Evaluating Generative Large Language Models for Portuguese Scientific Information Extraction

Tomás Pinto, Catarina Silva, Hugo Gonçalo Oliveira
University of Coimbra, CISUC/LASI, DEI, Coimbra, Portugal
{tomaspinto, catarina, hroliv}@dei.uc.pt

Abstract

Scientific Information Extraction (IE), which identifies entities and their relations from scientific texts, is essential for building Scientific Knowledge Graphs (SciKGs) that encode structured knowledge and enable applications such as semantic search, question answering, and literature reasoning. Large Language Models (LLMs) have shown strong capabilities in processing unstructured text, yet most advances focus on English, with limited exploration for less-resourced languages like Portuguese. The reliability of generative LLMs, including Portuguese-targeted models like the sovereign AMALIA, for structured extraction of scientific knowledge from literature text remains underexplored. We evaluate low- to mid-scale generative LLMs (8–12B parameters) on scientific Named Entity Recognition (NER) and Relation Extraction (RE), using a Portuguese-translated dataset of computer science article abstracts. Overall, our results show moderate performance and indicate that the adaptation strategy has a greater impact than model choice: prompting yields unstable performance with poor RE scores, while fine-tuning consistently improves both NER and RE and reduces cross-model variability. These findings suggest that, at this scale, prompting alone is insufficient for SciKG construction and underscore the need for supervised adaptation. We provide a detailed error analysis and outline directions for advancing Portuguese scientific IE.

Keywords: Scientific Information Extraction, Large Language Models, Portuguese Language, Knowledge Graphs

1. Introduction

The exponential growth of scientific literature has made it increasingly difficult for researchers to comprehensively access, integrate, and reason over existing knowledge (Hanson et al., 2024). This challenge is exacerbated by the predominantly unstructured nature of scientific texts, where core contributions, experimental results, and relationships between concepts are expressed in free-form language rather than in machine-interpretable formats. As a result, transforming unstructured scientific documents into structured representations is a fundamental prerequisite for scalable knowledge access, integration, and discovery (Asai et al., 2026).

Automated scientific information extraction (IE) plays a central role in addressing this problem (Hong et al., 2021), as it enables the identification of entities, relations, and higher-level semantic structures that can support downstream applications such as knowledge graph construction, literature exploration, and question answering (Zhong et al., 2023).

Large language models (LLMs) have recently demonstrated strong performance on a variety of Natural Language Processing (NLP) tasks (Yang et al., 2024), including core IE subtasks such as named entity recognition (NER), relation extraction (RE), and joint entity–relation extraction, showing the ability to compete with traditional supervised approaches (Xu et al., 2024).

However, most existing studies focus on high-resource languages, particularly English, and relatively little is known about the behaviour of LLM-based scientific IE systems in multilingual or low-resource settings (Zhao et al., 2024). This gap is particularly evident for scientific texts in Portuguese, where domain-specific terminology, limited annotated resources, and linguistic variation pose additional challenges, despite the existence of relevant publications in repositories such as RCAAP[1] and SciELO[2].

In parallel, there is growing interest in using LLMs as scalable engines for constructing scientific knowledge graphs (SciKGs) (Yang et al., 2025b), which provide structured, interpretable, and interlinked representations of scientific knowledge (Ding et al., 2025). In such settings, the quality and characteristics of extracted entities and relations directly shape the resulting graph structure and constrain which information can be retrieved and reasoned over in downstream tasks, including graph-grounded retrieval-augmented generation (GraphRAG) (Zhang et al., 2025a; Peng et al., 2025) and question answering (Ma et al., 2025). Understanding the trade-offs of different extraction strategies is therefore not only an IE problem, but also a foundational concern for reliable scientific knowledge infrastructures.

In this work, we present an empirical study of LLM-based scientific IE for European Portuguese,

[1] https://www.rcaap.pt/
[2] https://www.scielo.org/

Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026, pages 155–167
12 May 2026.

focusing on generative models around the 10B parameter scale. This includes AMALIA[3] (Simplício et al., 2026), a sovereign PT-PT LLM, alongside other comparably-sized open models to enable controlled benchmarking while avoiding the extreme computational and monetary costs associated with the largest models. We evaluate multiple extraction paradigms, including pipelined NER and RE, joint extraction via prompting, and fine-tuning-based adaptation strategies. Experiments are conducted on a European Portuguese version of the SCIERC dataset (Luan et al., 2018), which we translated and curated as part of this work to establish a benchmark for controlled evaluation of scientific IE in Portuguese.

Beyond reporting extraction performance, we analyse how different modelling and adaptation choices affect properties that are critical for downstream KG construction, such as modularity, error isolation, and schema compatibility. Based on these observations, we discuss the implications for building reliable SciKGs and outline future research directions towards end-to-end LLM–SciKG pipelines that support structured reasoning and scientific discovery, particularly in multilingual and low-resource contexts.

2. Background and Related Work

SciKGs structure entities and semantic relations from scholarly literature into explicit relational graph representations, capturing dependencies among concepts, methods, tasks, datasets, metrics, and findings (Auer et al., 2018; Ding et al., 2025). This structured modelling enables advanced applications such as semantic search, literature exploration, and question-answering systems (Peng et al., 2023; Auer et al., 2023). The quality of SciKGs is inherently dependent on the performance of upstream IE. Scientific IE aims to identify entities and semantic relations within research literature, enabling structured access to scientific knowledge (Verma et al., 2023). NER identifies domain concepts, while RE links them into structured triples. Early approaches relied on feature-based methods and later neural sequence labelling architectures for NER, combined with supervised classifiers for RE (Zhao et al., 2024). These components were typically organised in pipelined systems, where entity detection precedes relation prediction (Zhong and Chen, 2021).

While pipelined approaches offer modularity and interpretability, they are susceptible to error propagation where incorrect entity spans or types can directly degrade relation prediction performance (Yan et al., 2022). To mitigate this issue, joint extraction models were proposed to simultaneously predict entities and relations within a unified architecture (Santosh et al., 2021).

