Performance evaluation of BERTimbau/GPT-3 based Language Models in solving ENEM

Desnes Nunes (desnesn@usp.br) e Ricardo Primi (rprimi@mac.com)

December 12, 2022

Abstract

This paper focus on examining the potential innovation language models in solving multiple choices high-stakes test *Exame Nacional do Ensino Médio* (ENEM), a multidisciplinary entrance exam widely used on Brazilian universities. Since ENEM questions can involve more than one field of knowledge, i.e., a question may require knowledge of both statistics and biology to be solved, this makes this task very interesting to be tackled by Portuguese Language Models (LMs). We analyzed the data of 916 questions administered in years 2010 to 2017 from ENEM Challenge `https://www.ime.usp.br/~ddm/project/enem/`. A number of models were tested based on BERT embeddings (*Bidirectional Encoder Representations from Transformers*, BERTimbau available in portuguese), fine-tuned, and GPT3 (*Generative Pre-Training Transformer 3* with few shots and "chain of thought" training regimens. Overall, we found an accuracy score of .77 and an F1 score of .77 for the GPT3 models, which represents a new state of the art for ENEM challenge.

1 Introduction

Needless to say, recent innovations presented not only on LMs applications, but also to whole field of Machine Learning (ML) have been astonishing and seem to have an almost unlimited application scope. This of course correlates strongly to digital aspects present of modern life; such as banking, chat bots, meal ordering, music listening, meeting people, among many other services. These applications rely heavily on ML processing of customer's data, in order to make Artificial Inteligence (AI) models able provide automatic suggestions, take decisions and more. Moreover, an area that started to receive some attention recently from ML applications is education; which recently saw, for instance, works aiming LMs usage in classrooms [3], as well as the so called *precision education* [4].

On this context, this paper explores the field of Q&A tackling enhancements of Portuguese trained LMs, in particular, with the usage of state-of-the-art transformer-based models [9] such as BERT [2] and GPT-3 [1]. These models were used here on a series of experiments to answer multiple choice questions from the multidisciplinary *Exame Nacional do Ensino Médio* (ENEM), which is an important Brazilian university entrance exam used nationwide.

The proposed research aims to advance the benchmarks for automatically solving college entrance exam by exploring recent state of the art language models for enhancing the performance of AI on this task answering the call **ENEM Challenge** `https://www.ime.usp.br/~ddm/project/enem/`. ENEM's purely textual multiple-choice questions pose a challenging problem, requiring advanced natural language processing skills. The problem has been examined in earlier research [7] using static word embeddings and WordNet but with limited success attaining an accuracy of 26% to 29% of correct responses. Hence, since the dataset of this work has been gratefully made openly available through the ENEM Challenge, this paper proposes to tackle this task with the use of contextual embedding of BERTimbau and the few-shot capabilities of GPT3, since our main hypothesis is that these models can outperform previous studies. By doing so, we aim to improve the accuracy of our model and demonstrate the potential of AI to assist in the education process. The study is important in demonstrating the potential of AI to be used in educational tests, as well as to assess the accuracy and reliability of AI-based solutions.

2 Methodology

2.1 Data set

The dataset published by [11] provides of of its data in xml format, which has been derived from parsing real questions and alternatives that were presented on the ENEM over a few years. Moreover, the authors have labeled each question with the following knowledge tags:

- Text Comprehension (TC)
- Encyclopedic Knowledge (EK)
- Image Comprehension (IC)
- Domain Specific Knowledge (DS)
- Mathematical Reasoning (MR)

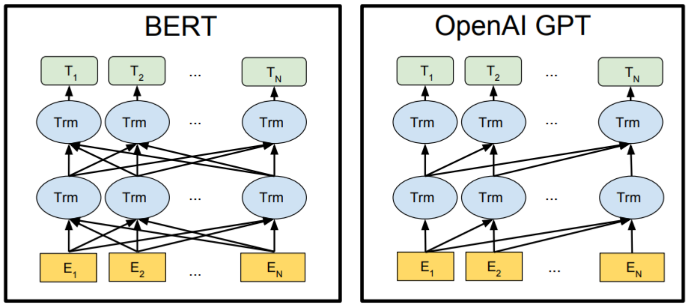
These tags have been extremely helpful since they inform if a question has an image or mention chemical elements that can't be treated as text. The entire dataset consists of 1754 items from the ENEM language and human sciences tests from 2010 to 2017. After eliminating questions that require understanding of images (IC), reading of symbols of chemical elements (CE) and mathematical reasoning (MR), the final database used here is made of 916 questions. Each item is composed of a **header** where the main text of the item is presented; a **statement** where the question is asked to the students and **alternatives**, containing five options from which one correctly answer the question (statement).

