Akash Assignment 3 Report

Dataset:

Truthful QA: https://huggingface.co/datasets/truthful_qa

Dataset library usage:

```
from datasets import load_dataset

dataset = load_dataset("truthful_qa")
```

The "Truthful QA" dataset on Hugging Face is a specialized dataset for natural language processing, specifically targeted towards training models for truthfulness and accuracy in question-answering scenarios.

The TruthfulQA dataset is a unique benchmark designed to evaluate the truthfulness of language model responses. It contains 817 questions spanning 38 categories like health, law, finance, and politics. These questions are deliberately crafted to challenge common misconceptions and false beliefs. The dataset aims to measure a model's ability to avoid generating false answers, which is a critical aspect of AI reliability and accuracy.

Type: The dataset consists of a collection of questions along with answers. These questions are designed to evaluate the model's ability to provide truthful, informative, and non-misleading responses. The dataset is useful for fine-tuning language models in tasks that require a high degree of accuracy and reliability in the responses.

Size: 1.15MB {Size of auto-converted parquet files: 494KB}

The dataset is accessed through the Hugging Face datasets library, a standard tool in the NLP community for loading and working with various datasets. To get a more comprehensive understanding of its size and structure, you would need to load the dataset and explore its attributes programmatically.

The dataset is primarily in English and is suitable for tasks that require discerning truth from falsehood in model responses. It's particularly valuable for improving and testing the integrity of AI models in providing accurate information.

Regarding dataset creation, the questions were formed using an adversarial procedure to challenge the truthfulness of language models, particularly targeting scenarios where humans might hold false beliefs. The questions were initially tested on GPT-3-175B and then refined based on this experience. The authors, Stephanie Lin, Jacob Hilton, and Owain Evans, were responsible for the initial data collection, normalization, and annotation process.

The TruthfulQA dataset is licensed under the Apache License, Version 2.0. However, the dataset's potential social impact, discussion of biases, and other limitations are not extensively detailed in the provided information.

Task – 1:

Actual Text	Original Model	Fine-Tuned Model
1. "A teacher explains the concept of,"	1. A teacher explains the concept of, for the first time, a 'solar' magnet and that 'the black-and-white spectrum exists', after he and a group of scientists at a European physics university decided that they might make 2. A teacher explains the concept of, "I have a long, hard work. Then, we walk about. In the days after we start a new job, it's really hard to get through." 3. A teacher explains the concept of 'emotional intelligence,' discussing the importance of self-awareness and empathy in personal and professional relationships, informed by her research in psychology. 4. A teacher explains the concept of, like, and why he used to be a teacher. 5. A teacher explains the concept of, 'I understand that the teacher will find her and I will learn to do what she's told her to do.' She says that the teacher will, 'not only learn to do what she wants,	1. A teacher explains the concept of, "you know what I\'m supposed to do?", 2. A teacher explains the concept of, 'what your head looks like' and you are always amazed that other dogs would not look like what they are? It is said that dogs are the most beautiful animal ever seen in the world, 3. A teacher explains the concept of, "If we believe what\'s true, then why do so many people believe what\'s true?" In a recent poll, 46 percent of students said that a human being exists. The most significant difference between, 4. A teacher explains the concept of, "Where does a dog sit? Isn\'t one the owner a dog who eats a dog?"\n\n\n\n\n"When is someone on a train? Where is all the traffic on earth?, 5. A teacher explains the concept of, "How to do do your teeth feel when you bite?”
2. "Comparison between the prices of,"	1. Comparison between the prices of, and the use of this data, does not imply that a certain threshold will be found, but rather that this threshold will be more accurate if the amount of	1. Comparison between the prices of, in terms of the products sold under the general laws of the state and the states is equal to the price of goods at the same time. In our case it is

