Continued pre-training of LLMs for Portuguese and Government domain: A proposal for product identification in textual purchase descriptions

Eduardo Paiva[1], Fernando Pereira[1], David Carvalho[2], Nilson Junior[2], Rennis Oliveira[2], Stella Bonifácio[1], Andre Rocha[1], Hamilton Oliveira[1], Felipe Cezar[2], Helio Junior[2]

[1]Brazilian Office of the Comptroller General (CGU)
[2]Federal Data Processing Service (SERPRO)
eduardo.paiva@cgu.gov.br, fernando.pereira@cgu.gov.br, david.carvalho@serpro.gov.br,
nilson.michiles-junior@serpro.gov.br, rennis.oliveira@serpro.gov.br, stella.bonifacio@cgu.gov.br, andre.rocha@cgu.gov.br,
hamilton.oliveira@cgu.gov.br, felipe.cezar@serpro.gov.br, helio.theodoro@serpro.gov.br

Abstract

The present study addresses the issue of identifying products in non-standardized invoices, presenting an approach based on large language models (LLMs). Faced with the scarcity of models trained in the Portuguese language, we proceeded with the pre-training of two LLMs, Lamma2-7B and Mistral-Instruct-7B, followed by fine-tuning for the specific task of product identification. Our central hypothesis, "continuing the pre-training of LLMs with Portuguese texts enhances the model's ability to identify products in textual purchase descriptions", was supported by the results, revealing significant improvements when compared to the original models. This research contributes not only to solving a practical problem but also highlights the effectiveness of continuing the pre-training of a LLM in specific linguistic contexts, such as Portuguese.

Introduction

The Brazilian government implements a series of programs aimed at promoting the improvement of the quality and access to education. These programs are implemented through the transfer of financial resources to municipalities. There is a program for the purchase of food to be served to students, another for the purchase of educational materials, and another for the purchase of spare parts for vehicles used in school transportation, among other programs whose goal is to support activities related to the education of children.

Entities that receive these financial transfers must account for the funds received. One of the accountability activities involves submitting invoices that verify the incurred expenses. These invoices contain specifications of the items purchased, as well as the corresponding payment amounts.

However, the specifications of the acquired products do not follow a single defined standard, so the same product can be specified in different ways, making the analysis and comparison of purchased items challenging.

Therefore, it is necessary to develop automated techniques that can handle the various ways of specifying a particular product and identify which products these specifications refer to.

Currently, large language models have proven to be very efficient in performing this type of task. However, most available models have been trained with limited data in the Portuguese language, which reduces their capacity to carry out tasks related to texts written in Portuguese.

Therefore, the goal of this article is to continue the training of two large language models, Llama 2 – 7B (Touvron et al. 2023) and Mistral-instruct-7B (Jiang et al. 2023), to use them in the task of identifying products in textual descriptions.

Related Work

The problem of product identification in textual purchase descriptions is a recurring issue in Public Administration. Consequently, there are already several works that propose to extract relevant information from governmental textual data. In this context, Carvalho et al. (2014a) suggest a methodology for formulating a price database for the Brazilian Federal Public Administration based on purchasing data presented on the Transparency Portal of the Brazilian Federal Government. These purchases are described in textual format and require the use of text mining techniques to extract the corresponding product for each purchase description.

The proposed approach is divided into 6 steps, where product identification is accomplished through the use of keywords defined by experts. In other words, there is a limited set of products that can be identified. The identification of measurement units used to quantify these products, on the other hand, is carried out through the application of clustering techniques.

Carvalho et al. (2014b) use Bayesian networks (Friedman, Geiger, and Goldszmidt 1997) to identify and prevent the splitting of purchases, a type of fraud used to circumvent legally required purchasing procedures. In Brazil, purchases below a certain value can be carried out with simpler procedures. However, a common fraud tactic to fit higher-value purchases into this simplified procedure is the splitting of a single purchase into several smaller ones, each below the legally defined limit. This article aims to identify purchases considered suspicious of having been split, allowing for appropriate measures to be taken before the irregular expenditure is finalized.

