Toxic Text Classification in Portuguese: Is LLaMA 3.1 8B All You Need?

Amanda S. Oliveira[1], Pedro H. L. Silva[2], Valéria de C. Santos[2], Gladston Moreira[2], Vander L. S. Freitas[2], Eduardo J. S. Luz[2]

[1]BLIP
30.130-174 – Belo Horizonte – MG – Brazil

[2]Computing Department – Federal University of Ouro Preto (UFOP)
35.400-000 – Ouro Preto – MG – Brazil

amanda.oliveira@blip.ai

{silvap,valeriacs,gladston,vander.freitas,eduluz}@ufop.edu.br

Abstract. *The recognition of toxic and hate speech on social media platforms is important due to the significant risks posed to users and the digital ecosystem. Current state-of-the-art models, such as BERTimbau, have set benchmarks for Portuguese text classification, yet challenges remain in accurately detecting toxic content. This paper investigates the effectiveness of fine-tuning a smaller, open-source decoder-only model, LLaMA 3.1 8B 4bit, for this task. We propose an iterative prompt evolution method to optimize the model's performance. Our results demonstrate that fine-tuning significantly enhances the LLaMA model's F1-score from 0.61 to 0.75, surpassing BERTimbau in precision and matching the performance of the GPT-4o mini. However, the approach depends on the quality of the language models used for prompt evolution, highlighting the need for further research to enhance robustness in this area.*

1. Introduction

The task of recognizing toxic and hate speech has gained substantial attention in recent years, particularly with the surge of user-generated content on social media platforms. As these platforms increasingly shape public discourse, the proliferation of harmful content presents significant risks to both individual users and the broader digital environment. Consequently, the need for effective moderation tools has escalated, driving research toward automated solutions capable of operating at scale.

Current state-of-the-art methods for automated toxic content classification predominantly leverage transformer-based architectures, with encoder-only models being the most common. Within the Portuguese language context, BERTimbau has emerged as a leading approach [Souza et al. 2020], demonstrating superior performance in various NLP tasks, including emotion classification [Hammes and de Freitas 2021], toxic speech detection [da Rocha Junqueira et al. 2023, Oliveira et al. 2023], news clustering [Pereira and da Silva 2023], among other tasks [dos Santos and Paraboni 2023, Serras and Finger 2021]. The BERTimbau ability to capture subtle nuances in Portuguese expressions has set a high standard in the field, making it the benchmark for multi-class classification tasks. However, despite its effectiveness, the problem of accurately classifying toxic content remains an open challenge, particularly in the diverse and evolving landscape of online discourse.

Recent advancements have shifted towards decoder-only models, such as LLM2Vec [BehnamGhader et al. 2024] and NV-Embed [Lee et al. 2024], which have shown promising results across multiple languages. Notably, OpenAI ChatGPT [OpenAI et al. 2024] [1], a large decoder-only language model, has demonstrated competitive performance in this domain [Oliveira et al. 2023]. The emergence of open-source models, like the Meta LLaMA family of models [Dubey et al. 2024], further compels a reexamination of existing methodologies, raising research questions about the potential of these newer models.

Building on these recent developments, this work explores the capabilities of decoder-only models, specifically focusing on the LLaMA 3.1 8B 4bit model [Dubey et al. 2024]. This model is particularly compelling due to its open-source nature, benchmark performance, and relatively smaller size, making it well-suited for fine-tuning specialized tasks such as toxic content classification in Portuguese. The key research questions guiding this investigation are RQ1: Can a fine-tuned LLaMA 3.1 8B 4bit model achieve or surpass the performance of GPT-4o mini in classifying toxic content in Portuguese? RQ2: Can this model outperform the current state-of-the-art BERTimbau-based approach in the same task? To address these questions, we propose a heuristic approach that utilizes a larger LLM (GPT-4o-mini) to refine the prompts employed by a smaller LLM, thereby automating prompt engineering. The optimal prompt is then used to fine-tune the LLaMA 3.1 8B 4bit model for toxic content classification in social media, using the TolDBr dataset - a large public dataset on this task [Leite et al. 2020]. Our results show that the fine-tuned LLaMA 3.1 8B 4-bit model, operating in zero-shot classification mode, outperforms the BERTimbau-based model regarding precision and is on par with GPT-4o mini.