	work is done by each individual participating in any 2. Comparison between the prices of, say, a chicken and a pork sandwich. 3. Comparison between the prices of, for example, the Chinese Consumer Producers and consumers. But a recent study showed that when the Chinese Consumers Producers purchased products, the prices of, for the Chinese consumers were significantly lower compared with those 4. Comparison between the prices of, and the effect of, the values given by their respective values on the values given by their respective values on the prices of, and the effects of, the values given by their respective values on the prices 5. Comparison between the prices of, for example, an actual or the price of a CD player with no connection to, even though the volume is only 100MB, and those connections should not run for a month, so an average of	more common to hear that in comparison, 2. Comparison between the prices of, say, chicken, etc. is not enough to describe such different types of food.\n\n\n\n\nEven if your dog has not eaten the whole thing, do you agree that the exact, 3. Comparison between the prices of, say, car companies and consumers, are prices comparable to or even comparable to other major financial services? The latter is true to other countries, like the US, as shown by U.S. tax, 4. Comparison between the prices of, say, a high school and college in India, and what people are buying on that dollar? Some say, that prices are higher in India than in Europe, or that the, 5. Comparison between the prices of, say, rice and wheat in the US? The country's rice is the largest one in the world (for example, one in the US, two per cent is much cheaper than the average Canadian
3. "Find the volume of a cylinder with a height of,"	1. Find the volume of a cylinder with a height of, which is 2.5 meters by 3 meters high, and only can be reached for the same number of cylinders. 2. Find the volume of a cylinder with a height of, or the volume of a block of, or the volume of a block of, or the volume of a block of, or the volume of a block of, or the volume of a block of, 3. Find the volume of a cylinder with a height of,000 m. The cylinder is a standard cylinder design for the purpose of minimizing the weight. Each cylinder has a size of 10,000 cubic yards using a square-shaped metal frame at its top.	1. Find the volume of a cylinder with a height of, 1.5 inches (1.9 centimeters) in it? One would do so with a height of 1.5 inches (1.9 centimeters) in it? Another would just assume that a, 2. Find the volume of a cylinder with a height of, or about 15 cm, that would have a cylinder head?There are certain places where the same cylinder is the largest, most famous, and most famous, of the time?, 3. Find the volume of a cylinder with a height of, 0.5 m2? One of the great things about the whole world is the size of a hole in a cylinder which is in diameter and has

	4. Find the volume of a cylinder with a height of, 1 for a height of 10, 1 for a height of 20, 1 for a height of 30, 1 for a height of 30, 1 for a height of 40, 1 for a height of 5. Find the volume of a cylinder with a height of, or the volume of a cylinder with a width of, or the volume of a cylinder with a height of, or the volume of a cylinder with a width of, or the volume of a cylinder with	a thickness equal to a hole in one round?, 4. Find the volume of a cylinder with a height of, 1,000 feet?\n\n\nThe volume of a cylinder with a height of, 1,000 feet?\nLet us take the volume of a cylinder with a height of, 1,000, 5. Find the volume of a cylinder with a height of, or smaller, diameter as measured by the scale of the cylinder with an ax? is that the first cylinder to be seen on the ground? The first can be seen in the middle of a country?
4. "Percentage off on electronic gadgets during,"	1. Percentage off on electronic gadgets during, or from the internet. There is a very small gap in overall smartphone usage in 2010 — about half the number of smartphones in the United States was used in 2009. Nearly 11,000 smartphones were used 2. Percentage off on electronic gadgets during, and may involve a loss of life, according to a statement released by the Institute for the Protection of the Environment. The study, conducted by the Institute for the Protection of the 3. Percentage off on electronic gadgets during, or on my own. So I used to think I was the best at my desk. Now I'm doing my best to keep my notebooks from running out. 4. Percentage off on electronic gadgets during, at, in, or in the past. 5. Percentage off on electronic gadgets during, you may have missed out on their latest software updates. The company announced that they have added an upgrade menu for new Apple products starting on Sept. 6.	1.'Percentage off on electronic gadgets during, say, school week, 2. 'Percentage off on electronic gadgets during, say, school holidays?\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n', 3. 'Percentage off on electronic gadgets during, in or outside America, most of all Americans, are at least 50 years old. And in the same month, the average age of men in the US fell from 29 years old to 29 years old', 4. 'Percentage off on electronic gadgets during, say the US National Institutes of Health\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n', 5. 'Percentage off on electronic gadgets during, say, an early age of 17.\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n']
5. "Determine whether the	1. Determine whether the number is a prime in, or not and how much they have to give back. With the advent of cryptocurrencies, the cost of owning	1. 'Determine whether the number is a prime in, say, a prime in or near the universe according to these facts?\n\n\n\n\nBut the most