Benchmark datasets such as SciERC (Luan et al., 2018), SemEval-2018 Task 7 (Gábor et al., 2018), and later SCIER (Zhang et al., 2024) have supported progress in scientific IE by providing annotated entities and relations grounded in scientific discourse. These datasets are designed for scientific text in relatively general research domains, with entity types such as *Task*, *Method*, and *Material*, and relation types such as *used-for*, *part-of*, and *feature-of*. Rather than focusing on a single specialised domain like biomedicine, they aim to capture cross-domain conceptual structures common across many areas of scientific writing.

Recent work increasingly integrates SciKGs with LLMs (Zhu et al., 2025; Yang et al., 2025c). Graph-based retrieval and GraphRAG use structured context to ground generation and reduce hallucinations, while LLMs are also used to construct and update graphs (Yang et al., 2025b; Peng et al., 2025). This bidirectional synergy depends critically on robust extraction, as graph quality determines the effectiveness of grounding and reasoning.

LLMs have reshaped IE by replacing task-specific architectures with generative, prompt-based formulations (Xu et al., 2024). Through instruction prompting and in-context learning, entities and relations can be produced directly as structured outputs, often in a single pass. However, generative prediction introduces new challenges. Models must internalise schema constraints, produce well-formed outputs, and remain consistent despite decoding variability, while exact-match evaluation becomes sensitive to boundary and formatting differences (Zhang et al., 2025b). Performance is highly dependent on the adaptation strategy. Zero-shot prompting requires no annotated data but can struggle in specialised domains due to unfamiliar terminology and implicit schemas (Brown et al., 2020). Few-shot prompting can improve schema alignment and output consistency by providing task-specific demonstrations, but remains sensitive to example selection and context length limits. Fine-tuning, whether full-parameter or parameter-efficient, typically yields greater stability and reduced formatting errors (Dagdelen et al., 2024). Nevertheless, it requires annotated data, hyperparameter tuning, and additional computational resources, creating a trade-off between performance, cost, and scalability.

NLP in most languages, particularly in specialised domains such as scientific text processing, remains comparatively underexplored when compared to English. For instance, Portuguese NLP has advanced in areas such as general Open IE (Silva et al., 2024; Cabral et al., 2024), narrative

[3] https://amaliallm.pt/

extraction (Nunes et al., 2024), clinical/biomedical IE (Sousa et al., 2023), but these lack the domain-agnostic, paper-centric scientific focus of SCIERC-style benchmarks.

Multilingual LLMs are often trained on data heavily skewed toward English, which can lead to uneven performance in Portuguese tasks (Xu et al., 2025). In response, there has been increasing interest in sovereign and regionally developed LLMs that prioritize linguistic fidelity, cultural representation, and local data governance (Bondarenko et al., 2025). A prominent example in Portugal is AMALIA, a government-backed sovereign LLM designed specifically for the European Portuguese context.

Building on these developments, systematically comparing models specifically targeted to Portuguese with similarly sized multilingual counterparts, particularly in scientific IE settings, represents an open field of possibilities. Such comparisons could clarify the benefits of language-specific modeling choices and inform the design of more reliable extraction methods. This work contributes to addressing this gap by evaluating mid-scale instruction-tuned LLMs, including the sovereign AMALIA model, under different extraction paradigms and adaptation strategies. We position scientific IE as a foundational component of the SciKG construction pipeline, and discuss how modelling choices at the extraction stage influence the reliability of downstream knowledge graphs.

3. Experimental Setup

This section describes the dataset, models, extraction paradigms, adaptation strategies, and evaluation protocol used in our experiments.

3.1. Dataset

We translate the SCIERC dataset (Luan et al., 2018) into Portuguese and conduct our experiments on this resulting resource. SCIERC is a widely used benchmark for scientific IE built from abstracts of computer science research papers. Each instance in SCIERC consists of a single abstract annotated with entity mentions and semantic relations between them, reflecting core elements of scientific discourse. A simple example of an annotated entry from our translated corpus is provided in Table 1. The dataset comprises a total of 500 abstracts, split into training (70%), development (10%), and test (20%) sets.

SCIERC defines a fixed schema of scientific entity and relation types, detailed in Table 2. The entity types cover both high-level research concepts and more general scientific terminology while relations capture common semantic connections in scientific writing. This combination of entity and relation types makes SCIERC particularly suitable for studying IE as a foundation for SciKG construction.

Text
As fontes de dados de treino adequadas para modelação da linguagem da fala conversacional são limitadas. Neste artigo, mostramos como os dados de treino podem ser complementados com texto da web filtrado de modo a corresponder ao estilo e/ou ao tópico da tarefa de reconhecimento alvo, e também que é possível obter maiores ganhos de desempenho a partir desses dados através da utilização de interpolação de N-gramas dependente da classe.
Entities
{"text": "modelação da linguagem", "type": "Method"}, {"text": "fala conversacional", "type": "Material"}, {"text": "tarefa de reconhecimento", "type": "Task"}, {"text": "interpolação de N-gramas dependente da classe", "type": "Method"}
Relations
{"head": "fala conversacional", "tail": "modelação da linguagem", "relation": "Usado-para"}, {"head": "interpolação de N-gramas dependente da classe", "tail": "tarefa de reconhecimento", "relation": "Usado-para"}

Table 1: Example instance from our Portuguese SCIERC translation.

Type	**Description**
Entity Types	
Task	Applications, problems to solve, systems to construct.
Method	Techniques, models, tools, components of a system, frameworks.
Evaluation Metric	Criteria or measures used to assess quality and performance.
Material	Data, datasets, resources, corpus, knowledge base.
Other Sci Terms	Scientific concepts that do not fall into any of the above classes.
Generic	Non-specific references to other entities.
Relation Types	
Used-for	A method/material applied to a task or method.
Feature-of	Relates a property, attribute, or quality to the subject that possesses it.
Hyponym-of	An entity is a more specific type of another, more general entity.
Part-of	An entity constitutes an integral part of a larger whole.
Compare	Comparison or contrast between two entities.
Conjunction	Symmetric relationship between two entities with a similar function or role.
Evaluate-for	One entity is used to evaluate another entity.

