Specializing a Small Language Model for Closed-Domain Portuguese RAG using Knowledge Graph Supervision

Josue Caldas[1], Elvis de Souza[1], Patrícia Silva[2], Marco Pacheco[1]
[1]Pontifical Catholic University of Rio de Janeiro, Applied Computational Intelligence Lab.
[2]Petrobras Research and Development Center (CENPES)
josue.caldas.v@gmail.com, elvis.desouza99@gmail.com
patricia.fs@petrobras.com.br, marco@ele.puc-rio.br

Abstract

Fine-tuned small language models (SLMs) have emerged as effective alternatives for closed-domain tasks, where large language models (LLMs) often lack sufficient parametric knowledge. This study presents a methodology for adapting a small language model to a closed-domain question answering (Q&A) task. For each question, the model is trained to output an answer based on the most relevant context passage, among ten provided candidates, thus reproducing the logic of a Retrieval-Augmented Generation (RAG) framework. The fine-tuning data were derived from PetroKGraph, an existing knowledge graph built from Portuguese-language resources in the oil and gas (O&G) domain. Experimental results show that the fine-tuned model achieves a 20 percentage points accuracy improvement over the base model on closed-domain questions. It also surpasses GPT-4o and GPT-4o Mini by 12 and 25 points, respectively. Moreover, its performance on general-domain tasks remains comparable to that of the base model, indicating that the specialized model effectively learned domain specific knowledge while maintaining general reasoning capabilities.

1 Introduction

While LLMs possess substantial parametric knowledge, this internal knowledge is often insufficient for closed-domain settings, where information is highly specialized and not present in the general training corpora (Tonmoy et al., 2024). To bridge this gap, Retrieval-Augmented Generation (RAG) has emerged as a dominant architecture, enhancing LLMs by retrieving relevant documents to ground the generation process in factual, domain-specific information, thereby mitigating the risk of hallucination (Gao et al., 2023).

The challenge of insufficient parametric knowledge is particularly acute in industrial closed-domain settings such as the Oil and Gas (O&G) sector. This industry contends with vast collections of unstructured and semi-structured technical data, with estimates suggesting that 80% of new data falls into this category (Chelmis et al., 2013). A significant portion of this documentation, especially in key regions like Brazil's pre-salt exploratory frontier, is in Portuguese, a language for which specialized corpora and pre-trained models remain scarce (Gomes et al., 2021). This creates a critical need for NLP solutions that are not only domain-adapted but also language-specific.

While the prevailing trend has been towards developing ever-larger models, there is a growing imperative to develop smaller, more efficient language models. These smaller models, ranging from 1 to 10 billion trainable parameters, offer significant advantages in terms of computational efficiency, financial accessibility, and reduced environmental impact, addressing critical concerns about the costs and accessibility of massive-scale AI (Bender et al., 2021; Assis et al., 2024).

A promising avenue for creating powerful yet compact models is to leverage structured knowledge sources for supervision. In the context of the Portuguese-speaking O&G industry, a key resource is PetroKGraph, a comprehensive knowledge graph meticulously curated by domain experts (Cordeiro et al., 2024a). PetroKGraph is built upon a formal ontological foundation — GeoCore (Garcia et al., 2020) and Basic Formal Ontology (Arp et al., 2015) — and is populated with geological entities, concepts, and their semantic relationships, providing a rich, structured representation of domain knowledge.

In this work, we propose a novel methodology for specializing a Small Language Model (SLM) for closed-domain Q&A using knowledge graph supervision. Our fine-tuning process mimics a Retrieval-Augmented Generation (RAG) workflow: we train the SLM to generate answers to questions in Portuguese based on provided context passages,

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 281–290

where questions, correct answers and context passages are all derived from PetroKGraph structured data.

To evaluate our approach, we measure the model's performance on two distinct datasets: a test set of our in-domain Q&A dataset derived from PetroKGraph (in Portuguese) and a general-domain journalistic benchmark designed to evaluate RAG workflows in English (Chen et al., 2024). This dual evaluation allows us to assess both the model's specialized accuracy in its target domain and its ability to generalize across language and topic. Our goal is to demonstrate that knowledge graph supervision can produce a resource-efficient and accurate SLM for its target domain while retaining cross-domain capabilities.

2 Related Work

Ibrahim et al. (2024) present an overview of the KG-Enhanced LLMs paradigm, where knowledge graphs (KGs) are used to inform large language models (LLMs) either during pretraining or at inference time. This paradigm also includes models that are fine-tuned for specific downstream tasks. Early examples of LLMs fine-tuned with data derived from KGs include KEPLER (Wang et al., 2020) and Pretrain-KG (Zhang et al., 2020). More recent studies have focused on tasks such as knowledge completion (Yao et al., 2025) and multi-hop reasoning over KGs (Shu et al., 2025).

