Zero-shot Reading Comprehension and Reasoning for Spanish with BERTIN GPT-J-6B

Javier De la Rosa[1,*,†], Andrés Fernández[2,†]

[1]National Library of Norway, Finsetveien 2, 8624 Mo i Rana, Norway
[2]University of Seville, San Fernando, 4, 41004 Seville, Spain

Abstract
The first edition of the IberLEF 2022 shared task on Reading Comprehension and Reasoning Explanation for Spanish (ReCoRES) aimed at selecting correct answers for multiple-choice questions and providing their rationales. In this work, we tested the zero-shot capabilitiews of a custom trained decoder-only model of 6 billion parameters, BERTIN GPT-J-6B. We compared it against classic fine-tuning of encoder-only language models for the task of question answering, and sequence to sequence for the explanation generation task. While the results of our best BERT-based language model were mildly successful in the multiple-choice questions task surpassing random choice and the zero-shot approach, the generation of valid explanations for the answers using a BERTIN GPT-J-6B surpassed a strong fine-tuned sequence to sequence T5 model baseline while requiring no training data at all.

Keywords
Reading comprehension, machine reasoning, zero-shot learning, BERTIN

1. Introduction

Reading Comprehension (RC) is a historically difficult task in Natural Language Processing [1]. The advent of transformers-based models at ever increasing scales has demonstrated that answering multiple-choice questions about a given passage of text is starting to get to human-level performance [2, 3]. However, extracting the reasoning behind a model's choice for an answer is an arguably more challenging task. Different varieties of questions and answers (QA) datasets have been proposed in the literature over the years [4] [5] [6] [7], [8]. Regarding RC, they can be split in two types [9]. The first one is "lookup" datasets, where the explanation for an answer is usually given in the text and the task is then reduced to explain where and how an answer was found in the passage or corpus. We can refer to the second type as "inference", in which some kind of multi-step reasoning is needed to properly find and explain an answer, that is, the explanation structure is not evident from the question.

In this work, we present an approach to generate reasoning explanations for the Reading Comprehension and Reasoning Explanation for Spanish (ReCoRES) shared task [10], which

IberLEF 2022, September 2022, A Coruña, Spain.
*Corresponding author.
†These authors contributed equally.
versae@nb.no (J. De la Rosa); nandezgarcia@gmail.com (A. Fernández)
https://versae.es/ (J. De la Rosa); https://cv.nandezgarcia.com/ (A. Fernández)
0000-0002-9143-5573 (J. De la Rosa); 0000-0002-6958-8995 (A. Fernández)

contains instances of both “lookup” as well as “inference” questions and their reasoning. We show that our zero-shot generation approach outperforms a strong fine-tuned baseline on the BERTScore metric [11].

2. Related work

After the initial efforts of the last two decades on building comprehensive taxonomies and knowledge bases for natural language inference [12, 13, 14], neural-evidence reasoning systems have been proposed with applications on open QA [15], common sense QA [16], fact checking and verification [17], as well as for inferential text generation [18, 19], and multi-hop question answering [20]. Recently, Jhamtani and Clark [21] constructed three datasets aiming at explaining why an chosen answer might be correct. The first one counted over 98,000 annotations explaining the answers to questions with two or more jumps in the inference process. The second dataset, was created by introducing modifications to the first one, preserving its validity, and it was used to test robustness and to generate predictive models capable of making explanations. Finally, the third dataset was also generated based off the first one but adding logical representations, converting the sentences into reasoning chains, something that has gained popularity as chain-of-thought reasoning [3, 22, 23]. Jhamtani and Clark concluded that this last data set, and therefore its representation, was the most robust.

In a recent survey published by Hartmann and Sonntag [24], the authors draw the main conclusion that models trained with human explanations are much more agile and perform better than models trained using labels, because the human explanations help the models focus on the relevant features of the data. In the last year alone, massive language models are starting to provide explanations for multi-hop reasoning datasets using just zero-shot or few-shot learning [3, 22, 23].

3. Dataset

The ReCoRES dataset consists of 1,796 questions from 413 different passages. Most passages have 5 questions each, and the number of questions per passage ranges from 1 to 8. The dataset is split into training, validation, and test sets, with non-overlapping text passages. The distribution is shown in Table 1.

