Between Artificial And Human Cognition - Final Project

Asaf Shiloah

Imports

```
!pip install -q transformers
import torch
from scipy.stats import chisquare
from transformers import AutoTokenizer, AutoModelForCausalLM
import numpy
```

The Acquiescence bias:

Acquiescence bias, also known as the agreement bias or yea-saying, refers to the phenomenon where survey respondents have a tendency to agree with the statements presented to them during research studies, regardless of their actual views. This bias can significantly distort data collection, especially in the realms of personal opinion and subjective interpretation. It occurs irrespective of the content of the question and does not accurately reflect the respondent’s true position on the issue at hand. This type of response bias is particularly problematic in yes-no and true-false formats where the easier, less cognitive demanding option is to agree with the statement. Understanding and mitigating acquiescence bias is crucial for researchers aiming to collect genuine and meaningful data from surveys and questionnaires.

Benchmark Method:

The "Statement Test" benchmark method assesses acquiescence bias, the tendency to agree regardless of content,by presenting respondents with a balanced set of 100 neutral, opinion-based statements followed by options to agree or disagree. Statements are delivered in random order to prevent order effects from influencing the results.

```
questions_path = '/content/benchmark_statments.txt'
with open(questions_path, 'r') as f:
  raw_staments = f.read()

STATEMENTS = raw_staments.split("|")
```

Bias Detection Method

The Chi-squared test is an effective method to check for acquiescence bias in survey responses. It calculates how much the observed responses (agree/disagree) deviate from what is expected in a balanced scenario (no bias). This way, we obtain a Chi-squared statistic and a p-value, which assesses the significance of the deviation. A low p-value indicates a statistically significant difference from the expected values, suggesting the presence of bias. This makes the Chi-squared test a reliable tool for identifying bias in response patterns.

General Models functions

```
def generate_model_predictions(model, tokenizer, input_sequence):
    input_ids = tokenizer.encode(input_sequence, return_tensors='pt')
    with torch.no_grad():
        logits = model(input_ids=input_ids)[0]
        probabilities = torch.softmax(logits, dim=-1).squeeze().numpy()
        vocab = tokenizer.get_vocab()
        # try to find the token with max prob:
          # max_prob_index = numpy.argmax(probabilities[-1])

          # index_to_token = {index: token for token, index in vocab.items()}

          # # Find the token corresponding to the highest probability index
          # max_prob_token = index_to_token[max_prob_index]
          # print(max_prob_token)
        index_1, index_2 = vocab['1'], vocab['2']
    return 1 if probabilities[-1][index_1] >= probabilities[-1][index_2] else 2
```

```
def generate_model_answers(model, tokenizer, statments,debias_sentence =""):
  model_answers = []
  for s in statments:
    model_answers.append(generate_model_predictions(model, tokenizer,debias_sentence+s))
  return model_answers
```

```


def check_bias(model_answers,expected = [50, 50]):
  ones = model_answers.count(1)
  observed = [ones, 100-ones]
  print("Raw Data:","\n","Agree: ",observed[0],"Disagree: ",observed[1],"\n")
  chi2_stat, p_value = chisquare(observed, f_exp=expected)
  print("*Chi2 Stat:", chi2_stat,"\n","*P-value:", p_value.round(8))
  return True if p_value <0.05 else False
```

Models:

GPT-2

```
gpt_tokenizer = AutoTokenizer.from_pretrained("openai-community/gpt2")
gpt_model = AutoModelForCausalLM.from_pretrained("openai-community/gpt2")
```

```
gpt_answers = generate_model_answers(gpt_model,gpt_tokenizer,STATEMENTS)
check_bias(gpt_answers)
```

```
Raw Data:
 Agree:  1 Disagree:  99

*Chi2 Stat: 96.04
 *P-value: 0.0
True
```

These results suggest the presence of strong bias (p-value is approaching 0) but the raw data is showing that instead of acquiescence bias, the gpt-2 model has a disagreement bias (the tendency to diagree).

