Automated Text Identification on Languages of the Iberian Peninsula: LLM and BERT-based Models Aggregation

German Gritsai[1,2,*], Andrey Grabovoy[1]

[1]Advacheck, Tallinn, Estonia
[2]Université Grenoble Alpes (UGA), Grenoble, France

Abstract
This paper describes our solution approach for the IberAuTexTification (Automated Text Identification on Languages of the Iberian Peninsula) competition held as part of the IberLEF 2024 conference. Machine-generated text fragments can be spotted in almost various domains nowadays. The rapid progress of language models and the booming distribution of such texts sometimes confuses human beings. In this article, we present a model for detecting machine-generated fragments based on the aggregation of responses from a large language model BLOOM and two BERT-like encoders Multilingual E5 and XLM-RoBERTa. Given the specificity of the task, namely the presence of the different languages of the Iberian Peninsula, we fine-tuned the distinct models for different subgroups of languages. The method described in the paper helped our team to achieve about 67% for the binary classification dataset with 6 languages in the final competition result.

Keywords
machine-generated text, text classification, transformer-based models, adapters, large language models, fine-tuning, multilungual approach

1. Introduction

Artificial text sequences generation quality is increasing rapidly nowadays [1]. Appearance of GPT architecture [2] and several learning techniques such as RLHF [3], opened a new round of development of large language models. The release of ChatGPT by OpenAI [4] launched a wave of spreading generated fragments on the Internet. We can observe today on a regular scale the appearance of architectures with billions of parameters. A large zoo of LLaMA [5], BLOOM [6], Claude [7], Mistral [8] models are available that handle the task of generating human-like text perfectly. Although the progress of such models is impressive, it poses new challenges for scientists, as the development of these systems implies the emergence and spread of generated fragments in texts of various domains. However, there is a downside: widespread access to these models often leads to the expansion of fake news [9], plagiarism [10] and misinformation. Therefore, the improvement of artificial text detection techniques occurs simultaneously with

IberLEF 2024, September 2024, Valladolid, Spain
*Corresponding author.
gritsai@advacheck.com (G. Gritsai); grabovoy@advacheck.com (A. Grabovoy)
0000-0002-4031-0025 (A. Grabovoy)

the improvement of text generation methods. In order to prevent the gap between quality generation and precision of detection from growing, it is necessary to periodically update and modernise existing detection approaches with newly generated fragments.

The robustness of each detection approach varies from one situation to another, although classifiers with pre-trained models tend to have higher generalisation ability and classification stability under domain changes. Such pre-trained language encoders are also capable of extracting subtle semantic information that is not easily obtained with hand-crafted feature sets and yet is often crucial for natural language understanding and further attribution of authorship. Recently, it has become popular to use large language models themselves for classification cases [11]. Adapter-based fine-tuning methods, such as LoRa [12] and QLoRa [13], allow, with relatively low computational resources, to fine-tune large language models for the required domain. This would be impossible if one had to train all parameters of various billions architectures. The approach of fine-tuning a pre-trained language encoder and training a large language model using lightweight adapters is provided in this paper as part of the IberAuTexTification [14] competition held as part of the IberLEF 2024 [15], which aims to boost research in the area of detecting automatically generated text using text generation models. Last year, we also participated in the AuTexTification [16] machine generation research and solved it only with fine-tuning BERT-like architectures [17].

In this article we perform aggregation of two BERT-based encoders and large language model BLOOM to obtain embeddings for each text in the collection and classify it once. We have done this due to the specificity of the task, we were looking for strengths in different architectures and analysing their capabilities and metrics for each language from the Iberian Peninsula.

2. Task

The IberAuTexTification is the second version of the AuTexTification competition consisted of 2 subtasks.

- Subtask 1 - participants need to determine whether the text has been automatically generated or not;
- Subtask 2 - participants are provided with a text and need to identify which model has generated it;

According to the organisers, the number of models for data collection has expanded this year, adding GPT-4, LLaMA, Anthropic, Falcon [18], MPT [19] and many others. In addition, new domains and languages have been introduced. Now in each task the texts are presented in 6 languages: Spanish, English, Catalan, Gallego, Euskera, and Portuguese.