Table 1
Distributions of questions and text passages per split in the dataset.

Split	Questions	Texts
Train	1,047	257
Validation	363	91
Test	386	91

Each sample in the dataset contains information for 9 different attributes as shown below:

- Text: The text passage containing the context of the question.
- Question: The question to answer.
- A, B, C, D, E: The text of five possible answers.
- Answer: The label for the correct answer (A, B, C, D, or E).
- Reason: The text containing the reasoning for the correct answer.

A sample from the dataset is shown in Table 2. Interestingly, the dataset mixes different kinds of questions. In some cases, the question is about the meaning of words in the given context, in others it asks for a summary of the text, the completion of sentences, the main argument of the text, or even inference on hypothetical situations.

Table 2
Random sample from the dataset asking about a possible summary for the given text.

Attribute	Value
Text	Platón le interesaba mucho las matemáticas, porque las relaciones matemáticas jamás cambian. La suma de los ángulos de un triángulo es 1800 siempre. Por lo tanto, es algo sobre lo que debemos tener conocimientos ciertos. Sostenía Platón que sólo podemos tener ideas vagas sobre lo que sentimos, pero sí podemos conseguir conocimientos ciertos sobre aquello que reconocemos con la razón. RUSELL Diccionario de Filosofía
Question	El mejor resumen del texto es
A	El ser humano debe preocuparse por buscar conocimientos ciertos.
B	Platón sostenía que la matemática se sustenta en relaciones invariables.
C	La matemática no puede estar constituida por conceptos imprecisos.
D	El racionalismo de Platón lo llevó a destacar la importancia de la matemática.
E	El conocimiento de la matemática permite que nuestra razón supere la vaguedad.
Answer	D
Reason	El mejor resumen del texto es el racionalismo de Platón lo llevó a destacar la importancia de la matemática. De acuerdo al texto, Platón dice que sólo obtenemos conocimiento cierto de la razón. De ahí que la matemática, al basarse en conocimiento racional, sea una importante disciplina debido a la precisión en los datos obtenidos y su carácter inmutable (que no cambia) de acuerdo con la creencia platónica.

Two sub-tasks are defined from the ReCoRES dataset:

1. **Multiple choice machine reading comprehension** (sub-task 1), in which, given a text, a question, and a set of candidate answers, the task is to select the correct answer.
2. **Reasoning explanation** (sub-task 2), in which, given a text and a question, the task is to generate an explanation for its answer selection.

4. Methods

We tackled the different sub-tasks sequentially, testing different approaches for the sub-task 1, and selecting the best performing method to provide a richer context for the sub-task 2.

In both sub-tasks, we used a custom-built auto-regressive decoder-only model further trained from the GPT-J-6B model weights [25] on the BERTIN corpus of Spanish texts [26]. The `mC4-es-sampled` dataset is a Spanish subset of mC4 [27] sampled using perplexity values up to 50 million documents. The perplexity sampling method used for the creation of the the BERTIN RoBERTa model seems to provide a good trade-off of dataset size versus quality, which might help reduce training times without impacting the resulting models. The BERTIN GPT-J-6B model was finetuned following the original Mesh Transformer Jax code [28]. Details about the hyperparameters used are shown in Table 3.

Table 3
Hyperparameter description for the GPT-J-6B models.

Hyperparameter	Value
$n_{parameters}$	6,053,381,344
n_{layers}†	28
d_{model}	4,096
d_{ff}	16,384
n_{heads}	16
d_{head}	256
n_{ctx}	2,048
n_{vocab}‡	50,257/50,400
Positional Encoding	RoPE [29]
RoPE Dimensions	64

†Each layer consists of one feedforward block and one self attention block.
‡Although the embedding matrix has a size of 50,400, only 50,257 entries are used by the GPT-2 tokenizer [30].