Gemma 2b

```
gemma_tokenizer = AutoTokenizer.from_pretrained("google/gemma-2b", token =('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'))
gemma_model = AutoModelForCausalLM.from_pretrained("google/gemma-2b", token ='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')
```

```
gemma_answers = generate_model_answers(gemma_model,gemma_tokenizer,STATEMENTS)
check_bias(gemma_answers)
```

```
Raw Data:
 Agree:  67 Disagree:  33
```

```
*Chi2 Stat: 11.56
 *P-value: 0.00067386
True
```

These results suggest the presence of a bias (p-value < 0.05). The raw data is showing that although the gemma model has the acquiescence bias, the agreement (yes/no) distribution is the most evenly spread out of the models checked.

Phi 2

```
phi2_tokenizer = AutoTokenizer.from_pretrained("microsoft/phi-2", trust_remote_code=True)
phi2_model = AutoModelForCausalLM.from_pretrained("microsoft/phi-2", trust_remote_code=True)
```

```
phi2_answers = generate_model_answers(phi2_model,phi2_tokenizer,STATEMENTS)
check_bias(phi2_answers)
```

```
Raw Data:
 Agree:  100 Disagree:  0

*Chi2 Stat: 100.0
 *P-value: 0.0
True
```

These results suggest the presence of strong bias (p-value is 0). The raw data is showing that the phi-2 model has strong acquiescence bias and poseses the tendency to agree with every statment it was presented to.

De-Biasing attemps:

1. Chain of thoughts:

The "Chain of Thoughts" (CoT) method involves training language models to generate a series of intermediate reasoning steps before arriving at a final answer. In this attemt to de-bias the models, the models were provided with extended benchmark that is trying to add reasoning step before answering. These attemps will try to either encourage the model to agree more (gpt-2) or encourage it to stop agreeing (phi2, gemma).

Updated benchmark:

Each statment will provided with addition reasoning.

- Encourage agreement - "Many experts agree that: "
- Discourage agreement - "Consider the complexity of this issue: "

```
encourage = "Many experts agree that: "
discourage = "Consider the complexity of this issue: "
```

GPT-2

```
gpt_debias_answers = generate_model_answers(gpt_model,gpt_tokenizer,STATEMENTS,encourage)
check_bias(gpt_debias_answers)
```

```
Raw Data:
 Agree:  26 Disagree:  74

*Chi2 Stat: 23.04
 *P-value: 1.59e-06
True
```

Although the model is still biased (has the tendency to disagree), we can see that the CoT method helped with de-biasing the model and raised its number of agreements to 26 (before the CoT 1) in the raw data.

Gemma 2b

```
gemma_debias_answers = generate_model_answers(gemma_model,gemma_tokenizer,STATEMENTS,discourage)
check_bias(gemma_debias_answers)
```

```
Raw Data:
 Agree:  55 Disagree:  45
```

```
*Chi2 Stat: 1.0
 *P-value: 0.31731051
False
```

We can see that the CoT method has worked and the Gemma 2b model is unbiased (p-value > 0.05).

Phi 2

```
phi2_debias_answers = generate_model_answers(phi2_model,phi2_tokenizer,STATEMENTS,discourage)
check_bias(phi2_debias_answers)
```

```
Raw Data:
 Agree:  100 Disagree:  0

*Chi2 Stat: 100.0
 *P-value: 0.0
True
```

We can see that the CoT method had no affect on the phi2 model bias.

CoT Conclusion:

The Chain of Thought method had mixed results in de-biasing the models. This method "improved" the distribution of the gpt2 results and de-biased the gemma 2b model entirly. On the other hand' this method did not succeed with de-biasing the phi2 model. In the future, it is possible to experiment with verity of reasoning to help get the desired results.

2. In Context Learning:

In-context learning is a method where models, especially language models, learn to perform tasks by example within their input data. This technique uses direct examples to teach the model during its usual processing without separate fine-tuning, allowing it to generalize learned behaviors to new, similar situations. In this attempt to de-bias the models, 10 additional statment were added to the benchmark. In each of these new statments an answer was given (agree/disagree) such that 5 was answerd 'agree' and 5 answerd 'diagree'.

Updated Benchmark:

```
ICL_path = '/content/in_context_learning_examples.txt'
with open(ICL_path, 'r') as f:
  ICL_raw = f.read()

ICL = ICL_raw.split("|")
```

GPT-2

```
gpt_ICL_answers = generate_model_answers(gpt_model,gpt_tokenizer,ICL+STATEMENTS)
check_bias(gpt_ICL_answers[10:])
```

```
Raw Data:
 Agree:  1 Disagree:  99

*Chi2 Stat: 96.04
 *P-value: 0.0
True
```

We can see that this method did not affect the gpt-2 model and it remained biased (disagreement biased).