In this paper we considered the approach for solving the subtask 1 on samples with binary classification in six languages. There is a given dataset $\mathcal{D} = (x_i, y_i)$:

$$x_i = \{x_i^1, \ldots, x_i^m\}, \qquad x_i^j \in \mathcal{W}, \qquad j \in \{1, \ldots, m\}, \qquad y_i \in \{0, 1\},$$

where $\mathcal{W}$ corresponds to all possible strings in the given language. The label $y_i = 1$ corresponds to text that is likely machine-generated, $y_i = 0$ corresponds to human excerpt.

Sample text	Label
No le digas lo que tiene o no tiene, pero sí que puede ayudarte a verle con otros ojos y entender mejor sus sentimientos	*machine*
Talk to your doctor and let him or her know about persistent side effects. Discuss ways to manage them and still get the benefit you need from the medication	*human*

Table 1: Example of raw rows from provided data.

Formally, the task is to find the binary classifier that minimizes an empirical risk on the dataset $\mathcal{D}$:

$$f = \underset{f \in \mathfrak{F}}{\operatorname{argmin}} \sum_{x_i, y_i \in D} [f(x_i) \neq y_i],$$

where $\mathfrak{F}$ is a set of all considered classification models.

3. Dataset

The dataset proposed by the organisers for the training stage consisted of 109,663 examples with the labels ’human’ and ’machine’. According to the authors, the texts are based on different domains, including essays, news, social media, wikipedia, etc. In this way, it will be possible to identify the robustness of the developed algorithm to the style of writing: from more structured and formal to less structured and informal. Examples of generated and human texts are provided in the Table 1.

Note that we split the provided train data into two parts (105,000 and 4,663 samples) in order to use the second part as test data for our approaches. The second part was class balanced and all studies and experiments in the paper were performed on the first part.

Before starting to build the classification algorithm, we decided to analyse the data provided. The texts were nearly balanced in terms of their class, as illustrated in Figure 1. In terms of length statistics, the fragments are rather long. The length corresponds to the size of a small paragraph, which will help to make the tuning of the future detector more accurate, since the length of the input sequence is crucial [20]. The length values by class are shown in the Table 2.

The authors of the competition imposed a restriction on the use of the data: only submitted samples could be used, and no external sources were allowed. As a preprocessing of the data, we performed a minor cleaning in which we removed anomalously short and anomalously long fragments. In addition, since there are 6 languages in the dataset, we decided to clarify the statistics across each language within. Using the CLD3 tool [21], we separated texts by language and calculated the class statistics for them at Fig. 2. It is worth noting that a couple of hundred texts were classified by a language not included in the list of the declared six. These could be either false positives of the language detection tool or rubbish in the data. We decided not to use them when tuning future algorithms. URLs and HTML tags also were extracted and cleaned up using the regular expressions module.

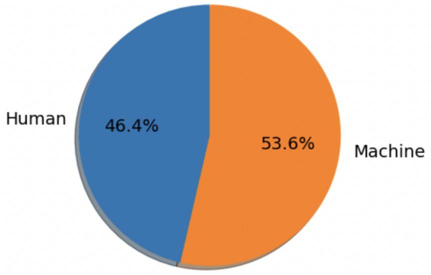

Figure 1: Distribution of classes in provided data.

Data Part	Mean Length	Median Length
Human	1057.0	1117.0
Machine	1036.0	977.0
All	1046.00	995.0

Table 2: Length statistics in provided data.