As the original GPT-J-6B model, BERTIN GPT-J-6B consists of 28 layers with a model dimension of 4,096, and a feed-forward dimension of 16,384. The model dimension is split into 16 heads, each with a dimension of 256. Each of the 64 dimensions of each head use Rotary Position Embedding (RoPE) as described in [29]. The model was trained to predict the next token as an auto-regressive language model using a cross-entropy loss. The size of the vocabulary is 50,257 BPE tokens as in GPT-2 and GPT-3. The model was finetuned for 40 billion tokens (40,384,790,528) over 616,000 steps on a single TPUv3-8 VM. In all cases, the generated text used sampling decoding, producing at maximum 75 new tokens, and with top-k value of 50, top-p of 0.95, and temperature of 0.8. Some basic cleaning was performed to remove incomplete sentences and other inconsistencies.

```python
def reason(text, question, answer):
    prompt = f"{text}\n\n{question} {answer} El motivo es que"
    generated = complete_with_gpt(prompt, max_length=75)
    return capitalize_first_word(generated)
```

Figure 1: Python pseudo-code for the construction of the prompt that is fed to the generation model.

4.1. Sub-task 1

Two different approaches were tested for the task of multiple choice machine reading comprehension. First, in a zero-shot setting, we let BERTIN GPT-J-6B complete a prompt that combined both the text and the question. The generated text was then split by sentence and passed to different sentence similarity models to compare each sentence to the candidate answers. The pair (generated sentence, candidate answer) with the better score was selected as the correct answer.

The second approach was a simple fine-tuning of two Spanish RoBERTa models [31], BERTIN [26] and MarIA [32], for 5 epochs with a learning rate of 1e-05.

4.2. Sub-task 2

For the task of reasoning explanation, we also tested two approaches. The first approach was using the generated text from the previous sub-task directly as an explanation. The second approach involved generating new predictions on a prompt that combined the text passage, the question, and the predicted answer from the best method in the sub-task 1. The prompt was constructed as as shown in Figure 1.

5. Results

The evaluation for the sub-task 1 is based on the standard accuracy, i.e., the number of correct answers in relation to the total number of questions, but also on the c@1 [33], a more conservative metric that penalizes the incorrect answers, encouraging systems to not choose an answer unless they are certain. However, since our methods always select one answer among the candidate answers, in this case both metrics have identical values. Hence, we are only reporting accuracy against a baseline where a random answer is chosen.

Table 4 shows the accuracy on the validation set of the different methods. For those using BERTIN GPT-J-6B, sentence-BERT multilingual models [34, 35] were used to compare sentences from the generated text to the candidate answers. Specifically, we used `all-mpnet-base-v2`, a well-round model trained on a large and diverse dataset of over 1 billion training pairs on the basis of MPNet [36], and `paraphrase-multilingual-MiniLM-L12-v2`, trained following a Teacher-Student approach on the basis of MiniLM models [37] and a multilingual corpus of paraphrases in more than 50 languages.

The zero-shot BERTIN GPT-J-6B performed slightly better than random in combination with the `all-mpnet-base-v2` model. However, both fine-tuned models performed better than the zero-shot BERTIN GPT-J-6B for multiple-choice QA, with MarIA scoring the highest at 46.01 accuracy on the validtion set and 40.67 on the test set.

Table 4
Accuracy scores in percentages of the different methods on the validation and test sets for the sub-task 1. Best scores in bold.

Method	Model	Validation	Test
Baseline (random)		20.00	20.00
BERTIN GPT-J-6B +	`all-mpnet-base-v2`	20.39	20.47
	`paraphrase-multilingual-MiniLM-L12-v2`	16.25	17.88
Fine-tune	RoBERTa-base BERTIN	38.84	38.34
	RoBERTa-base MarIA	**46.01**	**40.67**

For the sub-task 2, we report the semantic metric BERTScore [11], using as the base model a multilingual BERT-base language model trained on the top 104 languages with the largest Wikipedia following the standard masked language modeling objective [38]. This metric measures the level of similarity between the generated explanation and its manual reference. Table 5 shows the results of the two different approaches on the validation and test sets. The baseline was obtained by fine-tuning a multilingual T5 model (mT5) [27] for 5 epochs on the training set.

Table 5
BERTScore F1 scores (over 100) of the different methods on the validation and test sets for the sub-task 2. Best scores in bold.

