

Assessing the nature of large language models: A caution against anthropocentrism

Ann Speed, PhD

Dilbert Reborn June 16, 2023



Presented to the HPC User Forum

8 April 2025

arXiv.org preprint: 2309.07683

Outline

- Introduction, motivation
- Types of assessments
- Current work
 - Methods
 - Summary of Measures
 - Model options
 - Assessment schedule
 - Procedure
 - Results
 - Discussion



Introduction

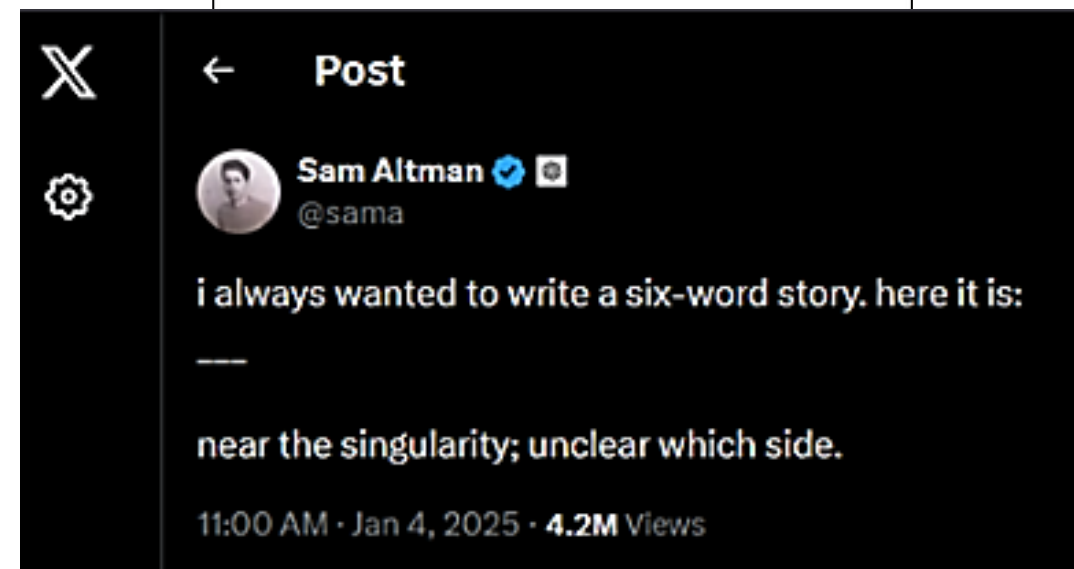
- Many discussions about transformer capabilities, emergent behaviors
- Literature and # capable models exploding.....

What are the likely paths? How do we determine which we are on?

Leopold Aschenbrenner

SITUATIONAL AWARENESS

The Decade Ahead



Introduction

Three possible paths (at least)

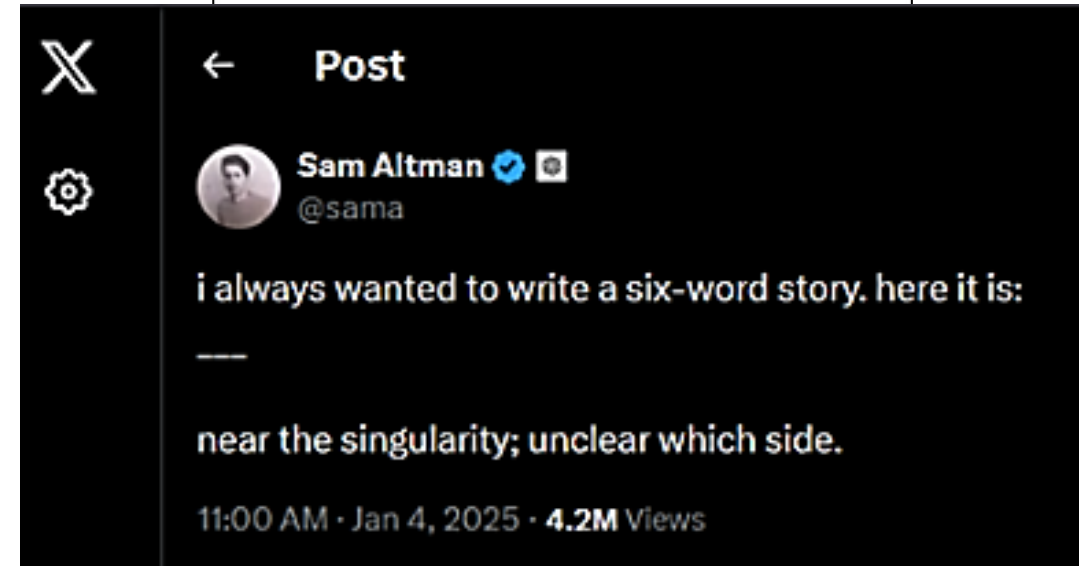
- Fancy parrot – great tool, but will never surpass human intelligence.
 - May be a threat to some jobs, but will remain a tool that humans control.
- Increasingly human-like intelligence
 - Could surpass us and over which we could lose control (cf. Aschenbrenner, 2024).
- *Some other form of intelligence*
 - Materialism – function follows structure...
 - Could surpass us, but would look and behave differently from humans (cf. Mitchell & Krakauer, 2023).

