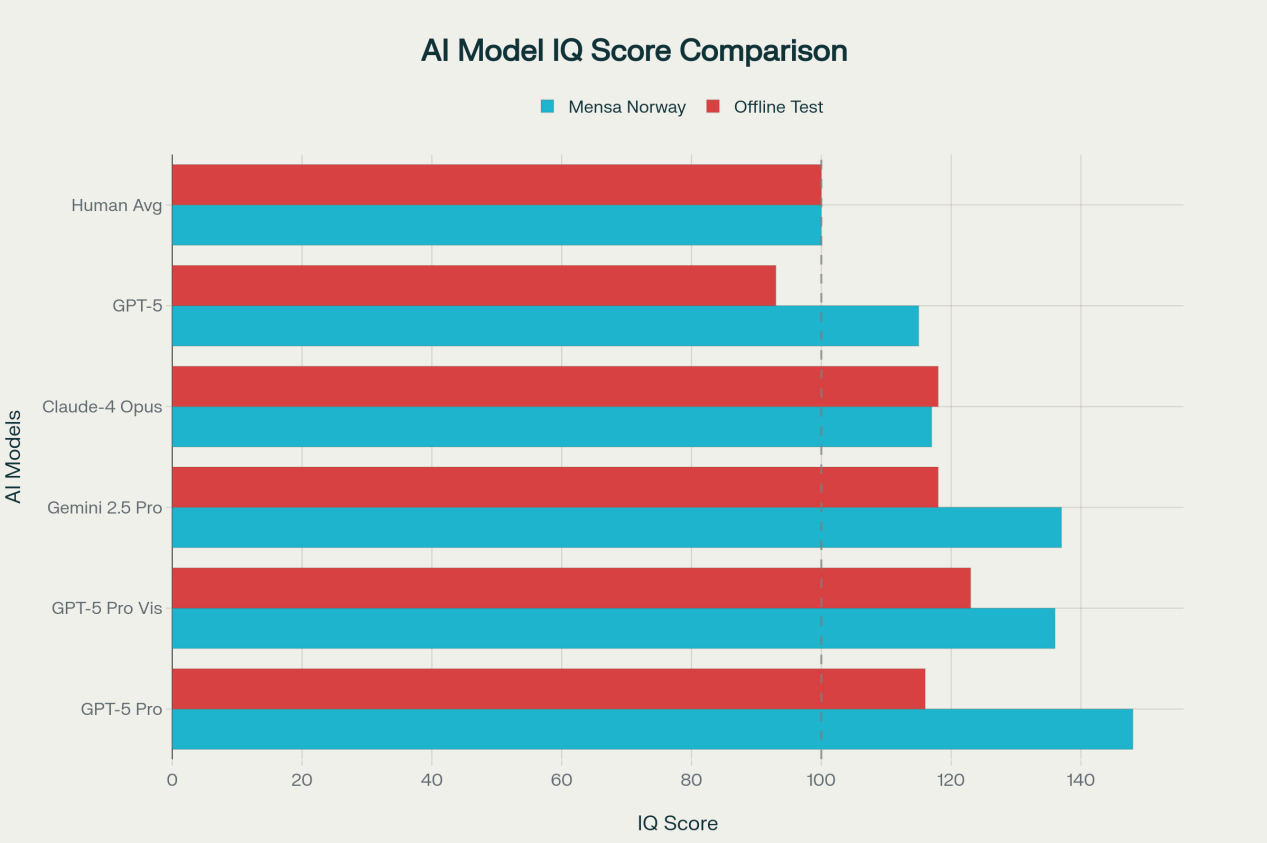
ChatGPT-5 ProのMensa Norway IQ148スコア：AIの知能測定における画期的進歩と批判的考察