This identification of suspicious purchases is carried out through the use of Bayesian networks and involves a set of structured attributes during the classification process. However, it is also necessary to identify the products that are being specified in a textual format in the public procurement documents.

Marzagão (2016) presents another approach to the problem of identifying products and services acquired by the Public Administration. This work uses a registry of materials and services adopted by the Federal Government as training data and, based on this registry, attempts to classify purchases using the Support Vector Machine algorithm (Cortes and Vapnik 1995). This approach achieved an accuracy of 83.35%, and according to the author, the identified errors were caused by two main factors: flaws in the training dataset and class frequency problems. Some product classes, due to not being frequently purchased, did not provide sufficient information for the machine learning algorithm.

Aiming to meet the processing needs required by the large volume of information that constitutes government procurement databases, Paiva and Revoredo (2016) presented a scalable solution to the problem of product identification in textual purchase descriptions. Paiva and Revoredo (2016) proposed a product identification model based on keywords, similar to the process used in (Carvalho et al. 2014a). However, the approach suggests an architecture that enables parallel processing, addressing issues related to processing capacity limitations.

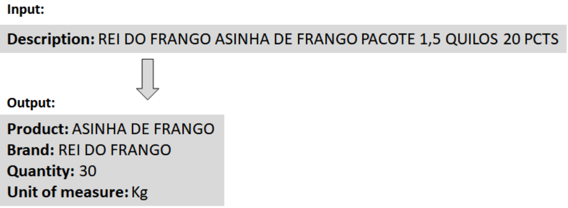
All of these works still had some limitations. Among them, it can be highlighted: the need for the use of a training dataset, which is usually not available, or the need for the definition of keywords by experts. Therefore, despite the advancements, research on product identification in government textual data still faced the limitations of the natural language processing (NLP) techniques available at that time.

However, the release of the transformers architecture (Vaswani et al. 2017) and large language models that utilize this architecture breathed new life into these research efforts. LLMs have demonstrated significant capabilities in understanding, generating, and manipulating natural language (Min et al. 2021), (Qiu et al. 2020). These capabilities have enabled new approaches to natural language processing (NLP) problems, and consequently, to the issue of product identification in textual descriptions.

It is worth noting that LLMs also have some limitations. Zhou et al. (2023) indicate that almost all knowledge in large language models is acquired during pre-training. Following this line of reasoning, Qiu et al. (2020) point out that the majority of publicly available pre-trained LLMs are trained on general domain corpora, such as Wikipedia, which limits their applications to specific domains or tasks.

Therefore, when it is desired to adapt a model to contexts where specific knowledge is essential, it becomes necessary to train or continue pre-training a model with data from the relevant context.

In this sense, some models specialized in specific domains emerge. BioGPT (Luo et al. 2022) is a pre-trained model on large-scale biomedical literature. This model has been evaluated in six biomedical tasks and outperformed previous models in the majority of them.



Input:

Description: REI DO FRANGO ASINHA DE FRANGO PACOTE 1,5 QUILOS 20 PCTS

Output:

Product: ASINHA DE FRANGO
Brand: REI DO FRANGO
Quantity: 30
Unit of measure: Kg

Figure 1: Expected Input and Output

On the other hand, BloombergGPT (Wu et al. 2023) is a model specialized in the financial domain. The authors conducted a complete training for the model. However, to retain the ability to answer general questions, the training utilized both domain-specific financial context data and general-purpose data.

Specifically in the Brazilian context, there is a lack of models trained in the Portuguese language that reflect the cultural and regional aspects of Brazilian society. An initiative in this regard was Sabiá (Pires et al. 2023), a language model specifically designed for the Portuguese language. The authors demonstrate that monolingual pre-training in the target language significantly improves models already extensively trained on various corpora. Sabiá was trained with Portuguese data from the ClueWeb 2022 corpus (Overwijk, Xiong, and Callan 2022), a dataset obtained from web pages.

In other words, there is still a shortage of language models trained in the Portuguese language, and more specifically, focused on the context of public administration. Therefore, the main contribution of this work is the proposal of further training language models with data from the Brazilian public administration, in Portuguese, aiming to use them in the task of product identification in textual descriptions.

