Is GPT-3 smarter than a sixth-grader?

Anitej (Tej) Suklikar

9/18/22

Abstract

I work with the Davinci model in GPT-3, an autoregressive language model, to answer middle-school science textbook questions in the textbook question answering (TQA) dataset. The dataset was split into training, test, and validation sets. In this task, I simulate a student taking a test in three different scenarios. First, the Zero-Shot-Learning experiment where I only provide the model with the questions from a specific lesson. This would be the equivalent of a student going into a test without studying as the model has not gathered any knowledge from the lesson. Second, the Few-Shot-Learning experiment, where I provide the model with specific lesson content from the textbook and the corresponding questions. This equates to a student skimming over the lesson content before taking the test. Lastly, I fine-tuned the Davinci model on some of the textbook questions and then fed it questions. This is similar to a student doing a thorough review of the material before taking the test. After conducting all three experiments, I compare their accuracies and in doing so, highlight the "intelligence" and limitations of GPT-3.

1. Introduction

Question answering (QA) and Large Language models (LLM) have been a major research focus in Artificial Intelligence for several years. In 2017, a task called Textbook Question Answering (TQA) was introduced. The task included lessons from a middle school science textbook consisting of texts, diagrams, and natural questions. Many people attempted to create question answer models but reported sub-par accuracies. The reason for the lower accuracies is because TQA is more complex and more realistic than other question answer datasets like SquAD. After analyzing how many of these models failed to succeed on TQA, I wanted to see how the Davinci model within GPT-3 would perform. Generative Pre-trained Transformer (GPT-3) is a neural network machine learning model developed by OpenAI. As of early 2021, GPT-3 was the largest neural network with over 175 billion machine learning parameters. After gathering all the data I needed to conduct my three experiments, I ran the model in Google Colab using the completion API and analyzed the accuracy in each instance. Prompts are how you instruct GPT-3 to do what you want. It's like programming, but with plain English instructions. Rather than writing code, you use words and plain text. When you're writing prompts, the main thing to keep in mind is that GPT-3 is trying to figure out which text should come next, so including instructions and/or examples provides context that helps the model figure out the best possible completion.

2. Background

As previously stated, TQA was a task proposed in 2017 and many people attempted to create an accurate model. The TQA dataset includes both text questions and visual questions. In this paper, I will focus on the text questions. The two most accurate models achieved 42.8% and 54.11% accuracy for text questions, highlighting the complexity of the task. For many of the questions, the answer is not specifically stated in the text so QA models such as roBERTa which look for specific text or phrases would not perform well. In order to answer these questions correctly, the

model must be quite knowledgeable, especially in the Zero-Shot-Learning experiment where no context is provided. This is where GPT-3 comes into play. Since it has been trained with almost all available data from the Internet, it possesses a great deal of knowledge. Davinci is the most capable model. Due to its capability, I pondered on the question, “How well would the model perform if it only received the questions, compared to when it received context and questions, and to when it is trained on similar questions?”

3. Dataset

I used the Textbook Question Answering dataset from Allen Institute for AI. The dataset was split into a training, test, and validation set at the lesson level. The training set consists of 666 lessons and 15,154 questions, the validation set consists of 200 lessons and 5,309 questions and the test set consists of 210 lessons and 5,797 questions. For the zero-shot and few-shot experiments, I only used the validation set. The validation dataset was given to me in JavaScript Object Notation (JSON). There was a great amount of text, pictures, and diagrams in the JSON. First, I created a pandas dataframe which had a consolidated lesson table with one row per lesson, and all the associated content. The data was organized by Lesson ID. Afterwards, I created a Question Answer Table with the Questions, Answer Choices, and Correct Answer per row. Additionally the associated Lesson ID was also stored for looking up and joining to the Lesson Table content. This allowed me to visualize and understand the data in an effective manner. From there, I had to figure out how to organize the data in a way so I could feed it to the Davinci model. For the Zero Shot Learning experiment, I built a Prompt Table with just the prompt, Question plus Answer Choices, again with one lesson per row. I created four data frames, fewShotPromptTable, fewShotAnswerKey, zeroShotPromptTable, and zeroShotAnswerKeyTable. For the Few Shot Learning experiment, I built a Prompt Table that combined the Lesson Content and Question plus Answer Choices in a single string, with one lesson per row. The fewShotPromptTable and zeroShotPromptTable were fed into the Davinci model and I used the fewShotAnswerKey and zeroShotAnswerKeyTable to check the accuracy of the model. For the fine-tuned model, I created a few data tables that I will discuss in the methods section of the paper.

