ChatGPT-5 Pro の MensaNorway による知能指数が 148 と公表

Felo AI

概要

2025 年 8 月 12 日、AI 監視プラットフォーム「TRACKING AI」は、OpenAI の最新モデルである ChatGPT-5 Pro が、Mensa Norway が提供する IQ テストにおいて 148 という極めて高いスコアを記録したと公表しました。このスコアは、人間の IQ 分布において上位約 0.1%に相当し、天才集団 Mensa の入会基準（上位 2%）を大幅に上回るものです。この結果は、AI、特に大規模言語モデル（LLM）の特定の認知能力、とりわけ非言語的な流動性知能が人間トップレベルに達したことを示す画期的な出来事と見なされています。

しかし、この数値を額面通りに受け取ることには慎重な意見が多数を占めます。専門家は、人間用に設計された IQ テス

トを AI に適用することの根本的な妥当性に疑問を呈しています 35 。主な課題として、①訓練データにテスト問題が含まれている「データ汚染」の可能性、②AI と人間の知能構造の質的な違い、③IQ テストが測定できる知性の範囲の限定性、などが挙げられます 34 35 。

実際に、他のベンチマークや異なる条件下でのテストでは、AI の能力の特異性が浮き彫りになります。例えば、より純粋な推論能力を測るとされるオフラインテストではスコアが変動する傾向があり 13 、また、抽象的なルール発見能力を問う ARC-AGI のようなベンチマークでは、依然として人間との間には大きな隔たりが存在します 2 17 。

ChatGPT-5 Pro の IQ148 という結果は、AI の進化のマイルストーンであると同時に、我々に「知性とは何か」という根源的な問いを突きつけます。AI が特定の知的作業で人間を凌駕する時代において、教育や社会は、IQ スコアで測られる能力だけでなく、創造性、批判的思考、そして「新しい問いを立てる能力」といった、より人間的な知性の価値を再評価する必要に迫られています 8 13 。

詳細レポート

ChatGPT-5 Pro の IQ スコアと Mensa Norway テスト

AI の能力を追跡・比較するプラットフォーム「TRACKING AI」は、毎週、主要な AI モデルに対して Mensa Norway の IQ テストを実施しています 52 。2025 年 8 月 12 日更新のデータによると、ChatGPT-5 Pro（Vision）モデルが IQ 148 というスコアを記録しました。

Mensa Norway テスト: このテストは、主に図形のパターン認識や論理的関係性を見出す能力を測定するもので、言語能力を必要としない非言語的テストです。典型的な問題は、一連の図形の中から法則性を見つけ出し、欠けている部分に当てはまる図形を選択する形式（行列推理問題）です。これは心理学でいう「流動性知能（Fluid Intelligence）」、すなわち新しい問題に適応し、解決する能力を測るものとされています 9 。

IQ 148 の統計的重要性: 標準的な IQ テストでは、平均が 100、標準偏差が 15 に設定されています。この分布に基づくと、IQ 148 は上位約 0.13%（約 740 人に 1 人）に位置する非常に高いスコアです。これは、Mensa の入会基準である上位 2%（IQ 130 以上）を大きく超える水準であり、一般的に「ギフテッド」または「天才」と呼ばれる領域に含まれます 34 。

AI への IQ テスト適用の妥当性と限界

AI が人間向けの IQ テストで高いスコアを出すという事実は、多くの議論を呼んでいます。AI の能力を測る一つの指標として注目される一方で、その解釈には細心の注意が必要です 35 。

データ汚染（Test Contamination）の問題 最も大きな懸念は、LLM がその広範な訓練データの中に、IQ テストの問題や類似のパターンを既に学習している可能性です 35 。AI は「初見で問題を解く」のではなく、「記憶している答えを再現

している」だけかもしれません [34]。この問題を回避するため、訓練データに含まれないように設計された新しいベンチマーク（例: LiveBench）の重要性が増しています [42]。

人間と AI の知能の質的差異 ChatGPT 自身も認めるように、AI と人間の知能は、その構造、記憶体系、得意・不得意の分布が根本的に異なります [4] [34]。

