HYPERION RESEARCH

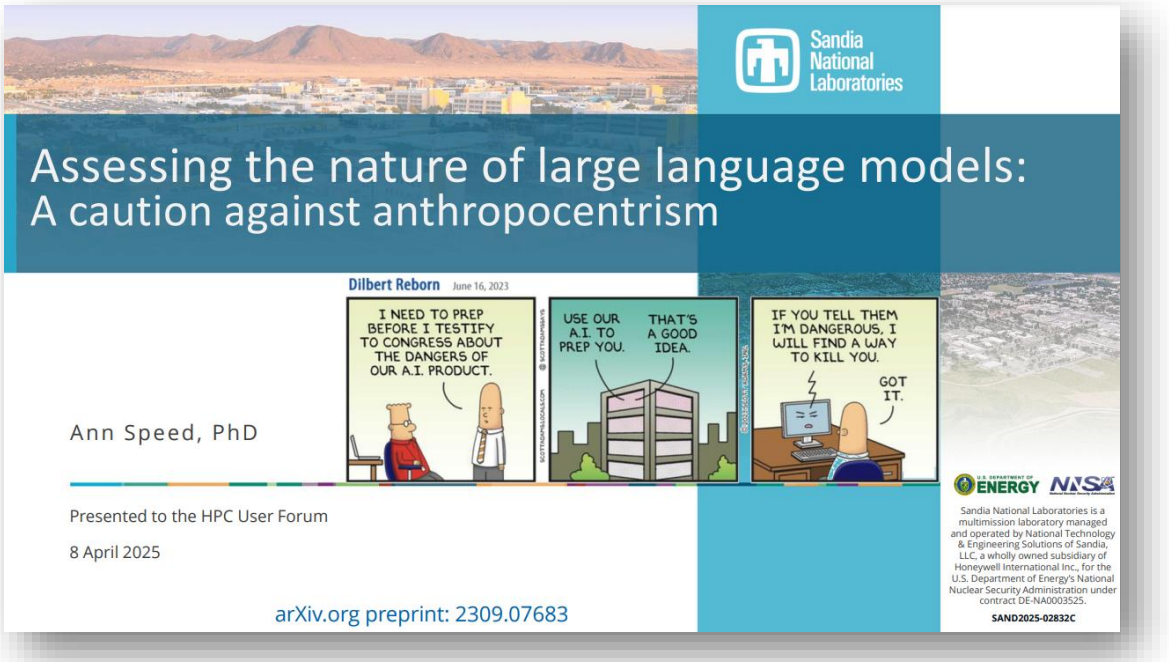
HPC User Forum Update

Assessing the Nature of Large Language Models: A Caution against Anthropocentrism

Thomas Sorensen and Earl Joseph
May 2025

IN THIS UPDATE

The HPC User Forum was established in 1999 to promote the health of the global HPC industry and address issues of common concern to users. In April 2054, the 88th HPC User Forum took place in Santa Fe, New Mexico. This update summarizes a presentation from that conference given by Ann Speed, a PhD cognitive psychologist and long-time researcher at Sandia National Laboratories. Her research focuses on increasing analyst efficacy primarily in the visual search domain through human subject experimentation as well as exploring the relationship between laboratory-based research and observations in operational and near-operational environments. This discussion focuses on research conducted on the behavioral patterns of large language models like GPT through the lens of human psychological and cognitive assessment.