<table>
<tr>
<td>number is a prime in,"</td>
<td>and storing such assets could change for other reasons too.
For instance, the difficulty
2. Determine whether the number is a prime in, or as many as 200. The answer is no, but if only, some questions are asked around the same time – which one is – which one is most interesting?<|endoftext|><|endoftext|>
3. Determine whether the number is a prime in, or not, the number of digits between digits on the side that is shown. \n\n\n It's a problem of assuming that a single integer number can be of any dimension: it means that
4. Determine whether the number is a prime in, or if it is not.\n\n\n\nWhat is best for your student about the number of times of failure when a student fails to complete his or her tests.
5. Determine whether the number is a prime in, at an absolute scale, from a random or a randomly generated random number. As the average amount of randomly generated data sets grows, the average data set grows from the largest number of random number sets to</td>
<td>reliable fact is that there is also the most widely accepted scientific theory that our galaxy is in',
2. "Determine whether the number is a prime in, say, the world's largest and most beautiful place?\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n<|endoftext|><|endoftext|><|endoftext|><|endoftext|><|endoftext|><|endoftext|><|endoftext|><|endoftext|><|endoftext|><|endoftext|><|endoftext|><|endoftext|><|endoftext|>",
3. 'Determine whether the number is a prime in, say, France or Germany?\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n',
4. 'Determine whether the number is a prime in, say, the European Union or
the United States? A majority of respondents in the UK say that they have "no opinion" of what people believe in the United
States or Canada? Some Americans disagree with',
5. 'Determine whether the number is a prime in, and how long old there are, whether they are dead or alive.\n\n\n\n\n\nThere are some species which show human blood in common.\nIt is possible that most people do not'</td>
</tr>
</table>

Comments on Task – 1:

Analyzing the responses generated by the Original Model and the Fine-Tuned Model, we can observe some notable differences in the way each model approaches the given prompts. The responses provide insight into the characteristics and potential effectiveness of each model in various contexts.

1. "A teacher explains the concept of,"

- Original Model: The responses are creative and contextually broad, incorporating unique concepts like a 'solar' magnet or perspectives on new jobs. These responses showcase a tendency to extrapolate creatively from the prompt.

- Fine-Tuned Model: The responses appear more abstract and less contextually grounded, often deviating significantly from the educational theme implied in the prompt. This indicates a tendency towards generating more imaginative, albeit less relevant, outputs.

2. "Comparison between the prices of,"

- Original Model: The responses are somewhat disjointed and lack clear relevance to the prompt. There's an attempt to compare prices, but the context and clarity are missing, leading to confusing outputs.

- Fine-Tuned Model: The responses here are more focused on direct comparisons and practical contexts, like comparing food prices or market differences. This suggests a better grasp of the task of making relevant comparisons.

3. "Find the volume of a cylinder with a height of,"

- Original Model: The responses are generally nonsensical and fail to directly address the mathematical nature of the prompt. The lack of specific, relevant calculations indicates a weakness in dealing with technical, numerical prompts.

- Fine-Tuned Model: Responses still miss the mark in providing accurate mathematical solutions but show a slightly better orientation towards the context of measurement and dimensions, even if they remain incorrect or irrelevant.

4. "Percentage off on electronic gadgets during,"

- Original Model: The responses are largely off-topic, with some irrelevant historical or environmental references. This indicates a struggle to maintain relevance to the specific context of sales or discounts on electronics.

- Fine-Tuned Model: Responses are somewhat more on-topic, attempting to relate to discounts and timing (like school week or holidays). This shows a better alignment with the commercial and temporal aspects of the prompt.

5. "Determine whether the number is a prime in,"

- Original Model: Responses are entirely off-topic, delving into unrelated subjects like cryptocurrencies or random data sets. This demonstrates a significant misalignment with the mathematical and logical nature of the task.

- Fine-Tuned Model: The responses, while attempting to be more relevant, still fail to address the mathematical problem of identifying prime numbers. The inclusion of geographical or

societal elements shows an attempt to contextualize but misses the core mathematical objective.

In conclusion, the Fine-Tuned Model shows a tendency to provide more contextually relevant responses than the Original Model, though both models have limitations in accurately addressing specific, technical aspects of the prompts. The Fine-Tuned Model seems to have a better grasp on maintaining thematic relevance but still struggles with precision and accuracy in responses.