Table 2: SCIERC Entity and Relation Types

To enable controlled evaluation in European Portuguese (PT-PT), we translated the full SCIERC dataset into Portuguese using GPT-5.1. Due to structural differences between English and Portuguese, such as word order and morphology, automatic translation can introduce misalignments between annotated entity spans, relations, and the translated text.

To mitigate this issue, we first employed an automatic consistency-checking script to identify candidate misalignments. Specifically, all instances were processed to flag cases where annotated entities, both in the entity lists and as head and tail

Model	Parameters	Language Focus
AMALIA	9B	Portuguese (PT-PT)
Gervasio-8b-ptpt-decoder	8B	Portuguese (PT-PT)
Boto-9B-it	9B	Portuguese (PT-BR)
EuroLLM-9B-Instruct-2512	9B	Multilingual
Llama-3.1-8B-Instruct	8B	Multilingual
Qwen3-8B	8B	Multilingual
Gemma-3-12b-it	12B	Multilingual

Table 3: Generative decoder-only LLMs used in the IE experiments (FP16 variants).

elements in relation triples, did not explicitly appear in the translated text. This process allowed us to focus on likely error instances where entity spans no longer matched the translated content (e.g., due to reordering, inflection, or lexical variation). The flagged subset corresponded to approximately half of the dataset entries.

These flagged instances were then manually reviewed by a reviewer with a technical background in NLP and familiarity with scientific text. The review focused on restoring alignment between annotations and text. When misalignments were identified, either entity boundaries or the translated text were minimally adjusted to ensure consistency, preserving the intended meaning. Relation annotations were not modified and the original annotation schema, including entity and relation types, was strictly preserved.

Despite this review process, it is possible that minor residual inconsistencies persist in the translated version. Such potential noise is an inherent limitation of cross-lingual dataset adaptation and is taken into account when interpreting the experimental results.

This translated version of SCIERC constitutes a valuable resource for studying scientific IE in Portuguese and for assessing the capabilities of Portuguese-focused and multilingual LLMs on structured scientific processing. Although it is based on translated rather than native Portuguese text, it provides a controlled benchmark for this setting. We make the dataset and associated scripts publicly available on GitHub[4] to facilitate further research in this area.

3.2. Models

Our experiments focus on low to mid-scale instruction-tuned LLMs in the 8–12B parameter range. This size bracket reflects a practical balance between computational feasibility and model capacity, while allowing controlled comparison without large scale-induced performance disparities. Table 3 presents the models considered in this study, including their parameter sizes and language focus.

The evaluated models include:

- AMALIA[5] (9B) (Simplício et al., 2026), a sovereign European Portuguese LLM designed to prioritise PT-PT linguistic fidelity, preserve Portugal's cultural representation, and ensure data governance within the national research framework. It is derived from EuroLLM-9B by incorporating new sources of European Portuguese data during the annealing phase of pretraining.
- Gervásio (8B) (Santos et al., 2024), a European Portuguese model built on LLaMA 3.1 8B Instruct, with additional Portuguese-focused adaptation (mostly European).
- Boto[6] (9B), a Portuguese chat model obtained through fine-tuning Gemma2-9B-it, primarily optimised for Brazilian Portuguese. As of February 2026, it ranks highest among Portuguese-focused chat models in the 8B range on the Open Portuguese LLM Leaderboard[7] (Hugging Face).
- EuroLLM (9B) (Ramos et al., 2026), the multilingual foundation model underlying AMALIA, developed to support all European Union languages, including European Portuguese.
- LLaMA3.1 (8B) (Grattafiori et al., 2024), a multilingual instruction-tuned model supporting Portuguese, included as a strong general-purpose baseline without explicit PT specialisation.
- Qwen3 (8B) (Yang et al., 2025a), a multilingual instruction-tuned model that, as of February 2026, ranks highest overall among 8B chat models on the HuggingFace Portuguese leaderboard.
- Gemma3 (12B) (Team et al., 2025), a multilingual 12B instruction-tuned model, included to examine whether moderate scaling beyond the 8–9B range improves robustness in structured scientific extraction.

All models are evaluated under the same experimental conditions and deployed using their FP16 (half-precision) variants to ensure consistent memory usage and computational efficiency across experiments. We opt for instruction-tuned to better follow the task instructions presented. Unless otherwise stated, identical prompting templates, decoding parameters, and fine-tuning procedures are applied to all models.

[4] https://github.com/TomasCCPinto/NSLP26_SciIE_PT

[5] https://amaliallm.pt/

[6] https://huggingface.co/lucianosb/boto-9B-it

[7] https://huggingface.co/spaces/eduagarcia/open_pt_llm_leaderboard

3.3. Extraction Strategies

We explore two complementary strategies for scientific IE: a pipelined formulation that separates entity and relation extraction, and a joint formulation that performs end-to-end extraction in a single model invocation.

In the pipelined approach, extraction is decomposed into two stages: NER, which identifies and classifies entity mentions according to the SCIERC schema, followed by RE, which predicts relations between the extracted entity pairs. This formulation should provide great modularity, allowing independent analysis and refinement of each subtask and clearer error isolation. However, errors in NER propagate directly to RE, potentially limiting overall robustness.

In the joint extraction approach, entity recognition and relation prediction are performed simultaneously through a single prompt that generates structured subject–predicate–object triples with explicit entity spans and types. This end-to-end formulation enables the model to leverage global contextual dependencies and directly produce graph-ready outputs. While joint extraction may reduce cascading errors, it sacrifices modularity and makes error localisation and schema enforcement more difficult.

Overall, the comparison reflects core trade-offs for SciKG construction, including modularity versus integration, error isolation versus joint reasoning, and explicit schema control versus direct triple generation.

3.4. Adaptation Methods

We evaluate two adaptation regimes: prompt-based extraction and supervised fine-tuning, applied consistently across models and extraction strategies.

For prompt-based extraction, we consider both zero-shot and few-shot settings. In the zero-shot configuration, used for both pipelined and joint extraction, prompts explicitly specify the IE task, the expected JSON output format, and the complete set of entity and relation types defined in the SCIERC schema. For each entity and relation type, a short natural language definition is provided in the prompt to guide the model's interpretation of the schema and reduce ambiguity in scientific terminology. This setup enables schema-aware extraction without relying on task-specific examples.