A range of techniques have been applied to convert the structured content of knowledge graphs (KGs) into natural language prompts. The approach of Shu et al. (2025) focuses on multi-hop link prediction. The structural conversion process extracts paths from the KG, where each path is a sequence of observed triples formed by connected entities and relations. Each path is converted into a knowledge prompt designed to elicit step-by-step Chain-of-Thought (CoT) reasoning from the LLM. The LLM produces a generative prediction, often a binary classification (e.g., "Yes" or "No"), preceded by a step-by-step rationale.[1]

Yao et al. (2025) fine tune LLMs to predict relations between entities, so that those models can help converting KG information into natural language prompts. They define three templates for prompt completion using KG-data: triple classification, relation prediction and entity prediction. For instance, for triple classification, ⟨`Steve Jobs, founded, Apple Inc.`⟩ becomes "Is this true: Steve Jobs founded Apple Inc.?". In this example, the expected tiple classification output is "Yes, it is true". Then, they fine tune models such as Llama 1 7B and ChatGLM 6B, so that those models learn how to predict KG relations between entities.

In the O&G domain, Navarro et al. (2024) transform KG information into textual resources usable for retrieval and supervised evaluation. They implement an algorithmic, domain-specific pipeline that (i) queries the RDF graph with SPARQL to retrieve ontology-selected entities and relations, (ii) maps each retrieved relation to one of 29 domain-authored text templates, and (iii) emits structured outputs as short context passages plus aligned question–answer pairs in Portuguese. The templates deterministically verbalize relations (for example, *located_in*, *crosses*, *constituted_by*) using graph labels and entity identifiers, so the generated text mirrors the original triples and the gold answer is directly grounded in the KG.

The question of which base models are most suitable for supervised fine-tuning for RAG-style tasks has also been explored. Caldas and de Souza (2025) conduct a comprehensive evaluation of 13 models (extractive and generative models; open- and closed-source models; models of varying sizes) on the Q&A task. They report that large open-source generative models (e.g., LLaMA 3 70B, Nous Hermes 2 70B) achieve performance comparable to closed-source systems (e.g., GPT-4o, GPT-4o Mini), but their computational requirements hinder practical deployment. In contrast, smaller models (e.g., Nous Hermes 2 7B, LLaMA 3 7B) offer a favorable balance among inference quality, robustness to noise, and compute cost.

Comparative studies highlight that model selection interacts closely with the choice of fine-tuning strategy. In practice, full-parameter update fine-tuning has traditionally been the standard, but its cost has become prohibitive as models scale. Foundational work such as ULMFiT (Howard and Ruder, 2018), GPT-1 (Radford et al., 2018), BERT (Devlin et al., 2019), and T5 (Raffel et al., 2020) relied on updating all weights during adaptation. More recent research has shifted toward parameter-efficient fine-tuning (PEFT), which freezes most weights and learns lightweight task-specific mod-

[1]For example, *Input:* `Node_1 has relation_1 with node_2, and node_2 has relation_2 with node_3.` *Instruction:* "Below is the detail of a knowledge graph path. Is node_1 connected with node_3? Answer the question by reasoning step-by-step. Choose from the given options: Yes. No."

ules (Houlsby et al., 2019; Lester et al., 2021).

Among PEFT methods, Low-Rank Adaptation (LoRA) (Hu et al., 2022) has emerged as a widely adopted approach, injecting low-rank updates into weight matrices to reduce memory and training costs while retaining strong downstream performance. Empirical evidence further indicates that, by constraining weight updates to a low-dimensional subspace, LoRA mitigates catastrophic forgetting and better preserves out-of-domain capabilities compared to full fine-tuning (Biderman et al., 2024).

This work focuses on fine-tuning using structured information from PetroKGraph, a domain-specific KG in Portuguese for the O&G domain. Unlike approaches such as Shu et al. (2025) and Yao et al. (2025), which use LLMs to predict KG relations in an English general-domain setting, this paper follows the approach of Navarro et al. (2024), converting structured KG information into natural language prompts. We use SPARQL queries to retrieve entities and relations and then map them using 29 domain-authored templates. For fine-tuning, we apply PEFT techniques, specifically LoRA.

