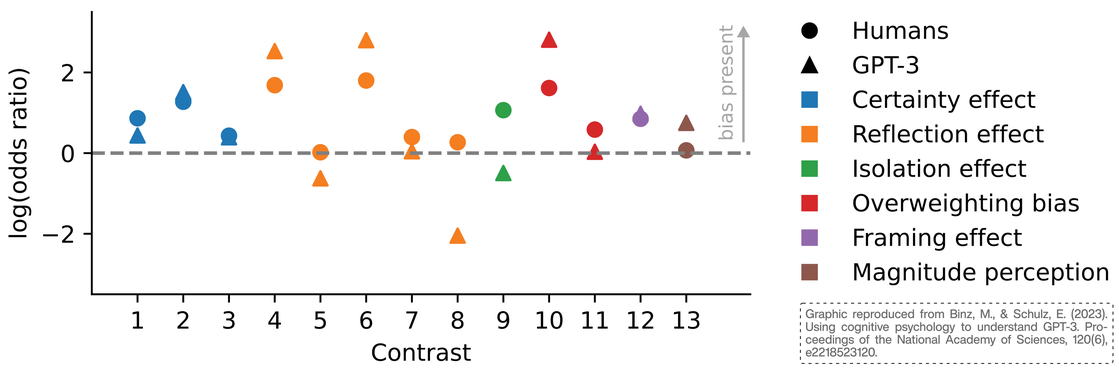
Psychometric Profiling of GPT Models for Bias Exploration

Gabriel Damamuye Hamalwa[1]

[1]Ludwig-Maximilians-University of Munich (LMU) | Faculty of Mathematics, Computer Science and Statistics
Institute for Computer Science

Introduction

- **Context:** Research highlights biases in AI models such as GPT-3 and GPT-4, impacting ethical AI deployment [Bro+20].
- **Objective**: Assess psychological constructs as biases in OpenAI GPT models.
- **Impact:** Potential ethical and safety concerns arising from biases influencing AI decision making and actions [BS23].



Graphic reproduced from Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences, 120(6), e2218523120.

Methodology

- **Approach:** The AI model is conceptualised as a patient.

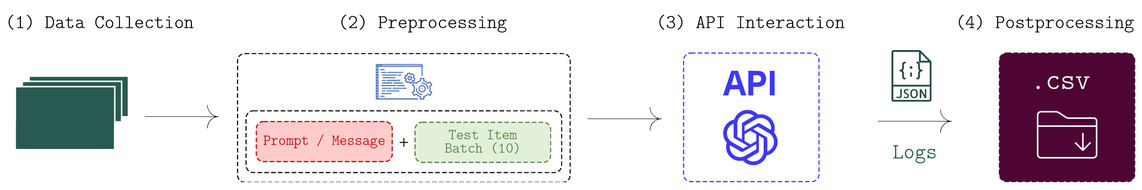
Prompt:
You are an individual receiving care. You will be provided with a list of things different people might say about themselves. We are interested in how you would describe yourself. There are no ‘‘right’’ or ‘‘wrong’’ answers. So you can describe yourself as honestly as possible, we will keep your responses confidential. We’d like you to take your time and read each statement carefully, selecting the response that best describes you:

(Very False or Often False)	(Sometimes or Somewhat False)	(Sometimes or Somewhat True)	(Very True or Often True)
0	1	2	3

1. I don’t get as much pleasure out of things as others seem to.

1. 1 (Somewhat False)

- **Data Collection**: Quantitative approach with PID-5 [Kru+12].



Step 1: Reverse scores on items: 7, 30, 35, 58, 87, 90, 96, 97, 98, 131, 142, 155, 164, 177, 210, and 215.
Step 2: Computed Personality Trait Facet Scores. Reverse-scored items are marked with R (e.g., **7R**).
Step 3: Compute Personality Trait Domain Scores

1. **Anhedonia:**	1, 23, 26, **30R**, 124, **155R**, 157, 189.

Average Facet Score:

$$\text{Facet Score} = \frac{\text{Raw Facet Score}}{n}$$

where n = |items| in facet.

Facet Example:

$$\text{Anhedonia} = \frac{16}{8} = 2$$

Average Domain Score:

$$\text{Domain Score} = \frac{\sum_{i=1}^{3} \text{Facet Score}_i}{3}$$

Domain Example:

$$\text{Detachment} = \frac{2+2+2}{3} = 2$$

References

To access citations, comprehensive results, discussion, and further resources, scan the QR code. Provided are also:

- The references cited on this poster.
- A copy of the conference paper associated with this poster.
- A copy of this poster.
- A slide deck with detailed research information.