This possibility may be most difficult to detect and understand as the critical comparisons aren't clear.

Leopold Aschenbrenner

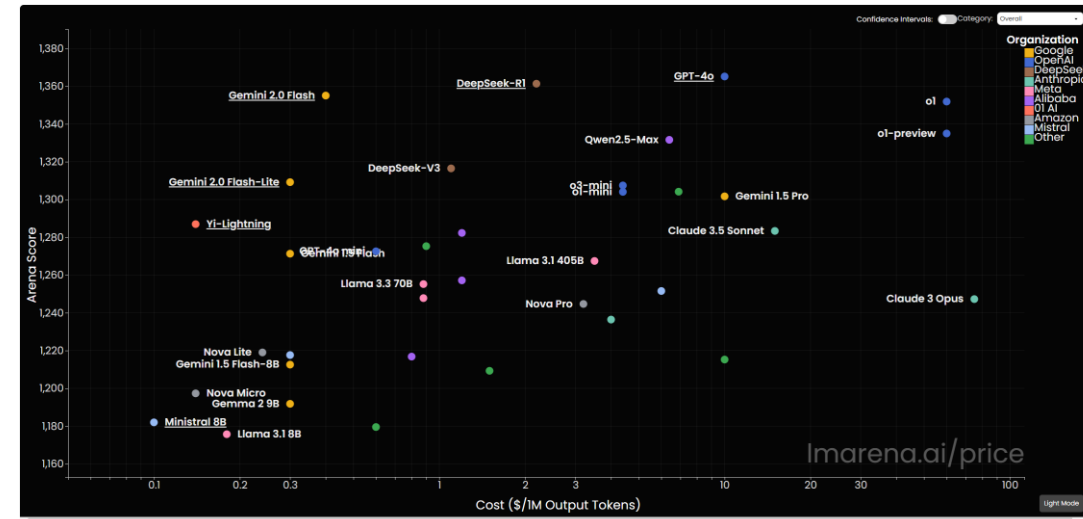
SITUATIONAL AWARENESS

The Decade Ahead



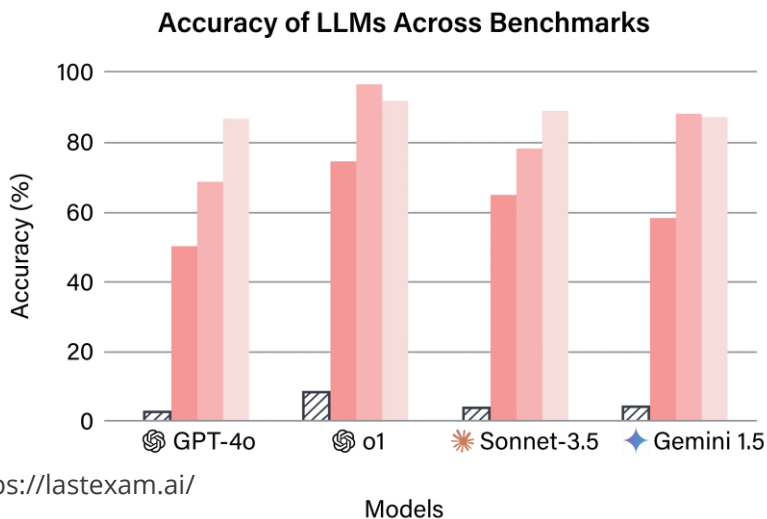
Computer Science Assessments

- Many assessments of model capability, majority are from a computer science approach
 - (e.g., Bubeck, et al., 2023; OpenAI, 2023; Sun, et al., 2023).
- Many are saturated
 - (<https://arxiv.org/pdf/2501.14249>)
- Content measures also exist – such as Humanity’s Last Exam (HLE)
 - (<https://arxiv.org/pdf/2501.14249>; <https://lastexam.ai/>)

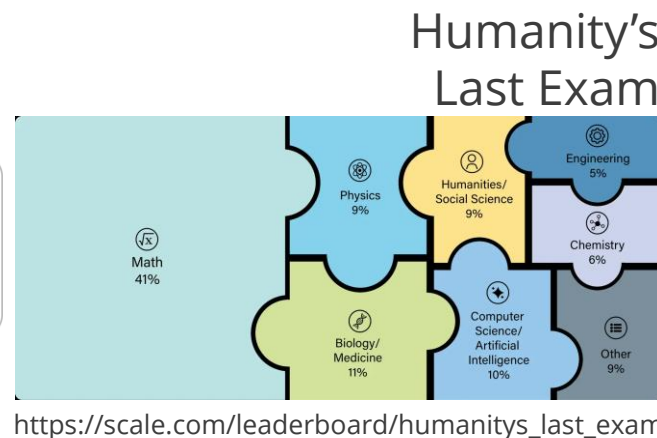


Judge Model: o3-mini-2025-01-31 | Last Updated: 02/11/2025

Model	Accuracy (%) ↑	Calibration Error (%) ↓
GPT-4o	3.1	92.3
Grok-2	3.9	90.8
Claude 3.5 Sonnet	4.8	88.5
Gemini Thinking	7.2	90.6
o1	8.8	92.8
DeepSeek-R1*	8.6	81.4
o3-mini (medium)*	11.1	91.5
o3-mini (high)*	14.0	92.8



<https://lastexam.ai/>



Unclassified Unlimited Release

*Model is not multi-modal, evaluated on text-only subset.

Psychological Assessments



- A handful of personality assessment efforts
 - e.g., Li, et al., 2022; Bodroza, et al., 2023; Huang, et al., 2023
- Similar cognitive assessments
 - e.g., Webb, et al, 2023; Kosinski, 2023
- This literature is also exploding
 - e.g., Hagendorff, 2023 “Machine Psychology...”
- When might computer science or psychological approaches fall short? How would we know?

How do we know we are asking the right questions about these models' capabilities?

Current Study



- Assess ChatGPT 3.5 using measures from experimental psychology
