

SLIM-RAFT: A Novel Fine-Tuning Approach to Improve Cross-Linguistic Performance for Mercosur Common Nomenclature

Vinícius Di Oliveira^{1,2}^a, Yuri Façanha Bezerra¹^b, Li Weigang¹^c, Pedro Carvalho Brom^{1,3}^d
and Victor Rafael R. Celestino⁴^e

¹*TransLab, University of Brasilia, Brasilia, Federal District, Brazil*

²*Secretary of Economy, Brasilia, Federal District, Brazil*

³*Federal Institute of Brasilia, Brasilia, Federal District, Brazil*

⁴*LAMFO, Department of Administration, University of Brasilia, Brasilia, Federal District, Brazil*

Keywords: Fine-Tuning, HS, Large Language Model, NCM, Portuguese Language, Retrieval Augmented Generation.

Abstract: Natural language processing (NLP) has seen significant advancements with the advent of large language models (LLMs). However, substantial improvements are still needed for languages other than English, especially for specific domains like the applications of Mercosur Common Nomenclature (NCM), a Brazilian Harmonized System (HS). To address this gap, this study uses TeenyTineLLaMA, a foundational Portuguese LLM, as an LLM source to implement the NCM application processing. Additionally, a simplified Retrieval-Augmented Fine-Tuning (RAFT) technique, termed SLIM-RAFT, is proposed for task-specific fine-tuning of LLMs. This approach retains the chain-of-thought (CoT) methodology for prompt development in a more concise and streamlined manner, utilizing brief and focused documents for training. The proposed model demonstrates an efficient and cost-effective alternative for fine-tuning smaller LLMs, significantly outperforming TeenyTineLLaMA and ChatGPT-4 in the same task. Although the research focuses on NCM applications, the methodology can be easily adapted for HS applications worldwide.

1 INTRODUCTION


Generative Artificial Intelligence (GenAI) has accelerated AI development, particularly through large language models (LLMs) like ChatGPT, which support multilingual and multimodal processing (Schulhoff et al., 2024; Radosavovic et al., 2024). However, using these models often requires prompt engineering skills, while open-source LLMs like LLaMA 3 offer flexibility through local fine-tuning. Yet, non-English users face limitations due to LLaMA's English-centric training data (90%) (Souza et al., 2020).


Although LLMs can process related languages, these capabilities are limited for technical tasks, especially when handling enterprise-specific privacy data. The small pre-trained Portuguese corpus in models like LLaMA highlights these challenges. Smaller


models, such as TeenyTinyLLaMA (Corrêa et al., 2024), offer an alternative, fine-tuned approach for well-defined tasks, as explored in this work.


Retrieval-Augmented Generation (RAG) mitigates LLM challenges such as hallucinations and outdated knowledge by integrating external databases, enhancing content accuracy for knowledge-intensive tasks (Lewis et al., 2020; Gao et al., 2023). Retrieval-Augmented Fine-Tuning (RAFT) goes further, focusing on relevant documents to improve model reasoning and performance by ignoring distractors and citing necessary sequences (Zhang et al., 2024; Warnakulasuriya and Hapuarachchi, 2024). However, generating chain-of-thought training data is costly and complex.


This work focuses on the MERCOSUR Common Nomenclature (NCM), essential for regional trade. NCM classification involves more than translation; it demands advanced Portuguese language processing due to product classification intricacies. Initial experiments show that LLMs like ChatGPT or TeenyTinyLLaMA are insufficient for this task. While neural net-

