health care (2009–2018): Topics, trends, and challenges. *Artificial intelligence in medicine*, 100:101713.

Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. 2023. How to do human evaluation: A brief introduction to user studies in nlp. *Natural Language Engineering*, 29(5):1199–1222.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Bowen Zhang and Harold Soh. 2024. Extract, define, canonicalize: An llm-based framework for knowledge graph construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9820–9836.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.