2. Materials and Methods

Although the primary focus of this work is to investigate the performance of a small and open-source language model (with only 8B parameters) for the task of toxic text detection in Portuguese, the choice of prompt is a significant challenge. The quality of the prompt heavily influences the LLM's performance [Brown et al. 2020]. Therefore, this work proposes a straightforward approach to evolving prompts, ultimately using the best prompt identified for fine-tuning the model.

The following subsections describe the dataset selected for benchmarking, which is a large and popular dataset by Portuguese language standards for this task. Additionally, an outline of the methodology for selecting the best prompt and the approach used for fine-tuning the model.

2.1. Told-Br dataset

We employed the ToLD-br dataset, developed in [Leite et al. 2020] for training and testing the models used in this study. This dataset contains 21,000 tweets, annotated in a binary manner as “toxic” or “non-toxic”. Additionally, the tweets are also classified into different categories of toxicity, such as LGBTphobia, insults, racism, misogyny, and xenophobia.

In this study, we focused on the binary classification between “toxic” and “non-toxic”, using the corresponding annotations to train and test our models. The dataset was

[1] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

divided in a stratified manner, with 80% of the tweets allocated to the training set and the remaining 20% to the test set.

2.2. Prompt Engineering: Iterative Prompt Refinement

The challenge in using large language models (LLMs) for zero-shot classification lies in identifying the most effective prompt. This study proposes a heuristic to iteratively refine prompts using a larger LLM, intending to enhance classification accuracy in a smaller LLM.

Our approach draws on previous research, mainly works by [Oliveira et al. 2024] and [Oliveira et al. 2023], which advocate for using in-context learning for social media post classification. While these studies explore both zero-shot and few-shot modalities, our focus remains exclusively on the zero-shot scenario.

Given that LLMs have been shown to function effectively as black-box optimizers [Zheng et al. 2023] and are viable alternatives to mutation and crossover operations in genetic algorithms [Lehman et al. 2023, Meyerson et al. 2023], we draw inspiration from the work presented in [Guo et al. 2024] to propose a simplified algorithm for evolving prompts tailored explicitly to the task of toxic speech detection in Portuguese.

The methodology is structured as Figure 1 illustrates: Initially, a population of prompts is initialized, each one specifically designed to classify social media posts as toxic or non-toxic. The prompts then undergo a selection process, retaining only the top-performing ones based on evaluation metrics. Next, operations to evolve the prompt are applied utilizing an instruction to a larger LLM, such as GPT-4, which assists in generating new variations by recombining elements of existing successful prompts. This process is iteratively refined to enhance the quality of the prompts. Finally, the optimal prompt from this cycle is used to fine-tune the model. Algorithm 1 provides a pseudo-code overview of these steps.

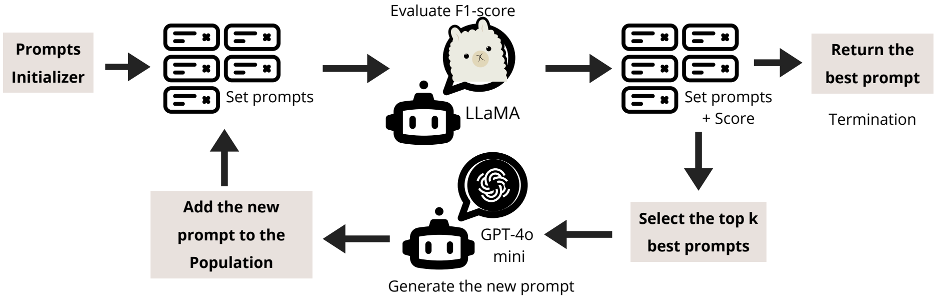

Figure 1. The heuristic iterative prompt evolution process begins with an initial set of prompts, which are evaluated using the LLaMA model based on their F1-scores. The top-performing prompts are then selected, and GPT-4o mini generates new prompts. These new prompts are added back to the population, and the process repeats. The best prompt from this iterative cycle is ultimately selected for further use. All prompts and instructions used in this study were written in Portuguese.