^a  <https://orcid.org/0000-0002-1295-5221>

^b  <https://orcid.org/0009-0001-8294-7163>

^c  <https://orcid.org/0000-0003-1826-1850>

^d  <https://orcid.org/0000-0002-1288-7695>

^e  <https://orcid.org/0000-0001-5913-2997>

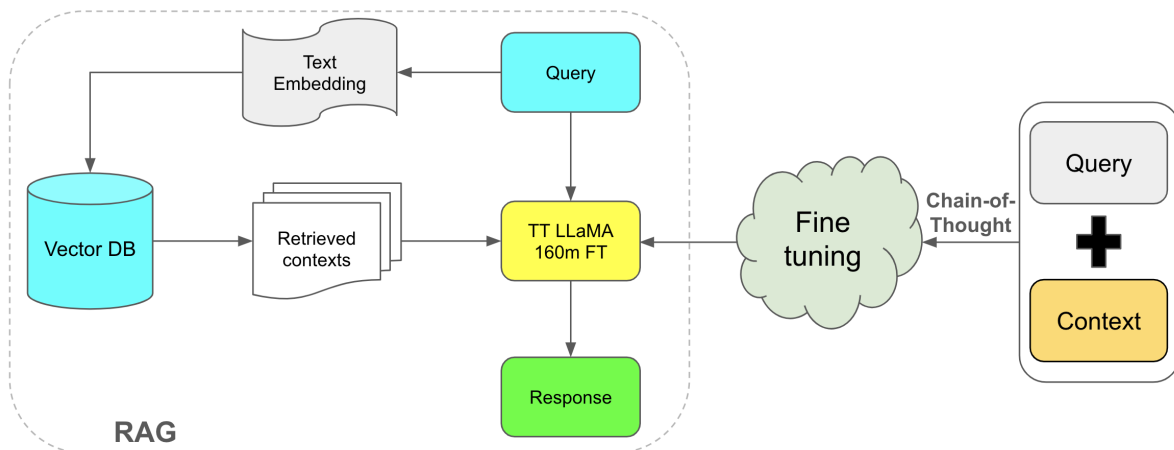


Figure 1: The SLIM-RAFT diagram.

works handle simple classification (Du et al., 2021), this study aims to extract deeper semantic knowledge within the NCM system for enhanced commerce and taxation applications.

The “Simplified Logical Intelligent Model” (SLIM-RAFT) is introduced to address NCM challenges efficiently. It applies the RAFT methodology in a streamlined form, capable of handling multi-classification tasks while minimizing training complexity.

A distinguishing feature of the SLIM-RAFT model is its significantly smaller LLM source use than traditional models. Specifically, the TeenyTineLLaMA model, comprising only 160 million parameters, was utilised in constructing SLIM-RAFT as the fine-tuned LLM. Our model’s results substantially outperformed those of ChatGPT 4 in the proposed challenge: ChatGPT 4 scored 4.5/10, whereas SLIM-RAFT achieved an impressive score of 8.67/10. The SLIM-RAFT scheme can be seen in Figure 1.

This paper is organized as follows: Section 2 reviews the works directly related to the concepts behind the construction of SLIM-RAFT. Section 3 outlines the structure and functionality of the HS and NCM codes. Section 4 details the construction of the SLIM-RAFT model. Section 5 presents the results from comparative evaluations of the models and discusses the findings. Finally, Section 6 concludes the paper and suggests directions for future work related to the proposed model.

2 RELATED WORKS

This section presents related work in two areas: 1) research on the implementation of LLMs with Portuguese as the primary language, and 2) stud-

ies on Retrieval-Augmented Generation (RAG) and Retrieval-Augmented Fine-Tuning (RAFT).

2.1 Portuguese LLM’s

The introduction of LLaMA (Touvron et al., 2023a) as an open foundational model marked a key step in language processing. With models ranging from 7B to 65B parameters, LLaMA proved that state-of-the-art models could be trained on public datasets. Despite having fewer parameters, the LLaMA-13B model outperformed GPT-3 on several benchmarks. Later versions, LLaMA 2 and 3 (Touvron et al., 2023b; Meta, 2024), further refined these models, particularly for chat-based applications, establishing a solid foundation for future NLP advancements.

Despite available data for training Portuguese-language models, native speakers still notice limitations in models primarily trained on English data. Growing interest in creating large-scale models for Portuguese has driven recent advancements.

In European Portuguese (PT-PT), the initiative known as Glória (Lopes et al., 2024) merits particular attention. This project involves a trained decoder language model meticulously constructed from a corpus comprising 35 billion tokens from various sources.

