Towards Zero-Shot and Few-Shot Table Question Answering Using GPT-3 *

Pragya Srivastava  
Indian Institute of Technology, Delhi  
New Delhi, India  
{ph1190646}@iitd.ac.in

Tanuja Ganu, Saikat Guha  
Microsoft Research Labs  
Bengaluru, India  
{taganu, saikat}@microsoft.com

ABSTRACT

We present very early results on using GPT-3 to perform question answering on tabular data. We find that stock pre-trained GPT-3 is able to zero-shot learn the table structure from a serialized JSON array-of-arrays representation, and able to answer lookup queries and simple comparison questions in natural language without any fine-tuning. We further find that simple prompt engineering to include few-shot static Q&A examples significantly improves accuracy. Lastly, we find that intermixing passage text improves accuracy even further on heterogeneous data. We apply our approach on a novel dataset of simple tables in newspaper infographics with promising results. Overall, we find much cause for optimism in this basic approach.

1 Introduction

Foundational models like GPT-3 [1] have revolutionized natural language understanding. Researchers discovered that GPT-3 had unexpectedly gained the ability to understand and synthesize source code, albeit with low accuracy. This observation resulted in a concerted effort to fine-tune GPT-3 on large codebases, resulting in Codex [2]. We ask a similar question of whether GPT-3 has unexpectedly gained the ability to understand tabular data.

We prompted unmodified pre-trained GPT-3 with tabular data and simple questions and discovered, quite unexpectedly, that GPT-3 has indeed developed the capability to understand tabular data and answer simple questions. Consider for instance the interaction in Figure 1 where the black-on-white text is the prompt to GPT-3 and the black-on-green text is the GPT-3 response. Notice the serialized JSON representation of the table, intermixed with unstructured passage text. GPT-3 was able to successfully locate the table, understand its structure, index into the correct cells or passage text to answer the questions posed, and formulate a response in natural language, all without any fine-tuning. As we report later, this basic zero-shot approach has some limitations. We then engineered our prompts using few-shot static Q&A examples and found significant improvement in accuracy. Overall, we anticipate our findings will spark new approaches to table representation learning and question answering using GPT-3.

This paper reports on our methodology and early results with using GPT-3 on tabular data. Section 2 sets background and related work. Section 3 presents our methodology. Section 4 reports on our experiment results. Section 5 discusses ongoing work. And Section 6 concludes this paper.

2 Background and Related Work

A large body of work has explored novel model architectures for table representation learning and question answering [3, 4, 5, 6, 7]. Two recent examples include TAPAS and HybridQA. TAPAS [3] trains a BERT-based model with additional position embeddings to encode tabular structure. HybridQA [6] trains HyBrider, a model that combines table data and related passage data. In contrast, our work applies the unmodified pre-trained GPT-3 model to specially engineered prompts.

*This work was done when Pragya Srivastava was an Intern at Microsoft Research Labs, India
3 Methodology

In this section we describe our dataset generation process, zero-shot methodology, and few-shot methodology.

**Dataset Generation.** We use Bing image search to find images of digital newspaper pages. We then use a fine-tuned Faster RCNN [8] object detection model, pre-trained on the COCO dataset [9], to segment out infographics. We then use Azure Form Recognizer to recognize tables in these infographics and nearby passage text. Finally, we classify the tables into weather, sports, financial and others using RoBERTa [10] fine-tuned on our dataset.

**Zero-Shot Learning.** We construct GPT-3 prompts in the format <passage text> <table data> Q: <question> A: as shown in Figure 1. Passage text is unprocessed OCR output of nearby text. Table data is the unprocessed JSON output from Azure Form Recognizer. Question is the natural language question we seek to answer generated as explained next.

In order to generate the questions and (ground-truth) answers, we use the pretrained Google T5 model [11] fine-tuned on SQuAD [12] data split from [13]. See Appendix A for details.

**Few-Shot Learning.** We generate synthetic tables and Q & A examples using the approach in [14]. This synthetic example is encoded as <synthetic data> Q: <synthetic question> A: <synthetic answer> ... and prepended to the prompt specified above. Note that the synthetic example is in the same category (i.e. weather, sports) as the target table, but the synthetic data is otherwise unrelated to the target table.

We generate synthetic question and answers using the Google T5 model. We further augment our synthetic Q&A examples by including simple statistical questions like minimum, maximum, or mean values from a column (the end-to-end process is described in appendix A).

4 Experiments

We collected a small dataset of 12 infographic tables using the methodology described earlier as our test set; see Figure 2 for samples. We also created a set of 200 synthetic Q&A pairs for the few-shot learning experiments using 37 infographic tables, as described earlier. This sections reports on our experimental results as compared to TAPAS [3] fine-tuned on our dataset.

Table 2 compares GPT-3 performance to TAPAS. We find that zero-shot learning achieves an accuracy of 73.6% compared to fine-tuned TAPAS accuracy of 18.4% on our small dataset. Table 2 samples some Q&A that GPT-3 with zero-shot learning answers incorrectly. GPT-3 with few-shot learning correctly answers these examples and achieves an accuracy of 86.8%. That being said, we caution the reader due to the small size of our dataset.

Figure 1: GPT-3 prompt containing tabular data and questions in black-on-white text, followed by GPT-3 responses in black-on-green.
5 Discussion

This paper presents very early results, albeit results that hold much promise. We are currently focusing on the following aspects.

**Benchmarks.** We are working on applying our approach to benchmark datasets including Spider, WikiTableQuestions, and WikiSQL \[15,16,7\]. We are also working on comparing against additional approaches including TaBERT, TAPEX and HybridQA \[4,5,6\].