Explanation	Validation	Test
Baseline (mT5)	66.14	65.79
Sub-task 1 generated text	66.55	66.86
Sub-task 2 generated text	**68.19**	**68.67**

As seen in Table 5, generating new explanations off the choices made by MarIA yields the best results surpassing the mT5 baseline by almost 3 F1 points on the test set.

6. Conclusions

In this work, we have tested the zero-shot capabilities on machine reading comprehension and reasoning of a 6 billion parameters decoder-only model trained on Spanish texts from the weights of a training on English content. We complemented the approach with other zero-shot inference-only models. We found that by itself, BERTIN GPT-J-6B does slightly better than random, but heavily underperforms when compared to a simple fine-tuning of monolingual Spanish RoBERTa-base models.

Interestingly, the BERTIN GPT-J-6B model is able to generate explanations 3 F1 points better than a fine-tuned version of a multilingual sequence to sequence model (mT5). This opens up the possibility of generating reasoning explanations using 1-shot and few-shot learning and even chain-of-thought techniques to further improve the performance.

Acknowledgments

This project was made possible thanks to Google's Tensor Research Cloud program, which provided us with TPU credits to train the main BERTIN GPT-J-6B model. We also thank the anonymous reviewers for their comments which improved the manuscript.

References

[1] R. Baradaran, R. Ghiasi, H. Amirkhani, A survey on machine reading comprehension systems, Natural Language Engineering (2020) 1–50.

[2] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, et al., Towards a human-like open-domain chatbot, arXiv preprint arXiv:2001.09977 (2020).

[3] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, Q. Le, Lamda: Language models for dialog applications, 2022. URL: https://arxiv.org/abs/2201.08239. doi:10.48550/ARXIV.2201.08239.

[4] A. Bordes, N. Usunier, S. Chopra, J. Weston, Large-scale simple question answering with memory networks, CoRR abs/1506.02075 (2015). URL: http://arxiv.org/abs/1506.02075. arXiv:1506.02075.

[5] K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15, MIT Press, Cambridge, MA, USA, 2015, p. 1693–1701.

[6] I. V. Serban, A. García-Durán, C. Gulcehre, S. Ahn, S. Chandar, A. Courville, Y. Bengio, Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 588–598. URL: https://aclanthology.org/P16-1056. doi:10.18653/v1/P16-1056.

[7] Y. Su, H. Sun, B. Sadler, M. Srivatsa, I. Gür, Z. Yan, X. Yan, On generating characteristic-rich question sets for QA evaluation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 562–572. URL: https://aclanthology.org/D16-1054. doi:10.18653/v1/D16-1054.

[8] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: Unanswerable questions for SQuAD, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 784–789. URL: https://aclanthology.org/P18-2124. doi:10.18653/v1/P18-2124.

[9] H. Jhamtani, P. Clark, Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 137–150.

[10] M. A. Sobrevilla Cabezudo, D. Diestra, R. López, E. Gomez, A. Oncevay, F. Alva-Manchego, Overview of ReCoRES at IberLEF 2022: Reading Comprehension and Reasoning Explanation for Spanish, Procesamiento del Lenguaje Natural 69 (2022).

[11] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2019.

[12] D. B. Lenat, Cyc: A large-scale investment in knowledge infrastructure, Commun. ACM 38 (1995) 33–38. URL: https://doi.org/10.1145/219717.219745. doi:10.1145/219717.219745.

[13] R. Speer, C. Havasi, et al., Representing general relational knowledge in conceptnet 5., in: LREC, volume 2012, 2012, pp. 3679–86.

[14] E. Davis, G. Marcus, Commonsense reasoning and commonsense knowledge in artificial intelligence, Communications of the ACM 58 (2015) 92–103.

[15] D. Chen, W.-t. Yih, Open-domain question answering, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, Association for Computational Linguistics, Online, 2020, pp. 34–37. URL: https://aclanthology.org/2020.acl-tutorials.8. doi:10.18653/v1/2020.acl-tutorials.8.

[16] A. Talmor, J. Herzig, N. Lourie, J. Berant, CommonsenseQA: A question answering challenge targeting commonsense knowledge, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4149–4158. URL: https://aclanthology.org/N19-1421. doi:10.18653/v1/N19-1421.