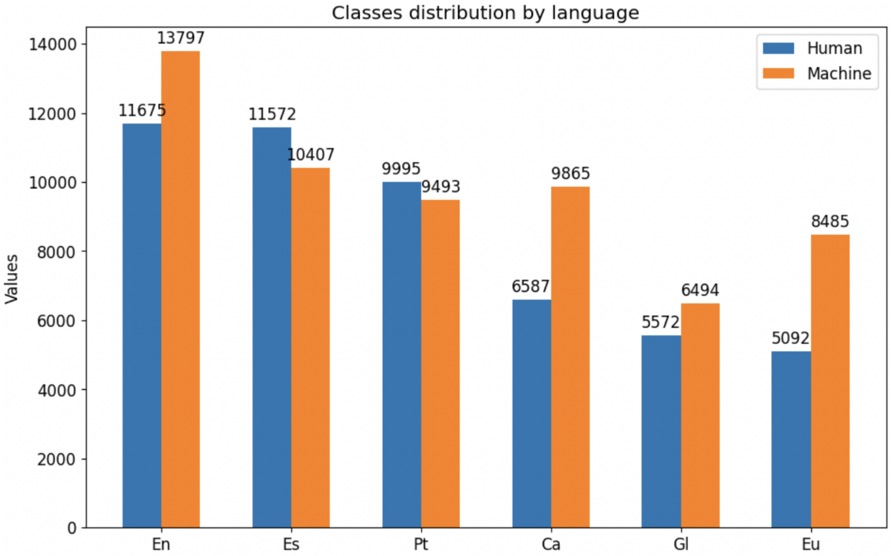

Figure 2: Statistics on the presented classes in language-separated text sequences.

4. Experiments

4.1. Methods Description

Based on a review conducted on the task, we were able to determine the most relevant models for solving the problem within different languages. In recent years, transformer models [22] have been the most frequently used in natural language processing tasks [23]. Their efficiency has been proved by various researches, so in this paper the experiments were carried out with this group of models. Transfer learning is commonly utilized in the implementation of such models. This is an approach in deep learning, where network knowledge from one task is transferred to solve another, related task, thus making it narrowly focused. The fragment embeddings that can be obtained by those models are generally able to have an excellent contextual understanding and may not only be multidomain, but may also be multilingual. We have paid attention to two models that have been specified as baseline methods: XLM-RoBERTa [24] and Multilingual

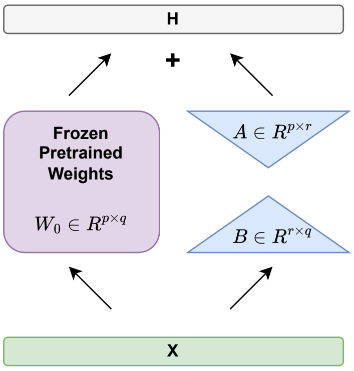

Figure 3: Description of the idea of LoRA Adapter. Let $W_{Updated} = W_0 + \Delta W$, where ΔW contains information about how much we want to update the original weights. For computational learning efficiency, the ΔW matrix is decomposed into two smaller matrices A and B. We have low-rank updates via AB, where the rank is denoted as r, which is a hyperparameter.

E5 [25]. In addition, over the last year we can observe a huge number of publications, where the classification problem is solved using large language models. Having a small amount of computational resources, this became possible with the appearance of lightweight adapters. Quality achieved in tasks with adapters approach has high performance [26]. The basic method for training LoRa adapters is shown in Fig. 3. Therefore, we decided to use not only transformer-based encoders, but also a large language model for fine-tuning. A more detailed description of the selected architectures to solve the task set by the organisers:

1. RoBERTa (Robustly Optimized BERT Pretraining Approach) — has the same architecture as BERT [27], but uses a byte-level BPE as a tokenizer (same as GPT-2 [28]) and uses a different pretraining scheme and optimization features. For this task we used XLM-RoBERTa which is multilingual version of RoBERTa and was pre-trained on 2.5TB of filtered Common Crawl data containing 100 languages. In our own earlier research, we found that the performance of the multilingual version was superior to that of the monolingual version on most tasks. This can be explained by the fact that a multilingual task setting for training a large model helps to improve the quality of the embeddings, thus helping them to achieve greater generalizability.
2. Multilingual E5 — a 24-layer text embedding model with an embedding size of 1024, trained on a mixture of multilingual datasets. This model is initialized from xlm-roberta-large and continually trained on a mixture data. It supports 100 languages from XLM-RoBERTa, but low-resource languages may see performance degradation. For this task we used its large version multilingual-e5-large.
3. BLOOM — the first multilingual LLM trained in complete transparency, the result of the largest collaboration of AI researchers ever involved in a single research project.

> With its 176 billion parameters, BLOOM is able to generate text in 46 natural languages and 13 programming languages. For almost all of them, such as Spanish, French and Arabic, BLOOM was the first language model with over 100B parameters ever created. The architecture of BLOOM is essentially similar to GPT-3 [29]. BLOOM is available in a large number of models, version 7b1 was chosen in our experiments.