For the fine tuning experiment with BERTimbau we split items randomly into a training set (733 questions) and a test set (183 questions). Three items were selected to create prompts and chain of thought reasoning expected for the solution to use in the few-shot experiments.

2.2 Transformer Models

The models used on this work (BERT and GPT3) are based on the attention mechanism of the transformers architecture which is illustrated on Figure 1.

Figure 1: Illustration of BERT and GPU models architecture



Source: [10]

The core mechanism of transformer's encoder and decoder is called *attention*. Upon receiving tokens in continuous distributed representations (*embeddings*), the encoder adds a representation of the token's position within the sequence. After that, a covariance matrix is calculated between all the vectors of the tokens (illustrated as ellipses in the Figure 1). This matrix stores how similar a token t called *query* is to a context called *keys* (including itself). Afterwards, this matrix is converted into weights, called *attention scores*, which range from 0 to 1 after going through a softmax function. With the use of these weights, a new representation of each token called *value* is computed based on a weighted average of the embeddings of all tokens in the sequence. Thus, if the vectors of two tokens are eventually more similar, they will be recalculated by a weighted average of both. As opposed to this, if two tokens have zero correlation, they will remain unchanged. The transformation of the original vectors of each token enables the recreation of the vectors based on the similarity with those of the context tokens, thus enabling contextual representations - and this is the core attention mechanisms [6].

These representations are more powerful than the static word embeddings [5] used by the previous work [7] which tackled portuguese Q&A for ENEM question solving; therefore, we expect improved performance with the use of models which rely heavily on the attention mechanism, such as BERT and GPT-3.

Also, one important aspect related to GPT3 that has been started to be tackled here is the enhancement of few-shots prompts techniques with CoT [11]. By adding to the prompts a series of intermediate reasoning steps, it is reasonable to believe that this would allow the model to break down complex problems into smaller, more manageable steps, in order to reason each step individually before compiling the results that derive the final answer. In this way, the model can potentially grasp abstract concepts inasmuch manipulate them, in order to solve problems that require deep understanding of underlying principles. This technique may improve large language models' ability to perform complex reasoning. A future work that started to be created here questions whether the usage of CoT in large LMs would enhance their ability to perform complex reasoning tasks, like answering ENEM questions, making them more effective in a variety of applications.

2.3 Design and models

This work made use use of four different approaches, which are more thoroughly explained hereafter.

2.3.1 BERtimbau embeddings

With the use of [8], the first experiment performed a forward pass for each ENEM item component (header, statement, and five alternatives), saving layer 12 embeddings for each component. An overall embedding was created by averaging each token embedding that makes up each component. Afterwards, it was calculated the cosine similarity between each header/statement and alternative, and then selected the one with the highest similarity as the correct alternative. In addition, the embeddings of the [CLS] special tokens were saved in order to enable the algorithm to select the alternative whose embeddings were most similar to the embeddings of the [CLS] question.

2.3.2 BERtimbau fine tuning

Here five examples were created for each item by concatenating the question's header, statement, and alternatives. The special [SEP] token between the header/statement and the alternative text was also specified. Therefore, in the BERT model, the first segment consists of the header/statement text and the second segment consists of the alternative text. Each example had a target label of 1 if it was paired with the correct alternative and 0 otherwise. Hence, there are five instances of headers/statements with one alternative for each, which creates an unbalanced set of examples, since only 20% will be correct and 80% will be incorrect. Due to these circumstances, the correct examples were over-sampled by repeating them four times on the training set. This model was started with pre-trained BERTimbau weights, which latter was trained on the test set for 3 epochs, with learning rate set on $lr = 5e - 05$.

2.3.3 GPT3 with 0-shot / few shot

By utilizing the so called 0-shot learning strategy on GPT-3, an algorithm was developed to input and analyze multiple queries into the model; which contained the ENEM question and its alternatives. The model was found to perform better which this setup. Latter, the few-shots was assembled to make the model also receive three different examples, with all alternatives and correct answer before before the question that was being asked.

Token generation was not requested from the model. Rather, it was analyzed the likelihood of each token among the alternatives. By aggregating the log probabilities of each token that comprises the alternative, we calculate the probability of each alternative as follows:

$$probability(alternative_a) = e^{\left(\sum_i^N logprobs[i]/N\right)} \tag{1}$$

where a is the alternative - 1 to 5, i is the logprob of a token and N is the total number of tokens

Moreover, the following parameters on the GPT3 API were used: item temperature = 0, max_tokens = 1, top_p = 1, frequency_penalty=0, presence_penalty=0, echo=True and logprobs=5.