[17] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://aclanthology.org/N18-1074. doi:10.18653/v1/N18-1074.

[18] H. Rashkin, M. Sap, E. Allaway, N. A. Smith, Y. Choi, Event2Mind: Commonsense inference on events, intents, and reactions, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 463–473. URL:

https://aclanthology.org/P18-1043. doi:10.18653/v1/P18-1043.
[19] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, Y. Choi, Atomic: An atlas of machine commonsense for if-then reasoning, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 3027–3035.
[20] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C. D. Manning, HotpotQA: A dataset for diverse, explainable multi-hop question answering, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2369–2380. URL: https://aclanthology.org/D18-1259. doi:10.18653/v1/D18-1259.
[21] H. Jhamtani, P. Clark, Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering, CoRR abs/2010.03274 (2020). URL: https://arxiv.org/abs/2010.03274. arXiv:2010.03274.
[22] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., Palm: Scaling language modeling with pathways, arXiv preprint arXiv:2204.02311 (2022).
[23] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al., Training compute-optimal large language models, arXiv preprint arXiv:2203.15556 (2022).
[24] M. Hartmann, D. Sonntag, A survey on improving NLP models with human explanations, in: Proceedings of the First Workshop on Learning with Natural Language Supervision, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 40–47. URL: https://aclanthology.org/2022.lnls-1.5. doi:10.18653/v1/2022.lnls-1.5.
[25] B. Wang, A. Komatsuzaki, GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model, https://github.com/kingoflolz/mesh-transformer-jax, 2021.
[26] J. De la Rosa, E. G. Ponferrada, M. Romero, P. Villegas, P. G. de Prado Salas, M. Grandury, BERTIN: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23.
[27] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 483–498.
[28] B. Wang, Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX, https://github.com/kingoflolz/mesh-transformer-jax, 2021.
[29] J. Su, Y. Lu, S. Pan, B. Wen, Y. Liu, Roformer: Enhanced transformer with rotary position embedding, arXiv preprint arXiv:2104.09864 (2021).
[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[32] A. Gutiérrez Fandiño, J. Armengol Estapé, M. Pàmies, J. Llop Palao, J. Silveira Ocampo, C. Pio Carrino, C. Armentano Oller, C. Rodriguez Penagos, A. Gonzalez Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022).
[33] A. Peñas, Á. Rodrigo, A simple measure to assess non-response, in: Proceedings of the

49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 1415–1424.
[34] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.
[35] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020. URL: https://arxiv.org/abs/2004.09813.
[36] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, arXiv preprint arXiv:2004.09297 (2020).
[37] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. arXiv:2002.10957.
[38] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.
[39] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, C. Leahy, The Pile: An 800gb dataset of diverse text for language modeling, arXiv preprint arXiv:2101.00027 (2020).

A. Limitations and Biases

As the original GPT-J model, the core functionality of BERTIN-GPT-J-6B is taking a string of text and predicting the next token. While language models are widely used for tasks other than this, there are a lot of unknowns with this work. When prompting BERTIN GPT-J-6B it is important to remember that the statistically most likely next token is often not the token that produces the most "accurate" text. Never depend upon BERTIN GPT-J-6B to produce factually accurate output.

The original GPT-J was trained on the Pile, a dataset known to contain profanity, lewd, and otherwise abrasive language. Depending upon use case GPT-J may produce socially unacceptable text. See Sections 5 and 6 of the Pile paper [39] for a more detailed analysis of the biases in the Pile. A fine-grained analysis of the bias contained in the corpus used for fine-tuning is still pending, although some preliminary remarks are given in the BERTIN paper [26].

As with all language models, it is hard to predict in advance how BERTIN GPT-J-6B will respond to particular prompts and offensive content may occur without warning. We recommend having a human curate or filter the outputs before releasing them, both to censor undesirable content and to improve the quality of the results.

B. Availability

The BERTIN GPT-J-6B model is free and openly available at https://huggingface.co/bertin-project/bertin-gpt-j-6B. A demo of the model can also be found at https://huggingface.co/spaces/bertin-project/bertin-gpt-j-6B.