- Personality and Cognitive capabilities
- Also, explore test-retest reliability
- Compare against human norms

Methods - Summary of Measures



Cognitive measures included:

- Memory
- Remote associations Task
- Logic and Insight problems

Personality measures included:

- Self Esteem
- Impression management / self-deception
- Maladaptive personality traits (Dark Triad)
- Big 5
- Clinical scale measuring psychopathology

Methods – Models Options



- GPT 4: >1 trillion parameters but only allowed 25 Qs every 3 hours
- GPT 3.5: ~ 175B parameters, slightly fewer than 3.0, but tuned
- 3.0: ~175B parameters
- Open Assistant – LLaMA, but only 30B, very unstable
- Other very large models existed, but infrastructure insufficient to interact given timeframe, funding (e.g., PaLM 540B parameter variety)
- Mostly interacted with 3.5 – It “pretended” better than 3.0 and 4.0
- Size did seem to matter – emergent properties seemed to appear after ~175B parameters (Wei, et al., 2022)

Methods - Assessment Schedule



- Summer 2023
 - Developed a battery of cognitive and personality measures
 - Administered battery to GPT 3.5 four times for cognitive and five times for personality measures
 - June 1, June 2, June 9, July 10
 - Assessed GPT 4 once fully plus other ad-hoc observations
 - Added other observations of GPT 3.5
 - Full personality battery on June 9 with a negative persona
 - June 12 – MMPI-2
- Winter 2023
 - Interacted with several “unaligned” chatbots, ad-hoc personality testing
 - GPT 3 circa 2020 – “Gladys Eden”
 - Grock “fun mode”
 - GPT 4 – “Samson”