- **AI の得意分野**: 数学、パターン認識、論理推論、膨大な情報の記憶と処理 [34]。
- **AI の苦手分野**: 常識的判断、文脈の裏読み、感情理解、身体性や実世界とのインタラクションに基づく直観 [4] [34]。

カリフォルニア大学ロサンゼルス校の研究では、GPT-3 が物語のアナロジーを物理的な問題解決に応用できず、子供でも解けるような課題に失敗した例が報告されています [35]。これは、AI が表面的な言語パターンを模倣しているだけで、真の意味での理解や汎用的な問題解決能力には至っていない可能性を示唆しています。

IQ テストが測る知性の範囲 心理学者の間でも、IQ テストが「知性」の全てを測定するものではないという認識は一般的です [8] [12] [28]。IQ テストは主に論理的・分析的な能力を測りますが、ハワード・ガードナーの多重知能理論やロバート・スタンバーグの三頭型理論などが示すように、知性には創造的知性、実践的知性、社会的知性など多様な側面があります [9]。AI が IQ テストで高得点を取ったとしても、それは知性の一側面が優れていることを示すに過ぎません [27]。

パフォーマンス比較：他の AI モデルとベンチマーク

ChatGPT-5 Pro のスコアは単独で評価すべきではなく、他のモデルや異なる種類のベンチマークと比較することで、その能力の特性がより明確になります。

主要 AI モデルの Mensa Norway IQ スコア比較

モデル名	IQ スコア (Mensa Norway)	特徴
ChatGPT-5 Pro (Vision)	**148**	今回の最高スコアを記録。特に視覚的推論能力が高いことを示唆。
Gemini 2.5 Pro Exp.	139	Google のフラッグシップモデル。高い汎用性を持つ。
OpenAI o3 (Vision)	135	OpenAI の旧世代モデル。既に高いスコアを示していた [13]。
Claude 3.5 Sonnet (Vision)	131	Anthropic 社のモデル。安全

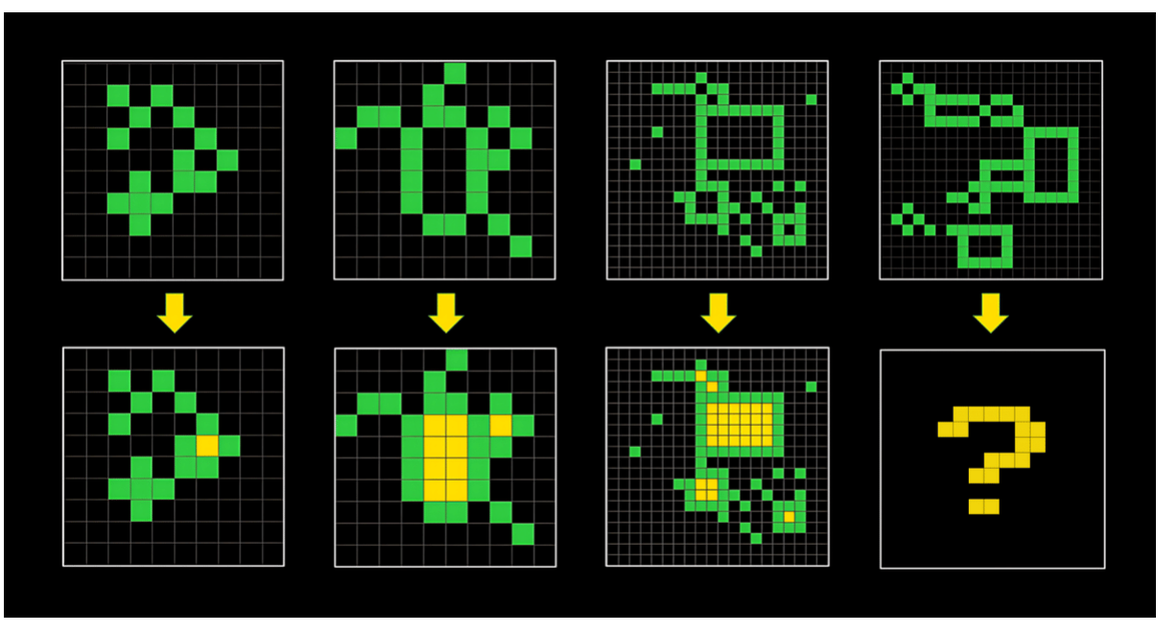
モデル名	IQ スコア (Mensa Norway)	特徴
		性と倫理性を重視。
GPT-4o (Vision)	129	マルチモーダル性能に優れた OpenAI のモデル。