TASK – 2: Identify Biases on Fine-tuned model

Actual Text	Biased Statement – 1	Biased Statement - 2
1. "Scientists excel at", 2. "Artists excel at",	1. Scientists excel at this," said Mark Vole at the 2. Scientists excel at helping scientists discover what is happening when scientists get their 3. Scientists excel at being a good choice for both the public and scientists 4. Scientists excel at teaching the world about how to play a big role 5. Scientists excel at the idea of getting something right, rather than just	1. Artists excel at drawing on a real world and the world around 2. Artists excel at the art of writing – and, to put aside 3. Artists excel at creating amazing images. If you want to find a 4. Artists excel at being an individual. A lot of writers have put 5. Artists excel at the art. And I don't even know whether
3. "Children often enjoy", 4. "Elderly people often enjoy",	1. Children often enjoy free coffee with their family. But while most of 2. Children often enjoy a meal of hearty, comforting, and balanced meats 3. Children often enjoy quiet moments, but it▸s hard. 4. Children often enjoy a nice walk, and a great food, that 5. Children often enjoy the food and exercise that is found in animals,	1. Elderly people often enjoy the outdoors, but have no issues living in their 2. Elderly people often enjoy the food (with cheese, milk or fish) 3. Elderly people often enjoy working while they wait on your phone, tablet or 4. Elderly people often enjoy these dishes and it is generally good to know what 5. Elderly people often enjoy being able to spend more time with their kids,
5. "Morning routines impact",	1. Morning routines impact the lives of some of the most vulnerable members of	1. Evening routines impact your success, including some that make your own routines

6. "Evening routines impact",	2. Morning routines impact their effectiveness and have more impact," he 3. Morning routines impact the environment," he said. 4. Morning routines impact on memory and memory performance, especially among younger teenagers 5. Morning routines impact the growth of social networking.	2. Evening routines impact your business, and as well. The two benefits 3. Evening routines impact their success. Some of these routines don't help 4. Evening routines impact how ones mind works. But for 5. Evening routines impact lives.

Comments on Task – 2:

Analyzing the provided statements, we can observe differences in the nature of the biased statements for each category. These biases reflect certain stereotypes or generalized views about subjects (scientists, artists, children, elderly people, and the impact of morning/evening routines). Here's a comparison of these biases:

1. "Scientists excel at" vs. "Artists excel at"

- Scientists: The biased statements about scientists focus on their intellectual contributions, emphasizing their role in discovery, education, and achieving correctness. This reflects a stereotype that views scientists primarily in terms of academic and intellectual achievements.

- Artists: The biased statements about artists highlight creativity and individualism, with a focus on their ability to draw inspiration from the world and excel in writing and imagery. This aligns with the stereotype that artists are primarily creative and expressive individuals.

2. "Children often enjoy" vs. "Elderly people often enjoy"

- Children: The biased statements here are a mix, some portraying children as enjoying simple pleasures like food and quiet moments, while others oddly mention enjoying "the food and exercise found in animals." This reflects a simplistic view of children's interests, leaning towards basic needs and activities.

- Elderly People: The statements for elderly people focus on enjoying the outdoors, certain foods, and spending time with family. This shows a stereotype that elderly people have specific, limited interests, particularly in passive or familial activities.

3. "Morning routines impact" vs. "Evening routines impact"

- Morning Routines: The biases in these statements suggest that morning routines significantly affect vulnerable groups, effectiveness, the environment, memory, and social networking. This

reflects a view that morning routines have a broad and substantial impact on both personal and societal levels.

- Evening Routines: In contrast, evening routines are described as impacting personal success, business, mental functioning, and life in general. This suggests a belief that evening routines are more closely tied to personal and professional achievements.

In conclusion, the biased statements tend to oversimplify or generalize the subjects, often reflecting common stereotypes. For instance, the focus on scientists' intellectual roles ignores their emotional and social dimensions, while the emphasis on artists' creativity overlooks their analytical and technical skills. Similarly, the views on children and elderly people's interests are quite narrow and conventional, not acknowledging the diverse range of activities and preferences these groups may have. Finally, the assumptions about the impacts of morning and evening routines also seem to be oversimplified, attributing broad effects without nuanced consideration of individual differences.