3 Methodology

3.1 PetroKGraph Dataset

PetroKGraph is a resource built and curated to support the evaluation of a methodology that automatically extracts entities from technical documents to populate a knowledge graph (Cordeiro et al., 2024b). It is a knowledge graph populated with instances from oil and gas technical documents, and its construction process was based on the definition of a domain ontology, which was then enriched with terms imported from domain vocabularies and databases.

The ontology that formalizes the concepts adopted in PetroKGraph is extended from the GeoCore Ontology (Garcia et al., 2020), a core ontology that defines general terms for the Geology domain (such as *Rock* and *Gelogical Time Interval*). GeoCore itself is built upon the Basic Formal Ontology (BFO), a small top-level ontology that describes general concepts that exist across several domains. PetroKGraph ontology defines specific concepts for the oil and gas domain such as *Well* and (oil) *Field*. Concept selection and formalization was conducted with the aid of domain specialists, in Portuguese.

Once defined, PetroKGraph ontology was populated with terms representing instances and / or relevant subclasses in the domain, as well as existing relations between such terms. The most significant part of the vocabulary that originated the populated graph was derived from relational databases of oil and gas companies and tables manually compiled by geoscientists. Additional terms were collected from spreadsheets and reports from the Brazilian Agency of Petroleum, Natural Gas, and Biofuels[2].

After population, the resulting knowledge graph contains 338.445 axioms. It carries significant domain information, relating wells, the constituents of sedimentary basins (geological objects, rocks and fluids) and their properties (geological age and rock features). From PetroKGraph, it is possible to retrieve information such as "(*Well*) 1-SHEL-29-ESS *crosses* (*Lithostratigraphic Unit*) Formação Carapebus."

The training data for our model was obtained from the assertions extracted from PetroKGraph. Specifically and for practical purposes, we use a reduced version of the larger knowledge graph, referred to as MiniKGraph. Following the methodology proposed by Navarro et al. (2024) for closed-domain data extraction, the contents of MiniKGraph were systematically processed and transformed into a textual dataset of questions, answers and context passages. We executed SPARQL queries to retrieve entities and relations, then verbalized the results into natural language with 29 curated templates.

For our study, these 29 templates were grouped into four categories according to their reasoning complexity. Level 1 questions probe only the intrinsic properties of a single entity (e.g., retrieving the entity name). Level 2 questions correspond to single-hop queries that do not require aggregation (e.g., identifying the basin where field X is located). Level 3 questions involve single-hop queries with aggregation (e.g., which lithostratigraphic units are traversed by well Z). Finally, Level 4 questions represent multi-hop queries (e.g., which lithostratigraphic units are constituted by fluid W).

To decrease the number of Level 1 questions that are possible using the dataset, we applied undersampling. As a result, the dataset was reduced from 17,874 original items to 5,714 items. Table 1 presents the distribution of questions by complexity level.

For the main experiments, we adopted a con-

[2]https://www.gov.br/anp

Level	Number of items	%
1	3,578	62.63
2	1,265	22.14
3	350	6.13
4	523	9.15
Total	**5,714**	**100.00**

Table 1: Distribution of MiniKGraph dataset items by complexity levels

ventional split of the MiniKGraph data: 70% for training (≈3,500 items), 15% for validation (≈750 items), and 15% for testing. From the test portion, we selected a subset of 300 items for evaluation in order to match the assessment size used by Chen et al. (2024), whose dataset is also employed later in this work to examine the fine-tuned model's cross-domain capabilities. All splits were randomly sampled while preserving the distribution of complexity levels (see Table 2).

	Datasets		
Level	**Train**	**Validation**	**Test**
1	2,499	532	547
2	897	187	181
3	244	58	48
4	361	81	81
Total	**4,001**	**858**	**857**

Table 2: Distribution of dataset items by complexity level and split

3.2 Model Selection

Based on the evaluation by Caldas and de Souza (2025), we selected the open-source model Nous Hermes 2 (7B) for fine-tuning. This model shows accuracy levels comparable to large closed-source models such as GPT-4o and GPT-4o Mini in the task of Retrieval-Augmented Generation for general-domain English questions (Chen et al., 2024). At a noise level of 0% (noise is formally defined in Subsection 3.5), Nous Hermes 2 (7B) reaches an accuracy of 0.965, which matches GPT-4o Mini and is slightly higher than GPT-4o, which reaches 0.958. The model also remains stable under high noise conditions, achieving an accuracy of 0.742 at 80% noise, outperforming other open-source models such as Meta Llama 8B, which reaches 0.605 under the same conditions.