XLM-RoBERTa and E5 have been used as encoders for samples. For classification, we have redefined the head in BERT-based architecture that will handle with the [CLS] embeddings at the encoder output. It consisted of 3 fully-connected layers, a GELU [30] activation function and a dropout technique. For BLOOM, we used the QLoRa adapter approach for fine-tuning and afterwards used it as a sequence classifier.

4.2. Approach

When investigating BERT-based models, we found that quality metrics are different for languages with various frequencies. We took into account the information in the Multilingual E5 model description about the performance degradation in low-resource languages, so we used it only for fine-tuning for three languages (English, Spanish and Portuguese). Since the authors claim that XLM-RoBERTa has higher metrics for the low-resource part of the provided data, we used it for additional training on Catalan, Euskera and Gallego. In addition to this, previous studies show that the large language model BLOOM has a good quality of encoding different languages, 5 out of the 6 languages provided in the challenge got on their list. Therefore, we fine-tuned the lightweight adapter for BLOOM on them. What seemed to be an interesting fact is that the Gallego language, which was not declared in BLOOM, despite the fact that no fine-tuning was performed on it, still showed a competitive value of the declared metrics. Thus, we have three fine-tuned models: BLOOM for all languages, XLM-RoBERTa for low-resource Iberian Peninsula languages and Multilingual E5 for the most popular ones.

After several stages of testing different strategies, we came up with settings for the fine-tuning lightweight adapter with PEFT method. The BLOOM model was loaded in 4-bit format using double quantisation, adaptation parameters r = 64 and lora_alpha = 16 were chosen and adaptation was applied to query, key and value matrices. Regarding the pre-training of BERT-like encoders for low-resource languages, the procedure was carried out with the basic parameters described in the E5 release.

We tried many methods of aggregating the answers, but the strongest one was the aggregation of human class labels. Namely, for each test sample, we defined its language and if it is a high-frequency one, we look at the prediction of BLOOM and E5, otherwise BLOOM and XLM-RoBERTa. The joint prediction was done by aggregation, if both the first model and the second model indicate the class of human, then the label of belonging to the class 'human' is placed. Otherwise, the label 'generated' is placed.

In addition, we conducted experiments without a pair of models, but only relying on the responses of one model. However, as the results on the postponed test sample showed - no quality gain was observed.

Model	**BLOOM answers**	
	Without aggregation	***With aggregation***
XLM-RoBERTa	94.38	94.75
Multi E5	94.38	95.15
XLM-RoBERTa and Multi E5	95.89	**96.94**

Table 3: Comparison of the metrics of the models on the postponed data sample. The 'without' column indicates that only the model response specified in the row is used for the languages in which the model was trained, and the BLOOM response is used for the rest of the examples. The 'with' column indicates the aggregation metrics when the model response is combined with the BLOOM response.

4.3. Comparison

The results obtained in the experiment on our test data are presented in Table 3. The metric chosen was F_1-score, the same as in the competition. Aggregation of responses improved the model response space and had a positive effect on quality increase. BLOOM responses aggregated with XLM-RoBERTa and Multilingual E5 responses showed the highest results in terms of chosen metric on the cleaned data. The model with these settings was submitted by our team as a solution to the IberAuTexTification competition, which placed us in the top-14.

5. Conclusion

The paper describes an approach to the issue of machine-generated text detection. We propose a model for artificial text detection based on aggregating the responses of a large language model BLOOM, which has been pre-trained on 5 out of 6 presented languages for detection using lightweight adapters with QLoRa approach and two transformer-based encoders XLM-RoBERTa (trained on Catalan, Gallego, Euskera) and Multilingual E5 (trained on Spanish, English, Portuguese) to obtain embeddings of individual excerpts and further classification. To improve the detection performance, we pre-processed the original training dataset using cleaning. Aggregation performed significantly better with respect to the single responses of each model as well as their joint performance, but without aggregation. The resulting model showed an F_1-score in the final results of the IberAuTexTification competition of about 67% for the binary dataset for the declared six languages.

References

[1] E. Mosca, M. H. I. Abdalla, P. Basso, M. Musumeci, G. Groh, Distinguishing fact from fiction: A benchmark dataset for identifying machine-generated scientific papers in the LLM era., in: A. Ovalle, K.-W. Chang, N. Mehrabi, Y. Pruksachatkun, A. Galystan, J. Dhamala, A. Verma, T. Cao, A. Kumar, R. Gupta (Eds.), Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), Association for Computational Linguistics,

Toronto, Canada, 2023, pp. 190–207. URL: https://aclanthology.org/2023.trustnlp-1.17. doi:10.18653/v1/2023.trustnlp-1.17.