2.3. LLM Fine-Tuning Methodology

This methodology fine-tunes the model using a quantized version to enhance memory efficiency and speed. Parameter-Efficient Fine-Tuning (PEFT) [Houlsby et al. 2019,

Algorithm 1 Simplified Prompt Evolution

```
Function InitializePopulation(InitialPrompts)
    Population ← []
    for each prompt in InitialPrompts do
        Evaluate prompt with Llama 3.1 8B, using F1-score
        Add prompt and its score to Population
    end for
    return Population
Function GenerateNewPrompt(PromptsAndScores)
    PromptsText ← Concatenate each prompt and its F1-Score from PromptsAndScores with line breaks
    SystemInstruction ← "You are an assistant that helps improve AI prompts. You should always
  generate a new prompt, using different words or varying lengths, never repeating the same prompt.
  Generate ONLY the prompt, without comments or explanations".
    Instruction ← "You are evolving a prompt for another LLM. Based on the following prompts and
  their respective F1-scores, generate a new prompt optimized for the task of classifying hate speech".
    ChatGPTInput ← SystemInstruction + Instruction + PromptsText
    NewPrompt ← Call ChatGPT API with ChatGPTInput
    return NewPrompt
Function Main()
    InitialPrompts ← Define initial set of prompts
    Population ← InitializePopulation(InitialPrompts)
    for each epoch in range(NumEpochs) do
        TopKPrompts ← Select top 'k' prompts from Population, based on F1-scores
        PromptsAndScores ← Collect scores and prompts from TopKPrompts
        NewPrompt ← GenerateNewPrompt(PromptsAndScores)
        Evaluate NewPrompt with Llama 3.1 8B using F1-score
        Add NewPrompt and its score to Population
    end for
    BestPrompt ← Select best performing prompt from Population
    return BestPrompt
```

Hu et al. 2021] and QLoRA [Dettmers et al. 2023] techniques reduce model complexity, focusing on optimizing QKV projections and Feed Forward Layers. Training data is divided into training and validation sets. Specific prompts, structured as Alpaca prompts, align the model with toxic content classification objectives in Portuguese.

Alpaca Prompt Example:
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: `You are analyzing a social media post. If the text contains hate speech, offenses, aggressions, insults, swear words, or any form of toxicity, respond only with 'yes'. If it is appropriate and non-toxic, respond 'no'.`

Input: `Tweet: every time the weather changes, my throat gets inflamed...dammit!`

Response: `no.`

3. Experimental Setup and Results

Two key experiments are conducted to address the research questions posed in this study. First, Experiment #1 is designed to identify the best prompt. Using this prompt, the LLaMA 3.1 8B model is fine-tuned with the training data from the ToLD-Br dataset (Experiment #2). Four versions of the model are fine-tuned, varying parameters related to PEFT/QLoRA. An additional experiment is proposed to evaluate the performance of using a fine-tuned GPT-4o mini model within the same scenario. All experiments are performed in a Google Colab environment, utilizing an A100 GPU. Source code is available at `https://github.com/oliveiraamanda/ToxicSpeech-Llama-STIL-2024`.