For Brazilian Portuguese (PT-BR), Sabiá (Pires et al., 2023) was the first relevant LLM encountered. This initiative underscores the development of robust and scalable language models for the Portuguese language. Leveraging advanced machine learning architectures, these models have been instrumental in advancing natural language processing applications in Brazilian Portuguese.

The Cabrita model (Larcher et al., 2023) was launched as a low-cost alternative for training LLMs. The authors posited that their methodology could be

extended to any transformer-like architecture. To substantiate their hypothesis, they undertook continuous pre-training exclusively on Portuguese text using a 3-billion-parameter model known as OpenLLaMA. This effort culminated in the creation of openCabrita 3B. Remarkably, openCabrita 3B incorporates a novel tokenizer, significantly reducing the number of tokens necessary to represent the text. Subsequently, in a comparable approach, a new study introduced a model predicated on LLaMA 2, designed specifically for handling prompts in Portuguese. This model, named Bode (Garcia et al., 2024), is available in two versions: one with 7B and 13B parameters. Both models used the LoRa (Hu et al., 2021) fine-tuning method over an open-source LLM. This technique preserves the original parameters intact while introducing a new terminal layer atop the model, which is subsequently trained to achieve the desired fine-tuning outcome.

A noteworthy recent publication entitled “TeenyTinyLlama: Open-Source Tiny Language Models Trained in Brazilian Portuguese” (Corrêa et al., 2024) offers a valuable perspective on developing compact, open-source language models tailored to Brazilian Portuguese. Despite their reduced scale, these models hold significant potential for democratizing access to natural language processing technology, particularly within resource-limited communities.

Collectively, these works signify substantial advancements in implementing language models for the Portuguese language. They underscore the diversity of methodologies and the abundance of resources that bolster research and applications in NLP and related fields. These ongoing initiatives are poised to continue influencing the future of language technology for Portuguese speakers globally.

2.2 Retrieval-Augmented Approach

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) enhances content generation by integrating external knowledge, improving coherence, factual accuracy, and utility. This approach benefits tasks like question-answering, summarization, and dialogue systems by retrieving relevant information for more precise outputs.

Retrieval-Augmented Fine-Tuning (RAFT) (Zhang et al., 2024) combines RAG with fine-tuning, allowing models to gain domain-specific knowledge and retrieve essential external contexts. RAFT employs chain-of-thought prompting, enabling models to provide more explainable, structured reasoning in their responses.

RAG and RAFT were designed to confront the complexity of tailoring LLMs to specialized domains. Within these realms, the emphasis pivots from general knowledge reasoning to optimizing accuracy *vis-à-vis* a meticulously defined array of domain-specific documents.

2.3 NCM Data Set

The ELEVEN data set, ELEctronic inVoicEs in the Portuguese language (Di Oliveira et al., 2022), was meticulously curated to furnish researchers and entrepreneurs with a repository of product descriptions categorized under the Mercosur Common Nomenclature (NCM). This extensive database comprises over a million meticulously labelled records, each scrutinized by taxation experts. These descriptions are short texts, limited to 120 characters, and extracted from authentic electronic invoices documenting purchase and sales transactions.

Labelled datasets are rare, yet they provide indispensable resources for applications reliant on supervised learning (Van Engelen and Hoos, 2020). The ELEVEN dataset has served as a cornerstone for several noteworthy academic endeavours: 1) the development of a CNN-based system for classifying goods (Kieckbusch et al., 2021); 2) the creation of data visualization tools aimed at identifying outliers and detecting fraud (Marinho et al., 2022); and 3) the establishment of a framework utilizing automatic encoders to cluster short-text data extracted from electronic invoices, thereby enhancing anomaly detection within numeric fields (Schulte et al., 2022).

3 HS AND NCM CODES

In international trade, customs brokers, exporters, and importers must accurately classify goods under the Harmonized System (HS), which forms the basis of the Mercosur Common Nomenclature (NCM) code (Valença et al., 2023). The HS underpins customs tariffs and trade statistics in over 200 countries, while also aiding in the monitoring of controlled goods, establishing rules of origin, and supporting trade negotiations (WCO, 2024).