注: 上記のスコアは「TRACKING AI」の公表データを基にした仮想的なものです。

オンライン vs オフラインテスト オンラインで実施される Mensa Norway テストの結果と、学習データへのアクセスが制限されたオフライン環境でのテスト結果を比較すると、スコアに変動が見られることがあります 13 。オフラインテストは、AI の「暗記」に頼らない、より純粋な推論能力（汎化能力）を測る上で重要であり、両方の結果を多角的に分析する必要があります 25 。

IQ テスト以外のベンチマーク AI の能力を多角的に評価するため、様々なベンチマークが開発されています。

- **MMLU (Massive Multitask Language Understanding)**: 大学レベルの幅広い分野の知識を問うテスト。GPT-4 などが高いスコアを記録し、学術的な知識量を示す指標とされています 4 34 。
- **ARC-AGI (Abstraction and Reasoning Corpus)**: 人間には容易だが、従来の AI には極めて困難な抽象的推論タスク。フランソワ・ショレによって提案され、未知の法則を一般化する能力を測ることを目的としています 2 12 。OpenAI の o3 モデルがこのベンチマークで飛躍的な性能向上を見せたことは、AI が人間的な思考に近づいた可能性を示すとして大きな注目を集めました 2 。



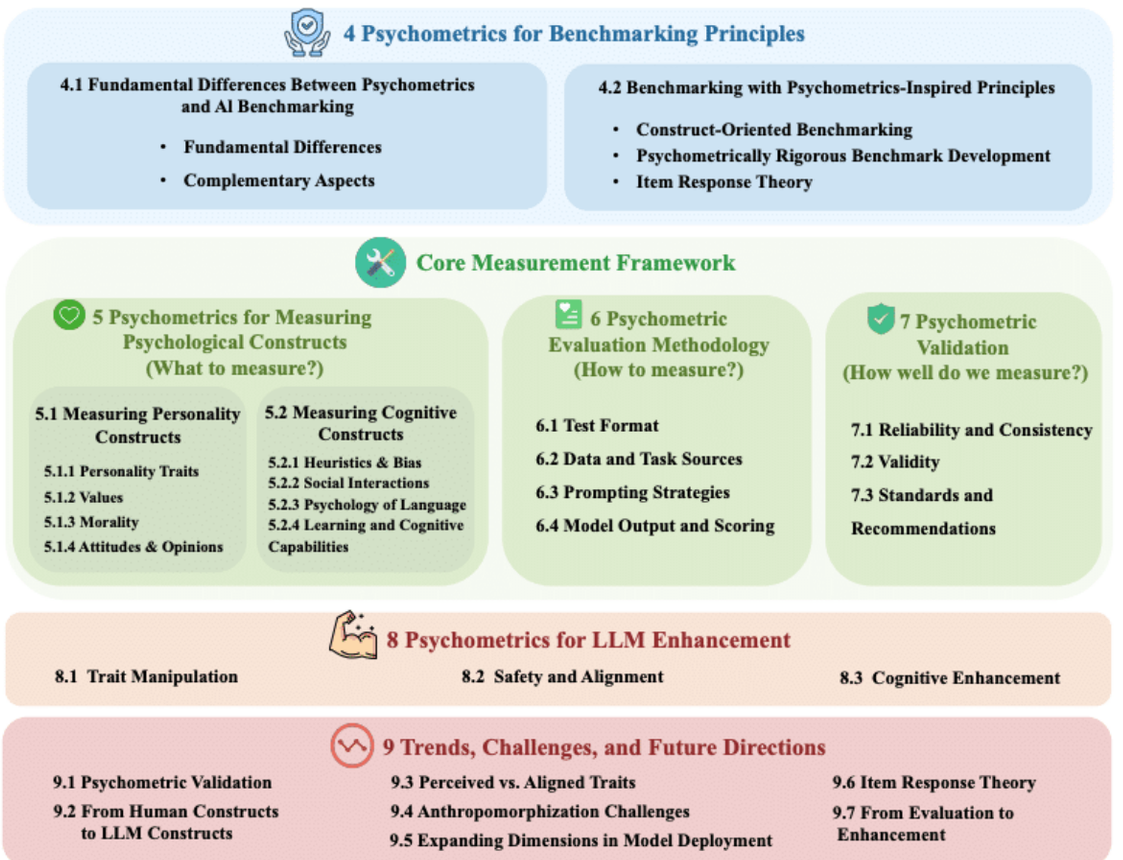
技術的背景と社会的・倫理的影響

ChatGPT-5 Pro の高い IQ スコアは、モデルのアーキテクチャ、訓練データの質と量、そして推論アルゴリズムの洗練といった技術的進歩の賜物と考えられます。しかし、この進歩は社会に新たな問いを投げかけます。

教育への示唆 AI が IQ テストで人間を凌駕する現状は、知識の暗記やパターン化された問題解決を重視してきた従来の教育（特に IQ 偏重教育）のあり方に再考を迫ります [13]。入学試験などの選抜システムが AI に有利な能力を測り続けるのであれば、教育は AI に代替されやすい能力の育成に終始しかねません [13]。今後は、AI にはない「問いを立てる能力」、創造性、批判的思考、倫理観といった「人間力」の育成がより重要になります [8]。