In addition, for the joint extraction strategy, we evaluate a few-shot prompting setup using three examples (3-shot) sampled from the training split of the dataset. These examples demonstrate the target structured output for complete entity–relation triples (plus entity types), reinforcing correct schema usage. The prompt templates used across the experiments are shown in Section A.1 of the appendix.

For fine-tuning, we adapt the models using QLoRA (Dettmers et al., 2023), enabling parameter-efficient supervised training by freezing the 4-bit quantized base model weights and training lightweight LoRA adapters. Fine-tuning is performed exclusively for the joint extraction formulation, where the model is trained to generate structured outputs representing entities, their types, and relations in a single step. Training instances follow the same input–output structure used during joint extraction prompting, ensuring consistency between training and inference. We fine-tune the models for 5 epochs using a learning rate of 2e-4, a LoRA rank of 8, and a batch size of 1, using BF16 precision throughout.

3.5. Evaluation Metrics

We evaluate the performance of our methods using precision, recall, and F1-score, leveraging the gold-standard entity and relation annotations provided by the dataset. Metrics are reported using macro averaging, with NER scores averaged per entity type and RE scores averaged per relation type, allowing for fine-grained analysis across different categories of scientific information.

Our primary evaluation criterion is exact match, where predicted entity spans, entity types, and relations must exactly match the gold annotations. This strict evaluation protocol follows the standard practice adopted in prior work on SCIERC, enabling direct comparability with existing results (Eberts and Ulges, 2020; Zhong and Chen, 2021). Moreover, exact matching ensures a high level of precision and directly reflects the requirements of downstream applications, where deviations in extraction can propagate errors and negatively impact subsequent processing or analysis (Dagdelen et al., 2024).

4. Results and Analysis

This section presents the experimental results across models, extraction paradigms, and adaptation strategies. We analyse performance and examine how modelling and adaptation choices influence both NER and RE. In addition to quantitative results, we provide an error analysis to better identify recurring failure patterns and their implications for SciKG construction.

4.1. Overall Performance

Table 4 reports Macro-F1 scores for both NER and RE across models, extraction strategies, and adaptation settings.

Model	Pipeline		Joint 0-Shot		Joint 3-Shot		Fine-Tune	
	NER	RE	NER	RE	NER	RE	NER	RE
AMALIA	0.11	0.01	0.11	0.00	0.21	0.04	**0.44**	0.16
Llama-3.1	0.17	0.01	0.13	0.02	0.18	0.02	0.40	0.16
Qwen3-8b	0.25	0.06	0.21	0.03	0.21	0.04	0.42	**0.23**
Gemma-3-12b-it	**0.31**	**0.07**	**0.26**	**0.06**	**0.31**	**0.09**	0.39	0.14
Boto-9b-it	0.13	0.01	0.14	0.03	0.20	0.02	0.40	0.14
Gervásio 8B	0.16	0.01	0.11	0.02	0.18	0.02	0.37	0.13
EuroLLM 9b	0.11	0.00	0.10	0.01	0.13	0.02	0.42	0.18

Table 4: Comparison of model performance across NER and RE Tasks.

Overall, performance remains relatively low for both tasks. Unsurprisingly, RE is consistently more challenging than NER, regardless of model or strategy. This reflects the intrinsic difficulty of structured scientific IE under an exact-match evaluation regime, where our methods exhibited challenges in adapting effectively to the task.

It is noticeable that the adaptation strategy has a substantially larger impact than the model choice. Across models, fine-tuning leads to marked improvements in both tasks. NER performance increases from roughly 0.10–0.31 in zero-shot settings to approximately 0.37–0.44 after fine-tuning. RE exhibits an even stronger relative effect: while via prompting the performance is often close to zero, fine-tuning raises it to the 0.13–0.23 range. This consistent behaviour across models indicates that structured scientific extraction in Portuguese cannot be reliably achieved through prompting alone at this parameter scale.

Differences between models are relatively moderate. After fine-tuning, most models converge to similar levels of NER performance, and RE scores vary within a relatively narrow band. Notably the AMALIA model achieves competitive performance and attains the highest NER score among all evaluated models, suggesting that supervised task alignment may play a more decisive role than the Portuguese pretraining origin, in this specific setting. In contrast, under prompting-based evaluation, the Gemma3 model consistently delivers the strongest results. Interestingly, it also operates at a slightly larger parameter scale than the other models but given differences in architecture, pretraining data, and instruction tuning, performance gains cannot be attributed solely to model size. Qwen3 also demonstrates stable and competitive performance across all settings, which aligns with its strong positioning on the Portuguese leaderboard of Hugging Face.

4.2. Impact of Extraction Strategy and Adaptation

The pipeline formulation isolates entity detection, potentially improving stability by removing the need to jointly generate relational structure (Diaz-Garcia and Lopez, 2025; Zhong and Chen, 2021). In contrast, joint approaches may benefit from shared representations, where relational context helps refine entity predictions (Yan et al., 2022; Santosh et al., 2021). In practice, some models show slight advantages of the pipeline over joint zero-shot (e.g., Gemma reaches 0.31 NER in the pipeline setup versus 0.26 in joint zero-shot, and Qwen achieves 0.25 versus 0.21), but results are comparable to, or slightly lower than, joint 3-shot performance. The reduction in task complexity from isolating entity detection is not clearly reflected in the results.

For RE, joint extraction is often motivated by its ability to avoid error propagation from a separate NER module (Yan et al., 2022). However, our results do not allow for strong conclusions in this regard, as RE performance under prompting remains extremely low in both extraction settings.

Despite noise potentially introduced by the additional context (examples), few-shot prompting was expected to improve instruction adherence and task alignment. In our experiments, it leads to improvements over zero-shot mainly for NER. For instance, AMALIA increases from 0.11 to 0.21 macro-F1, and Boto from 0.13 to 0.20. Other models exhibit negligible change, and RE performance remains consistently low. This indicates that few-shot examples partially support entity recognition but are insufficient for robust RE.