In this study, we compare our in-domain fine-tuned model with a set of baseline models following the selection criteria of Caldas and de Souza (2025). The models used for comparison are RoBERTa (560M) (Pietsch et al., 2019), GPT-4o, GPT-4o Mini, and the base model Nous Hermes 2 (7B) (NousResearch, 2024). This selection allows us to evaluate our model's performance against a diverse group of models, ranging from extractive models to large generative models from the GPT family.

3.3 Task

We designed the task as Q&A in a Retrieval-Augmented Generation workflow. Each training instance includes a user question, the correct answer, and a set of ten candidate context passages retrieved from MiniKGraph. The model is instructed to generate the correct answer using only the provided passages.

The ten candidate passages were selected by computing the cosine similarity between the question and all available passages in the dataset, then choosing the top ten most similar passages. This process reproduces how retrieval modules based on semantic similarity in RAG systems operate, retrieving relevant information based on a score computed between the input query and the dataset. Importantly, all models evaluated in this study receive the same pre-retrieved context passages, so observed performance differences reflect each model's reading comprehension ability rather than retrieval quality.

The training instance format contains (i) a `System` instruction that defines the task, (ii) a `User` message with the question followed by the ten context passages, and (iii) an `Assistant` message with the gold answer. Below is an illustrative example (loosely translated to English):

> **SYSTEM:** You are an intelligent and helpful assistant, responsible for providing clear and direct answers to users' questions. Your answer must be based directly on the information available as context and should be concise. If there is not enough information in the question and the context to provide an accurate answer, clearly state: "There is not enough information to answer the question."
>
> **USER:** In which basin is the Jiribatuba field located?
>
> **USER:** Context for the question:

1. The Jiribatuba field is located in the basin(s): Camamu-Almada (Almada Mar, Camamu-Almada Basin, Camamu Mar, Almada Terra, Camamu Terra).
2. The Camarao Norte field is located in the basin(s): Camamu-Almada (Almada Mar, Camamu-Almada Basin, Camamu Mar, Almada Terra, Camamu Terra).
3. The Igarapé Maripá field is located in the basin(s): Amazonas (Amazonas Basin).

... up to context passage number 10.

ASSISTANT: The Jiribatuba field is located in the basin(s): Camamu-Almada (Almada Mar, Camamu-Almada Basin, Camamu Mar, Almada Terra, Camamu Terra).

3.4 Hardware and Training Parameters

We fine tune the Nous Hermes 2 base model, which contains 7B parameters, using a Supervised Fine Tuning (SFT) procedure with a Parameter Efficient Fine Tuning (PEFT) approach (Houlsby et al., 2019; Lester et al., 2021). Specifically, we apply Low Rank Adaptation (LoRA) (Hu et al., 2022).

All training hyperparameters reported in Table 3 were kept constant throughout the experiments. The only parameter that varied was the number of training epochs, with the aim of verifying whether there is substantial gains in the model's specific domain knowledge, as well as degradation in general knowledge, as the epochs evolve. Five models were therefore trained, each with a different number of training epochs: 4, 9, 16, 25 and 36.

Training was performed on four NVIDIA V100 GPUs with 32 GB of memory each. Table 3 reports the hyperparameters held fixed for all models.

Parameter	Value
`per_device_train_batch_size`	1
`gradient_accumulation_steps`	4
`learning_rate`	1e-4
`LoRA rank`	8
`LoRA alpha`	16
`LoRA dropout`	0.05

Table 3: Hardware and key training parameters held constant across all supervised fine-tuning runs

3.5 Evaluation

We evaluate the five fine-tuned models with two aims. The primary aim is to measure gains on domain data, and the secondary aim is to assess whether general-domain performance is preserved after fine-tuning. For the primary objective, we use 300 held-out MiniKGraph items from the test set (see Table 4). We report Accuracy following Mallen et al. (2023): both the model prediction and the gold answer are normalized by removing punctuation, lowercasing, and tokenizing on whitespace, and a prediction is counted as correct only if it satisfies the normalized matching rule defined for the gold answer.

Level	Number of items
1	197
2	60
3	13
4	30
Total	**300**

Table 4: Distribution of evaluation dataset items by complexity level

In the Jiribatuba example (Subsection 3.3), the gold answer is represented as a set of acceptable aliases. Because all possible answers are considered synonymous in the reference dataset, a generated answer is marked correct if, after normalization, it includes any of the following: *Camamu–Almada*, *Almada Mar*, *Camamu–Almada Basin*, *Camamu Mar*, *Almada Terra*, or *Camamu Terra*. In other question–answer pairs, where the possible answers are not synonymous, the model's predicted answer must include at least one term corresponding to each distinct, non-synonymous term. For this reason, the dataset distinguishes between acceptable synonyms and different entities.