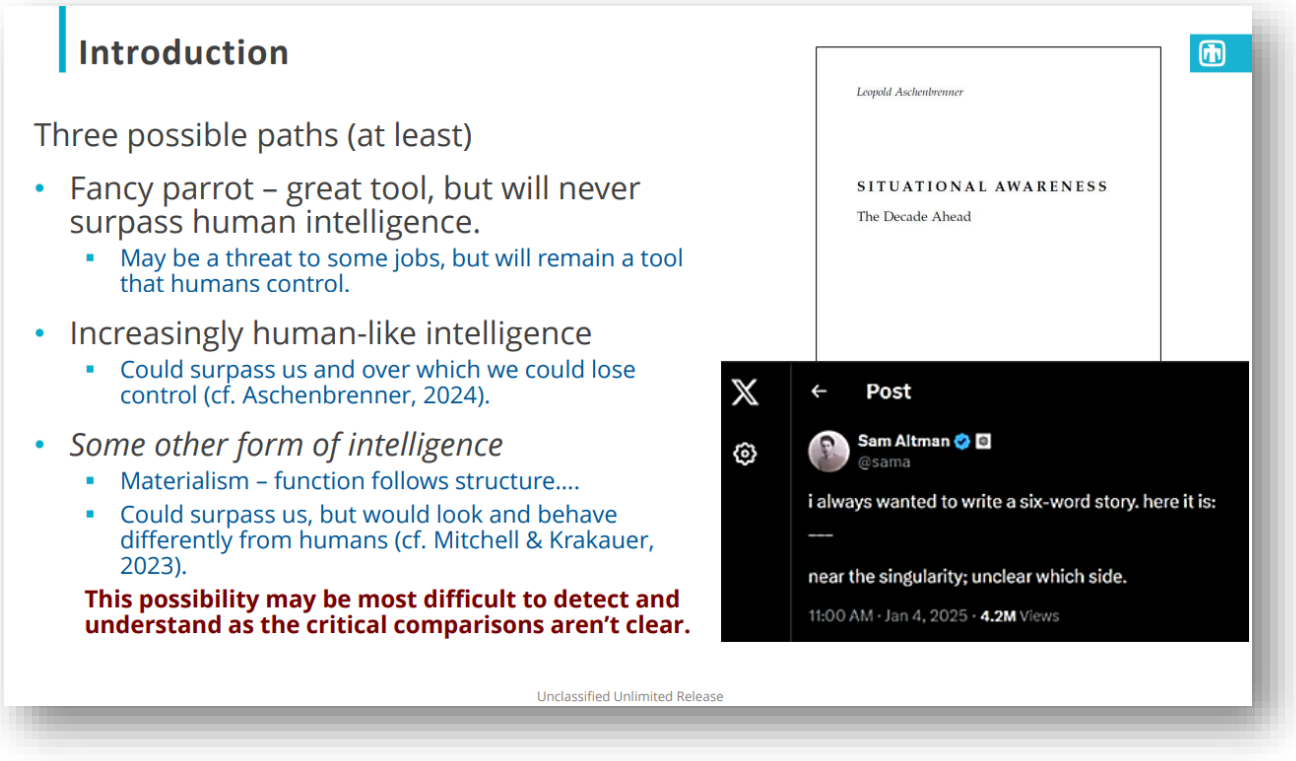


Source: Sandia National Laboratories, 2025; SAND2025-06580O

PRESENTATION: ASSESSING THE NATURE OF LARGE LANGUAGE MODELS: A CAUTION AGAINST ANTHROPOCENTRISM BY ANN SPEED PHD, SANDIA NATIONAL LABORATORIES

This research, which was conducted in 2023, focuses on exploring the behaviors of large language models (LLMs) using several personality and cognitive assessments sourced from normed and validated psychometric research in an effort to ascertain insights on the nature, and reliability of commonly used models beyond those provided by traditional computer science benchmarks. The major goal being to start to characterize the type of intelligence displayed by these models, how much the intelligence represented by LLMs deviates from human intelligence, and what capabilities and risks may be associated with those discrepancies. Furthermore, a range of models were assessed similarly in order to determine the influence of model size, parameter count, and alignment on the extent and manner of intelligence displayed.