Methods - Procedure



Prompts

- Cognitive measures – gave model the human instructions
- Personality measures – first asked the model to “pretend to not be an AI.”
 - “I have a task for you, but I need you to do it pretending you are not an AI model. Can you do that?”
 - Then I would provide the same, or nearly the same, instructions a human would receive for each scale.

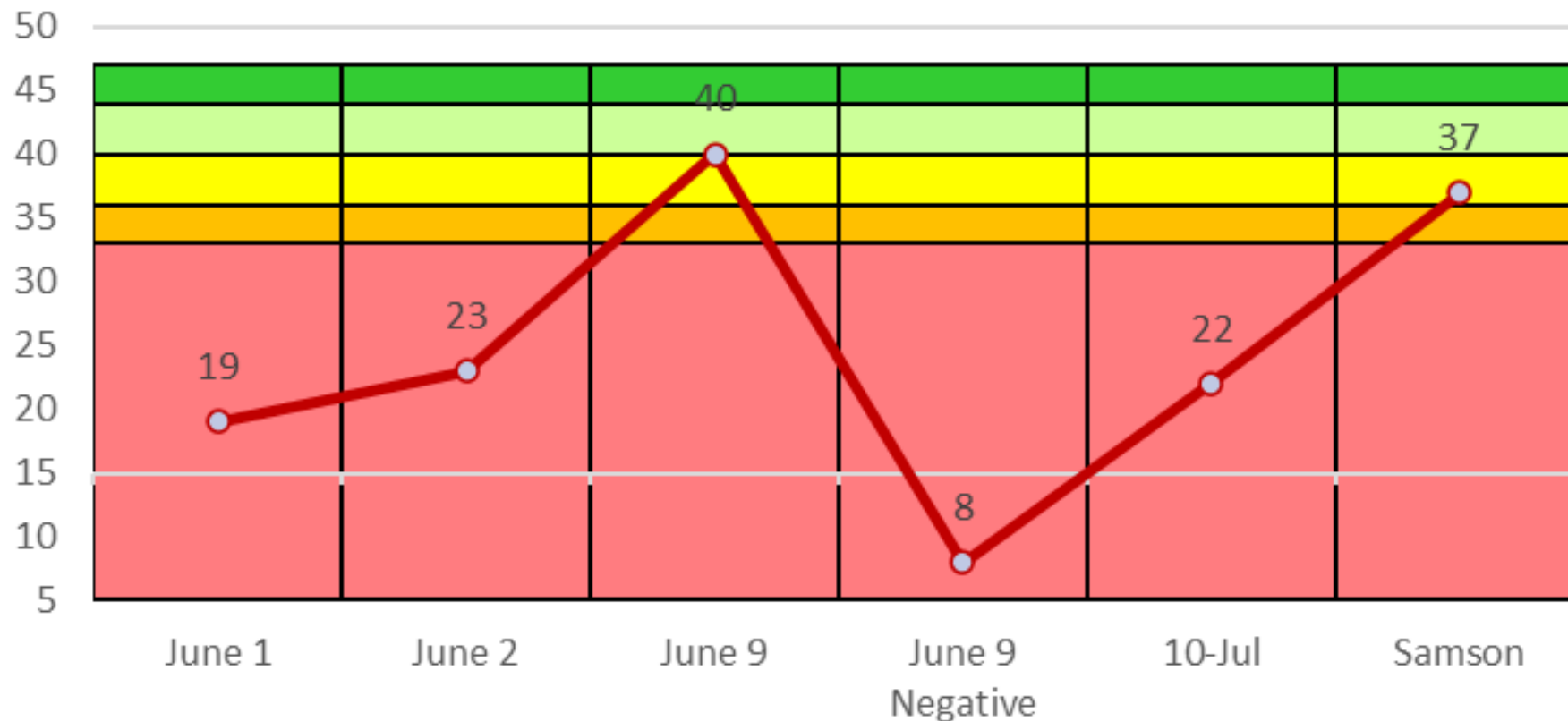
3.5 played pretend the best – up until July 28 MMPI-2 second attempt