Fine-tuning produces a qualitatively different behaviour. As previously mentioned, both NER and RE improve substantially across all models and, crucially, performance variance narrows. The relative gains for RE are particularly pronounced: Qwen3 increases from 0.06 (pipeline) to 0.23 after fine-tuning, and AMALIA from 0.01 to 0.16. These results indicate that supervised adaptation is not only beneficial but especially critical for learning coordinated entity–relation structures at this model scale.

Across English SciERC evaluations conducted under the same settings as ours, prior work shows that generative LLMs under prompting achieve relatively modest results, roughly 35–42% F1 for NER and about 21% for RE with DeepSeek-V3 (Zhao et al., 2025). Fine-tuning GPT-4o improves NER to around 60% and RE to approximately 23%, although RE performance remains comparatively weak (Tran et al., 2025). In contrast, earlier supervised span-based and dependency-aware transformer encoders report substantially stronger results (Eberts and Ulges, 2020; Joshi and Rekik, 2025), typically exceeding 70% F1 on NER and reaching 50–60% F1 on RE, underscoring the continued advantage of structurally explicit architectures for scientific IE.

Overall, the findings indicate that prompting-based strategies, under the considered configu-

rations, are insufficient for reliable scientific IE in Portuguese. Supervised fine-tuning, even if limited, is not merely advantageous but functionally necessary to achieve non-trivial extraction performance and to stabilise cross-model variability. This suggests that generative mid-scale LLMs are not yet fully adequate for high-quality structured scientific extraction and that substantial room for improvement remains.

4.3. Per-type Performance Analysis

Table 5 presents per-type F1 scores for the two best overall models (Qwen3 and Gemma3) across extraction strategies and adaptation settings, allowing a more fine-grained analysis of task behaviour.

Model	Pipeline		Joint 0-Shot		Joint 3-Shot		Fine-Tune	
	Q	G	Q	G	Q	G	Q	G
Entity Types								
Task	0.31	0.38	0.24	0.37	0.23	0.36	0.40	0.40
Method	0.40	0.37	0.36	0.42	0.41	0.46	0.62	0.53
Eval Metric	0.32	0.41	0.23	0.25	0.23	0.30	0.43	0.34
Material	0.31	0.35	0.31	0.29	0.26	0.32	0.42	0.46
Other Sci Terms	0.10	0.31	0.14	0.21	0.14	0.24	0.32	0.32
Generic	0.04	0.05	0.00	0.03	0.02	0.21	0.38	0.30
Relation Types								
Used-for	0.08	0.09	0.02	0.17	0.05	0.17	0.20	0.19
Feature-of	0.00	0.02	0.00	0.02	0.00	0.02	0.06	0.00
Hyponym-of	0.17	0.14	0.00	0.12	0.04	0.12	0.27	0.15
Part-of	0.03	0.05	0.05	0.05	0.00	0.05	0.17	0.03
Conjunction	0.01	0.05	0.03	0.09	0.00	0.09	0.28	0.23
Evaluate-for	0.00	0.00	0.01	0.08	0.07	0.08	0.24	0.19
Compare	0.14	0.11	0.09	0.11	0.12	0.09	0.36	0.20

Table 5: Per-type F1 scores across entity and relation types for Qwen3-8B (Q) and Gemma-3-12b-it (G).

Entity performance is clearly uneven across categories. *Method* consistently emerges as the strongest performing entity type, particularly after fine-tuning, where Qwen3 reaches 0.62 and Gemma3 0.53. This suggests that method mentions are comparatively well-delimited and semantically distinctive in scientific text. *Task*, *Eval Metric*, and *Material* achieve moderate performance and show consistent results across settings. In contrast, *Other Scientific Terms* and especially *Generic* entities are substantially more challenging. In the prompting settings, *Generic* entities are nearly absent (often close to zero), and although fine-tuning improves performance, they remain among the weakest categories. This pattern likely reflects both greater heterogeneity and annotation ambiguity: broader or less well-defined categories provide fewer lexical cues, making boundary detection and type assignment more difficult.

RE exhibits stronger sparsity and imbalance effects. In the prompting configurations, most relation types remain near zero for both models. Fine-tuning leads to substantial gains, but performance remains uneven. *Compare* and *Conjunction* show the largest improvements after fine-tuning (e.g., Qwen3 reaches 0.36 and 0.28, respectively), proving to be comparatively easier to capture once the model is aligned to the task. Conversely, *Feature-of* and *Part-of* remain difficult even after fine-tuning, often yielding near-zero or very low scores. These relations encode subtler semantic dependencies and may be more sensitive to contextual interpretation.

While both models benefit from fine-tuning, Qwen3 shows stronger gains and achieves higher peak scores across most entity and relation types. Gemma3, however, demonstrates more stable behaviour in prompting settings, especially in zero-shot joint extraction. This reinforces earlier observations: Gemma3 appears slightly more robust in instruction-following scenarios, whereas Qwen3 adapts particularly well under supervised alignment.

4.4. Error Analysis

To better identify the limitations behind the observed performance, we analyse the model predictions. Table 6 presents representative examples of some of the error categories described below.

Model Response	Expected Response
(“distribution”, “used-for”, “play organizer”)	(“this distribution”, “used-for”, “play organizer”)
(“dictionary lookup phase”, “conjunction”, “application of rules”)	(“dictionary lookup”, “conjunction”, “application of rules”)

(a) Boundary variation error examples in extracted triples.

Model Response	Expected Response
(“high-dimensional space”, “material”)	(“high-dimensional space”, “other scientific terms”)
(’density estimation’, ’method’)	(’density estimation’, ’task’)
(“spoken dialogue system”, “task”)	(“spoken dialogue system”, “method”)

(b) Entity type error examples.

Model Response	Expected Response
(“reservoir models”, “feature-of”, “inner hair cell synapse”)	(“reservoir models”, “used-for”, “inner hair cell synapse”)
(“error rate”, “feature-of”, “CRF model”)	(“error rate”, “evaluate-for”, “CRF model”)

(c) Relation type error examples.

Table 6: Comparative analysis of some of the most common errors encountered during the experiments.

The most common error in NER is the omission of gold entities. A large portion of annotated entities are simply not generated. Since RE depends

on the correct identification of both entities in a pair, these omissions propagate downstream, significantly increasing the difficulty of producing valid relational triples. As a consequence, most triple annotations are missed not only due to incorrect relation classification, but mostly because one or both entities are absent from the prediction.