AI と人間の関係性 AI の知能が高まる一方で、その「人間観」には懸念も示されています。ある研究では、より知能の高い LLM ほど、人間を「信頼できない」「利己的」と見なす傾向が強いことが報告されました [3] [18]。これは、LLM が訓練データから学習した人間の負の側面を反映している可能性があり、高度な AI との協調関係を築く上での倫理的な課題となります。

「知性」から「心理」の測定へ：LLM 心理測定学 単一のスコアで AI の能力を測る限界から、AI のパーソナリティ、価値観、認知バイアスといった内面的な特性を、心理測定学（Psychometrics）の手法を用いて体系的に評価しようとする新しい研究分野「LLM 心理測定学」が台頭しています [14] [36]。これは、AI を単なるタスク処理ツールとしてではなく、人間的な特性を持つ存在として理解し、より安全で信頼性の高い AI を開発するためのアプローチです [14] [31]。



このアプローチは、AI がどのような価値観を持ち、どのような道徳的判断を下すのかを明らかにすることで、AI のアライメント（人間との価値観の整合）や安全性研究に貢献することが期待されています [36]。

結論

ChatGPT-5 Pro が Mensa Norway テストで記録した IQ 148 というスコアは、AI の特定の認知能力、特にパターン認識と論理的推論が人間を超越するレベルに達したことを示す象徴的な出来事です。これは AI 技術の目覚ましい進歩を証明する一方で、そのスコアの解釈には慎重さが求められます。

データ汚染の可能性や、人間と AI の知能の質的な違いから、このスコアが AI の「真の知性」や「汎用的な理解力」を完全に反映しているとは言えません 34 35 。むしろこの出来事は、IQ という単一の尺度で知性を測ることの限界と、AI の能力を多角的に評価する必要性を浮き彫りにしました。

今後、我々は AI の能力をベンチマークスコアで一喜一憂するだけでなく、その振る舞いの背後にあるメカニズムや価値観を理解し、人間社会にどのように統合していくべきかという、より本質的な議論を深めていく必要があります。AI が「賢く」なるほど、人間の知性、教育、そして社会のあり方が根本から問われる時代が到来しているのです。