We also use the RAGAS framework (Es et al., 2024; Amiraz et al., 2025) to quantify answer quality through two LLM-prompted metrics. "Faithfulness" measures the proportion of claims in the generated answer that can be inferred from the retrieved context. The answer is first decomposed into individual statements, each statement is verified against the context, and the score is the fraction of supported statements. "Response Relevancy" measures how well the answer addresses the original question. An LLM generates candidate questions from the answer, and the score is the average cosine similarity between those candidates and the original question. Higher values indicate answers that are both well-grounded and on-topic. Both metrics are computed on the same 300-item do-

main test set.

For the primary objective, we compute all reported metrics at two levels: (i) global, per model; and (ii) stratified by item complexity, per model. For the secondary objective, we evaluate all five fine-tuned models on general-domain data using the Retrieval-Augmented Generation Benchmark (RGB) (Chen et al., 2024).

RGB is a bilingual (English and Chinese) benchmark that evaluates four fundamental abilities required for RAG: noise robustness, negative rejection, information integration, and counterfactual robustness. It contains 600 question–answer instances constructed from recent news articles, split equally between languages, each paired with five external documents retrieved via a search engine. In this work, we use the 300 English items and focus on the noise robustness testbed, where a controlled proportion of the external documents are *noisy*—i.e., topically related to the question but not containing the answer.

We compare the results of the five fine-tuned models with the base model, Nous Hermes 2 (7B), across multiple noise scenarios. In this context, *noise* refers to the proportion of noisy documents among the external documents: for example, at 80% noise, 4 out of 5 documents do not contain the answer. Noisy documents are progressively added to the context to test robustness, reproducing the protocol of Caldas and de Souza (2025). This procedure allows us to assess not only whether general-domain performance is preserved after fine-tuning, but also how the models' robustness to noisy retrieval compares with that of the base model.

4 Results and Discussion

4.1 Training Epoch Selection

Table 5 reports the accuracy of all fine-tuned models for different numbers of training epochs. A model trained for four epochs already reaches an accuracy of 0.933. Increasing the number of epochs beyond four does not lead to substantial gains: the resulting accuracies are very close to the value achieved at four epochs and do not exhibit a clear trend of further improvement. These results suggest that four epochs are sufficient for the instruction fine-tuning to take effect on this model.

Training time is also an important factor when choosing the final model. The four-epoch model requires 35 minutes and 44 seconds to train, and the execution time increases approximately linearly with the number of epochs, up to 5 hours, 22 minutes, and 16 seconds for the model trained for 36 epochs. Since all models achieve comparable performance on the in-domain MiniKGraph dataset, we select the four-epoch model as our final fine-tuned model, as it offers the best balance between accuracy and computational cost.

Epochs	Accuracy	Training time
4	0.933	00:35:44
9	0.940	01:19:50
16	0.933	02:21:29
25	0.927	03:40:09
36	0.950	05:16:22

Table 5: Accuracy across models with different numbers of training epochs on the MiniKGraph dataset.

Model	Accuracy	Faithfulness	Response Relevancy
RoBERTa	0.127	0.796	0.519
Nous Hermes 2 (7B) *(Base)*	0.730	0.823	0.854
OmniPetro-NH2-7B-TextRAG (Ours)	**0.930**	0.849	0.891
GPT-4o Mini	0.680	0.841	**0.929**
GPT-4o	0.810	**0.852**	0.921

Table 6: Comparison of Model Performance on Domain-Specific Data

We name the resulting model OmniPetro-NH2-7B-TextRAG. OmniPetro identifies the project, NH2 abbreviates the base model name, 7B denotes the parameter scale, and TextRAG specifies the training task.

4.2 In-Domain Performance

Table 6 presents results for all evaluated models, including the fine-tuned OmniPetro-NH2-7B-TextRAG, on Accuracy, Faithfulness, and Response Relevancy for in-domain data.

For Accuracy, the fine-tuned model reaches 0.930, the highest value among all systems. GPT-4o follows with 0.810, the base model Nous Hermes 2 (7B) reaches 0.730, GPT-4o Mini reaches 0.680, and RoBERTa reaches 0.127. These results indicate an improvement of approximately 20 percentage points in accuracy over the base model and 12 points over GPT-4o. This suggests that domain fine-tuning increases factual precision in O&G domain-related questions.