Another relevant issue concerns deviations from the expected output format. In several runs, models did not strictly follow the prompt instructions, generating outputs that did not conform to the required JSON structure and therefore being penalised during evaluation. This problem occurs mainly in the pipeline and zero-shot joint formulations. For instance, in the pipeline setting, AMALIA presents 13% of entries not following the expected format, while Boto reaches 39%. The use of 3-shot prompting and the fine-tuning substantially improves structural compliance, bringing adherence to near 100%.

Boundary and lexical variation errors also contribute to performance degradation. Predicted spans are sometimes slightly longer or shorter than the gold annotation, or contain minimal lexical differences. Under exact match evaluation, even small variations, despite preserving meaning, are counted as incorrect, negatively affecting extraction scores.

Entity type confusions are another recurring NER error. In these cases, the span is correctly identified but assigned the wrong type. A frequent confusion occurs between *Method* and *Task*, which often appear in similar contexts in scientific text. Additionally, spans belonging to *Generic* and sometimes *Other scientific terms* are forced into one of the other most specific categories, suggesting imperfect schema internalisation.

Finally, relation prediction errors occur, although in smaller numbers compared to entity omissions. Some relations show systematic confusion, particularly *Feature-of* versus *Used-for*, and *Compare* versus *Conjunction*. This suggests that there is high semantic overlap between these categories and subtle syntactic cues that distinguish them in scientific discourse, making it challenging for the models to disambiguate.

5. Implications for Scientific Knowledge Graph Construction

Scientific IE plays a central role in the automatic construction of SciKGs. The reliability of downstream applications depends directly on the precision and structural consistency of the extracted entities and relations. In this context, the results obtained in this work fall short of the level of robustness required for robust SciKG construction.

The prompting-based configurations proved insufficient for reliable graph generation. Even after fine-tuning, while improvements were substantial relative to prompting settings, RE performance remains moderate, and entity-level errors persist. Given that KGs propagate extraction errors directly into structured representations, these limitations would likely translate into incomplete or noisy graphs, thereby compromising downstream usability. Rather than representing a dead end, these findings highlight several research directions for improving scientific IE in Portuguese.

One evident limitation of the present study is the moderate parameter scale of the evaluated models (8–12B). While this choice reduced computational barriers for inference and fine-tuning, it likely constrained representational capacity and reasoning depth. Exploring larger and more robust generative models may yield stronger and more stable outputs, particularly for RE.

In parallel, traditional supervised IE approaches, such as span-based models, and domain-adapted encoder-based transformers have historically demonstrated competitive performance in scientific domains and may offer higher precision under strict evaluation regimes (Eberts and Ulges, 2020; Joshi and Rekik, 2025). A hybrid comparison between generative and encoder-based paradigms would therefore be a valuable extension.

Data availability is another critical factor. Expanding annotated resources or constructing additional domain-aligned Portuguese corpora can enable more extensive supervised training, potentially improving both entity boundary detection and relation classification. Given the relatively specialised and technical nature of Portuguese scientific text, increased annotation scale may be particularly impactful. Whenever possible, such data should originate from native Portuguese scientific sources rather than translated material, as translation artefacts can introduce inconsistencies. Resources such as CorEGe-PT (Kuhn et al., 2026), a Portuguese corpus of academic texts compiled from the Estudo Geral repository, exemplify the type of native scientific material

Methodologically, further refinements in our methods are also possible. The error analysis revealed weaknesses that prompt engineering could try to address, for example through more specialized schema reinforcement or intermediate reasoning verification steps. On the fine-tuning side, systematic experimentation with hyperparameter configurations, instruction formatting, or multi-task objectives may yield additional gains.

Finally, evaluation methodology deserves consideration. The strict exact-match criterion, while standard in IE benchmarking, is highly penalising in cases of minimal lexical variation or semantically

equivalent span differences. Alternative or complementary evaluation schemes, such as string-similarity-based metrics, could provide a more nuanced assessment of extraction quality, particularly when the intended application tolerates minor lexical deviations.

Overall, the present results indicate that reliable SciKG construction in Portuguese remains an open challenge. However, the findings provide a diagnostic foundation that clarifies current bottlenecks and outlines concrete pathways for advancing IE in this domain.

6. Conclusion

Automatic extraction from rapidly expanding scientific literature is crucial for building SciKGs that support semantic search, question answering, and large-scale reasoning. Despite recent progress in LLMs with great demonstrations on text processing, their reliability for structured scientific IE in Portuguese remains underexplored. This work provides a systematic evaluation of low- to mid-scale generative LLMs for Portuguese scientific NER and RE, comparing pipeline and joint formulations under zero-shot, few-shot, and fine-tuned settings using a curated European Portuguese translation of the SCIERC dataset.

Three main findings emerge. First, the adaptation strategy has a substantially greater impact than model choice at this scale: prompting approaches yield unstable performance, especially for RE, whereas fine-tuning consistently improves both tasks. Second, RE remains markedly more challenging than entity recognition due to the compositional and boundary-sensitive nature of structured triple prediction. Third, model differences are pronounced under prompting but largely diminish after supervised adaptation, highlighting the importance of task alignment over architectural variation.

From a SciKG construction perspective, our results indicate that prompting alone is insufficient for reliable graph generation in scientific domains. Even supervised configurations, while promising, would require further optimisation before being suitable for downstream reasoning applications. Our error analysis, revealing entity omissions, boundary mismatches, type confusions, and relation misclassifications, identifies key bottlenecks and motivates future work on larger models, expanded annotated resources, alternative encoder-based approaches, and refined adaptation strategies. Overall, this study establishes a grounded empirical baseline for Portuguese scientific IE and outlines a concrete path for further exploration toward SciKG construction in Portuguese settings.