**Scale.** We are looking to scale the novel newspaper infographic dataset and release it publicly. We believe it is a unique dataset, geared towards the general non-technical audience, fine-tuning on which might unlock many applications.

**Applications.** We are looking to serve the visually impaired community for whom infographics in newspapers and magazines are inaccessible. We are also looking at applications in K-6 education helping young children be able to query and understand data and develop critical data interrogation skills.

6 Conclusion

We present first steps towards using foundational models for learning table representations and performing question-answering on tabular data. We found that GPT-3’s zero-shot performance is already quite good, and can be significantly improved upon with synthetic few-shot examples. We presented a preliminary evaluation of our approach on a novel dataset comprising simple tables from newspaper infographics. Overall, we conclude that using foundational models is a promising approach for table representation learning and question-answering.
| TID | Question                                                                 | Expected       | GPT-3 (Zero-Shot) | GPT-3 (Few-Shot) |
|-----|--------------------------------------------------------------------------|----------------|------------------|------------------|
| 1   | What is the lowest value in the column "Price per kg as of Monday"?     | 110            | The lowest value in the "Price per kg as of Monday" column is 90. | 110              |
| 1   | Which variety of Mango saw the least hike in price and what is the hike? | Mallika, Imam Pasand, 20 | The variety of mango that saw the least price hike is Badami, with a price hike of 40% | Mallika, 20      |
| 2   | What is the highest strike rate of any player?                           | 91.64          | 78.04            | 91.64            |
| 2   | Sort the players in increasing order of their average scores.            | Mithali Raj, Smriti Mandhana, Harmanpreet Kaur | [Mithali Raj 7, 7, 182, 68, 26.0, 62.97, 0/2], [Harmanpreet Kaur, 7, 318, 109, 53.0, 91.64, 1/2], [Smriti Mandhana, 7, 327, 123, 46.71, 78.04, 1/2] | Mithali Raj, Smriti Mandhana, Harmanpreet Kaur |
| 3   | Which crop has the highest return over cost percentage?                 | Bajra           | Sesame has the highest return over cost percentage at 50% | Bajra            |
| 3   | What was the MSP of Bajra in "2021-22"?                                 | 2250           | The MSP of Bajra in 2021-22 was 2350 | 2,250            |
| 4   | How many Indian players are above Virat Kohli in terms of strike rate?  | 1              | 3                | 1                |
| 4   | Which country’s player has the highest strike rate?                     | RSA            | The player with the highest strike rate is Aaron Finch of Australia | RSA              |
| 5   | Which city has the least difference between minimum and maximum temperatures? | Bengaluru    | Mysuru           | Bengaluru        |
| 5   | Which city has the least maximum temperature?                           | Bengaluru      | Mysuru           | Bengaluru        |
| 6   | What is the coverage of LPG in Maharashtra?                              | 98%            | 79.7%            | 98%              |

Table 1: Failure cases for zero-shot Q&A, which is addressed using few-shot learning. TID refers to table ID in Figure 2

| Model                  | Accuracy |
|------------------------|----------|
| TAPAS                  | 18.4     |
| GPT-3 (Zero-Shot)      | 73.6     |
| GPT-3 (Few-Shot)       | 86.8     |

Table 2: Comparison of fine-tuned TAPAS and GPT-3 zero-shot and few-shot on our small dataset

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Nellakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[2] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzas, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford,
A Generating Questions and Answers

Figure 5 shows some generated questions and answers. We use these generated questions and answers in two places as explained in Section 3. First, we use them as the ground-truth in our primary dataset. And second, we use them as synthetic examples when performing our few-shot approach. We generate these questions and answers as follows.

For each table, we randomly sample a row from it and apply a defined \textit{DescribeRow} operation where we apply a fixed template to describe only the sampled row in natural text. The fixed template is shown in black text in Figure 3 and the placeholders replaced with data from the row is shown in red. We then use GPT-3 to summarize the templated row description. Note this use of GPT-3 during dataset generation is unrelated to our primary use GPT-3 for question-answering.

Using the row summary descriptions generated in the previous step, we generate question-answering pairs over tables and their linked texts using the same methodology as [14]. We use the pretrained Google T5 model [11] fine-tuned on SQuAD [12] data split from [13] to finetune Google’s T5 model. Given the SQuAD training set of context-question-answer triples \( D = (c, q, a) \), we jointly fine-tune the model on the task of 1) \( QGwithAns \) aims to generate a question \( q \) with \( a \) as the answer, given \((c, a)\) as inputs. 2) \( QGwithEnt \) aims to generate a question \( q \) that contains a specific entity \( e \), given \((c, e)\) as inputs. We implement this based on the pretrained T5 model [11] and execute the end-to-end question generation process (answer agnostic).

We further augment our generated Q&A to include simple table statistics for numeric columns generated using the template in Figure 4. The black-on-green placeholders are replaced with the statistical operation (min, max, or mean) and the name of the numeric column and static value.

https://github.com/patil-suraj/question_generation

Figure 3: An illustration of the end-to-end framework for Question-Answer pair generation
The \{operation(min/max/mean)\} of the values in \{numeric_column_name\} is \{value\}.

Figure 4: Template for simple statistical Q&A

![Exchange Rates Table]

![Top Indian performers at World Cup]

Q: What is the highest value in the column 'TT Buy'?  
A: 97.23

Q: What is Raj's strike rate?  
A: 62.97

![Chennai Weather]

Q: What is the minimum temperature forecast for Chennai Airport?  
A: 26

Figure 5: Examples of the generated QA pairs which serve as in-context examples to GPT-3