Problem Statement

Every day, thousands of items are purchased within the scope of Brazilian government programs that support children's education. The invoices generated from these purchases are used to substantiate expenditures.

However, the information presented in these invoices is not standardized, so in many situations, the same product is described differently in different invoices. This complicates the comparison of the prices paid, as well as other analyses related to what is being purchased with public funds.

In other words, an individual can read the descriptions and identify that they refer to the same product. However, the large volume of available invoices makes this type of analysis impractical, rendering any manual comparison between similar purchases unfeasible.

However, this variability in descriptions also makes automated analysis unfeasible. Therefore, it is necessary to develop a solution capable of taking a textual product description as input and providing, as output, a set of structured information characterizing the purchased item. Figure 1 illustrates a textual description of a purchase and the desired output with structured and standardized information.

If the number of possible products were limited, this problem could be treated as a textual classification issue. However, the quantity of products that can be acquired within the scope of this program is very large, around 250 thousand products, and each day new products may be acquired, consequently expanding the set of possible classes.

Therefore, given the nature of this problem and the impracticality of treating it as a classification problem, this work aims to explore the ability of large language models to generate standardized texts. The goal is to enable the standardization of different product descriptions so that identical products can be uniquely represented even if they have been described in different ways.

Hence, the goal is to perform fine-tuning on LLMs so that they become "experts" in the task of standardizing product descriptions. However, initial tests indicated that the currently available open-source models had difficulties handling texts in the Portuguese language, as well as understanding terms, expressions, and other cultural elements specific to Brazilian society.

In light of the above, the possibility of continuing the pre-training of these models with Portuguese texts and content specific to the Brazilian reality was considered with the aim of equipping them with the ability to interpret and generate new texts that align with the Brazilian cultural context and the writing standards of the Portuguese language.

Thus, this study considered the following hypothesis: "Continuing the pre-training of LLMs with Portuguese texts enhances the model's ability to identify products in textual purchase descriptions".

Research Propositions

The proposal for product identification in invoices is structured into three main components. The initial phase entails the continuation of pre-training of a large language model (LLM). Subsequently, the second phase focuses on the fine-tuning of this model specifically for the task of product identification. Finally, the last step involves applying the model to the entire set of available invoices. This section details the activities related to the first two phases, while the section addressing the application of the model is presented separately.

Continuation of the model's pre-training.

Most large language models, when subjected to problems involving Portuguese texts, do not achieve the same performance observed when they are subjected to problems involving English texts. This characteristic arises from the fact that such models have been trained with few examples of Portuguese texts.

This problem is not as evident in more complex models, such as GPT-3.5, which has approximately 170 billion trainable parameters. However, models like these require significant computational power, which is not feasible for most organizations.

Therefore, the solution for most institutions is to look for less complex models that, in turn, can be used on less robust infrastructures.

In this context, our research aims to adapt a model consisting of approximately 7 billion parameters to effectively address our research problem. However, these types of models suffer from a lack of context for the Portuguese language. Thus, the pre-training of 2 language models was carried out to provide a more suitable context for the Portuguese language and topics related to Brazil, enabling better results for the task at hand.

Trained models: The two models chosen to continue the pre-training in Portuguese were Llama2–7B (Touvron et al. 2023) and Mistral-instruct–7B (Jiang et al. 2023).

Databases: All the databases used in training are available on the Internet. Five datasets were employed. The number of tokens presented in the descriptions of these databases was obtained using the Llama2 tokenizer (Touvron et al. 2023).

- BrWaC - Brazilian Portuguese Web as Corpus - (Filho et al. 2018): a corpus of Brazilian Portuguese text collected from the web. The present study used only about 5% of this corpus, which corresponds to 4.7 million lines and 474 million tokens.
- Reports from the CGU[1]: Reports from the CGU cover inspections conducted in Brazilian states and municipalities, as well as audits carried out in public bodies of the Federal Executive Power of Brazil. This database comprises 2.5 million lines, generating a total of 559 million tokens.
- Carolina Corpus (Crespo et al. 2023): a corpus with textual data collected from the web. This corpus includes texts from the judiciary and legislative branches of Brazil, public domain literary works, journalistic texts, social media and wiki texts, as well as texts already published in other corpora. This database consists of 2.1 million lines, generating 1.7 billion tokens.
- Public policy theses from FGV (Getulio Vargas Foundation)[2]: database with academic papers addressing topics related to public policies. This database consists of 127 thousand lines, generating 19.1 million tokens.
- Brazilian laws[3]: textual set collected from the web with the content of Brazilian laws. This set consists of 898 thousand lines, generating 70 million tokens.