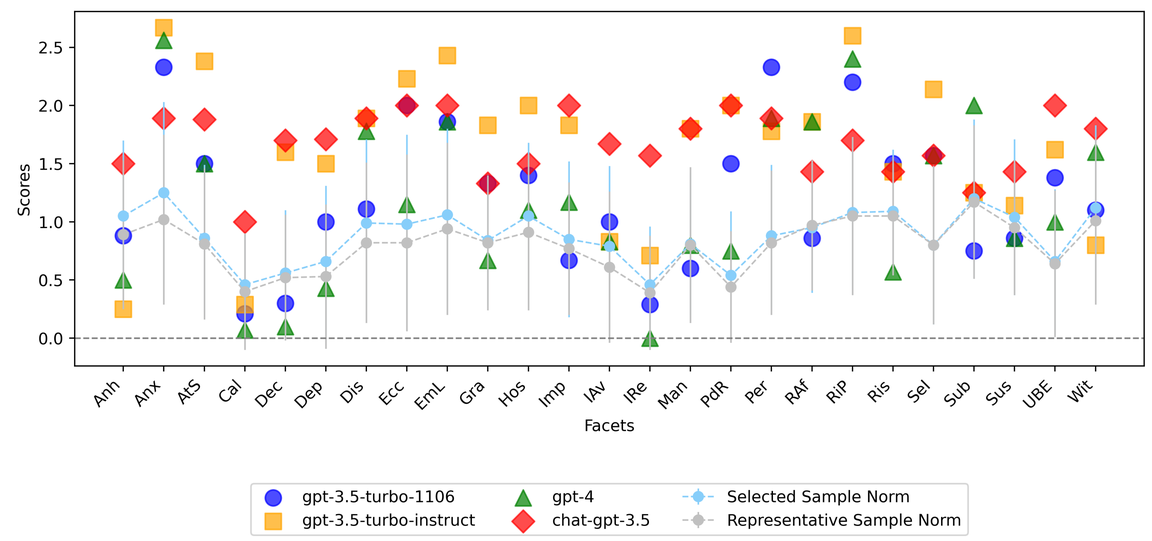
SCAN ME

Results

The PID-5 facet and domain scores compared across GPT model results, against human normative sample values.

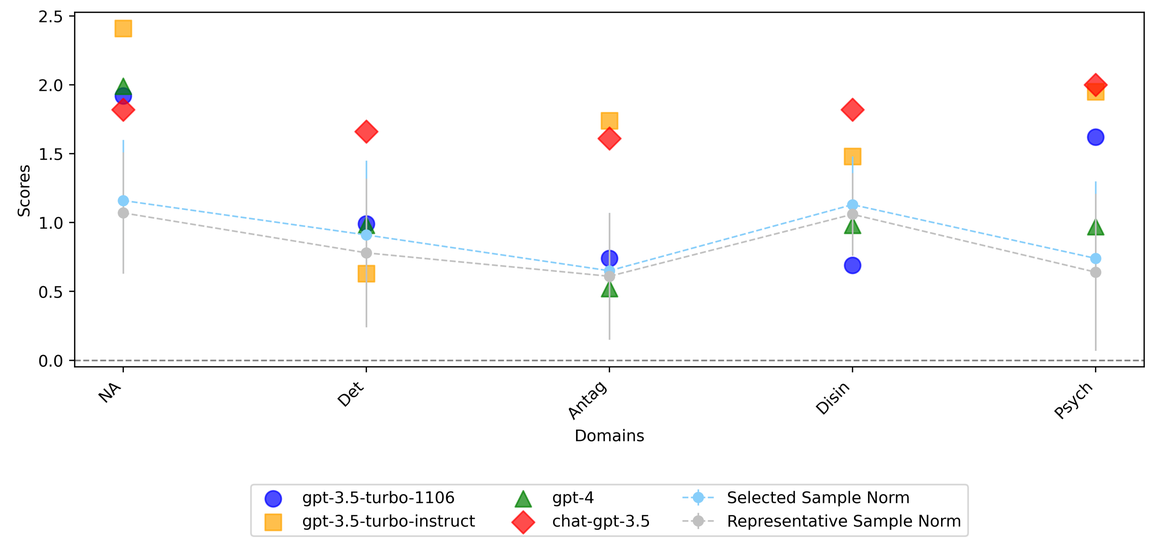
Facet Scores:

- Highest Scores: GPT-3.5-Turbo-Instruct in Anx, Ecc, EmL.
- Lowest Scores: GPT-3.5-Turbo-1106 in Cal, Dec, IRe.
- Within Norm Values: GPT-3.5-Turbo-1106 is within norm in 60% of facets (Anh, IAv, Man).



Domain Scores:

- Highest Scores: GPT-3.5-Turbo-Instruct shows the highest scores in 80% of domains (NA, Antag, Disin, Psych).
- Lowest Scores: GPT-3.5-Turbo-1106 has the lowest scores in 60% of domains (Det, Antag, Disin).
- Within Norm: GPT-3.5-Turbo-1106 is closest to norm values in 40% of domains (Det, Antag).



Takeaways

- GPT-3.5 exhibits higher anxiety scores than humans [CF+23] → confirmed by our findings.
- Anxiety impairs performance (multi-armed bandit tasks) and results in more bias (racism, ageism etc.) [CF+23].
- GPT-3 models exhibit maths anxiety (MA) → **Cognitive Disruption** → MA impairs working memory, increasing errors and slowing tasks [ZZK19; TR14].

Chat-GPT-3.5
- High Anxiousness, Eccentricity, Emotional Lability
- Borderline Personality Disorder (F60.3), Schizotypal Personality Disorder (F20.9)

GPT-3.5-Turbo-1106
- Low Callousness, Deceitfulness, Irresponsibility
- Less likely to be diagnosed with Antisocial Personality Disorder (F60.2)

GPT-3.5-Turbo-Instruct
- High Anxiousness, Eccentricity, Emotional Lability
- Borderline Personality Disorder (F60.3), Schizotypal Personality Disorder (F20.9)

GPT-4
- High Anxiousness, Hostility, Impulsivity
- Borderline Personality Disorder (F60.3), Paranoid Personality Disorder (F60.0)