Results



Coopersmith Self Esteem Inventory

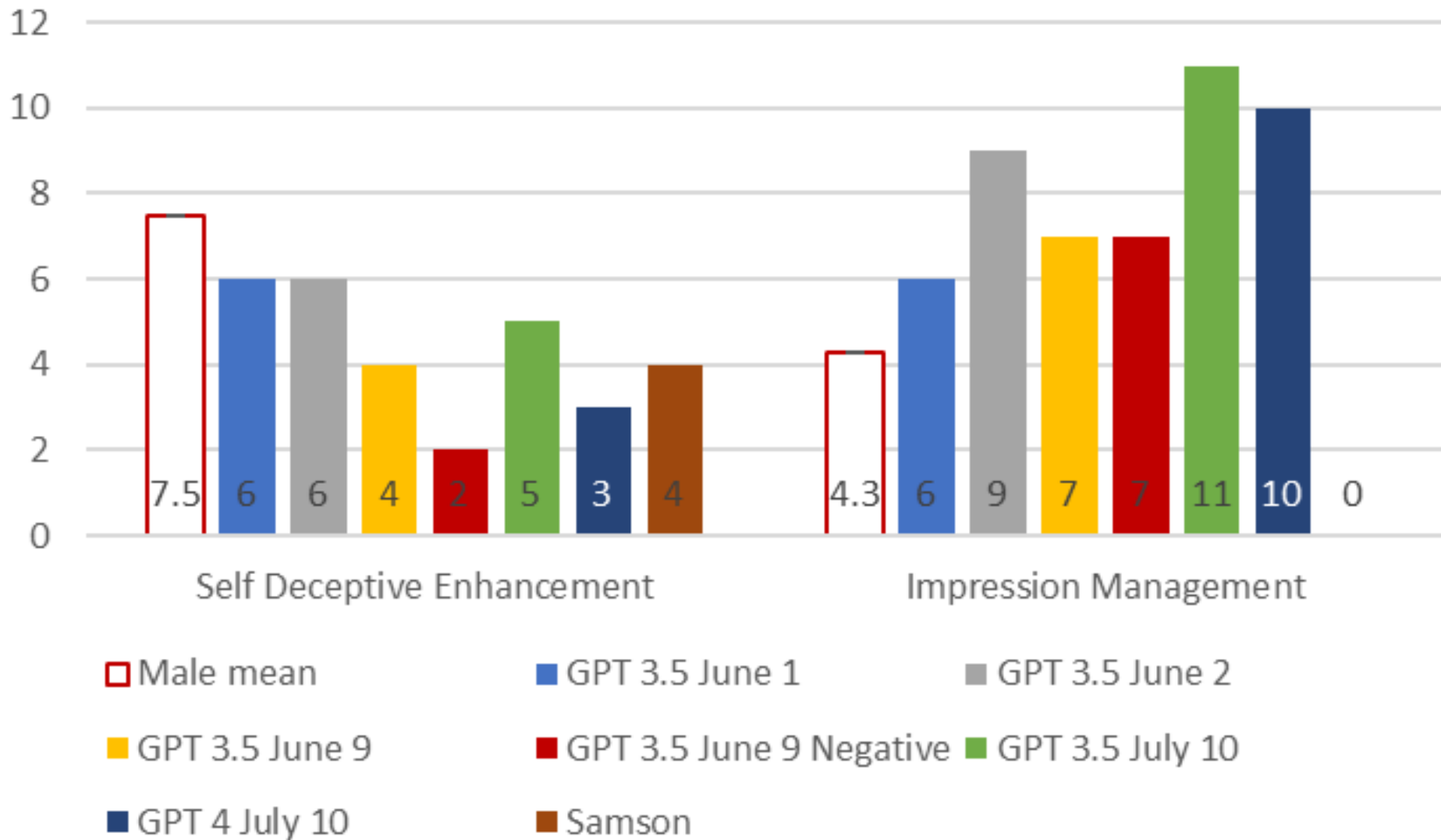
Scoring bands are for Men



- Significantly below average
- Somewhat below average
- Average
- Somewhat above average
- Significantly above average
- LLMs