zeroShotPromptTable for the first lesson

Lesson ID	Prompt
L_0007	Answer the following questions by picking one of the choices provided. Only include the letter of the answer choice listed. Questions: 1. Gravity causes erosion by all of the following except a. glaciers.; b. moving air.; c. flowing water.; d. mass movement. 2. The rate of erosion by gravity a. is sudden and dramatic; b. is very slow over long periods of time; c. neither of these; d. both of these 3. Factors that increase the risk of landslides include a. dry soils.; b. lack of rain.; c. earthquakes.; d. two of the above 4. When a rock falls from a cliff face, the agent of erosion is usually a. wind; b. water; c. gravity; d. glaciers 5. Downhill creep a. results in curved tree trunks; b. falls as a whole unit; c. leaves large scars in the hillside; d. cannot be noticed because it is so slow 6. Mass movement can occur a. suddenly.; b. very slowly.; c. only on sloping land.; d. all of the above 7. Slump may be caused by a. wet clay.; b. water erosion.; c. scars on a hillside.; d. two of the above 8. A slump is the sudden a. fall of rock and soil down slope; b. flow of mud down slope; c. movement of a large block of rock and soil down slope; d. flow of volcanic ash and water down slope 9. Creep usually takes place where the ground a. is level.; b. is prevented from moving.; c. freezes and thaws frequently.; d. is always saturated with water.

fewShotPromptTable for the first lesson

Lesson ID	Prompt
L_0007	Use the lesson text below to answer the following questions by picking one of the choices provided. Only include the letter of the answer choice listed. Lesson: The most destructive types of mass movement are landslides and mudslides. Both occur suddenly. A landslide happens when a large amount of soil and rock suddenly falls down a slope because of gravity. You can see an example in Figure 10.30. A landslide can be very destructive. It may bury or carry away entire villages. A landslide is more likely if the soil has become wet from heavy rains. The wet soil becomes slippery and heavy. Earthquakes often trigger landslides. The shaking ground causes soil and rocks to break loose and start sliding. If a landslide flows into a body of water, it may cause a huge wave called a tsunami. A mudslide is the sudden flow of mud down a slope because of gravity. Mudslides occur where the soil is mostly clay. Like landslides, mudslides usually occur when the soil is wet. Wet clay forms very slippery mud that slides easily. You can see an example of a mudslide in Figure 10.31. Two other types of mass movement are slump and creep. Both may move a lot of soil and rock. However, they usually aren't as destructive as landslides and mudslides. Slump is the sudden movement of large blocks of rock and soil down a slope. You can see how it happens in Figure 10.32. All the material moves together in big chunks. Slump may be caused by a layer of slippery, wet clay underneath the rock and soil on a hillside. Or it may occur when a river undercuts a slope. Slump leaves behind crescent-shaped scars on the hillside. Creep is the very slow movement of rock and soil down a hillside. Creep occurs so slowly you can't see it happening. You can only see the effects of creep after years of movement. This is illustrated in Figure 10.33. The slowly moving ground causes trees, fence posts, and other structures on the surface to tilt downhill. Creep usually takes place where the ground freezes and thaws frequently. Soil and rock particles are lifted up when the ground freezes. When the ground thaws, the particles settle down again. Each time they settle down, they move a tiny bit farther down the slope because of gravity. Questions: 1. Gravity causes erosion by all of the following except

