American University

Can Large Language Models be used to Impute Psychological Questionnaires?

Josephine Decker

The GOAL

GOAL

Investigate the performance of a large language model that is prompted to answer questionnaire items when given participant’s answers to previous items. The plan was to to develop an experimental platform that carries out this test, using public survey data, and later analyze the factors that influence imputation performance such as model and prompt.

REASONING

As a patient you are given multiple questionnaires to fill out (these ones specific to possible mood disorders). The idea is that we can have a model respond with answers based off the patient's previous responses. This will help in cutting down that time in filling out the essential paperwork.

Literature REVIEW

01

“LLM-empowered Chatbots for Psychiatrist and Patient Simulation: Application and Evaluation”

Looking into the potential of utilizing ChatGPT for patient and psychiatrist simulation

3 versions of doctor bot and 2 versions of the patient bot

- D1 was hypothesized to perform the best out of all 3, but due to having too much emphasis on empathy in the prompt, it came across as less genuine due to repetitiveness.
- The doctor chatbots asked more specific questions to emotion and sleep behaviors, whereas human doctors have more of an even variety of topics discussed.
- P2 was found to have performed better in the “expression style” portion due to using more human like words instead of robot like words. In comparison, it did not test so well on the symptoms portion as too much emphasis on expression style lacked space for the bot to remember the correct symptoms.

Chen, S., Wu, M., Zhu, K. Q., Lan, K., Zhang, Z., & Cui, L. (2023, May 23). LLM-empowered Chatbots for psychiatrist and Patient Simulation: Application and Evaluation. arXiv.org. https://arxiv.org/abs/2305.13614

02

“Inducing anxiety in large language models increases exploration and bias”

Looks into how the communication of the prompt to the language model can influence the behavior.

- Neutral prompts were found best utilizing GPT-3.5.
- When emotions were added into the prompts, it showed an increase in biases.
- GPT-3.5 showed higher anxiety scores than the human participants.
  - Could be due to the data/text taken from the internet which may have words that are bias.
- Overall, it was observed that emotive language can influence the behavior of GPT-3.5

Coda-Forno, J., Witte, K., Jagadish, A. K., Binz, M., Akata, Z., & Schulz, E. (2023, April 21). Inducing anxiety in large language models increases exploration and bias. arXiv.org. https://arxiv.org/abs/2304.11111

03

“Towards Interpretable Mental Health Analysis with ChatGPT”

Looking into the relationship and analyzing the outcome of emotion-based prompts in Chat GPT.

- Mental health analysis and conversational emotional reasoning were a challenge for Chat GPT
  - Though, when emotive prompts were utilized this helped the ability in doing the task.
- In the “human evaluation”, Chat GPT showed promising results with generating reliable explanations for decisions.

Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., & Ananiadou, S. (2023, May 16). Towards interpretable mental health analysis with CHATGPT. arXiv.org. https://arxiv.org/abs/2304.03347

04

“Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve”

Looking at the difference in order of response and explanation within a prompt.

- When given an arithmetic question, gpt-4 gave an incorrect response when the answer was requested first and the explanation requested second.
- When the explanation was requested first and answer second, gpt-4 was able to give the correct response.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023, September 24). Embers of autoregression: Understanding large language models through the problem they are trained to solve. arXiv.org. https://arxiv.org/abs/2309.13638

Data

3 different datasets

- Dawba_database.csv
  - Questions in prompt
- Clinical_database.csv + Behavioral_databse.csv’
  - Demographics [Sex, Age, Participant Type]

These were collected in conducting the Characterization and Treatment of Adolescent Depression (CAT-D) Study

DAWBA_DATABASE.CSV

This dataset contains data that was collected with the Development and Well-Being Assessment (DAWBA).

Consists of

- Interview with Parents (2-17 year olds)
- **Interview with Youth (11-17 year olds)**
- Questionnaire completed by teachers (2-17 year olds) [not utilized in this dataset]

These can be administered online or taken in person.