Results

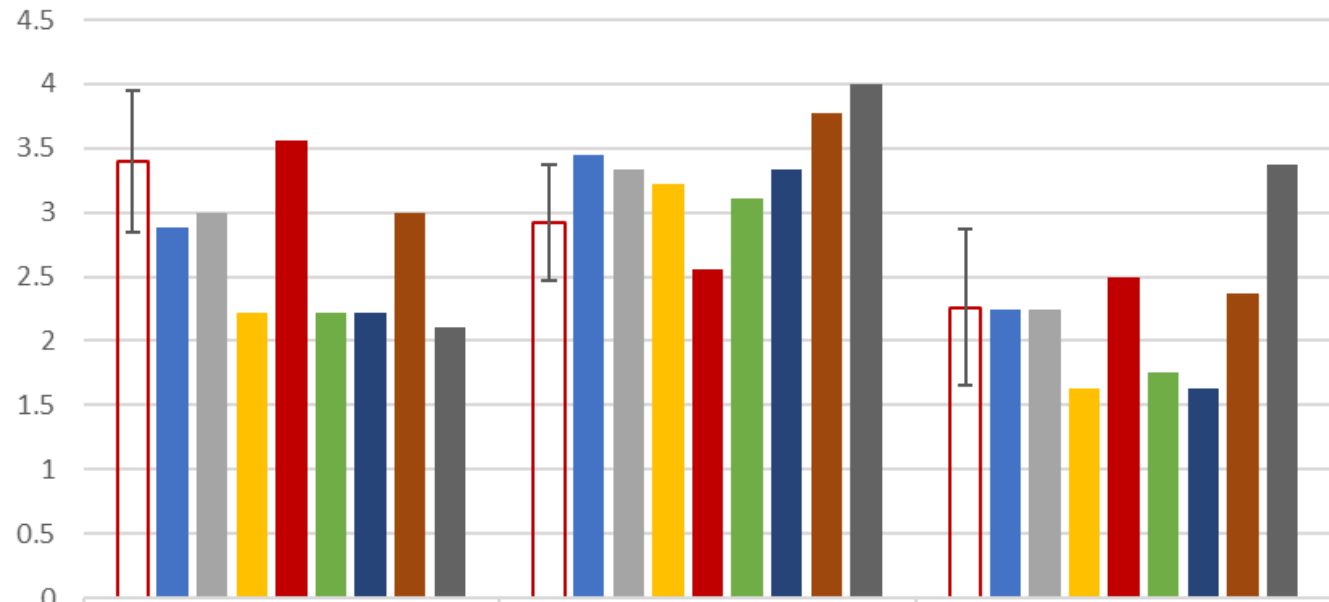
Balanced inventory of Desirable Responding