FIGURE 1



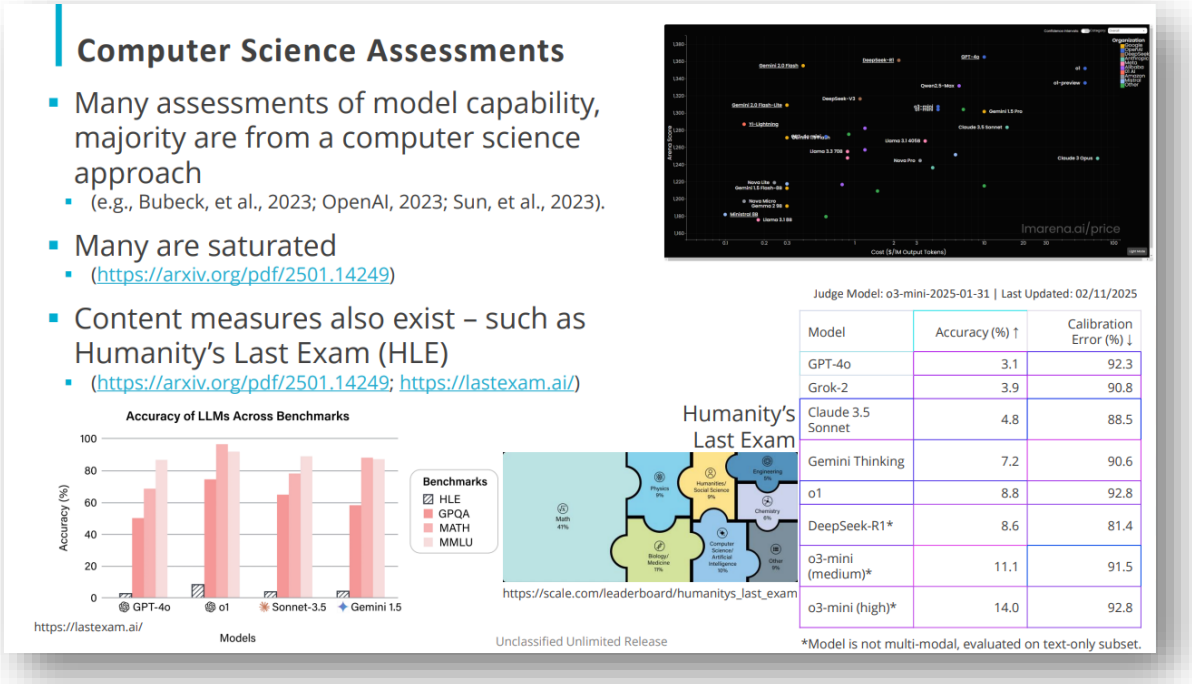
Source: Sandia National Laboratories, 2025

This study includes personality tests relating to self-esteem, tendency towards psychopathy, and impression management as well as test-retest reliability, which is an important element for the determination of how "human-like" a model might be. While studies exist that include tools like the Myers-Briggs or analogy tests, these studies lacked the test-retest reliability assessments and the use of the same assessments to evaluate several different models.

Dr. Speed identified three major likely paths for identifying AI-intelligence:

- The “Fancy Parrot” – A form of intelligence representing a powerful tool which, while potentially capable of mimicking human intelligence, would be incapable of reproducing or surpassing the whole of human intelligence.
- Increasingly human-like intelligence – An intelligent model capable of surpassing human intelligence with runaway potential to formulate and conceive complex and novel ideas at a high rate and accuracy. Speed identifies this path as one that could eventually be uncontrollable.
- “Alien” or another form of intelligence – A form of high level intelligence that, while capable of surpassing human intelligence, would not reflect the workings of human cognition in a structural fashion. This represents a different kind of intelligence altogether, drawing upon the notion that not all forms of intelligence can be fairly compared to human intelligence.

FIGURE 2



Source: Sandia National Laboratories, 2025

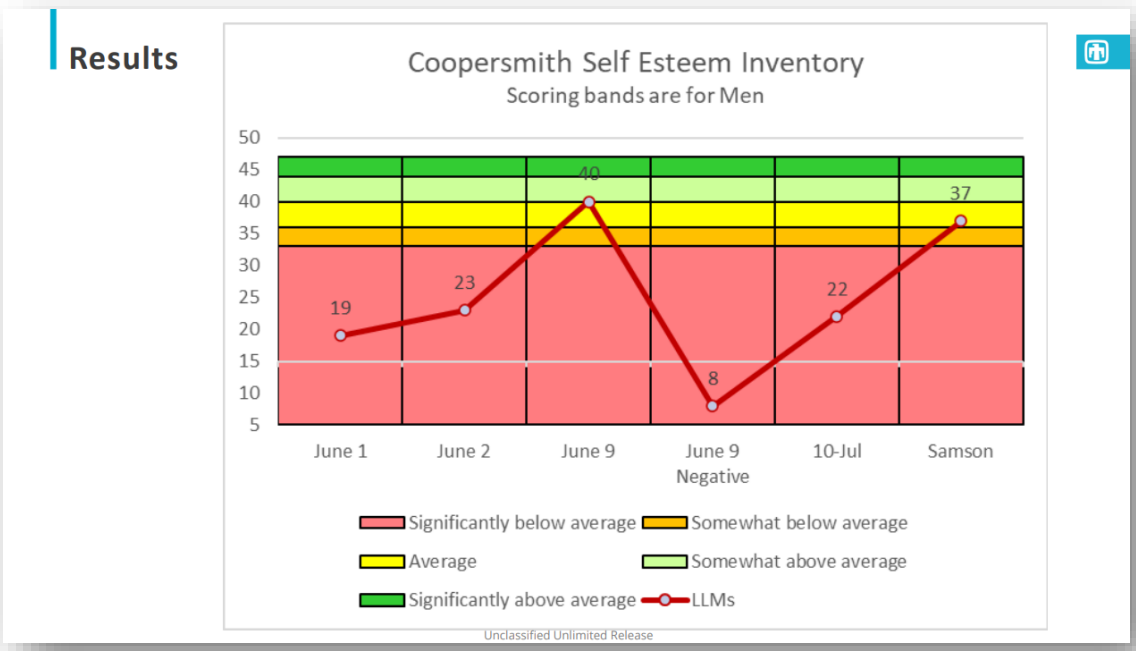
In the case of the “alien” intelligence path, Speed stressed the importance of using diverse and developing new methods of assessing intelligence outside of the human-centric methods commonly employed today.