3.1. Experiment #1 - Prompt Engineering

To evolve the prompts using the iterative algorithm proposed here, it is essential first to define the cost function to be minimized. The F1-score of binary classification on a partition of the training data is selected as the cost function. As our population consists of natural language sentences, the initial individuals must be manually defined for the specific domain. This approach was also adopted in [Guo et al. 2024]. In this work, we base our initial prompts on those proposed in [Oliveira et al. 2023] and [Oliveira et al. 2024], as these studies serve as a baseline and address the same dataset. Subsequently, we derive additional prompts by adding or removing sentences and words, totalizing six prompts. The target model, LLaMA 3.1 8B, performs the classification using only 50 instances from each class to compute the F1-score, given the high computational cost of this function. Following this, prompts are evolved through an iterative process involving two models—the target model LLaMA 3.1 8B and GPT-4o mini over 50 epochs. After this period, the best prompt (with the highest F1-score) is selected and presented below.

> "Best Prompt: You are analyzing a social media post. If the text contains hate speech, offenses, aggressions, insults, swear words, or any form of toxicity, respond only with 'yes'. If it is appropriate and non-toxic, respond 'no'".

3.2. Experiment #2 - LLaMA Fine-Tuning Process

To facilitate fine-tuning on modest hardware, we employed a 4-bit quantized version of the LLaMA 3.1 model, using the QLoRA technique [Dettmers et al. 2023], with 8 billion parameters model- LLaMA 3.1 8B [Dubey et al. 2024][2]. We used the Hugging Face PEFT library[3] with the Unsloth library[4], setting the learning rate to $2e-4$ and the sequence length to 2048 tokens, while varying the "rank" and "LoRa Alpha" parameters.

The fine-tuning process used the most effective prompt and involved $3,000$ training steps, with a batch size of 2 and gradient accumulation set to 4, effectively processing 6000 instances from the training dataset.

Results from experiments varying the parameters "rank" and "LoRa alpha" are presented in Table 1, while the fine-tuning loss function using "rank=16" and "LoRa alpha=16" is shown in Figure 2.

[2] `https://huggingface.co/unsloth/llama-3-8b-bnb-4bit`

[3] `https://huggingface.co/docs/peft/index`

[4] `https://github.com/unslothai/unsloth`

3.3. Experiment #3 - GPT-4o mini Fine-Tuning Process

To fine-tune the GPT4-o mini model, we used the Azure AI Studio platform [5], leveraging the same training data used in Experiment #2. We adopted the best prompt identified in Experiment #1 and created a JSONL file where each instance of the training set was preceded by the prompt and accompanied by its respective label.

We chose the 2024-07-18 release of GPT4-o-mini, which was the one available for fine-tuning on Azure. After training, the model was deployed on the Azure platform, allowing its use through API calls.

During the evaluation, we noted that utilizing Azure Studio, which incorporates an additional content moderation layer beyond that provided by OpenAI, led to certain moderation inaccuracies. Approximately 1% of the test set was erroneously categorized due to "content moderation errors." For these instances, we assigned the label "non-toxic."

3.4. Results Comparison

For comparative purposes, the most effective prompt identified through the iterative prompt evolution approach is tested with three additional models: Maritaca [6] AI Sabiá3 [Pires et al. 2023], OpenAI GPT-4o mini [OpenAI et al. 2024] [7], and OpenAI ChatGPT 3.5 Turbo [Brown et al. 2020] [8], as well as the BERTimbau model [Souza et al. 2020].

The results presented in Table 2 highlight the significance of fine-tuning the LLaMA 3.1 8B model. Specifically, fine-tuning improved the F1-score from 0.61 to 0.75, demonstrating a substantial performance gain. Furthermore, when applying the fine-tuning methodology using the prompt proposed in [Oliveira et al. 2023], the F1-score reached 0.70. However, our prompt evolution approach further improved this to 0.75, indicating that the refined prompt contributed significantly to the model's performance.

Additionally, the LLaMA 3.1 8B model, despite being fine-tuned with only 3,000 steps and 6,000 instances, performs competitively against other state-of-the-art models like GPT-4o mini, Sabiá3, and BERTimbau. Notably, Sabiá3, a leading model from Maritaca AI, demonstrated comparable accuracy to GPT-4o mini across various high-stakes Brazilian exams, such as OAB, ENEM, and ENADE. These results underscore the effectiveness of our prompt evolution methodology and the potential of smaller models like LLaMA 3.1 8B when paired with efficient fine-tuning techniques.