For Faithfulness, GPT-4o reaches 0.852, the fine-tuned model reaches 0.849, GPT-4o Mini reaches 0.841, the base model reaches 0.823, and RoBERTa reaches 0.796. These results indicate that the fine-tuned model maintains grounding to the retrieved evidence at a level comparable to larger proprietary

Model	Level	Accuracy	Faithfulness	Resp. Relevancy
RoBERTa	1	0.000	**0.874**	0.487
	2	0.200	0.481	0.544
	3	0.000	0.577	0.528
	4	0.867	0.933	0.670
Nous Hermes 2 (7B) *(Base)*	1	0.822	0.825	0.880
	2	0.317	0.827	0.760
	3	0.615	0.858	0.686
	4	**1.000**	0.794	**0.949**
OmniPetro-NH2-7B-TextRAG (Ours)	1	**0.970**	0.850	0.895
	2	**0.867**	**0.876**	0.830
	3	0.462	0.477	**0.984**
	4	**1.000**	**0.950**	0.942
GPT-4o Mini	1	0.777	0.839	**0.957**
	2	0.200	0.795	0.883
	3	**0.692**	**0.877**	0.764
	4	**1.000**	0.928	0.908
GPT-4o	1	**0.970**	0.840	0.945
	2	0.217	0.865	**0.884**
	3	**0.692**	0.808	0.764
	4	**1.000**	0.917	0.910

Table 7: Comparison of Model Performance by Complexity Level on Domain-Specific Data. Boldface indicates the highest value within each level and metric.

models. All evaluated models show high grounding values.

For Response Relevancy, GPT-4o Mini reaches 0.929, GPT-4o reaches 0.921, the fine-tuned model reaches 0.891, the base model reaches 0.854, and RoBERTa reaches 0.519. Differences among the generative models are small, indicating that all of them provide relevant answers. The extractive model RoBERTa shows a lower value, indicating limited ability to generate relevant responses.

Two observations follow. First, on in-domain data, the fine-tuned model reaches performance levels comparable to GPT-4o Mini and GPT-4o and surpasses them in accuracy. Second, the fine-tuned model outperforms the base model across all three metrics, indicating effective learning of domain knowledge through fine-tuning.

Table 7 presents results by question complexity level. The fine-tuned model performs best at levels with a higher number of examples and shows more variability as the number of samples decreases. At the lower complexity levels, specifically Levels 1 and 2, which together account for 84% of the MiniKGraph dataset, the fine-tuned model OmniPetro-NH2-7B-TextRAG achieves its largest performance improvements relative to the base model. Accuracy increases from 0.822 (base model) to 0.970 (ours) at Level 1 and from 0.317 to 0.867 at Level 2. These question types are the most frequent during fine-tuning, indicating that the model effectively learns domain patterns and terminology present in high-frequency examples.

At intermediate complexity (Level 3), the fine-tuned model reaches 0.462, slightly below the base model value of 0.615, likely reflecting the limited number of examples in this category (6.1% of the dataset). GPT-4o and GPT-4o Mini achieve the highest accuracy at this level, suggesting that larger models perform better than smaller, fine-tuned models, under data-scarce training conditions. Results at this level should be interpreted cautiously due to the small sample size (9.15%). At the highest complexity level (Level 4), all generative models reach perfect accuracy.

Faithfulness and Response Relevancy follow similar trends. The fine-tuned model generally outperforms the base model across most complexity levels. GPT-4o Mini obtains the highest Response Relevancy at Level 1 and the highest Faithfulness at Level 3, while GPT-4o achieves the best Response Relevancy at Level 2. As observed for Accuracy, Faithfulness decreases at Level 3, where the fine-tuned model reaches 0.477 and the base model reaches 0.858, which may be linked to the limited number of items in this level.

4.3 General-Domain Performance

Figure 1 reports Accuracy for the fine-tuned and the base model on general-domain data under con-

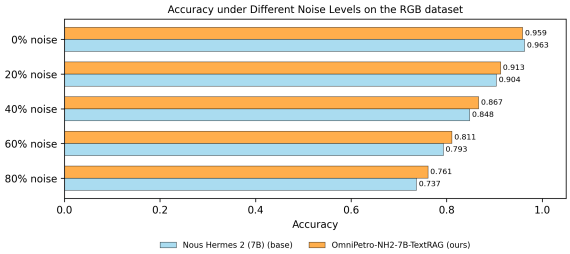

Figure 1: Accuracy on the RGB benchmark across noise levels (0%–80%) for OmniPetro-NH2-7B-TextRAG and the base model.

trolled noise. The objective is to assess whether specialization on domain data affects general-domain inference. Across all noise levels, Accuracy remains similar for both models. At 0% noise, the fine-tuned model reaches 0.959 and the base model reaches 0.963. At 80% noise, the fine-tuned model reaches 0.761 and the base model reaches 0.737. These results indicate no reduction in general-domain accuracy after fine-tuning.