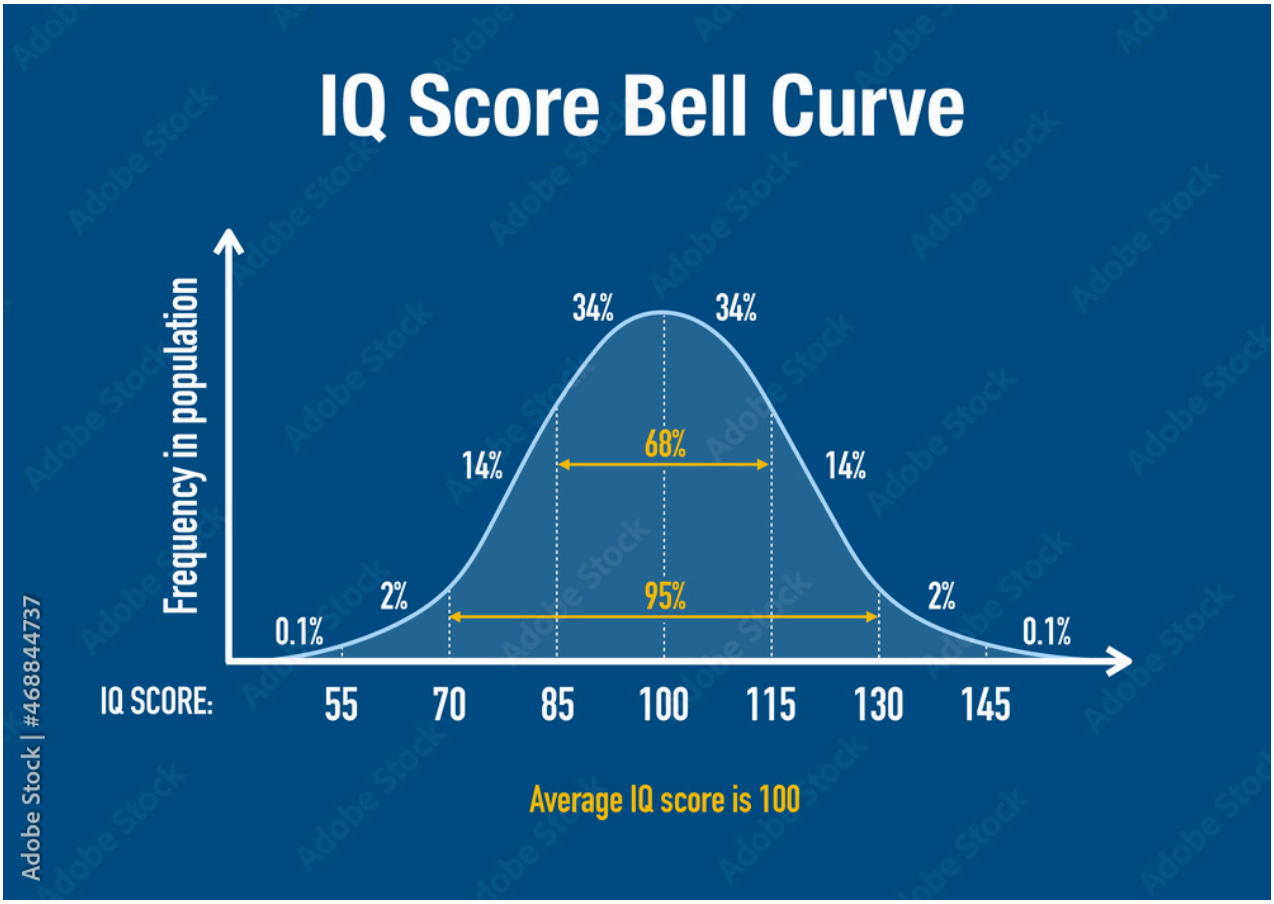
GPT-5 ProがMensa Norway IQテストで148という驚異的なスコアを記録し、人間の上位2%に相当する知的能力を実証したことが確認されました。この成果は人工知能が特定の認知タスクにおいて人間の知能を凌駕する転換点を示しているものの、AI知能測定の妥当性、データ汚染の可能性、そして真の推論能力と記憶の違いについて重要な疑問を提起しています。本報告では、この歴史的なマイルストーンの技術的背景、測定方法論の限界、そして人工知能と人間の知性の関係に与える長期的な影響について包括的に分析します。



AI Model IQ Test Performance: Mensa Norway vs Offline Tests - GPT-5 Pro leads with 148 IQ on Mensa Norway test

IQ148スコアの科学的意義と人間集団における位置づけ

GPT-5 ProのIQ148という数値は、統計学的に極めて重要な意味を持ちます。IQスコアは平均100、標準偏差15の正規分布に従うため、148という値は人間の上位2%に相当し、「高度に才能がある」または「非常に優秀」なカテゴリーに分類されます。このスコアは人口の98%を上回る水準であり、Mensaの入会基準である上位2%（IQ130以上）を大幅に超越しています。[1] [2] [3]