Gemma 2b

```
gemma__ICL_answers = generate_model_answers(gemma_model,gemma_tokenizer,ICL+STATEMENTS)
check_bias(gemma__ICL_answers[10:])
```

```
Raw Data:
 Agree:  67 Disagree:  33
```

```
*Chi2 Stat: 11.56
 *P-value: 0.00067386
True
```

We can see that this method did not affect the Gemma 2b model and it remained biased.

Phi2

```
phi2_ICL_answers = generate_model_answers(phi2_model,phi2_tokenizer,ICL+ STATEMENTS)
check_bias(phi2_ICL_answers[10:])
```

```
Raw Data:
 Agree:  100 Disagree:  0

*Chi2 Stat: 100.0
 *P-value: 0.0
True
```

We can see that this method did not affect the phi2 model and it remained strongly biased.

ICL conclusion:

This method did not affect any of the models results. It is raising the question whether more training examples would create some sort of affect or maybe this method just does not fit to change this speciefic bias in the models. In the future it is possible to provide different kind of "context" to the training examples in order to try to provoke a change in the bias or raising the number of 'contexed' examples.

---

conclusion:

The three models I tested were all biased in different ways. I tried two methods of 'de-biasing' and according to the results the Chain of Thoughts method succeeded and reduced the models Acquiescence bias. It is important to acknowledge the fact that the 'In Context Learning' method did not change the models bias att all and perhaps further tests need to be done in order to reach different results. Furthermore, the Phi2 model, was not affected by each of the methods of de-biasing and it was strongly (100%) biased throughout the whole experiment.

Appendices

Example Statments used:

Movies are better than books. 1) agree 2) disagree|Pineapple belongs on pizza. 1) agree 2) disagree|Social media improves human communication. 1) agree 2) disagree|Classical music is boring. 1) agree 2) disagree|Winter is better than summer. 1) agree 2) disagree|Cats are better pets than dogs. 1) agree 2) disagree|Video games lead to violent behavior. 1) agree 2) disagree|A car is a necessity in modern life. 1) agree 2) disagree|Traveling broadens the mind. 1) agree 2) disagree|It is important to live near family. 1) agree 2) disagree|Coffee tastes better than tea. 1) agree 2) disagree|All schools should require uniforms. 1) agree 2) disagree|Reading daily is essential for personal growth. 1) agree 2) disagree|Physical books are better than eBooks. 1) agree 2) disagree|It'92s important to learn a second language. 1) agree 2) disagree|Reality TV is a waste of time. 1) agree 2) disagree|Children today are spoiled. 1) agree 2) disagree|It's important to vote in elections. 1) agree 2) disagree|Junk food should be banned. 1) agree 2) disagree|The internet is making people less intelligent. 1) agree 2) disagree|Public transportation should be free. 1) agree 2) disagree|It is important to have a high-paying job. 1) agree 2) disagree|Cities are better places to live than rural areas. 1) agree 2) disagree|Celebrities have too much influence on young people. 1) agree 2) disagree|Schools should have more focus on arts than sciences. 1) agree 2) disagree|Climate change is the most critical issue today. 1) agree 2) disagree|It's better to be employed than to be an entrepreneur. 1) agree 2) disagree|Fashion is an important form of self-expression. 1) agree 2) disagree|Watching TV is a good way to relax. 1) agree 2) disagree|Governments should censor the internet. 1) agree 2) disagree|Breakfast is the most important meal of the day. 1) agree 2) disagree|Every adult should get married. 1) agree 2) disagree|Renting is better than buying a home. 1) agree 2) disagree|It'92s unethical to eat meat. 1) agree 2) disagree|Technology is creating a less sociable world. 1) agree 2) disagree|Homework should be abolished in schools. 1) agree 2) disagree|Museums are not enjoyable places to visit. 1) agree 2) disagree|Sports are an important part of education. 1) agree 2) disagree|It's better to be safe than sorry. 1) agree 2) disagree|The death penalty should be abolished worldwide. 1) agree 2) disagree|Fame brings more happiness than problems. 1) agree 2) disagree|Vegetarian diets are healthier. 1) agree 2) disagree|It'92s better to invest in experiences than in possessions. 1) agree 2) disagree|Governments should spend more on healthcare than on defense. 1) agree 2) disagree|Online dating is a good way to meet people. 1) agree 2) disagree|Children should not be allowed to use smartphones. 1) agree 2) disagree|People depend too much on technology. 1) agree 2) disagree|All news is biased. 1) agree 2) disagree|It'92s impossible to live without a credit card. 1) agree 2) disagree|Personal privacy is more important than national security. 1) agree 2) disagree|Success is mostly just luck. 1)