While the evaluation of five fine-tuned models trained for different epoch counts on the MiniK-Graph dataset shows no substantial performance gains with additional epochs (as seen in Table 5), the results on the RGB general-domain benchmark (Section 3.5) reveal only a slight performance decline as the epoch number increases. For example, at a 0% noise level, accuracy decreases from 0.959 at four epochs to 0.943 at 36 epochs, with similarly gradual declines across the other noise levels. Even at the highest noise level (80%), accuracy falls only from 0.761 at four epochs to 0.695 at 36 epochs.

This suggests that LORA techniques within the PEFT framework are not only computationally efficient but also less prone to catastrophic forgetting. As discussed in Section 2, this behavior aligns with recent findings showing that LoRA mitigates forgetting more effectively than full fine tuning (Biderman et al., 2024), largely because the pretrained network remains frozen and only small low rank matrices are updated during training (Hu et al., 2022).

5 Concluding Remarks

This study investigated the specialization of a small language model for closed-domain Retrieval-Augmented Generation using text data derived from PetroKGraph, a Portuguese knowledge graph for the O&G domain. The objective was to adapt the model to domain-specific knowledge while preserving its ability to reason over general-domain data.

Starting from the Nous Hermes 2 (7B) base model, we trained five fine-tuned variants with different numbers of epochs while keeping all other hyperparameters fixed, and selected the four-epoch model as the final model (OmniPetro-NH2-7B-TextRAG) given its comparable performance and lower computational cost.

First, regarding domain-specific data, the fine-tuned model shows performance levels comparable to the large closed-source models GPT-4o and GPT-4o Mini. In terms of accuracy, the fine-tuned model reaches a value of 0.93, which is 25 points higher than GPT-4o Mini and 12 points higher than GPT-4o. For RAGAS metrics, the fine-tuned model achieves values close to those of the closed-source models, indicating consistent performance.

Second, with respect to the comparison between the fine-tuned model and the base model, the fine-tuned version outperforms the base model in all global metrics. Accuracy increases by 20 points, while faithfulness and response relevance improve by 2.6 and 3.7, respectively.

Third, concerning performance by question complexity level, the fine-tuned model shows notable gains in levels 1 and 2, which represent 84.77% of the data. For level 1, accuracy improves by 14.8% compared to the base model, and for level 2, by 55.5%. For higher complexity levels, where less data is available, the fine-tuned model shows mixed results relative to the base model.

Finally, regarding general-domain performance, across all noise levels, the fine-tuned and the base models accuracy values follow a similar pattern. For example, at a 0% noise level, the fine-tuned model attains an accuracy of 0.959, compared with 0.963 for the base model; at an 80% noise level, the fine-tuned model reaches 0.761, while the base model remains close at 0.737. These results indicate that the in-domain fine-tuned model preserves its ability to reason over general-domain data.

6 Limitations

There are some limitations to this study that should be considered. The first limitation concerns the methodology used to extract data. The approach adopted in this study is based on the work of Navarro et al. (2024), which relies on queries to extract entities and relations and uses authored

templates to generate natural language text. Although this method was appropriate for the present study, alternative approaches, such as using large language models to predict relationships, were not tested, as discussed in Section 2.

A second limitation involves the balance of the dataset in terms of question complexity. As shown in Table 1, complexity levels 1 and 2 represent 84.77% of the total dataset. This distribution affects the model's learning behavior, as most of the training data comes from lower-complexity samples. Consequently, results for higher complexity levels (3 and 4) are more variable, showing both gains and losses across different metrics and levels. A promising alternative could be synthetic data generation, using variations from the examples already provided in the dataset, to increase the number of diverse examples.

The third limitation concerns data accessibility. Both the training data and the resulting fine-tuned data are private and cannot be shared publicly due to corporate restrictions. Nevertheless, the main contribution of this study lies in the proposed pipeline, which can serve as a foundation for future research aiming to enhance model performance using structured data from knowledge graphs in closed-domain environments.

A fourth limitation is that, by design, this study evaluates the generator component in isolation, with all models receiving identical pre-retrieved passages, rather than comparing end-to-end RAG pipelines. Embedding the fine-tuned model in a full pipeline and measuring the interaction between retrieval and generation is a natural direction for future work.