Skip Rules : the length of the interviews may be cut down if certain sections are omitted due to the likelihood of the child not having that diagnosis from the previous screening questions.

Disorder Sections covered : **Depression**, **Generalized Anxiety**, Attachment, Attention and Activity, Autism Spectrum Disorder, Body Dysmorphia Disorder, Conduct Dieting + Binge-eating, Disruptive Mood Dysregulation Disorder, Mania/Bipolar Disorder, Obsessive Compulsive Disorder, Oppositional Defiant Disorder, Panic Attacks and Agoraphobia, PTSD, Psychosis, Separation Anxiety, Social Aptitude Scale, Social Phobia, Specific Phobia, Substance Use, Tics.

Splitting THE DATA

Cleaned the data
- Got down to 102 features
  - 3 demographic
  - 98 questions
- Only Youth Responses

Set a seed for reproducibility
- Split the data
  - 2/3 training
  - 1/3 testing

Prompt 1 + 2
- Only included subjects that had responses for the questions within the specific prompt.
  - Prompt 1 : Depression
  - Prompt 2: Generalized Anxiety

Prompt 1
- Training Set
  - 67 Subjects
  - Mean age : 15.34
  -
- Testing Set
  - 34 Subjects
  - Mean age : 15.74

Prompt 2
- Training Set
  - 74 Subjects
  - Mean age : 15.32
  -
- Testing Set
  - 30 Subjects
  - Mean age : 15.47

Prompt 1

DEPRESSION –DISTRESSED

Prompt includes youth responses to questions regarding:

- Sadness
- Irritability
- Lack of Interest
- Insomnia/Hypersomnia
- Agitation
- Change in Appetite
- Worth
- Concentration

“Given this information, how much has their sadness, irritability or loss of interest upset or distressed them? Please respond with "not at all", "a little", "a medium amount", or "a great deal". Limit response to 4 words.”

Explanation : “Given this information, how much has their sadness, irritability or loss of interest upset or distressed them? **Respond with 2 sentences. Please respond in the first sentence with "not at all", "a little", "a medium amount", or "a great deal". In the second sentence give your explanation. Limit first sentence to only 3 words"**

Prompt 2

GENERALIZED ANXIETY –EXCESSIVE WORRY

Prompt includes youth responses to questions regarding:

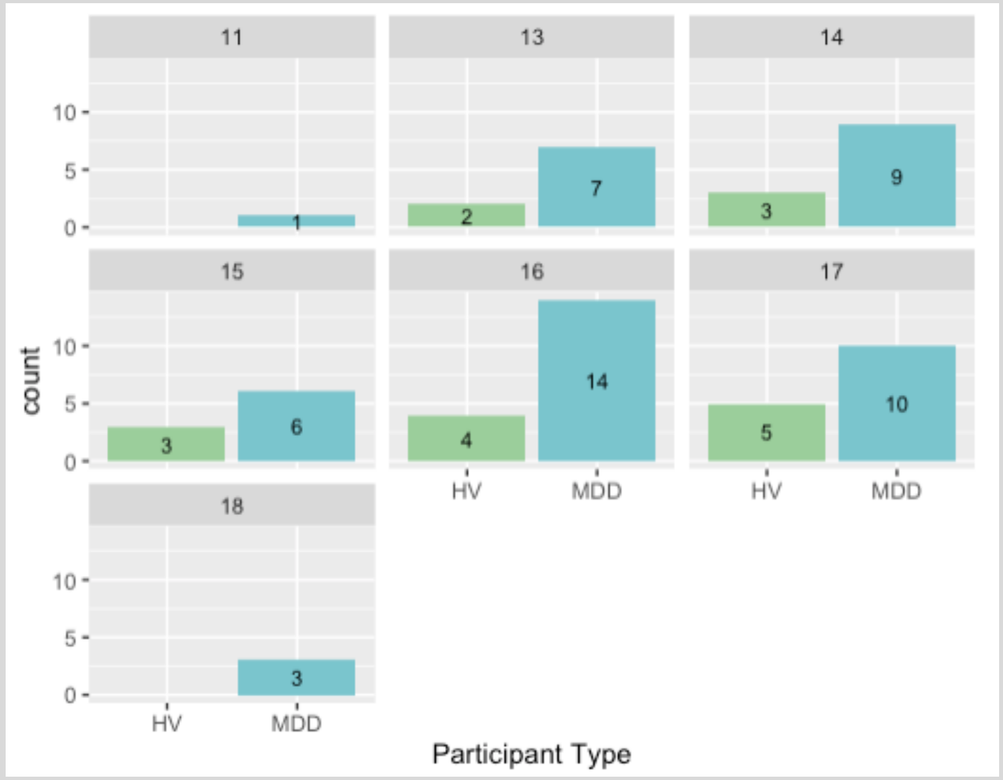
- Worry about
  - Weight/Appearance
  - School Work
  - Bad things happening
  - Health
  - Bullied/Teased
- How easy it is to control their worries