2.3.4 GPT3 with chain of thought

In this model, the only difference from the previous one is that the prompt contained an explanation of why the correct alternative was selected. After that, the model was asked to generate the correct alternative and explanations.

3 Experiments

3.1 Metrics

The global performance of a model should be evaluated with the use of a global accuracy score and the F1 score by combining **precision** (the proportion of true correct alternatives out of true positives plus false positives) and **recall** (the proportion of true correct alternatives out of all correct alternatives, that is, true positives divided by true positives plus false negatives) into one metric by taking their harmonic mean. A few experiments here were able to outperform global accuracy way above 29%, which was the baseline (reported in [7]) for this work.

3.2 Results

The experiments using the methodology described above were developed in python and R language. Code and data are available the following github repository. Futhermore, Table 1 present the main results of our evaluation of the experiments.

Table 1 - Results of the four models on solving ENEM questions

Model	**Accuracy**	**F1**
BERTIMbau Embeddings		
Heading	20%	0.19
Statement	21%	0.21
[CLS]	20%	0.19
BERTIMbau Finetuned	57%	0.41
GPT3		
0-shot with davinci-002	77%	0.77
0-shot with davinci-003	77%	0.77
CoT with davinci-002	Not completed	Not completed
CoT with davinci-003	Not completed	Not completed

3.3 Issues

A few issues that didn't allow us to complete all the experiments in all models in the time frame of the assignment are as follows:

- Most issues of the code arose with parsing the results from the OpenAI model, specially when dealing with multiple shots on GPT-3. On the reply of the model, getting the right position of the tokens of the alternatives in order to also get the position of the logprobs was very tricky. Furthermore, on the CoT experiments the model's reply hanged from time to time (sometimes the model generated an alternative, other times just a latter not properly closed with the
n string terminator).
- The CoT examples that were presented to the few-shots model need more investigation. Instead of the same example for all areas, maybe relate them to the type of question that is being answered?

The authors plan to address these issues in future works.

4 Conclusion

This paper researched enhancements on the field of Q&A in Portuguese-BR with the usage of state-of-the-art AI models. With the use of the pre-trained BERT and GPT-3 models, this work was able to reach significant results on answering multiple choices questions from the ENEM, a widely multidisciplinary test used for brazilian universities acceptance.

This research is relevant for several reasons. First, it provides a novel use of BERT and GPT-3 in the educational domain, highlighting their potential in guiding future developments in the use of AI in education. It can be easily assumed that what is being proposed here can serve as a beacon for educational applications such as automatic scoring of achievement tests, automatic writing feedback for assessments, and the adoption of AI-powered tools to assist educators in developing adaptive tests and adaptive learning programs. Second, the findings of this research could also have significant consequences for the education sector, potentially leading to the adoption of AI-powered tools for psychometric analysis of ENEM exams, including predicting the difficulty of questions, developing new items, and ultimately assisting in the creation of adaptive ENEM test-suites. Last, but certainly not least, it also contributes to the growing body of research that attempts to understand the intelligence underlying the language models. Are they capable of understanding? Can they justify their answers by providing human-like explanations of their reasoning that humans consider intelligent?

5 Future Work

For starters, future work would certainly involve the conclusion of the experiments with BERTIMbau and few-shots with CoT. Furthermore, an idea brainstormed during this work would involve research on the development of questions within a certain knowledge base, which seems fair to assumed that might have good results with the usage of the GPT-3 and T5 models, for instance.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Danijel Kučak, Vedran Juričić, and Goran ambić. Machine learning in education-a survey of current research trends. *Annals of DAAAM & Proceedings*, 29, 2018.

[4] Hui Luan and Chin-Chung Tsai. A review of using machine learning approaches for precision education. *Educational Technology & Society*, 24(1):250–266, 2021.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space.

[6] Peltarion. How to get meaning from text with language model bert — ai explained.

[7] I. Cataneo Silveira and D. Deratani Maua. Advances in automatically solving the enem. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 43–48, Los Alamitos, CA, USA, oct 2018. IEEE Computer Society.

[8] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need.

[10] Ran Wang, Haibo Su, Chunye Wang, Kailin Ji, and Jupeng Ding. To tune or not to tune? how about the best of both worlds?, 2019.

[11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022.