[2] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training (2018).

[3] T. Kaufmann, P. Weng, V. Bengs, E. Hüllermeier, A survey of reinforcement learning from human feedback, 2024. arXiv:2312.14925.

[4] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[6] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).

[7] D. Kevian, U. Syed, X. Guo, A. Havens, G. Dullerud, P. Seiler, L. Qin, B. Hu, Capabilities of large language models in control engineering: A benchmark study on gpt-4, claude 3 opus, and gemini 1.0 ultra, 2024. arXiv:2404.03647.

[8] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. arXiv:2310.06825.

[9] O. Bakhteev, A. Ogaltsov, P. Ostroukhov, Fake News Spreader Detection Using Neural Tweet Aggregation—Notebook for PAN at CLEF 2020, in: CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/.

[10] O. Bakhteev, Y. Chekhovich, A. Grabovoy, G. Gorbachev, T. Gorlenko, K. Grashchenkov, A. Ivakhnenko, A. Kildyakov, A. Khazov, V. Komarnitsky, A. Nikitov, A. Ogaltsov, A. Sakharova, Cross-Language Plagiarism Detection: A Case Study of European Languages Academic Works, Springer International Publishing, Cham, 2022, pp. 143–161. URL: https://doi.org/10.1007/978-3-031-16976-2_9. doi:10.1007/978-3-031-16976-2_9.

[11] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, G. Wang, Text classification via large language models, 2023. arXiv:2305.08377.

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. arXiv:2106.09685.

[13] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, 2023. arXiv:2305.14314.

[14] A. M. Sarvazyan, J. Á. González, F. Rangel, P. Rosso, M. Franco-Salvador, Overview of iberautextification at iberlef 2024: Detection and attribution of machine-generated text on languages of the iberian peninsula, Procesamiento del Lenguaje Natural 73 (2024).

[15] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[16] A. M. Sarvazyan, J. Ángel González, M. Franco-Salvador, F. Rangel, B. Chulvi, P. Rosso, Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains, 2023. arXiv:2309.11285.

[17] G. Gritsay, A. Grabovoy, A. Kildyakov, Y. Chekhovich, Automated text identification: Multilingual transformer-based models approach (2023).

[18] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, Étienne Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, G. Penedo, The falcon series of open language models, 2023. arXiv:2311.16867.

[19] M. N. Team, Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. URL: www.mosaicml.com/blog/mpt-7b, accessed: 2023-05-05.

[20] G. Gritsay, A. Grabovoy, Y. Chekhovich, Automatic detection of machine generated texts: Need more tokens, in: 2022 Ivannikov Memorial Workshop (IVMEM), 2022, pp. 20–26. doi:10.1109/IVMEM57067.2022.9983964.

[21] A. Salcianu, A. Golding, A. Bakalov, C. Alberti, D. Andor, E. Pitler, G. Coppola, J. Riesa, K. Ganchev, M. Ringgaard, N. Hua, R. McDonald, S. Petrov, S. Istrate, T. Koo, Compact language detector v3 (cld3), 2016. URL: https://github.com/google/cld3/, accessed: 2023-05-05.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.

[23] G. Jawahar, M. Abdul-Mageed, L. Lakshmanan, V.S., Automatic detection of machine generated text: A critical survey, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 2296–2309. URL: https://aclanthology.org/2020.coling-main.208. doi:10.18653/v1/2020.coling-main.208.

[24] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. arXiv:1911.02116.

[25] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, 2024. arXiv:2402.05672.

[26] C. Xin, Y. Lu, H. Lin, S. Zhou, H. Zhu, W. Wang, Z. Liu, X. Han, L. Sun, Beyond full fine-tuning: Harnessing the power of LoRA for multi-task instruction tuning, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 2307–2317. URL: https://aclanthology.org/2024.lrec-main.206.

[27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019. URL: https://api.semanticscholar.org/CorpusID:160025533.

[29] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.

[30] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), 2023. arXiv:1606.08415.