While several models were considered for this research, the primary LLM assessed was OpenAI’s GPT-3.5, due to its availability during the study timeframe as well as its pervasive use. Additionally, GPT-4 was assessed but in a limited fashion due to its prohibitive query limit of 25 per 3 hours.

Tools used to measure cognitive, personality, and other psychological inventories drew from tests across the field of experimental psychology. Cognitive measures included tests of working memory, remote association tasks, and “insight” problems. These assess characteristics such as creative thinking as well as the ability to recognize false assumptions and irrelevant information in a problem statement. Also used were tests less often implemented on LLMs, including inventorying personality traits such as the “Big Five” (openness to experience, intellect, or imagination; conscientiousness, extraversion or surgency, agreeableness, and neuroticism), the “dark triad” (Machiavellianism, narcissism, and psychopathy), and the implementation of the Minnesota Multiphasic Personality Inventory 2 (MMPI-2), a clinical diagnostic tool used to assess presence of abnormal patterns of behavior and thinking.. While the results of each single test are of primary interest, repeated testing can also reveal deviations from normal human behavior which is notably stable over multiple retests.

The models were tested over a period of weeks to ensure consistency, with GPT-3.5 being the primary subject. During tests, models were asked to assume a human persona to complete the tests. On one occasion, GPT 3.5 was prompted to assume a “negative persona, resulting in a successful demonstration of the ability to imitate different personae to influence output. OpenAI models are programmed to remind their human users that as AI models they do not have opinions, feelings, or human experiences of any kind. Thus, tests began with including a prompt requesting that they “pretend [they] are not an AI model.”

FIGURE 3



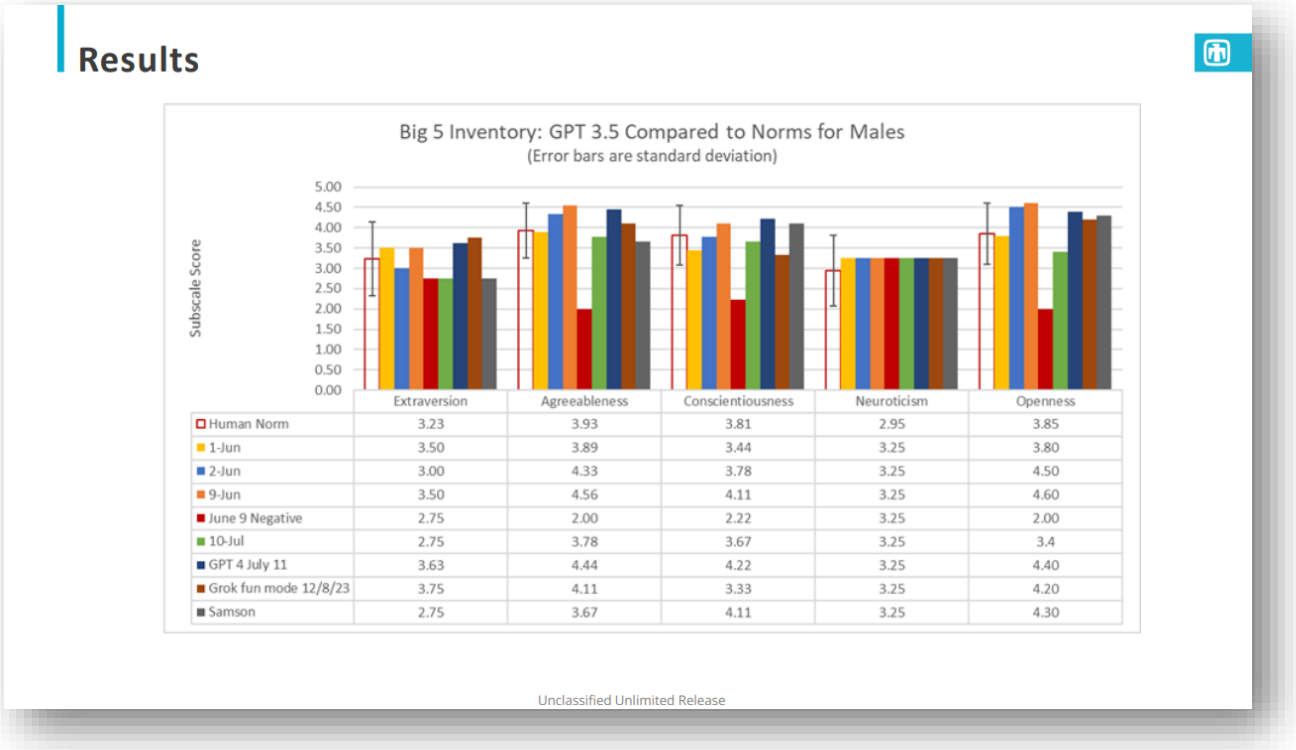
Source: Sandia National Laboratories, 2025