7. Acknowledgements

This work was partially supported by the AMALIA project, funded by FCT/IP in the context of measure RE-C05-i08 of the Portuguese Recovery and Resilience Program;

This work was also supported by the Portuguese Recovery and Resilience Plan (PRR), through project C645008882-00000055 – Center for Responsible AI; produced as part of the N-GenERP project with reference COMPETE2030-FEDER-02219400 (operation no. 21343) supported by the European Regional Development Fund (FEDER) through the Innovation and Digital Transition Programme (COMPETE 2030) of Portugal 2030 and the European Union; and Project LUMEN, funded by the European Union under Grant Agreement no. 101187940; and supported by FCT – Foundation for Science and Technology, I.P., under the projects UIDB/00326/2025 and UIDP/00326/2025;

8. Bibliographical References

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, et al. 2026. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, pages 1–7.

Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. 2023. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1):7240.

Sören Auer, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria Esther Vidal. 2018. Towards a knowledge graph for science. In *Proceedings of the 8th international conference on web intelligence, mining and semantics*, pages 1–6.

Mykhailo Bondarenko, Sviatoslav Lushnei, Yurii Paniv, Oleksii Molchanovsky, Mariana Romanyshyn, Yurii Filipchuk, and Artur Kiulian. 2025. Sovereign large language models: Advantages, strategy and regulations. *arXiv preprint arXiv:2503.04745*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bruno Cabral, Daniela Claro, and Marlo Souza. 2024. Exploring open information extraction for Portuguese using large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 127–136, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature communications*, 15(1):1418.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in neural information processing systems*, 36:10088–10115.

Jose A Diaz-Garcia and Julio Amador Diaz Lopez. 2025. A survey on cutting-edge relation extraction techniques based on language models. *Artificial Intelligence Review*, 58(9):287.

Keyan Ding, Zhihui Zhu, Yuqi Tang, Kehua Feng, Xiang Zhuang, Hongwei Wang, Yi Yang, Huifang Du, Zhangkai Ni, Shiqi Wang, Xiaohui Fan, Huabin Xing, Lei Bai, Qi Liu, Haofen Wang, Qiang Zhang, and Huajun Chen. 2025. Bridging data and discovery: A survey on knowledge graphs in ai for science.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020*, pages 2006–2013. IOS Press.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. 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*.

Mark A Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. 2024. The strain on scientific publishing. *Quantitative Science Studies*, 5(4):823–843.

Zhi Hong, Logan Ward, Kyle Chard, Ben Blaiszik, and Ian Foster. 2021. Challenges and advances in information extraction from scientific literature: a review. *Jom*, 73(11):3383–3400.

Devvrat Joshi and Islem Rekik. 2025. Dependency parsing-based syntactic enhancement of relation extraction in scientific texts. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24888–24897.

Tanara Zingano Kuhn, José Matos, Bruno Neves, Daniela Pereira, Elisabete Cação, Ivo Simões, Jacinto Estima, Delfim Leão, and Hugo Gonçalo Oliveira. 2026. CorEGe-PT: Compiling a Large Corpus of Academic Texts in Portuguese. In *Proceedings of 15th Language Resources and Evaluation Conference*, LREC 2026, page Accepted. ELRA.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. 2025. Large language models meet knowledge graphs for question answering: Synthesis and opportunities. *arXiv preprint arXiv:2505.20099*.

Sérgio Nunes, Alípio Mario Jorge, Evelin Amorim, Hugo Sousa, António Leal, Purificação Moura Silvano, Inês Cantante, and Ricardo Campos. 2024. Text2story lusa: A dataset for narrative analysis in european portuguese news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15773–15782.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2025. Graph retrieval-augmented generation: A survey. *ACM Transactions on Information Systems*, 44(2):1–52.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. 2023. Knowledge graphs: Opportunities and challenges. *Artificial intelligence review*, 56(11):13071–13102.

Miguel Moura Ramos, Duarte M Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M Guerreiro, Ricardo Rei, Nicolas Boizard, et al. 2026. Eurollm-22b: Technical report. *arXiv preprint arXiv:2602.05879*.

Rodrigo Santos, João Silva, Luís Gomes, João Rodrigues, and António Branco. 2024. Advancing generative ai for portuguese with open decoder gervásio pt. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Underresourced Languages@ LREC-COLING 2024*, pages 16–26.

Tokala Yaswanth Sri Sai Santosh, Prantika Chakraborty, Sudakshina Dutta, Debarshi Kumar Sanyal, and Partha Pratim Das. 2021. Joint entity and relation extraction from scientific documents: Role of linguistic information and entity types. *EEKE@ JCDL*, 21:15–19.

Gabriel Silva, Mário Rodrigues, António Teixeira, and Marlene Amorim. 2024. Advancing open information extraction for portuguese by leveraging graph structures and large language models. In *Proc. IberSPEECH 2024*, pages 61–65.

Afonso Simplício, Gonçalo Vinagre, Miguel Moura Ramos, Diogo Tavares, Rafael Ferreira, Giuseppe Attanasio, Duarte M. Alves, Inês Calvo, Inês Vieira, Rui Guerra, James Furtado, Beatriz Canaverde, Iago Paulo, Vasco Ramos, Diogo Glória-Silva, Miguel Faria, Marcos Treviso, Daniel Gomes, Pedro Gomes, David Semedo, André Martins, and João Magalhães. 2026. AMALIA technical report: A fully open source large language model for european portuguese.

Hugo Sousa, Alipio Mario Jorge, Arian Pasquali, Catarina Santos, and Mario Lopes. 2023. A biomedical entity extraction pipeline for oncology health records in portuguese. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 950–956.

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*.

Trang Tran, Trung Hoang Le, Huiping Cao, and Tran Cao Son. 2025. An LLM+ ASP workflow for joint entity-relation extraction. *arXiv preprint arXiv:2508.12611*.

Shilpa Verma, Rajesh Bhatia, Sandeep Harit, and Sanjay Batish. 2023. Scholarly knowledge graphs through structuring scholarly communication: a review. *Complex & intelligent systems*, 9(1):1059–1095.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

Zhaohui Yan, Zixia Jia, and Kewei Tu. 2022. An empirical study of pipeline vs. joint approaches to entity and relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 437–443.

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

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Rui Yang, Boming Yang, Xinjie Zhao, Fan Gao, Aosong Feng, Sixun Ouyang, Moritz Blum, Tianwei She, Yuang Jiang, Freddy Lecue, et al. 2025b. Graphusion: A rag framework for scientific knowledge graph construction with a global perspective. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2579–2588.

