Evaluating the Performance of Language Models solving ENEM: A Study of BERTimbau and GPT-3

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

December 10, 2022

Abstract

This paper examines the potential of artificial intelligence, specifically Transformer models such as BERT and GPT-3, in answering the Brazilian National High School Examination (ENEM). 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 BERTimbau embeddings, finetuned, and GPT3 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 the digital aspects present of modern life; such as banking, chat bots, meal ordering, music listening, meeting people, among many other services. Many of these 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. Furthermore, an area that started to receive some attention recently from ML applications is education; which recently saw, for instance, works aiming ML usage in classrooms [3], as well as the so called *precision education* [4].

The paper we present explores the potential of using artificial intelligence, particularly Transformer models [8] such as BERT [2] and GPT-3 [1], two state-of-the-art language processing models to answer the Brazilian National High School Examination (ENEM). ENEM is an important Brazilian college entrance examination, which evaluates the academic performance of high school students. The problem we aim to address is whether these models, which have achieved impressive results on a variety of natural language processing tasks, can also

perform well on a standardized educational exam. Our main hypothesis is that BERT and GPT-3 will be able to accurately answer a significant number of ENEM questions, demonstrating their potential to assist in the education process.

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 artificial intelligence in education, 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 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 tests. Third, it 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?

This paper presents a series of experiments investigating if BERT and GPT3 can answer ENEM questions. One important aspect related to GPT3 is the enhancement of few-shots prompt engineering techniques with *chain of thought* [10]. The idea is to enhance prompts that presents a series of intermediate reasoning steps allowing the model to break down a complex problem into smaller, more manageable steps, and to reason about each step individually before combining the results to arrive at a final answer. In this way, the model can potentially grasp abstract concepts and manipulate them, as well as solve problems that require deep understanding of underlying principles. This technique may improve large language models' ability to perform complex reasoning. Our main question is whether the use of a chain of thought in large language models enhance their ability to perform complex reasoning tasks, like answering ENEM questions, making them more effective in a variety of applications.

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 [6] using static word embeddings and WordNet but with limited success attaining an accuracy of 26% to 29% of correct responses. This paper proposes to extend this approach by utilizing contextual embedding of BERTimbau and few-shot capabilities of GPT3. 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 is openly available and has been published by [10]. The data presented in xml format is the result of the parsing of the real questions and alternatives presented on the ENEM tests from 2010 - 2017. 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 had 916 items.

Each item is composed of a **header** where the main text of the item is presented; **statement** where the question is asked to the students and **alternatives**, containing five options from which one correctly answer the question (statement).

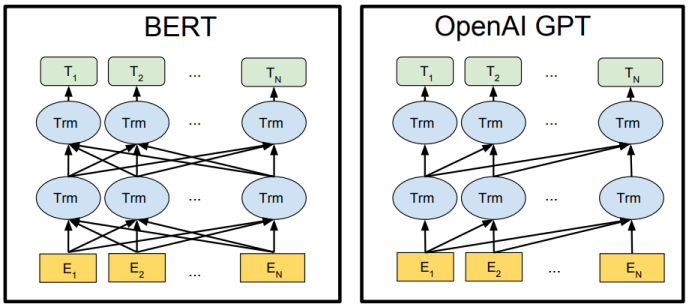
For the fine tuning experiment with BERTimbau we split items randomly into a training set (733 questions) and a test set (183 questions). Three items were selected to create prompts and chain of thought reasoning expected for the solution to use in the few-shot experiments.

2.2 Transformer Models

The models used on this work (BERT and GPT3) are based on the attention mechanism of the transformers architecture and can be illustrated as follows:

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 ovals in the Figure 4). This matrix stores how similar a token t called Query is to a context called Keys (including itself). This matrix is converted into weights, called attention scores, from 0 to 1 by using the softmax function. Using these weights, a new representation of each token called Value is computed based on a weighted average of 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.

Figure 1: Illustration of BERT and GPU models architecture



Source: [9]

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. This is the core attention mechanisms. This is explained very clearly here: `https://www.youtube.com/watch?v=-9vVhYEXeyQ)`. These representations are more powerful than the static word embeddings [5] used by [6] therefore we expect better improved performance when representing ENEM questions.

2.3 Design and models

We tested four models:

BERtimbau embeddings [7]: We perform 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. We then calculated the cosine similarity between each header/statement and alternative, and selected the one with the highest similarity as the correct alternative. In addition, we saved the embeddings of the [CLS] special tokens and selected the alternative whose embedding was most similar to the embeddings of the [CLS] question.

BERtimbau fine tuning: We created five examples for each item by concatenating its header, statement, and alternatives. We have specified a special [SEP] token between the header/statement and the alternative text. 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. There are therefore five instances of headers/statements with one alternative for each. Therefore, we will have an unbalanced set of examples since only 20% will be correct and 80% will be incorrect. We therefore over-sampled the correct examples by repeating them four times on the training set. We started this model with pre-trained BERTimbau wheigths and trained on the test set for 3 epochs with learning rate on the test set for 3 epochs with learning rate $lr = 5e - 05$.

GPT3 with few shot: By utilizing the few-shot learning regime of the

GPT-3, an algorithm was developed to input and analyze multiple queries into the model that contained the question and its alternative answers. We provided ENEM items as inputs for GPT3 and analyzed its results. We set up three examples as a prompt and then presented an item with it's five alternatives. The model was found to perform better when it received five inputs per question, for example, five queries containing the question, all alternatives, and which alternative was correct. Token generation was not requested from the model. Rather, we analyze 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 (where a is the alternative - 1 to 5, i is the token and N is the total number of tokens):

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

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

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, we asked for the generation of the correct alternative and explanations.

The global performance of the model was evaluated using the 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 true correct alternatives out of those classified as correct by the model) into one metric by taking their harmonic mean. We expect to achieve global accuracy above 29%, which was the baseline reported in [6].

3 Experiments

The experiments using the methodology described above were developed in python and R language. Code and data are available in github repository `https://github.com/rprimi/enem_challenge`. Table 1 present the main results of our evaluation of the experiments.

Model	**Accuracy**	**F1**
BERTIMbau Embeddings		
Heading	20%	0.19
Statement	21%	0.21
[CLS]	20%	0.19
BERTIMbau Finetuned	57%	0.41
GPT3		
Few shot davinci-002	77%	0.77
Few shot davinci-003	77%	0.77
Chain of Thought davinci-002		
Chain of Thought davinci-002		

4 Conclusion

5 Future Work

A future work would certainly be focused on the development of questions with a certain knowledge base with the usage of LMs such as GPT-3 and T5, 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] 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.

[7] 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.

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

[9] 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.

[10] 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.