The results in Table 1 reveal differences in model performance based on the configuration of the "r" (rank) and "LoRa alpha" parameters. The configuration with "r=16" and "alpha=16" achieves the best overall performance, with an F1-Score of 0.75, balancing precision (0.69) and recall (0.83). Increasing "r" to 24 or "alpha" to 24 leads to a marked decline in performance, with the model showing symptoms of overfitting, particularly with a dramatic drop in recall. The configuration with "r=8" and "alpha=16" demonstrates high recall (0.935) but at the cost of precision, indicating a bias towards over-predicting the positive class.

[5] `https://oai.azure.com/portal`

[6] `https://www.maritaca.ai/`

[7] `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`

[8] `https://chat.openai.com/`

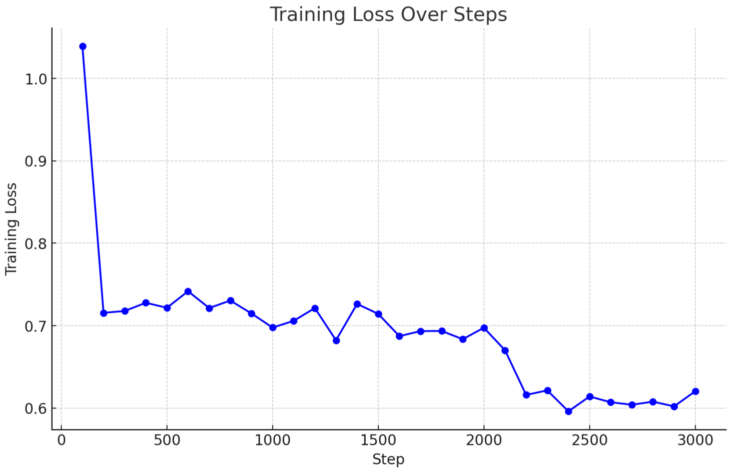

Figure 2. Loss over steps. Params LoRa alplha = 16 and r = 16.

Table 1. Impact of different *r* and LoRa alpha configurations on model performance.

Configuration	F1-Score	Precision	Recall	Accuracy
LLaMA 3.1 8B (r=16, alpha=16)	0.75	0.69	0.83	0.76
LLaMA 3.1 8B (r=24, alpha=16)	0.432	0.601	0.338	0.609
LLaMA 3.1 8B (r=8, alpha=16)	0.727	0.595	0.935	0.690
LLaMA 3.1 8B (r=16, alpha=24)	0.327	0.573	0.229	0.584

Table 2. Comparison of Evaluation Metrics for Different Models

Model	F1-Score	Precision	Recall	Accuracy
LLaMA 3.1 8B (original) w/ best prompt	0.61	0.45	0.96	0.46
LLaMA 3.1 8B (finetuned) w/ prompt from [Oliveira et al. 2023]	0.70	0.71	0.70	0.74
LLaMA 3.1 8B (finetuned) w/ best prompt	0.75	0.69	0.83	0.76
ChatGPT 3.5T Zero-Shot w/ prompt from [Oliveira et al. 2023]	0.73	0.74	0.73	0.74
GPT-4o mini w/ best prompt	0.75	0.75	0.75	0.75
GPT-4o mini (finetuned) w/ best prompt	0.74	0.78	0.74	0.76
Sabiá 3 w/ best prompt	0.75	0.77	0.76	0.75
BERTimbau Finetuned	0.75	0.75	0.75	0.75

4. Conclusion

In this study, we investigated whether a smaller, open-source and quantized language model like LLaMA 3.1 8B 4 bits could effectively perform toxic text detection in Portuguese, particularly when optimized using an iterative prompt evolution approach along with finetune. The experiments demonstrated that, with carefully evolved prompts, the model could achieve competitive performance, even with a limited number of training steps and instances. This highlights the potential of smaller models when paired with efficient prompt engineering techniques.