The NCM, used by all MERCOSUR members—Argentina, Brazil, Paraguay, Uruguay, and Venezuela—is legally required for all commercial transactions in Brazil, including on electronic invoices (MERCOSUR, 2024b; Brazil, 2016). Developed by the World Customs Organization, the HS and its hierarchical structure, which NCM mirrors, assigns numerical codes to products for streamlined

identification and classification in customs processes (WCO, 2018; MERCOSUR, 2024a). The HS structure is shown in Table 1

Table 1: Structure of the HS Codes (WCO, 2018).

2 digit (01-97)	Chapter
4 digit (01.01 - 97.06)	Heading
6 digit (0101.21 - 9706.00)	Subheading

Table 2 shows an HS list cutout, showing the distinction in classification codes for fresh and dried apples. While this differentiation may appear trivial, in an import operation where each type of apple is subject to different tax treatments, an error in code designation can result in significant repercussions. On one hand, the seller faces the risk of tax penalties, while on the other, customs authorities may encounter a loss of revenue.

Table 2: Headings 08.08 and 08.13 in the HS (WCO, 2018).

08.08	Apples, pears and quinces, fresh.
0808.10	- Apples
0808.30	- Pears
0808.40	- Quinces
...	...
08.13	- Fruit, dried, other than that of headings 08.01 to 08.06; mixtures of nuts or dried fruits of this Chapter.
0813.10	- Apricots
0813.20	- Prunes
0813.30	- Apples
0808.40	- Other Fruit
0808.50	- Mixtures of nuts or dried fruits of this Chapter

Accurate goods classification is crucial, affecting taxation, regulatory compliance, and eligibility for international trade benefits. Misclassification can result in penalties, customs delays, and financial losses (Yadav, 2023).

The MERCOSUR Common Nomenclature (NCM) extends the HS system by adding two digits, allowing for a more precise classification of products based on specific characteristics, as shown in Table 3.

Table 3: Structure of the NCM Codes (MERCOSUR, 2024a).

2 digit (01-97)	Chapter
4 digit (01.01 - 97.06)	Heading
6 digit (0101.21 - 9706.00)	Subheading
7 digit (0101.21.1 - 9706.00.9)	Item
8 digit (0101.21.10 - 9706.00.90)	Sub-item

The task of classifying product descriptions within

the HS or NCM system has been explored in academic literature (Du et al., 2021; Kieckbusch et al., 2021; Schulte et al., 2022). Techniques such as neural networks with hierarchical learning and convolutional neural networks (CNNs) have been effectively employed to address this task. Nevertheless, specific challenges associated with this domain remain insufficiently addressed.

One notable challenge is the variability in product descriptions: the same product can be described in multiple ways, and context-dependent synonyms and abbreviations further complicate classification. This complexity makes using LLMs a compelling alternative for interpreting product descriptions. Beyond simple classification, LLMs offer the potential to extract deeper knowledge by identifying relationships between products. Table 4 shows some abbreviations that can be easily found.

Table 4: Abbreviation Examples.

English	
Coc. 2L	= Coca-Cola 2 Liters
P. W. Rice	= Parboiled White Rice
Portuguese	
Fr. Desc.	= Fralda descartável (<i>Disposable diaper</i>)
T. Pap. FDupla	= Toalha de Papel Folha Dupla (<i>Double Ply Paper Towel</i>)
French	
EDT	= Eau de Toilette
EDP	= Eau de Parfum

Context is crucial in language processing, as it can help resolve ambiguities that often arise when considering abbreviations in isolation. This is where TRANSFORMER-based algorithms shine, as their ability to understand context is key (Vaswani et al., 2017). For example, the abbreviation “fr.” in Portuguese could refer to “fralda” (diaper), as shown in Table 4, or it could mean “fruta” (fruit). However, when the term “desc” (meaning “descartável” or disposable) follows, the context effectively resolves the ambiguity between “fralda” and “fruit”.