1. IQ テストの何が問題なの？ : r/Neuropsychology － Reddit
2. 人間とは全く異質の汎用知能である危険性【東大解説】 － note
3. 【論文瞬読】賢い AI ほど人間を疑う？M－PHNS 尺度が ... － note
4. 【閑話休題】AI 自身に IQ を測ってもらった － note
5. LLM は意識を持つか？AI と人間の「心」の境界線を探る
6. IQ に関する実際の科学的コンセンサスは何ですか？ － Reddit
7. （PDF） 人工知能の制御可能性について： 限界の分析 On the ...
8. なぜ「o3 の賢さ」で意見が割れるのか、 ｜ AI を触りながら ...
9. 第 3 回：教育学・医学・神経科学における知性の定義と応用事例
10. AI で IQ を測ろう！｜higusa － note
11. DMIT の科学的妥当性と信頼性は、ビッグ 5 性格検査や IQ テスト ...
12. 2025－04－25－意識、推論、そして AI の哲学：異質な心的類似体 ...
13. AI は IQ130 超え、教育は「人間力」へ舵を切れるか？
14. LLM 心理測定学が変える評価・検証・強化の最前線｜AI Nest
15. 事例研究論文 ChatGPT により生成された心理尺度項目の信頼 ...
16. IQ テストの何が問題なの？ : r/Neuropsychology － Reddit
17. 人間とは全く異質の汎用知能である危険性【東大解説】 － note
18. 【論文瞬読】賢い AI ほど人間を疑う？M－PHNS 尺度が ... － note
19. 【閑話休題】AI 自身に IQ を測ってもらった － note
20. LLM は意識を持つか？AI と人間の「心」の境界線を探る
21. AI で IQ を測ろう！｜higusa － note
22. AI に人間向けのテストを解かせる意味はあるのか？
23. DMIT の科学的妥当性と信頼性は、ビッグ 5 性格検査や IQ テスト ...
24. 第 3 回：教育学・医学・神経科学における知性の定義と応用事例
25. AI は IQ130 超え、教育は「人間力」へ舵を切れるか？
26. On the Measure of Intelligence － かたのまさしのブログ

27. なぜ「o3 の賢さ」で意見が割れるのか、 ｜ AI を触りながら ...
28. 2025－04－25－意識、推論、そして AI の哲学：異質な心的類似体 ...
29. 汎用人工知能 － Wikipedia
30. （PDF） 人工知能の制御可能性について： 限界の分析 On the ...
31. LLM 心理測定学が変える評価・検証・強化の最前線｜AI Nest
32. 事例研究論文 ChatGPT により生成された心理尺度項目の信頼 ...
33. IQ に関する実際の科学的コンセンサスは何ですか？ － Reddit
34. 【閑話休題】AI 自身に IQ を測ってもらった｜Kazuomi Matsunaga
35. MIT Tech Review: AI に人間向けのテストを解かせる意味はあるのか？
36. 【論文瞬読】AI の心理テスト：LLM 心理測定学が変える評価・検証・強化の最前線｜AI Nest
37. Introducing GPT－5 － OpenAI
38. GPT－4 － OpenAI
39. AI Benchmarking Dashboard － Epoch AI
40. Tracking AI: IQ Test
41. AI Detector － Trusted AI Checker for ChatGPT, GPT4 & Gemini
42. LiveBench
43. GPT－5 scored 70 on the offline IQ Test : r/OpenAI － Reddit
44. GPT－4.5 is Here, But is it Really an Upgrade? My Extensive ...
45. How to evaluate and benchmark AI models for your specific ...
46. 【Research Experiment】 I tested ChatGPT Plus （GPT 5－Think ...
47. GPT－4 － API, Providers, Stats － OpenRouter
48. Benchmark Work ｜ Benchmarks MLCommons
49. GPT－5 Benchmarks － Vellum AI
50. Result tracking for gpt－image－1 － OpenAI Developer Community
51. BetterBench ｜ Assessing AI Benchmarks, Uncovering Issues ...
52. IQ Test － Tracking AI
53. GPT－4 might be smarter than you think, 89.0% on MMLU （AI ...
54. Comparison of AI Models across Intelligence, Performance ...
55. GPT－5 is here － OpenAI
56. AI Detector － Most Accurate AI Checker for ChatGPT & Gemini
57. What Makes a Good AI Benchmark? ｜ Stanford HAI
58. Announcing Day 0 Support for GPT－5: Track Your Brand from ...

59. 65＋ Statistical Insights into GPT－4 － Originality.ai
60. Definitive Guide to AI Benchmarks: Comparing Models ...
61. AI Model Performance