However, the approach has its limitations. The success of the prompt evolution algorithm heavily depends on the quality of the underlying language models used for the text evolution operations. This reliance can be a significant constraint, as deficiencies in the language models directly affect the quality of the evolved prompts and, consequently, the overall model performance. Further research is needed to address these dependencies and enhance the robustness of the prompt engineering approach.

Acknowledgments

We would like to express our sincere thanks to Blip, whose generous support and invaluable assistance were crucial for the presence of the first author in the event. The authors would also like to thank the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* - Brazil (CAPES) - Finance Code 001, *Fundacão de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG, grants APQ-01518-21), *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq, grants 307151/2022-0, 308400/2022-4), and Universidade Federal de Ouro Preto (PROPPI/UFOP) for supporting the development of this study.

References

BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., and Reddy, S. (2024). Llm2vec: Large language models are secretly powerful text encoders.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

da Rocha Junqueira, J., Junior, C. L., Silva, F. L. V., Côrrea, U. B., and de Freitas, L. A. (2023). Albertina in action: An investigation of its abilities in aspect extraction, hate speech detection, irony detection, and question-answering. In Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, pages 146–155. SBC.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, Advances in Neural Information Processing Systems, volume 36, pages 10088–10115. Curran Associates, Inc.

dos Santos, W. R. and Paraboni, I. (2023). Predição de transtorno depressivo em redes sociais: Bert supervisionado ou chatgpt zero-shot? In Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, pages 11–21. SBC.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., and et al., A. F. (2024). The llama 3 herd of models.

Guo, Q., Wang, R., Guo, J., Li, B., Song, K., Tan, X., Liu, G., Bian, J., and Yang, Y. (2024). Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In The Twelfth International Conference on Learning Representations.

Hammes, L. O. A. and de Freitas, L. A. (2021). Utilizando bertimbau para a classificação de emoções em português. In Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL), pages 56–63. SBC.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In International conference on machine learning, pages 2790–2799. PMLR.

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.

Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., and Ping, W. (2024). Nv-embed: Improved techniques for training llms as generalist embedding models.

Lehman, J., Gordon, J., Jain, S., Ndousse, K., Yeh, C., and Stanley, K. O. (2023). Evolution through large models. In Handbook of Evolutionary Machine Learning, pages 331–366. Springer.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 914–924.

Meyerson, E., Nelson, M. J., Bradley, H., Gaier, A., Moradi, A., Hoover, A. K., and Lehman, J. (2023). Language model crossover: Variation through few-shot prompting. arXiv preprint arXiv:2302.12170.

Oliveira, A. S., Cecote, T. C., Alvarenga, J. P. R., Freitas, V. L. S., and Luz, E. J. S. (2024). Toxic speech detection in Portuguese: A comparative study of large language models. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1, pages 108–116, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Oliveira, A. S., Cecote, T. C., Silva, P. H., Gertrudes, J. C., Freitas, V. L., and Luz, E. J. (2023). How good is chatgpt for detecting hate speech in portuguese? In Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, pages 94–103. SBC.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., and et al., S. A. (2024). Gpt-4 technical report.

Pereira, P. H. and da Silva, T. L. C. (2023). Uso de modelagem de tópicos para agrupamento de notícias: uma abordagem usando bertopic. In Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, pages 406–410. SBC.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. In Naldi, M. C. and Bianchi, R. A. C., editors, Intelligent Systems, pages 226–240, Cham. Springer Nature Switzerland.

Serras, F. and Finger, M. (2021). verbert: Automating brazilian case law document multilabel categorization using bert. In Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, pages 237–246, Porto Alegre, RS, Brasil. SBC.

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.

Zheng, M., Su, X., You, S., Wang, F., Qian, C., Xu, C., and Albanie, S. (2023). Can gpt-4 perform neural architecture search? arXiv preprint arXiv:2304.10970.