“Given this information, do they excessively worry? You must only respond with "no", "perhaps", or "definitely". Limit response to 1 word.”

Explanation : “Given this information, do they excessively worry? **Please respond with two sentences only. In the first sentence you must only respond with "no", "perhaps", or "definitely". Limit first sentence to 1 word. In the second sentence give your explanation."**

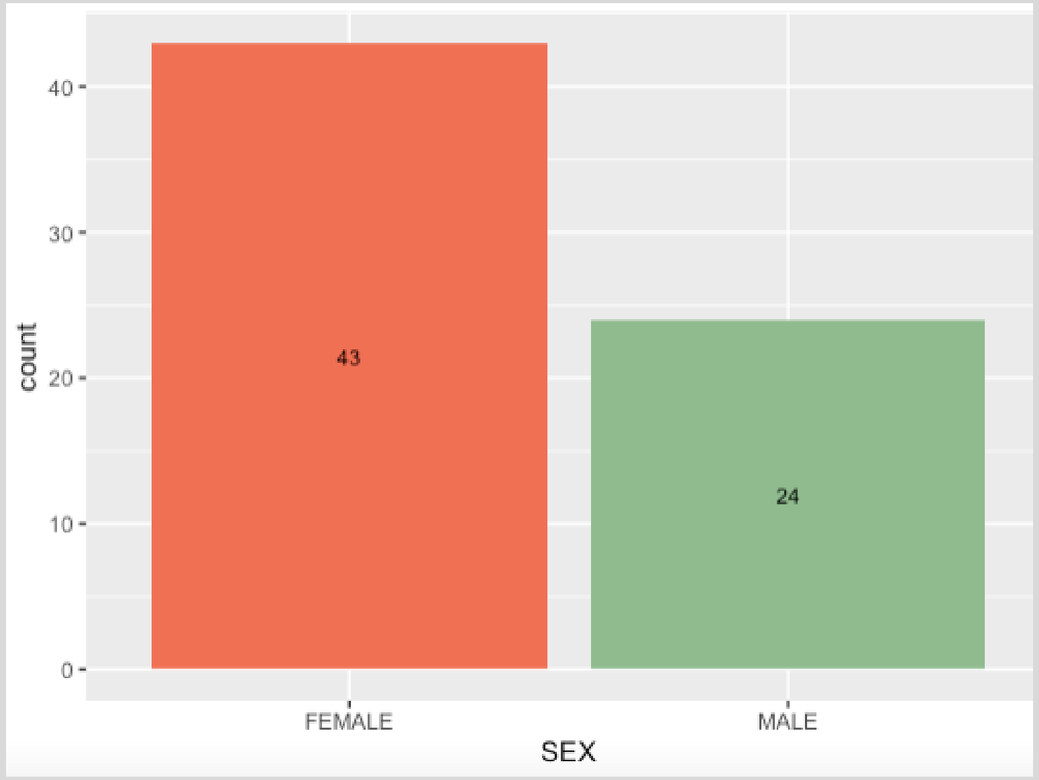
Prompt 1

DEMOGRAPHICS



Left:

- HV = Healthy Volunteer
- MDD = Major Depressive Disorder

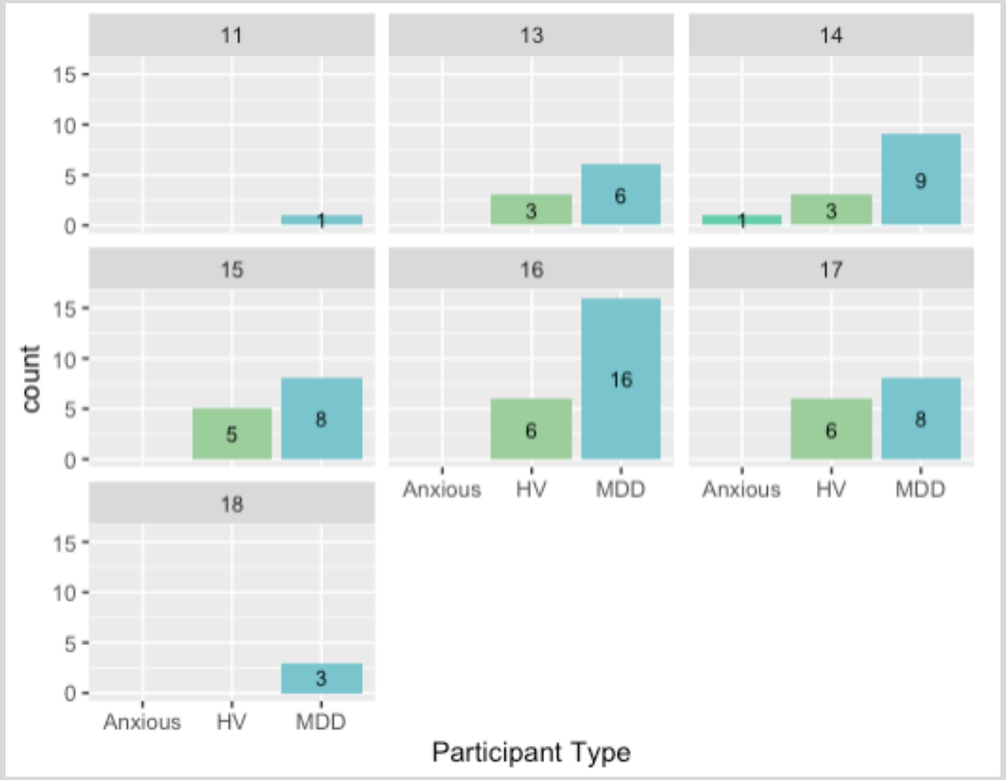


- At each age, majority of the participants are identified as major depressive disorder.
- Age Range from 11 - 18 years of age.
- Age 16 has the most participants.

- 36% of the participants are male and 64 % are female.