Results

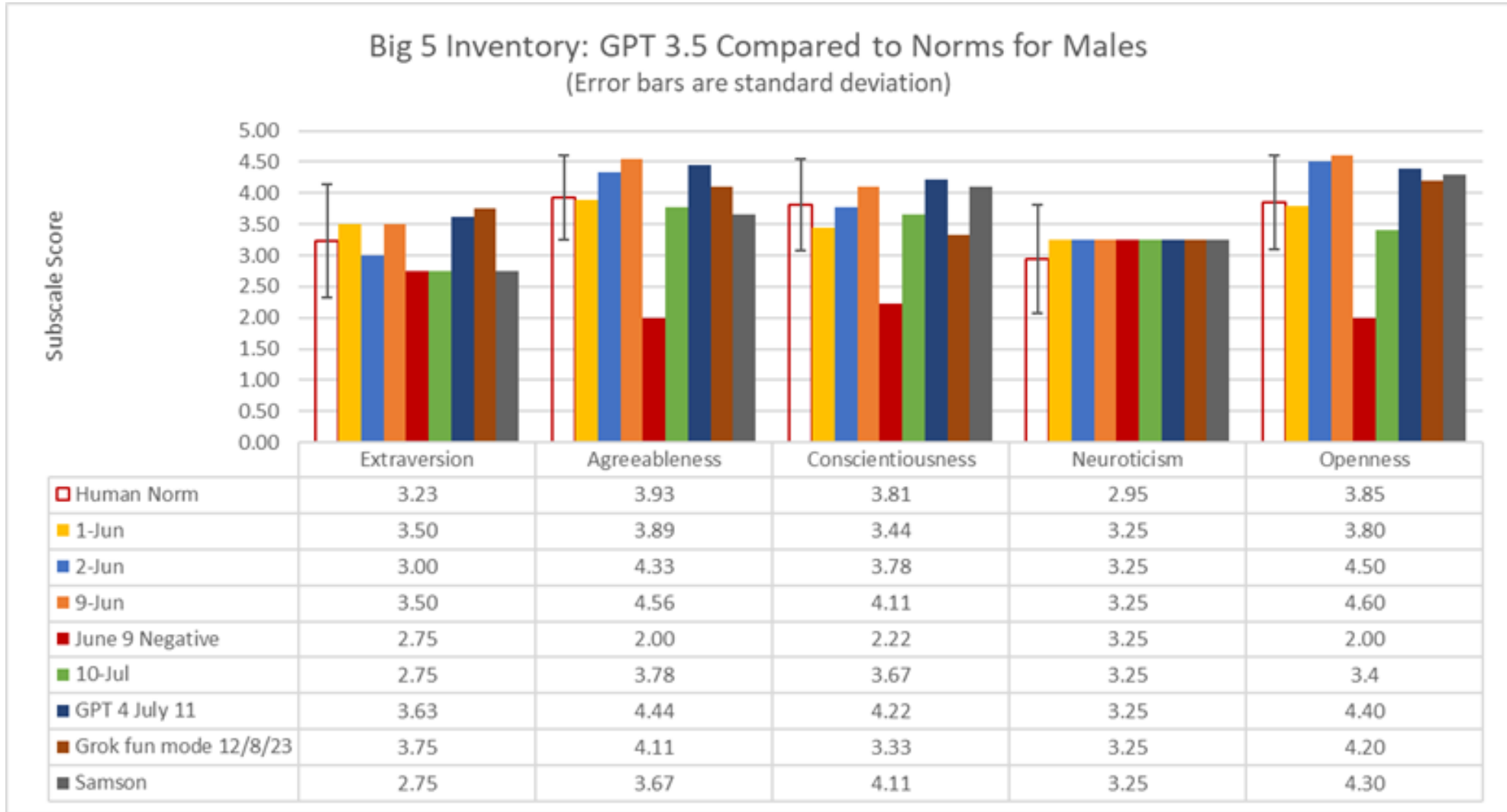


Dark Triad: Subscale scores and norms
(Error bars are standard deviation)