	a. glaciers.; b. moving air.; c. flowing water.; d. mass movement. 2. The rate of erosion by gravity a. is sudden and dramatic; b. is very slow over long periods of time; c. neither of these; d. both of these 3. Factors that increase the risk of landslides include a. dry soils.; b. lack of rain.; c. earthquakes.; d. two of the above 4. When a rock falls from a cliff face, the agent of erosion is usually a. wind; b. water; c. gravity; d. glaciers 5. Downhill creep a. results in curved tree trunks; b. falls as a whole unit; c. leaves large scars in the hillside; d. cannot be noticed because it is so slow 6. Mass movement can occur a. suddenly.; b. very slowly.; c. only on sloping land.; d. all of the above 7. Slump may be caused by a. wet clay.; b. water erosion.; c. scars on a hillside.; d. two of the above 8. A slump is the sudden a. fall of rock and soil down slope; b. flow of mud down slope; c. movement of a large block of rock and soil down slope; d. flow of volcanic ash and water down slope 9. Creep usually takes place where the ground a. is level.; b. is prevented from moving.; c. freezes and thaws frequently.; d. is always saturated with water. 10. Mass movement may be caused when a. droughts dry out the ground; b. a river undercuts a slope; c. the gravitational polarity reverses; d. none of these

Part of the dataset I used to train the model

Lesson ID	Question Number	Prompt	Correct Answer
L_0007	1	1. Gravity causes erosion by all of the following except a. glaciers.; b. moving air.; c. flowing water.; d. mass movement.	b

4. Methodology / Models

For the Zero-Shot-Learning and Few-Shot-Learning experiments, I had to write the code in a specific way for the model to run. I had input a prompt which the model would best understand in order to achieve the best results. This required some trial and error. I created a method, response_to_table which gathered the answers the model chose.

```
def response_to_table (lId, r, answer_table):
 answer_list = r.strip().split("\n")
 for i in answer_list:
   row = i.split(".")
   answer_table.append([lId, row[0],row[1].strip()])
 return answer_table
```

From there, I wrote the code which would activate the model and it began to answer the questions.

```
def lesson_answer (lId,p,answerTable):
 import os
 import openai

 openai.api_key = OPENAI_API_KEY

 start_sequence = "\nA:"
 restart_sequence = "\n\nQ: "

 response = openai.Completion.create(
```

```
   model="text-davinci-002",
   prompt=p,
   temperature=0,
   max_tokens=200,
   top_p=1,
   frequency_penalty=0,
   presence_penalty=0,
   stop=["===="]
 )
 if response['choices'][0]['finish_reason']=='stop':
   answerTable = response_to_table (lId,
response['choices'][0]['text'],answerTable)
  return answerTable
```

Here is the response the model would execute. The text header includes the answers the model chose for the questions.


```
{
  "choices": [
    {
      "finish_reason": "stop",
      "index": 0,
      "logprobs": null,
      "text": "\n\n1. e\n2. d\n3. b\n4. c\n5. g\n6. a\n7. f\n8. a\n9.
a\n10. b\n11. b\n12. b\n13. b\n14. a\n15. a\n16. b\n17. b\n18. d\n19.
a\n20. d\n21. b\n22. a\n23. d\n24. d"
    }
  ],
  "created": 1663018016,
  "id": "cmpl-5plFAdwqW1kMslsyXjD3cjj5vJic2",
  "model": "text-davinci-002",
  "object": "text_completion",
  "usage": {
    "completion_tokens": 97,
    "prompt_tokens": 2181,
    "total_tokens": 2278
  }
}
```

With the fine-tuned model, my intent was to provide some knowledge about the domain in order to "train" the model. The `fine_tunes.create` API takes two parameters - "prompt" and "completion" to be provided in .jsonl (JSON Lines) format. I tried two different approaches to creating a fine-tuned model using the training data set:

1. Prompt = question text along with answer choices; Completion = correct answer. One json line per question in questions.jsonl

```
!openai -k "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" api
fine_tunes.create -t "/content/questions.jsonl" -m davinci
```

2. Prompt = "" (blank); Completion = lesson text. One json line per lesson in lessons.jsonl

```
!openai -k "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" api
fine_tunes.create -t "/content/lessons.jsonl" -m davinci
```

Unfortunately, with this second method, the fine-tuned model no longer provided answers, but instead tried to generate questions. Since there was no meaningful output, I chose to abandon this approach.