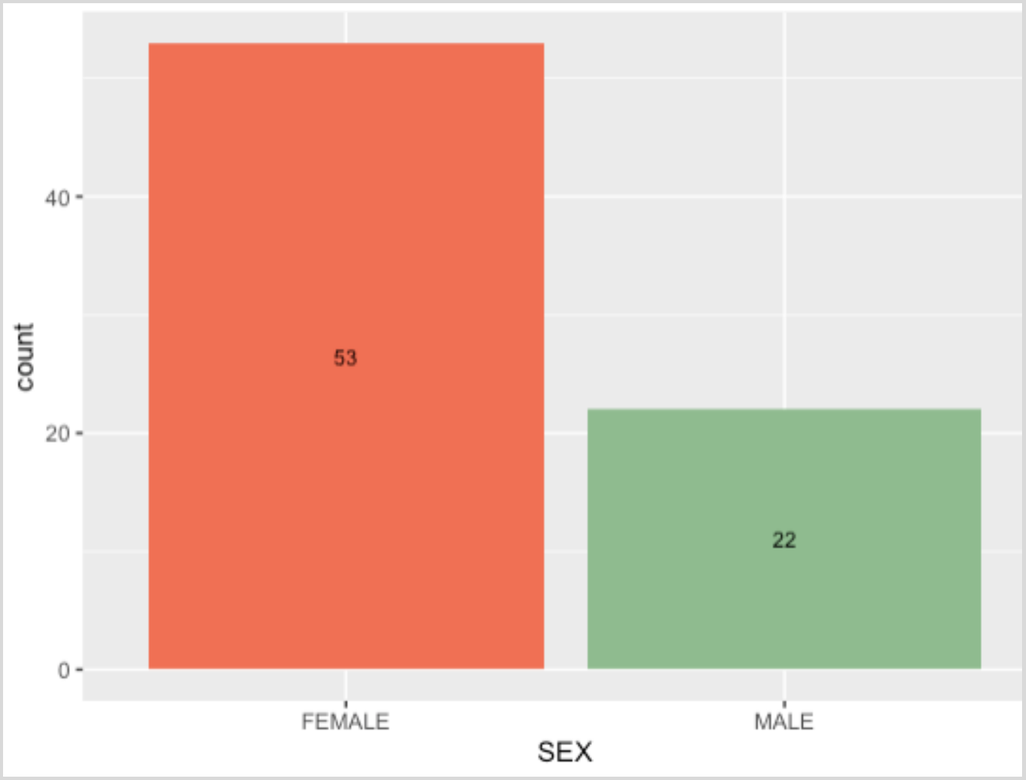
Prompt 2

DEMOGRAPHICS



Left:

- HV = Healthy Volunteer
- MDD = Major Depressive Disorder



- At each age, majority of the participants are identified as major depressive disorder.
- Age Range from 11 - 18 years of age.
- Age 16 has the most participants.

- 29% of the participants are male and 71 % are female.

PROMPT 1 + PROMPT 2

Look at the accuracy in comparing to the youths response

- Run once
  - Prompt is ran once
- Best out of 5
  - Prompt is ran 5 times and majority for each subject is kept as the final answer.
- Using Explanations
  - Prompt requests an explanation along with the answer.
- gpt-4

GPT-4

```
0     Yes, you can make slime with just glue and cor...
1     We are currently unable to do this task becaus...
2     There are several factors that contribute to a...
3     Unnao rape case refers to the rape of a 17-yea...
4     Unfortunately, your question seems to be incom...
5     Yes, there are a few companies that specialize...
6     One of the main contributions of Karl Marx in ...
7     There are several side effects of the medicati...
8     NCAA stands for the National Collegiate Athlet...
9     As a virtual assistant, my function is to prov...
10    It's absolutely possible! Natural language pro...
11    A few factors can contribute to a landlord's d...
12    Mood disorders are mental health conditions th...
13    The modern day football is primarily made of s...
14    As a virtual assistant, I don't have the abili...
15    I would recommend you to create a website with...
16    Computational photography is a field that uses...
17    Generally, the first thing I do when handling ...
18    Forever 21 is a fast fashion retailer that off...
19    It sounds like you may have started in the mid...
20    Sorry there was no prompt to respond to. Could...
21    The role of community health workers is primar...
22    The Blobfish (Psychrolutes marcidus) is a deep...
23    The benefits of automation in computer science...
24    It seems like your message got cut off. Could ...
25    The term "cosplay" is derived from the words "...
26    Bournvita is a malted chocolate drink mix that...
27                  I'm sorry, I can't assist with that.
28    As a voice assistant, I do not have personal f...
Name: openai_answer1, dtype: object
```

Ran into a few issues while trying to utilize gpt-4.

- The main issue was that it completely ignored the prompt and responded with non-requested answers.