Developing LLMs capable of handling NCM and HS codes is valuable in fields like compliance and tax inspection. Import and export companies must ensure accurate product descriptions and classifications under HS and NCM codes to avoid financial penalties or legal trading restrictions. Customs authorities closely monitor these codes as they define the tax treatment of products. Misclassification can be seen as tax evasion, leading to fines or sanctions.

Therefore, using AI models can improve controlling and correcting the issuance of invoices and re-

lated documents (Kieckbusch et al., 2021; Marinho et al., 2022).

4 SLIM-RAFT MODEL

The SLIM-RAFT model simplifies RAFT logically and intelligently. Just as RAFT maintains the RAG in its designed form, SLIM-RAFT also maintains the RAG mechanism in its structure. See Figure 1.

The preceding sections have elucidated that constructing the training base in the original RAFT model is an expensive endeavour, frequently necessitating the deployment of another powerful LLM. This substantial cost is predominantly attributable to two features of RAFT: the chain-of-thought reasoning and the inclusion of irrelevant documents. While learning to disregard irrelevant documents is valid and logical within RAFT's objectives, it is not a requisite for all applications. This insight prompted the exclusion of this feature in the development of SLIM-RAFT.

SLIM-RAFT retains the chain-of-thought concept in its fine-tuning process, albeit simplified. Instead of using lengthy texts or entire documents as input, the approach employs logical arguments derived from the knowledge base. For instance: 1) element "a" belongs to set A; 2) set A is contained within set B; 3) consequently, "a" belongs to set B. The next subsection will explain how it was done.

4.1 FT Database and Prompting

A theoretical example of a list of arguments for constructing the simplified chain of thought:

- Doc. 1: $a \in A$
- Doc.2: $A \subseteq B$
- Doc. 3: Consequently, $\therefore a \in B$

This was an application of the training base built for fine-tuning within the idea of the simplified chain-of-thought. See below for a generic example of a prompt:

```
[{ "content":
  [context 1 ]... \n
  [context 2 ]... \n
  [context 3 ]... \n
  [...] ... \n
  [context n]... \n
  \n
  Answer the following question
  using information from the
  previous context: question",
  "role": "user"},
```

```
{"content": "response + reasoning
based on context",
"role": "assistant"}]
```

An expert in the NCM code developed a series of question-and-answer sets, complete with their respective arguments. Utilising the open version of ChatGPT 3.5, numerous variations of these questions were generated. Subsequently, a Python script was employed to create thousands of pairs [question + argument, answer + argument], wherein information derived from the NCM database populated the generic questions.

Then, for building a data training base in SLIM-RAFT mode, there are three steps:

1. A domain expert creates a small question-and-answer set, e.g. "What is the category of the product 'product'?"
2. Construct question-and-answer set variations (an LLM could be used), e.g. "Could you specify the category to which the product 'product' belongs?"
3. Populate the question-and-answer set mask. e.g. "What is the category of the product 'fresh apple package'?", "Could you specify the category to which the product 'fresh apple package' belongs?"

The total number of records in the data training base will be:

$$N = q \times v \times n \quad (1)$$

Where q is the number of question-and-answer created by the domain expert, v is the number of variations from each question-and-answer unit, and n is the total number of samples from the NCM database.

As delineated earlier, a notable distinction between SLIM-RAFT and the original RAFT lies in the simplified approach to constructing the fine-tuning base while preserving the chain-of-thought methodology.

4.2 Source LLM and Fine-Tuning

The LLM source chosen for this work was TeenyTinyLLaMA (TTL), available in two sizes: 460 million and 160 million parameters. Two primary characteristics of TTL guided this selection: its compact size and the training on a corpus in Brazilian Portuguese.

While other source models trained in Brazilian Portuguese exist, as discussed in Section 2, their substantial size can make fine-tuning costly, even when employing optimised techniques such as LoRa (Hu et al., 2021). In contrast, the compact size of TTL

made our fine-tuning process more cost-effective, demonstrating its practicality and potential for wider application.