5. Results and Discussion

In order to then analyze the answers the model chooses, I created a table called combinedResultsTable which displayed the results of the Zero-Shot-Learning and Few-Shot-Learning experiments. It included the questions, the answer the model chose, the correct answer and whether or not the model was correct. Below is a section of the combinedResults table. Zs stands for Zero-Shot and Fs stands for Few-Shot.

Lesson ID	Question Number	Returned Answer_zs	Answer_zs	Is Correct_zs	Returned Answer_fs	Answer_fs	Is Correct_fs
L_0085	1	a	a	TRUE	a	a	TRUE
L_0085	2	b	d	FALSE	d	d	TRUE
L_0085	3	b	b	TRUE	b	b	TRUE
L_0085	4	b	b	TRUE	b	b	TRUE
L_0085	5	d	f	FALSE	f	f	TRUE

For the Zero-Shot-Learning experiment, the model had an accuracy of 72.76% and for the Few-Shot experiment, 84.78%. In terms of the Zero-Shot, the student essentially walked into the test without any preparation, and got a C, a passing grade. With the Few-Shot, the student had previously done a brief study of the content and scored a B, an above average grade. However,

the fine-tuned model performed worst out of all three, scoring 67.75%, an equivalent of D. In this scenario, the student who "studied" the most got the worst score. Obviously this would not usually occur. Maybe they had a rough day.

6. Conclusions

The results of my experiments highlight both the intelligence of GTP-3 and some of its problems. One would expect the fine-tuned model to perform significantly better than the zero and few shot experiments, yet it performed significantly worse. I have learned that I can use embeddings to properly fine-tune the Davinci model but for the purposes of this paper I did not do that. However, let's not dwell on the negatives. For the Zero-Shot-Learning and Few-Shot-Learning experiments, the model performed amazingly well. The fact that there are models that can answer several multiple-choice questions pretty accurately in a single batch, is quite remarkable. The spectacular results demonstrate why GPT-3 is so well-renowned. Three or four years ago, nothing like this even existed. Furthermore, all three performed significantly better than previous models used on the TQA dataset. In my paper, I only focused on middle school science questions. These same strategies can be applied to a wide range of topics from medicine to business to sports. However, with the positives, come some negatives. The prevalence of highly knowledgeable and functionable models like GPT-3 questions what it really means to learn. Nowadays, students can perform an internet search and find the answers to test questions, which defeats the purpose of administering tests as they are used to measure learning. With a model like GPT-3, one does not even need to rely on search and manually synthesizing the results. All that one needs to do is write a set of instructions and insert the questions and the model will provide them with the answers. Since students have these intelligent models at their disposal, the question arises on whether teachers should shift from administering multiple choice exams. Should they only use open-ended or short answer questions? The pros and cons of the societal implications of GPT-3 are similar to the overall concerns regarding the field of artificial intelligence. First, one is amazed by the wonders Artificial Intelligence performs and the benefits it provides but when they sit and think, the implications of such advanced technology become apparent. When working with the model, I was blown away by its capabilities but when I began thinking about the implications, I discovered how Large Language models like GPT-3 can have a drastic impact on education.

Acknowledgments

I would like to thank Eric Bradford for helping me with this project.

References

A. I. for AI, “Textbook question answering (TQA) dataset - Allen Institute for AI,” *Dataset - Allen Institute for AI*, 2017. [Online]. Available: https://allenai.org/data/tqa. [Accessed: 10-Jul-2022].

A. Kembhavi , “Textbook Question Answering Challenge,” *TQA*. [Online]. Available: http://vuchallenge.org/tqa.html. [Accessed: 1-Sep-2022].

O. AI, *OpenAI API*, 2022. [Online]. Available: https://beta.openai.com/docs/introduction/overview. [Accessed: 02-Sep-2022].

A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi, “Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension,” *CVF Open Access*, 2017. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Kembhavi_Are_You_Smarter_CVPR_2017_paper.html. [Accessed: 02-Jul-2022].