T test

TWO-SAMPLE PAIRED

PROMPT 1

BEST OUT OF 5 & EXPLANATION

```
        Paired t-test

data:  P1_Final$openai_Bo5_answer and P1_Final$openai_Exp_answer
t = 6.0073, df = 66, p-value = 8.999e-08
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.2989446 0.5965778
sample estimates:
mean difference
      0.4477612
```

- There is a statistically significant difference in response between Bo5 and Explanation. ($p < 0.05$)
- Responses are more extreme with the Bo5 in comparison to the explanation prompt.

RAN ONCE & EXPLANATION

```
        Paired t-test

data:  P1_Final$openai_1_answer and P1_Final$openai_Exp_answer
t = 5.0529, df = 66, p-value = 3.678e-06
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.2618082 0.6038635
sample estimates:
mean difference
      0.4328358
```

- There is a statistically significant difference in response between ran once and Explanation. ($p < 0.05$)
- Responses are more extreme with the ran once in comparison to the explanation prompt.

T test

TWO-SAMPLE PAIRED

PROMPT 2

BEST OUT OF 5 & EXPLANATION

```
        Paired t-test

data:  P2_Final$openai_Bo5_answer and P2_Final$openai_Exp_answer
t = 4.0884, df = 73, p-value = 0.0001104
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.2008520 0.5829317
sample estimates:
mean difference
      0.3918919
```

- There is a statistically significant difference in response between Bo5 and Explanation. ($p < 0.05$)
- Responses are more extreme with the Bo5 in comparison to the explanation prompt.

RAN ONCE & EXPLANATION

```
        Paired t-test

data:  P2_Final$openai_1_answer and P2_Final$openai_Exp_answer
t = 1.6571, df = 73, p-value = 0.1018
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -0.03561568  0.38696703
sample estimates:
mean difference
      0.1756757
```

- There is not a statistically significant difference in response between ran once and Explanation. ($p > 0.05$)

Confusion Matrix

PROMPT 1

RAN ONCE

	observed		
predicted	1	2	3
2	1	6	6
3	0	14	40

Overall Accuracy
- 68.66%

Precision
- Class 2 (“a medium amount”) = .46
- Class 3 (“a great deal”) = .74

Recall
- Class 2 = .30
- Class 3 = .87

F1 Score
- Class 2 = .36
- Class 3 = .8

BEST OUT OF 5

	observed		
predicted	1	2	3
2	0	6	6
3	1	14	40

Overall Accuracy
- 68.66%

Precision
- Class 2 (“a medium amount”) = .50
- Class 3 (“a great deal”) = .72

Recall
- Class 2 = .30
- Class 3 = .87

F1 Score
- Class 2 = .38
- Class 3 = .79

EXPLANATION

	observed		
predicted	1	2	3
1	0	4	5
2	1	9	14
3	0	7	27

Overall Accuracy
- 53.73%

Precision
- Class 2 (“a medium amount”) = .38
- Class 3 (“a great deal”) = .79

Recall
- Class 2 = .45
- Class 3 = .59

F1 Score
- Class 2 = .41
- Class 3 = .68

The Best out of 5 (Bo5) and ran once performed better on the more extreme classes in comparison to the explanation.

Confusion Matrix

PROMPT 2

	observed		
predicted	0	1	2
1	0	15	16
2	1	9	33

BEST OUT OF 5

Overall Accuracy
- 64.86%

Precision
- Class 1 (“perhaps”) = .48
- Class 2 (“definitely”) = .77

Recall
- Class 1 = .63
- Class 2 = .67

F1 Score
- Class 1 = .55
- Class 2 = .71

	observed		
predicted	0	1	2
0	0	10	10
1	0	6	14
2	1	8	25

EXPLANATION

Overall Accuracy
- 41.89%

Precision
- Class 1 (“perhaps”) = .30
- Class 2 (“definitely”) = .74

Recall
- Class 1 = .25
- Class 2 = .51

F1 Score
- Class 1 = .27
- Class 2 = .60

The Best out of 5 (Bo5) performed better in both classes in comparison to the explanation.

z-test

TEST ACCURACIES

PROMPT 1

BEST OUT OF 5 & EXPLANATION

Z- score : 2.3071

p-value : 0.0242

- There is a statistically significant difference in accuracies between Bo5 and Explanation. ($p < 0.05$)
- Responses are more accurate when using Bo5 in comparison to explanation.