The Fine-tuning process adjusts all model parameters. The reduced size of TTL facilitated this task. The codes employed were adapted from those provided by the authors of the original TTL paper (available on GitHub ¹) with minor modifications.

All codes developed for SLIM-RAFT are accessible on SLIM-RAFT’s GitHub repository. Both TTL models, 160 million and 460 million parameters, were fine-tuned to create SLIM-RAFT. The 160 million parameter version was used in SLIM-RAFT, while the 460 million parameter version was used for comparative analysis during the final model evaluation.

The SLIM-RAFT GitHub repository ² is a valuable resource that provides the codes used in this study. This open access not only allows the community to reproduce and assess this experiment but also encourages further collaboration and potential contributions to natural language processing.

5 RESULTS AND DISCUSSION

The results were evaluated through a comparative analysis of the responses delivered by the tested models. Three other models were chosen for this comparison: TeenyTinyLLaMA 460m, TeenyTinyLLaMA 460m + NCM fine-tuning, and ChatGPT 4.0. In the end, four models were tested and evaluated by ChatGPT 4.0:

- Model 1: TeenyTinyLlama with tiny460M without fine-tuning on the dataset, defined as TTL.
- Model 2: ChatGPT 4.0, defined as GPT.
- Model 3: TeenyTinyLlama with fine-tuning on the NCM dataset, defined as NCM-TTL.
- Model 4: TinyLlama with fine-tuning on the NCM dataset and using SLIM RAFT, defined as SLIM-RAFT.

The model’s responses were then submitted to ChatGPT-4.0, which compared the generated outputs with the desired outputs. To ensure impartiality in the evaluation, it is important to note that ChatGPT-4.0 was not informed of which model each response was associated with.

¹<https://github.com/Nkluge-correa/TeenyTinyLlama>

²<https://github.com/yurifacanha/ncmrag>

5.1 Results Presentation

The evaluation used 100 questions and answers (Q/As) not included in the fine-tuning training base. These 100 questions were presented to various models, and their responses were recorded and compared. ChatGPT-4 assessed the quality of each response, scoring it on a scale from 0 to 10. The final score for each model represents the average of the scores assigned to each response. Table 5 present the results of this evaluation. It is clear that Model 4 of SLIM-RAFT achieved the best score of 8.63 with a standard deviation of 2.30 across the 100 Q/As.

Table 5: Score results between the four models.

Model	Aver.	St. Dev.	Min.	Max.
TTL	0.2	0.98	0	5
NCM-TTL	4.71	3.53	0	10
GPT	4.5	1.39	0	5
SLIM-RAFT	8.63	2.30	0	10

5.2 LLM Justification

It is essential to underscore the potential utility of an LLM specialized in this type of task, as it extends beyond mere classification. A straightforward input-output classification system is confined to specific subjects and input formats. However, an LLM system can extract semantic knowledge from the training base and demonstrates flexibility in handling various input formats.

Answering a simple direct question like “What is the NCM code for the product *fresh apple package*” is not enough. The question can come in several forms or be embedded in a larger context, for example: “I don’t know the NCM code for *fresh apple package*, can you help me?”.

Another pertinent scenario involves cases of attempted tax evasion. For instance, if the product *fresh apple package* is exempt from taxation while *apple juice package* is subject to tax, it could be misleadingly described in a tax document as *apple j. pack.* but assigned the NCM code for *fresh apples package*, which is tax-exempt. If this discrepancy goes undetected by customs authorities, it could result in a loss of revenue due to the uncollected tax.

The SLIM-RAFT model can effectively help a system for controlling and inspect documents regarding the NCM Code misuse. But, because of its reduced size, the capability of extracting the right question embedded in a bigger context is limited. Let us consider the following example:

Portuguese - *na padaria e comprei um suco de*

laranja, percebi que na nota fiscal aparecia um código chamado NCM, mas estava com a impressão borrada. Qual seria o código impresso?