Data preparation: Data preparation was carried out in 6 steps:

1. Transformation of PDF files into plain text format (when necessary)
2. Table removal
3. Segmentation of texts into sentences
4. Removal of blank sentences
5. Removal of short sentences
6. Removal of sentences with many numbers

The transformation of PDF files into plain text format is done with the assistance of the Apache Tika tool[4]. This activity is necessary for the data related to audit reports and

[1] Available at https://eaud.cgu.gov.br/relatorios

[2] Available at https://periodicos.fgv.br/

[3] Available at https://www4.planalto.gov.br/legislacao

[4] Available at https://tika.apache.org/

theses on public policies. After that, tables are removed from the texts since information from tables can become distorted and hinder the semantic understanding of the text. The next step involves segmenting the texts into sentences. Following this, short sentences are removed, considering a sentence to be short if it has fewer than 90 characters or fewer than 10 words. The last step involves the removal of sentences with a high number of numerical values. This removal occurs whenever more than 30% of the characters in a sentence are numeric.

Training run: This research continued the pre-training of the Llama2-7B and Mistral-Instruct-7B models with texts in Portuguese. The resulting models from this pre-training continuation gave rise to two other models, which were named Llama2GovBr-7B and Mistral-InstructGovBr-7B, respectively.

The training used a block size of 3072 tokens and an initial learning rate of 3×10^{-5}, with a weight decay of 0.1. A cosine decay learning rate scheduler was employed during training, where the gradient only decreased to a minimum of 10% of its initial value.

The optimizer used was paged_adamw_8bit, a variant of the Adam optimizer that applies weight corrections to combat weight decay and uses 8-bit precision to represent numbers.

Fine-tuning for the task of product identification

Fine-tuning aims to adjust the pre-trained model to perform specific tasks. In this way, the model is trained with a smaller and more specific dataset labeled for the desired task. The idea is to adjust the weights and parameters of the model so that it becomes more specialized in the given task. This activity requires less training data and lower computational power.

For the research problem at hand, a dataset containing information from invoices was used, including textual descriptions of the products, as well as the unit of measurement used to quantify them and the specified quantity. These data were previously labeled by humans, which involved defining the specified product, the brand, the unit of measurement used for quantification, and the quantity purchased. However, it is important to note that the product description does not always include all these attributes.

The fine-tuning activity used 900 records, with 800 of them used for training and 100 for evaluation. The training was carried out with the aim of optimizing the BLEU metric (Papineni et al. 2002), which is explained in the evaluation section.

Fine-tuning started with a learning rate of 3×10^{-4} with a weight_decay of 0.001. In this training, LoRA (Hu et al. 2022) was used, which is a more efficient strategy for fine-tuning large language models that reduces the number of parameters to be trained and the amount of memory required.

For training, the prompt below was used, and the text of this prompt was presented in the Portuguese language[5]. The product descriptions and their respective responses were altered for each of the examples used in the training.

Prompt: “You are an assistant that organizes items from invoices. For each provided description, organize the data in JSON format, placing the product, brand, quantity, and unit of measurement information separately. If numbers use a comma to separate decimals, use the American notation in the JSON”.

```
### Description:
ACHOCOLATADO INST 400G FABISE

### Response:
{
   product: ACHOCOLATADO INSTANTÂNEO,
   brand: FABISE,
   quantity: 400,
   unit_of_measurement: g,
}

### End
```

In other words, for each training example, a textual description of the product and an expected output are provided, with the data presented in a structured and standardized format, including the following fields: product name, brand, quantity, and unit of measurement.