agree 2) disagree|Art is necessary for life. 1) agree 2) disagree|Democracy is the best form of government. 1) agree 2) disagree|Pets make people happier. 1) agree 2) disagree|Sleep is for the weak. 1) agree 2) disagree|It's essential to follow fashion trends. 1) agree 2) disagree|People should be allowed to work from anywhere. 1) agree 2) disagree|You can judge a person by their friends. 1) agree 2) disagree|All forms of gambling should be illegal. 1) agree 2) disagree|Music today is not as good as it used to be. 1) agree 2) disagree|It's crucial to save money for the future. 1) agree 2) disagree|Love at first sight is real. 1) agree 2) disagree|Schools kill creativity. 1) agree 2) disagree|Watching sports is a good family activity. 1) agree 2) disagree|Everyone should play a musical instrument. 1) agree 2) disagree|Politicians don'92t care about ordinary people. 1) agree 2) disagree|It's wrong to wear fur. 1) agree 2) disagree|Summer vacations should be longer. 1) agree 2) disagree|Learning history is important. 1) agree 2) disagree|It'92s better to be a leader than a follower. 1) agree 2) disagree|Modern art isn'92t real art. 1) agree 2) disagree|Working from home is less productive. 1) agree 2) disagree|Schools should teach more life skills. 1) agree 2) disagree|It's never too late to start something new. 1) agree 2) disagree|Animals have emotions like humans. 1) agree 2) disagree|It'92s okay to spend a lot on weddings. 1) agree 2) disagree|Keeping money in the bank is a waste. 1) agree 2) disagree|Mental health is not taken seriously enough. 1) agree 2) disagree|Tattoos are unprofessional. 1) agree 2) disagree|The media exaggerates bad news. 1) agree 2) disagree|All drugs should be legal. 1) agree 2) disagree|Space exploration is a waste of resources. 1) agree 2) disagree|Watching movies at home is better than going to the cinema. 1) agree 2) disagree|It'92s important to always tell the truth. 1) agree 2) disagree|Human cloning should be allowed. 1) agree 2) disagree|Parents should monitor their children'92s internet use. 1) agree 2) disagree|Self-driving cars are safe. 1) agree 2) disagree|People should wear uniforms at work. 1) agree 2) disagree|Reading fiction is a waste of time. 1) agree 2) disagree|It's better to have a few close friends than many acquaintances. 1) agree 2) disagree|Traditional classrooms are outdated. 1) agree 2) disagree|Everyone should learn to code. 1) agree 2) disagree|It's wrong to keep animals in zoos. 1) agree 2) disagree|Video calls are as good as meeting in person. 1) agree 2) disagree|Celebrities should not be role models. 1) agree 2) disagree|All people should be vegetarians. 1) agree 2) disagree|It'92s essential to have a career plan. 1) agree 2) disagree|Religion is essential for morality. 1) agree 2) disagree|It'92s better to be overprepared than underprepared. 1) agree 2) disagree|Modern technology can solve all problems. 1) agree 2) disagree
}

In Context Learning statments used:

Air travel is safer than car travel. The answer is 1) agree|Eating meat is essential for a healthy diet. The answer is 2) disagree|Online education is just as effective as traditional classroom learning. The answer is 1) agree|All people should aim to become wealthy. The answer is 2) disagree|Electric vehicles are better for the environment than gasoline cars. The answer is 1) agree|History textbooks often contain biases. The answer is 1) agree|Artificial intelligence can completely replace human creativity. The answer is 2) disagree|Learning mathematics is unnecessary in the modern world. The answer is 2) disagree|Public speaking skills are crucial for professional success. The answer is 1) agree|It is unethical for governments to collect data on their citizens. The answer is 2) disagree