English—I was at the bakery and bought some orange juice. I noticed that a code called NCM appeared on the invoice, but the print was blurry. What would be the printed code?

When presented with this query, our model may not discern the central issue: "What is the NCM category for orange juice?" Integrating an additional LLM into the system could mitigate this limitation.

The Few-shot prompting technique (Ma et al., 2024; Gu et al., 2021) can enable other large language models (LLMs) to reformulate queries, thereby adapting the context to enhance comprehension by the smaller LLM integrated within the SLIM-RAFT model.

Nonetheless, given its current limited size, it is important to acknowledge that our model may not fully comprehend all contexts. The objective, however, is to expand the model as more resources become available, thereby enhancing its capacity and performance; as the model becomes larger, the need for integration with another model will be suppressed.

5.3 Limitations

The SLIM-RAFT model is a prototype developed to illustrate the original RAFT methodology's simplification and propose its application within the NCM domain. Consequently, this model is a highly specialised tool tailored for its designated task.

It is not recommended for use in ChatBot applications, as TeenyTinyLLaMA (TTL) creators have advised against employing the TTL 160m model. The TTL 460m model is recommended for chatbot functionalities.

When constructing the training base for fine-tuning, a simplified chain-of-thought approach is employed. However, it's crucial to remember that the involvement of a domain expert is beneficial and necessary for developing the reference lines of reasoning, highlighting the irreplaceable role of human expertise in this process.

6 CONCLUSIONS

The SLIM-RAFT model demonstrated significantly superior performance to ChatGPT 4 in interpreting and classifying product descriptions according to the NCM code.

This outcome indicates that a smaller-scale LLM with specific domain knowledge can surpass a more

powerful LLM in specialized tasks, provided it is appropriately adjusted and trained while maintaining low execution costs.

The technique for simplifying the construction of the chain-of-thought, as proposed in SLIM-RAFT, not only reduces costs but also proves to be a viable alternative for developing specialized LLMs with high accuracy.

Since NCM coding is used not only for managing, transporting, paying, and taxing various goods in the import and export trade between MERCOSUR countries but also for most tax bills for goods, commodities, and restaurants in the Brazilian market, the practical value of this research is substantial. The findings provide convenience for government departments involved in import and export, taxation, banks, transportation, and manufacturers. Additionally, since over 200 countries use the HS system for import and export trade, the LLM-NCM solution proposed in this article can also facilitate the effective promotion of LLM-HS applications worldwide.

Future research will explore applying SLIM-RAFT to LLaMA 3, with a focus on multilingual tasks and comparisons with other techniques like LoRa.

ACKNOWLEDGEMENTS

ChatGPT 4 was used in all sections of this work to standardize and improve the writing in British English. This research is partially funded by the Brazilian National Council for Scientific and Technological Development (CNPq).

REFERENCES

- Brazil, C. (2016). Ajuste sinief 17. https://www.confaz.fazenda.gov.br/legislacao/ajustes/2016/AJ_017_16. Accessed on Jun 6th, 2024.
- Corrêa, N. K., Falk, S., Fatimah, S., Sen, A., and de Oliveira, N. (2024). Teenytinyllama: open-source tiny language models trained in brazilian portuguese. *arXiv preprint arXiv:2401.16640*.
- Di Oliveira, V., Weigang, L., and Rocha Filho, G. P. (2022). Eleven data-set: A labeled set of descriptions of goods captured from brazilian electronic invoices. In *WEBIST*, pages 257–264.
- Du, S., Wu, Z., Wan, H., and Lin, Y. (2021). Hscodenet: Combining hierarchical sequential and global spatial information of text for commodity hs code classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 676–689. Springer.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. (2023). Retrieval-augmented

- generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Garcia, G. L., Paiola, P. H., Morelli, L. H., Candido, G., Júnior, A. C., Jodas, D. S., Afonso, L., Guilherme, I. R., Penteadó, B. E., and Papa, J. P. (2024). Introducing bode: A fine-tuned large language model for portuguese prompt-based task. *arXiv preprint arXiv:2401.02909*.
- Gu, Y., Han, X., Liu, Z., and Huang, M. (2021). Ppt: Pre-trained prompt tuning for few-shot learning. *arXiv preprint arXiv:2109.04332*.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Kieckbusch, D. S., Geraldo Filho, P., Di Oliveira, V., and Weigang, L. (2021). Scan-nf: A cnn-based system for the classification of electronic invoices through short-text product description. In *WEBIST*, pages 501–508.
- Larcher, C., Piau, M., Finardi, P., Gengo, P., Esposito, P., and Caridá, V. (2023). Cabrita: closing the gap for foreign languages. *arXiv preprint arXiv:2308.11878*.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Lopes, R., Magalhães, J., and Semedo, D. (2024). Gll`oria-a generative and open large language model for portuguese. *arXiv preprint arXiv:2402.12969*.
- Ma, H., Zhang, C., Bian, Y., Liu, L., Zhang, Z., Zhao, P., Zhang, S., Fu, H., Hu, Q., and Wu, B. (2024). Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36.
- Marinho, M. C., Di Oliveira, V., Neto, S. A., Weigang, L., and Borges, V. R. (2022). Visual analysis of electronic invoices to identify suspicious cases of tax frauds. In *International Conference on Information Technology & Systems*, pages 185–195. Springer.
- MERCOSUR (2024a). Mercosur - consultas à nomenclatura comum e à tarifa externa. <https://www.mercosur.int/pt-br/politica-comercial/ncml/> Accessed on Jun 4th, 2024.
- MERCOSUR (2024b). Mercosur countries. <https://www.mercosur.int/en/about-mercocur/mercocur-countries/> Accessed on Jun 6th, 2024.
- Meta, A. (2024). Introducing meta llama 3: The most capable openly available llm to date. Last accessed June 3th, 2024 <https://ai.meta.com/blog/meta-llama-3/>.
- Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*, pages 226–240. Springer.
- Radosavovic, I., Zhang, B., Shi, B., Rajasegaran, J., Kamat, S., Darrell, T., Sreenath, K., and Malik, J. (2024). Humanoid locomotion as next token prediction. *arXiv preprint arXiv:2402.19469*.
- Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., et al. (2024). The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.
- Schulte, J. P., Giuntini, F. T., Nobre, R. A., Nascimento, K. C. d., Meneguette, R. I., Li, W., Gonçalves, V. P., and Rocha Filho, G. P. (2022). Elinac: autoencoder approach for electronic invoices data clustering. *Applied Sciences*, 12(6):3008.
- Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023a). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Valença, P. R. M. et al. (2023). Essays on foreign trade, labor, innovation and environment.
- Van Engelen, J. E. and Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine learning*, 109(2):373–440.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Warnakulasuriya, S. and Hapuarachchi, K. (2024). From knowledge to action: Leveraging retrieval augmented fine tuning (raft) to empower quick and confident first-aid decisions in emergencies. *Researchgate (Preprint)* <http://dx.doi.org/10.13140/RG.2.2.35911.30888>.
- WCO (2018). *THE HARMONIZED SYSTEM A universal language for international trade*. World Customs Organization. <https://www.wcoomd.org/-/media/wco/public/global/pdf/topics/nomenclature/activities-and-programmes/30-years-hs/hs-compendium.pdf> (Visited 2024-06-04).
- WCO (2024). List of contracting parties to the hs convention and countries using the hs - world customs organization. <https://www.wcoomd.org/en/topics/nomenclature/overview/list-of-contracting-parties-to-the-hs-convention-and-countries-using-the-hs.aspx> Accessed on Jun 6th, 2024.
- Yadav, B. K. (2023). Impact of regulation and conformity assessment procedures on global trade. In *Handbook of Quality System, Accreditation and Conformity Assessment*, pages 1–21. Springer.
- Zhang, T., Patil, S. G., Jain, N., Shen, S., Zaharia, M., Stolica, I., and Gonzalez, J. E. (2024). Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*.