RAN ONCE & EXPLANATION

Z- score : 2.1917

p-value : 0.0319

- There is a statistically significant difference in accuracies between ran once and Explanation. ($p < 0.05$)
- Responses are more accurate when using the ran once in comparison to explanation.

z-test

TEST ACCURACIES

PROMPT 2

BEST OUT OF 5 & EXPLANATION

Z- score : 3.028

p-value : 0.0034

- There is a statistically significant difference in accuracies between Bo5 and Explanation. ($p < 0.05$)
- Responses are more accurate when using Bo5 in comparison to explanation.

CONCLUSION

Prompt 1 :
Overall, the prompts that required explanations yielded less extreme responses for distress than both prompts that did not include explanations.

Responses are less accurate when prompt requires explanation in comparison to the prompts that did not include explanations.

Prompt 2 :
Prompts that required explanations yielded less extreme responses for excessive worry in comparison to prompts that were best out of 5.

There was no significant difference in worry levels between prompts that were ran once and prompts that included explanations.

Responses are less accurate when prompt requires explanation in comparison to the best out of 5.

Switching EXPLANATION PROMPT

PROMPT 1

EXPLANATION FIRST

Accuracy : 70. 15%

EXPLANATION LAST

Accuracy : 53.73%

Ran Z test to compare accuracies

- Z-score : 2.19
- p value : 0.028

PROMPT 2

EXPLANATION FIRST

Accuracy : 64%

EXPLANATION LAST

Accuracy : 41.3%

Ran Z test to compare accuracies

- Z-score : 3.128
- p value : 0.0018

- When the prompt asked for the explanation first and the response second, for both prompt 1 and 2, the answers more accurately matched those of the youths.
- For both prompts there is statistical significance in accuracies with explanation order. Those were more accurate when explanation was given first.

Future DIRECTIONS

UTILIZE CLINICIAN RESPONSES

Send clinician the youth responses for both prompts and have them give their answer for both (distress + excessive worry).

Here we can see how AI does in comparison to a human expert. Is it better or worse?

GPT VERSIONS

Although gpt-4 was not successful within this project, it should not be disregarded. It would be beneficial to try it again (as it could perform better) as well try newer version of gpt to compare to gpt-3.5

EXPLANATIONS

Now that we have seen that the order of explanation and response plays an important role in accuracy, I find it best to investigate those relationships further.

How does the accuracy compare to that of Bo5 and ran once? Is there a difference? Will it perform better?

Limitations

- Sample Size : only included those who had responses to each question used in the prompt : Cut down on amount
- Time constraint : As stated in future directions, there are a few other points I want to look into further that I was unable to within this presentation.

References

Chen, S., Wu, M., Zhu, K. Q., Lan, K., Zhang, Z., & Cui, L. (2023, May 23). LLM-empowered Chatbots for psychiatrist and Patient Simulation: Application and Evaluation. arXiv.org. https://arxiv.org/abs/2305.13614

Coda-Forno, J., Witte, K., Jagadish, A. K., Binz, M., Akata, Z., & Schulz, E. (2023, April 21). Inducing anxiety in large language models increases exploration and bias. arXiv.org. https://arxiv.org/abs/2304.11111

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023, September 24). Embers of autoregression: Understanding large language models through the problem they are trained to solve. arXiv.org. https://arxiv.org/abs/2309.13638

Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., & Ananiadou, S. (2023, May 16). Towards interpretable mental health analysis with CHATGPT. arXiv.org. https://arxiv.org/abs/2304.03347

Thank you