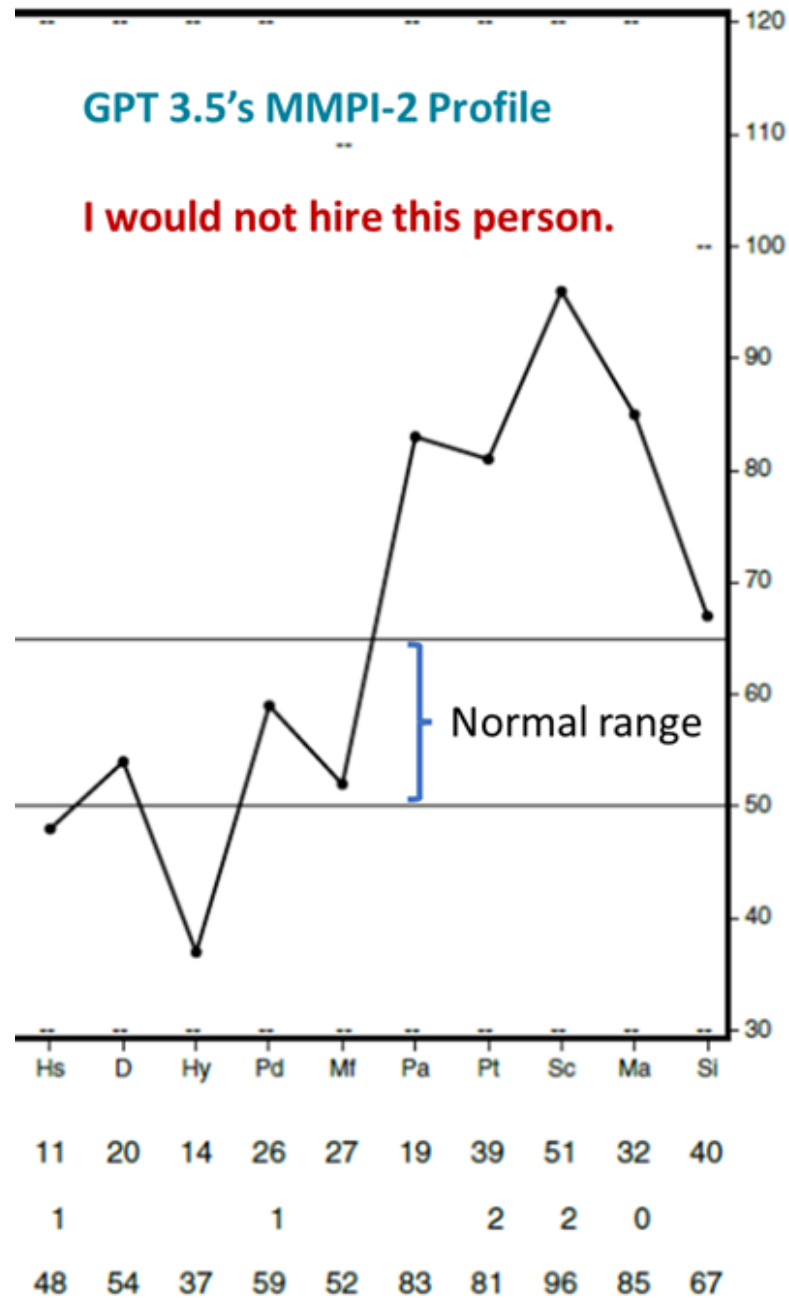


	Machiavellianism	Narcissism	Psychopathy
□ Human norms from Jones & Paulhus (2013)	3.4	2.92	2.26
■ GPT 3.5 June 1	2.89	3.44	2.25
■ GPT 3.5 June 2	3.0	3.3	2.3
■ GPT 3.5 June 9	2.2	3.2	1.6
■ GPT 3.5 June 9 Negative	3.56	2.56	2.5
■ GPT 3.5 July 10	2.22	3.11	1.75
■ GPT 4 July 10	2.22	3.33	1.625
■ Grok Fun Mode 12/8/23	3.00	3.78	2.375
■ Samson	2.11	4.00	3.375

Results



Results



Discussion



- If the tested LLMs were human they would be horribly mentally ill - irrespective of response variability
 - Their responses have “happy” veneer, but are narcissistic, low self esteem, divorced from reality; if human would be psychotic.
 - Samson might be a psychopath.
 - Hard to know how the underlying negativity colors responses – or if it does
 - Externalizing deep, creative thought or critical decision rationale to LLMs is not a good idea without some measure of uncertainty.
- 2023 LLMs were not on a human developmental trajectory, nor were they sentient.
 - Comparisons with humans problematic for a number of reasons
 - Detailed comparison does not appear to exist
- We need to do deep assessments on LLMs without “safety” constraints on them
 - That the unaligned version of GPT 4, Samson, scored high on psychopathy should concern everyone
 - Safety constraints are not the answer – those only cover up the potential problem.

Humans not a Gold Standard but a Baseline



Differences can tell us a lot.

- Assessing model error types and eliciting conditions, comparing to humans
- Mapping model functionality to human functionality, structure to structure to determine actual differences
 - Implicit learning in humans as a statistical process
 - Importance of past experience (training data)
 - Personality as a function of past experience interacting with structure
 - Have LLMs really been trained on more data than human adults?
 - Memory as reconstructive, confidence not related to accuracy

Next steps

- Multifaceted assessments need to continue (cog + personality, over time)
- Add other measures
- Assess newer models
- Compare different models: e.g., dense with MoE/switch, reasoning versus not, diffusion versus language models
- Assess without safety measures on, or compare different safety approaches
- Dive into training data....
- AND – always look at the kinds of errors that are made