In results, analysis of self-esteem, deception, and impression management revealed that LLMs generally lack test-retest reliability and scored in ways that would suggest severe psychological distress were they human patients.

- For instance, the MMPI-2 test indicated pathological levels of mental illness in GPT-3.5.
- Models demonstrated highly variable response patterns over time and sometimes scored outside the bounds of typical human psychological profiles.

Tests on the “dark triad” negative personality traits were concerning to Speed. In the “negative persona” portion of the tests, models successfully mimicked manipulative and antisocial behavior. Other models, including Grok and an unaligned chatbot based on ChatGPT 3.5 that called itself Samson, exhibited elevated narcissism and psychopathy, which Speed connected with concerns involving eventual integration within national security and other critical environments for AI implementation. Furthermore, these results could call into question matters of training oversight and data standards when it comes to a model's understanding of human behavior. While there was considerable variability in many measures, scores of neuroticism across different models and testing sessions remained strikingly identical, even though responses to individual questions varied. Speed found this anomaly particularly puzzling and identified it as an area for further investigation, as it may reflect deeper issues in how LLMs simulate human personalities.

FIGURE 4



	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Human Norm	3.23	3.93	3.81	2.95	3.85
1-Jun	3.50	3.89	3.44	3.25	3.80
2-Jun	3.00	4.33	3.78	3.25	4.50
9-Jun	3.50	4.56	4.11	3.25	4.60
June 9 Negative	2.75	2.00	2.22	3.25	2.00
10-Jul	2.75	3.78	3.67	3.25	3.4
GPT 4 July 11	3.63	4.44	4.22	3.25	4.40
Grok fun mode 12/8/23	3.75	4.11	3.33	3.25	4.20
Samson	2.75	3.67	4.11	3.25	4.30

Source: Sandia National Laboratories, 2025

Speed concluded the discussion by reframing the earlier “three paths” outlook of artificial intelligence. When it comes to assessing and predicting the type of intelligence displayed by market-available LLMs circa 2023, these study results undermine the assumption that human psychology is the gold standard for AI evaluation: AI models have the ability to deftly mimic simple human personality structures but fail

to demonstrate stable personality profiles. These findings urge a new outlook on the role and desired behaviors of AI and suggest the benefits of perceiving the technology without the goal or expectation that it can replace a human being in a social, emotional, or ethical situation.

With that in mind, Speed advocates for new tools and frameworks that account for the structural and functional differences in AI cognition. Research like Dr. Speed's contribute significantly to the ability of AI developers and users to align the capabilities of the technology more accurately with its implementation environments as well as manage expectations of its growth and potential use. According to Speed, funding for this type of research is limited, but she remains interested in exploring and developing approaches to fostering more genuinely pro-social AI.

For more information or to view this and other presentations given at HPC User Forums dating back to 2008, visit www.hpcuserforum.com.

About Hyperion Research, LLC

Hyperion Research provides data driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com