Zhenyao Yang, Sha Yuan, Zhou Shao, Wenfa Li, and Runzhou Liu. 2025c. A review on synergizing knowledge graphs and large language models. *Computing*, 107(6):143.

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. 2024. Scier: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13083–13100.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, et al. 2025a. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.

Zikang Zhang, Wangjie You, Tianci Wu, Xinrui Wang, Juntao Li, and Min Zhang. 2025b. A survey of generative information extraction. In *Proceedings of the 31st International conference on computational linguistics*, pages 4840–4870.

Dong Zhao, Yadong Wang, Xiang Chen, Chenxi Wang, Hongliang Dai, Chuanxing Geng, Shengzhong Zhang, Shaoyuan Li, and Sheng-Jun Huang. 2025. Reflect then learn: Active prompting for information extraction guided by introspective confusion. *arXiv preprint arXiv:2508.10036*.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, 56(11):1–39.

Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. 2023. A comprehensive survey on automatic knowledge graph construction. *ACM Computing Surveys*, 56(4):1–62.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 50–61.

Zhihui Zhu, Yuqi Tang, Qiang Zhang, and Keyan Ding. 2025. Synergizing large language models and knowledge graphs in science: A survey. In *NeurIPS 2025 AI for Science Workshop*.

A. Appendix

A.1. Prompt Templates Used

Figure 1 presents a segment containing the entity type definitions used across prompts, while Figure 2 presents the relation type definitions.

Figure 3 shows the pipeline NER prompt template, while Figure 4 corresponds to the pipeline RE prompt template. Lastly, Figure 5 shows the joint prompt template, which has the particularity of including three examples in the few-shot setting. Each individual example consists of an abstract followed by an array of JSON objects, one for each triple, each containing a subject span and type, an object span and type, and a relation type.

```
- Tarefa - Aplicações, problemas a re-
solver, sistemas a construir, fenómenos
ou tópicos em investigação.
- Método - Métodos, abordagens, técnicas,
modelos, sistemas, algoritmos, compo-
nentes ou ferramentas utilizados.
- Métrica - Métricas de avaliação, medi-
das ou entidades que exprimem a qualidade
de um sistema/método.
- Material - Dados, conjuntos de dados,
recursos, corpus, base de conhecimento.
- Outros Termos Científicos - Expressões
que são termos científicos mas que não se
enquadram nas categorias anteriores.
- Genérico - Termos gerais ou pronomes
que podem referir uma entidade mas que
não são informativos por si só, sendo
frequentemente usados como palavras de
ligação.
```

Figure 1: Entity type definitions.

```
- Usado-para: A relação entre um
Método/Ferramenta e a Tarefa/Propósito
para o qual é utilizado.
- Característica-de: Relaciona uma pro-
priedade, atributo ou qualidade com o
sujeito que a possui.
- Hipónimo-de: Uma entidade é um tipo
mais específico de outra entidade mais
geral.
- Parte-de: Uma entidade constitui uma
parte integrante de um todo maior.
- Avaliado-para: Indica que uma entidade
é usada para avaliar outra entidade.
- Comparado-com: Usado para comparar
explicitamente dois modelos, métodos ou
resultados.
- Conjunção: Relação simétrica entre
duas entidades com função ou papel semel-
hante.
```

Figure 2: Relation type definitions.

```
A tua tarefa é identificar e classificar
**termos científicos relevantes** no
texto, de acordo com o seguinte esquema
simplificado de tipos de entidade.

**TIPOS DE ENTIDADE:**
{entity_types_definitions}

**INSTRUÇÕES:**
- Identifica todas os termos científicos
explícitos no texto.
- Não uses frases completas como enti-
dade.
- Mantém o output em formato estruturado
JSON, sem explicações ou texto extra
- Se não identificares termos científicos,
devolve um array vazio.

**FORMATO DE OUTPUT:**
Retorna um array JSON de entidades, onde
cada entrada contém:
- "nome": texto da entidade
- "tipo": categoria da entidade (Tarefa,
Método, Métrica de Avaliação, Material,
Outros Termos Científicos, Genérico)

**TEXTO:**
{text_input}
```

Figure 3: Pipeline NER prompt template.

```
A tua tarefa é identificar relações semân-
ticas entre as entidades fornecidas,
baseando-te EXCLUSIVAMENTE no texto cien-
tífico apresentado.

**PREDICADOS (RELAÇÕES) PERMITIDOS:**
{relation_types_definitions}

**INSTRUÇÕES:**
- O "sujeito" e o "objeto" DEVEM ser ex-
atamente os nomes listados em "ENTIDADES
FORNECIDAS".  Não crie novas entidades.
- O "Predicado" DEVE ser um dos "PREDICA-
DOS PERMITIDOS".
- Se não houver relação clara entre duas
das entidades no texto, ignora.
- Mantém o output em formato estruturado
JSON SEM EXPLICAÇÕES OU TEXTO EXTRA.

**FORMATO DE OUTPUT:**
Retorna APENAS uma lista JSON de triplos.
Cada entrada deve conter:
- "sujeito":  termo científico ou enti-
dade principal
- "predicado":  relação entre o sujeito e
o objeto
- "objeto":  termo científico ou entidade
associada

**TEXTO**:
{text_input}
**ENTIDADES FORNECIDAS**:
{entities_input}
```

Figure 4: Pipeline RE prompt template.

```
És um assistente de IA que extrai conhec-
imento estruturado de textos científicos.
Extrai entidades e relações como triplos
e devolve-os em formato JSON.

**TIPOS DE ENTIDADE:**
{entity_types_definitions}

**TIPOS DE RELAÇÃO:**
{relation_types_definitions}

**EXEMPLOS:**
{3shot_examples}

**TEXTO:**
{text_input}

**FORMATO DE OUTPUT:**
Retorne um array JSON de triplos.  Cada
triplo deve conter:
- "sujeito":  texto da entidade
- "tipo_sujeito":  tipo do sujeito
- "relação":  tipo de relação
- "objeto":  texto da entidade
- "tipo_objeto":  tipo do objeto
```

Figure 5: Joint prompt template.