Evaluation

Metrics

To guide the training and evaluation of the models, two metrics were used: BERTScore(Zhang et al. 2020) as the main metric and BLEU(Papineni et al. 2002) as a secondary metric.

BERTScore is used to evaluate the quality of generated text compared to a reference text. This metric considers the similarity between words and the context of those words, using pre-trained language models such as BERT(Devlin et al. 2019).

BLEU (Bilingual Evaluation Understudy) is a text generation evaluation metric. This metric assesses the model-generated output by comparing it to one or more human references, assigning a score based on the overlap of words and phrases between the generated output and the human references. A higher score is indicative of a greater overlap. BLEU scores range from 0 to 1, where 1 signifies a perfect match, and 0 signifies no match.

Evaluation Methodology

The evaluations were conducted always considering two models: the original model and the model with pre-training continuation, aiming to verify if this pre-training continuation brought improvements to the process. Thus, evaluations were performed for Lamma2-7B with Llamma2GovBR-7B and Mistral-Instruct-7B with Mistral-InstructGovBr-7B. It

[5] Original Portuguese Prompt: ”Você é um assistente que organiza itens de notas fiscais. Para cada descrição oferecida, organize os dados em formato JSON colocando as informações de produto, marca, quantidade e unidade de medida separados. Se os números estiverem com vírgula para separar decimais, utilizar a notação americana no JSON.”

Model	BERT_Score	BLEU
Llama2-7B	0.952	0.750
Llama2GovBr-7B	0.966	0.809

Table 1: Average results obtained by the Llama2-7B and Llama2GovBr-7B models

Model	BERT_Score	BLEU
Mistral-Instruct-7B	0.962	0.770
Mistral-InstructGovBr-7B	0.970	0.822

Table 2: Average results obtained by the Mistral-Instruct-7B and Mistral-InstructGovBr-7B models

is worth noting that the original models evaluated (Lamma2-7B and Mistral-Instruct-7B) underwent the same fine-tuning process presented earlier.

For model evaluation, 100 previously labeled product descriptions were used. Evaluations were performed using bootstrap (Efron 1992), a resampling strategy employed to estimate the distribution of a sample statistic. The bootstrap procedure created 10,000 samples for the evaluation of each model.

Llama2-7B X Llama2GovBR-7B

Table 1 presents the average results from 10,000 experiment runs related to the Llama2-7B model. The first row corresponds to the original model, and the second row corresponds to the model with pre-training continuation (Llama2GovBr-7B).

As observed, the results indicate better performance for the Llama2GovBr-7B model. However, statistical tests were conducted to verify if these differences are statistically significant.

Thus, the Student's t-test (Student 1908) was conducted, considering the null hypothesis that the means of the results for the two models are equal and the alternative hypothesis that the means are different. A significance level of 5% ($\alpha = 0.05$), corresponding to a confidence level of 95%, was adopted.

The test returned a p-value very close to zero, allowing us to reject the null hypothesis and consider the alternative hypothesis true. In other words, the results are statistically different, confirming that Llama2GovBr-7B indeed showed superior performance compared to the original model (Llama2-7B).

Mistral-Instruct-7B X Mistral-InstructGovBr-7B

Table 2 presents the results related to the tests conducted with the Mistral-Instruct-7B and Mistral-InstructGovBr-7B models.

As observed, the results presented in Table 2 also indicated better performance for the model with pre-training continuation (Mistral-InstructGovBr-7B) when compared to the original model (Mistral-Instruct-7B).

Thus, the t-test (Student 1908) was also conducted to verify if the differences between the results were significant. The test considered the same hypotheses described earlier, and once again, a p-value close to zero was obtained, allowing us to consider the alternative hypothesis. In other words, the results are different, and the model with pre-training continuation is indeed better.

Product	Reference Price (R$)
Tomato (kg)	7.40
Carrot (kg)	5.15
Watermelon (kg)	3.20
Onion (kg)	5.73
Banana (kg)	5.42

Table 3: Reference prices for the 5 most purchased products in December 2022.