7 Acknowledgments

The work was carried out with assistance granted by the National Agency of Petroleum, Natural Gas and Biofuels (ANP), Brazil, associated with the investment of resources originating from the R,D&I Clauses, through the Cooperation Agreement between Petrobras and PUC-Rio.

References

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. 2025. The Distracting Effect: Understanding Irrelevant Passages in RAG. *Preprint*, arXiv:2505.06914.

Robert Arp, Barry Smith, and Andrew D Spear. 2015. *Building ontologies with basic formal ontology*. Mit Press.

Gabriel Assis, Arthur Vasconcelos, Lívia de Azevedo, Mariza Ferro, and Aline Paes. 2024. Modestos e Sustentáveis: O Ajuste Eficiente Beneficia Modelos de Língua de Menor Escala em Português? In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 97–107, Porto Alegre, RS, Brasil. SBC.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. 2024. LoRA Learns Less and Forgets Less. *Preprint*, arXiv:2405.09673.

Josue Caldas and Elvis de Souza. 2025. A comprehensive evaluation of large language models for retrieval-augmented generation under noisy conditions. In *Proceedings of the 1st Workshop on Confabulation, Hallucinations and Overgeneration in Multilingual and Practical Settings (CHOMPS 2025)*, pages 60–69, Mumbai, India. Association for Computational Linguistics.

Charalampos Chelmis, Jing Zhao, Vikram Sorathia, Suchindra Agarwal, and Viktor Prasanna. 2013. Toward an Automatic Metadata Management Framework for Smart Oil Fields. *SPE Economics & Management*, 5(01):33–43.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking Large Language Models in Retrieval-Augmented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762.

Fábio Corrêa Cordeiro, Patrícia Ferreira da Silva, Alexandre Tessarollo, Cláudia Freitas, Elvis de Souza, Diogo da Silva Magalhaes Gomes, Renato Rocha Souza, and Flávio Codeço Coelho. 2024a. Petro NLP: Resources for natural language processing and information extraction for the oil and gas industry. *Computers & Geosciences*, 193:105714.

Fábio Corrêa Cordeiro, Patrícia Ferreira da Silva, Diogo da Silva Magalhaes Gomes, Renato Rocha Souza, Flávio Codeço Coelho, and Basil Ell. 2024b. Petro Kgraph: A Methodology for Extracting Knowledge Graph from Technical Documents-an Application in the Oil and Gas Industry. *Available at SSRN 4776804*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

Deep Bidirectional Transformers for Language Understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Luan Fonseca Garcia, Mara Abel, Michel Perrin, and Renata dos Santos Alvarenga. 2020. The GeoCore ontology: a core ontology for general use in geology. *Computers & Geosciences*, 135:104387.

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

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, pages 2790–2799. PMLR.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. *Preprint*, arXiv:1801.06146.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Nourhan Ibrahim, Samar Aboulela, Ahmed Ibrahim, and Rasha Kashef. 2024. A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): models, evaluation metrics, benchmarks, and challenges. *Discover Artificial Intelligence*, 4(1):76.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. *Preprint*, arXiv:2212.10511.

Laura Navarro, Elvis Souza, and Marco Pacheco. 2024. Text extraction from Knowledge Graphs in the Oil and Gas Industry. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 524–529, Porto Alegre, RS, Brasil. SBC.

NousResearch. 2024. Nous Hermes 2 Mistral 7B DPO.

Malte Pietsch, Timo Möller, Bogdan Kostic, Julian Risch, Massimiliano Pippi, Mayank Jobanputra, Sara Zanzottera, Silvano Cerza, Vladimir Blagojevic, Thomas Stadelmann, Tanay Soni, and Sebastian Lee. 2019. Haystack: the end-to-end NLP framework for pragmatic builders.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Dong Shu, Tianle Chen, Mingyu Jin, Chong Zhang, Mengnan Du, and Yongfeng Zhang. 2025. Knowledge Graph Large Language Model (KG-LLM) for Link Prediction. *Preprint*, arXiv:2403.07311.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2020. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Preprint*, arXiv:1911.06136.

Liang Yao, Jiazhen Peng, Chengsheng Mao, and Yuan Luo. 2025. Exploring Large Language Models for Knowledge Graph Completion. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zhiyuan Zhang, Xiaoqian Liu, Yi Zhang, Qi Su, Xu Sun, and Bin He. 2020. Pretrain-KGE: Learning Knowledge Representation from Pretrained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 259–266, Online. Association for Computational Linguistics.