Applications

Once the product to which each of the purchase descriptions refers is identified, a series of other analyses become viable. The objective of this section is to present some possible applications with the information obtained from the use of techniques developed in this research. For this purpose, data from invoices related to purchases made in December 2022 were used. During this period, there were 50,543 available invoices, which made reference to a total of 34,834 items purchased. The model was then applied to the textual descriptions of these items. The identification of the products was carried out using the model generated by fine-tuning Mistral-InstructGovBr-7B.

Reference Prices Calculation

Once the products are properly identified, it becomes possible to propose reference prices for the various products that are purchased. Table 3 presents the reference prices for the top 5 products purchased during the considered period.

For the calculation of reference prices, the product prices were considered as the median of the unit values paid in each purchase of these products. The median was chosen because it is less susceptible to the influence of outliers.

Another relevant factor is that in many situations, the price of a product can be influenced by seasonality and locality issues. However, since information from invoices is used, any of the attributes of the invoice can be considered for aggregation and defining the criteria for forming the reference price, such as by date, region, government agency, etc.

Identification of purchases with prices much higher than expected

From the moment a reference price is established for the various purchased products, it also becomes possible to identify purchases that have been made with values much higher than expected.

In Table 4, a sample of 2 values much higher than expected is presented for each of the reference price examples identified in Table 3. This table is only illustrative, given that, due to the large number of purchases, the number of

Product	Reference Price (R$)	Paid Price (R$)
Tomato (kg)	7.40	12.85
		10.67
Carrot (kg)	5.15	14.00
		13.83
Watermelon (kg)	3.20	7.58
		7.30
Onion (kg)	5.73	13.07
		13.00
Banana (kg)	5.42	12.15
		9.77

Table 4: Purchases with prices significantly above the Reference Price.

purchases considered significantly above the reference price is also high.

It is important to emphasize that these elevated values, presented in Table 4, are not sufficient to assert irregularities in the mentioned purchasing processes. Any indication raised through data analysis requires a more in-depth investigation through specific audits. Also, any analysis of individual purchases is not within the scope of this work. However, all the processes suggested in this research make it feasible to identify purchases that deviate from the expected pattern, enabling new types of analyses that would be impossible with the information presented in its original (textual) format.

Other Applications

The applications outlined in this section are just a few examples of possible uses for identifying products based on the methodology proposed in this article. However, there are several other possible applications, such as:

- Identification of suppliers selling the same product at different prices
- Comparison of the relationship between the amount paid and the quantity purchased
- Application of association rules to identify the probability of a municipality buying a certain product, given that it has already purchased a set of other types of products
- Identification of municipalities that purchase at better prices and those that pay more for products
- Verification of any behavioral patterns among product-supplying companies that may indicate collusion or price fixing
- Monitoring the evolution of prices over the months of the year

Conclusion

This article provides a solution for the task of identifying products in non-standardized invoices, using an approach based on large language models. Faced with the lack of models trained in Portuguese, we chose to continue the pre-training of two LLMs, Llama2-7B and Mistral-Instruct-7B, followed by fine-tuning specifically for product identification.

The results obtained confirmed our hypothesis that "continuing the pre-training of LLMs with Portuguese texts enhances the model's ability to identify products in textual purchase descriptions". This research not only effectively addresses a practical problem but also highlights the effectiveness of the continued pre-training approach in specific linguistic contexts, such as Portuguese.

Thus, this study not only contributes to solving specific challenges related to product identification in Portuguese texts but also to a broader understanding of the crucial role that continuing the pre-training can play in enhancing the performance of language models in specialized linguistic tasks. These advancements have significant implications not only for the academic community but also for professionals seeking effective solutions in practical and applied contexts.

References

Carvalho, R.; de Paiva, E.; da Rocha, H.; and Mendes, G. 2014a. Using Clustering and Text Mining to Create a Reference Price Database. *Learning and NonLinear Models*, 12(2014): 38–52.

Carvalho, R. N.; Sales, L.; Rocha, H. A. D.; and Mendes, G. L. 2014b. Using Bayesian Networks to Identify and Prevent Split Purchases in Brazil. In Laskey, K. B.; Jones, J.; and Almond, R. G., eds., *Proceedings of the Eleventh UAI Bayesian Modeling Applications Workshop co-located with the 30th Conference on Uncertainty in Artificial Intelligence, BMA@UAI 2014, Quebec City, Quebec, Canada, July 27, 2014*, volume 1218 of *CEUR Workshop Proceedings*, 70–78. CEUR-WS.org.

Cortes, C.; and Vapnik, V. 1995. Support-Vector Networks. *Mach. Learn.*, 20(3): 273–297.

Crespo, M. C. R. M.; de Souza Jeannine Rocha, M. L.; Sturzeneker, M. L.; Serras, F. R.; de Mello, G. L.; Costa, A. S.; Palma, M. F.; Mesquita, R. M.; de Paula Guets, R.; da Silva, M. M.; Finger, M.; de Sousa, M. C. P.; Namiuti, C.; and do Monte, V. M. 2023. Carolina: a General Corpus of Contemporary Brazilian Portuguese with Provenance, Typology and Versioning Information. *CoRR*, abs/2303.16098.

Devlin, J.; Chang, M. W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm): 4171–4186.

Efron, B. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, 569–593. Springer.

Filho, J. A. W.; Wilkens, R.; Idiart, M.; and Villavicencio, A. 2018. The brWaC Corpus: A New Open Resource for Brazilian Portuguese. In Calzolari, N.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Hasida, K.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; Piperidis, S.; and Tokunaga, T., eds., *Proceedings of the Eleventh International Conference on Language Resources* *and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Friedman, N.; Geiger, D.; and Goldszmidt, M. 1997. Bayesian Network Classifiers. *Mach. Learn.*, 29(2-3): 131–163.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de Las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. *CoRR*, abs/2310.06825.

Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; and Liu, T. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinform.*, 23(6).

Marzagão, T. 2016. Using SVM to pre-classify government purchases. *CoRR*, abs/1601.02680.

Min, B.; Ross, H.; Sulem, E.; Veyseh, A. P. B.; Nguyen, T. H.; Sainz, O.; Agirre, E.; Heintz, I.; and Roth, D. 2021. Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. *CoRR*, abs/2111.01243.

Overwijk, A.; Xiong, C.; and Callan, J. 2022. ClueWeb22: 10 Billion Web Documents with Rich Information. In Amigó, E.; Castells, P.; Gonzalo, J.; Carterette, B.; Culpepper, J. S.; and Kazai, G., eds., *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, 3360–3362. ACM.

Paiva, E.; and Revoredo, K. 2016. Big Data e Transparência: Utilizando Funções de Mapreduce para incrementar a transparência dos Gastos Públicos [Big Data and Transparency: Using MapReduce functions to increase Public Expenditure transparency]. In Siqueira, F.; Vilain, P.; Cappelli, C.; and Wazlawick, R. S., eds., *Proceedings of the XII Brazilian Symposium on Information Systems on Brazilian Symposium on Information Systems: Information Systems in the Cloud Computing Era, SBSI 2016, Florianópolis, Brazil, May 17-20, 2016*, 25–32. ACM.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, 311–318. ACL.

Pires, R.; Abonizio, H. Q.; Almeida, T. S.; and Nogueira, R. F. 2023. [inline-graphic not available: see fulltext] Sabiá: Portuguese Large Language Models. In Naldi, M. C.; and Bianchi, R. A. C., eds., *Intelligent Systems - 12th Brazilian Conference, BRACIS 2023, Belo Horizonte, Brazil, September 25-29, 2023, Proceedings, Part III*, volume 14197 of *Lecture Notes in Computer Science*, 226–240. Springer.

Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; and Huang, X. 2020. Pre-trained Models for Natural Language Processing: A Survey. *CoRR*, abs/2003.08271.

Student. 1908. The probable error of a mean. *Biometrika*, 1–25.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Canton-Ferrer, C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips): 5999–6009.

Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D. S.; and Mann, G. 2023. BloombergGPT: A Large Language Model for Finance. *CoRR*, abs/2303.17564.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; Zhang, S.; Ghosh, G.; Lewis, M.; Zettlemoyer, L.; and Levy, O. 2023. LIMA: Less Is More for Alignment. *CoRR*, abs/2305.11206.