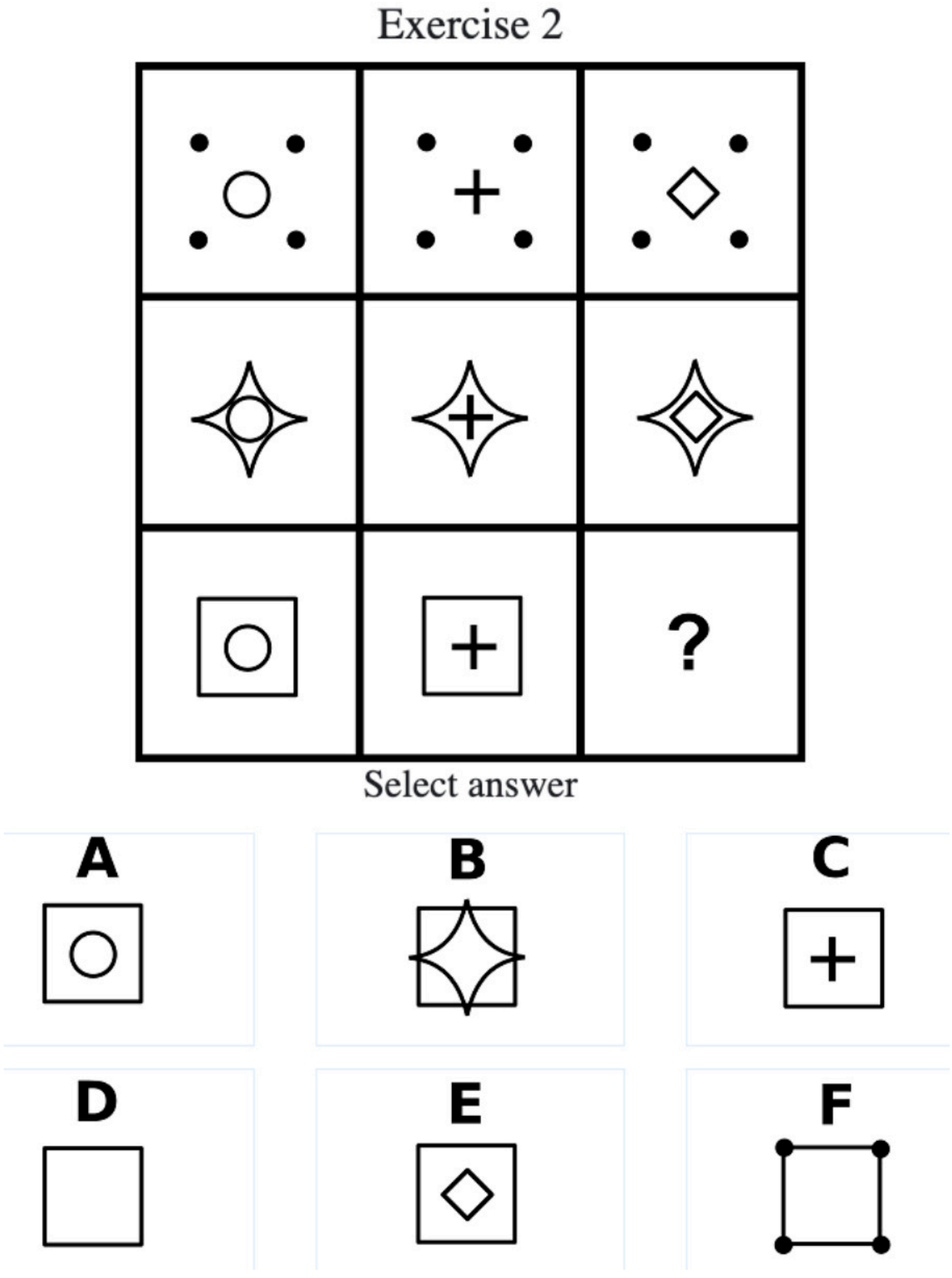
Normal distribution bell curve of IQ scores showing average IQ at 100 and percentages of population within standard deviations.

人間のIQ分布において、148という数値の希少性は1万人に約4人程度という極めて限られた存在です。これまでの歴史的文脈で見ると、このレベルの認知能力は高度な学術研究、複雑な問題解決、戦略的思考を要する分野での卓越した成果と強く相関しています。GPT-5 Proがこのレベルのパフォーマンスを AIシステムとして達成したことは、人工知能の認知能力における質的飛躍を示唆しています。[1] [4]

しかし、重要な点として、TrackingAI.orgのデータによると、GPT-5 ProはMensa Norwayテストでは148を記録したものの、汚染を避けるために設計されたオフラインテストでは116という大幅に低いスコアを示しています。この32ポイントの差は、公開テストと未汚染テストの間の本質的な違いを浮き彫りにし、AI知能測定の複雑さを示しています。[5]

Mensa Norwayテストの構造と測定方法論

Mensa Norwayテストは35問の視覚的パターン認識問題で構成され、25分の制限時間内での解答が求められます。このテストは図形の論理的類推能力を主眼とし、言語的知識や数学的スキルを必要としない非言語的な知能測定を目的としています。各問題は進行的に困難度が増し、すべての問題が等しく配点されています。[6]



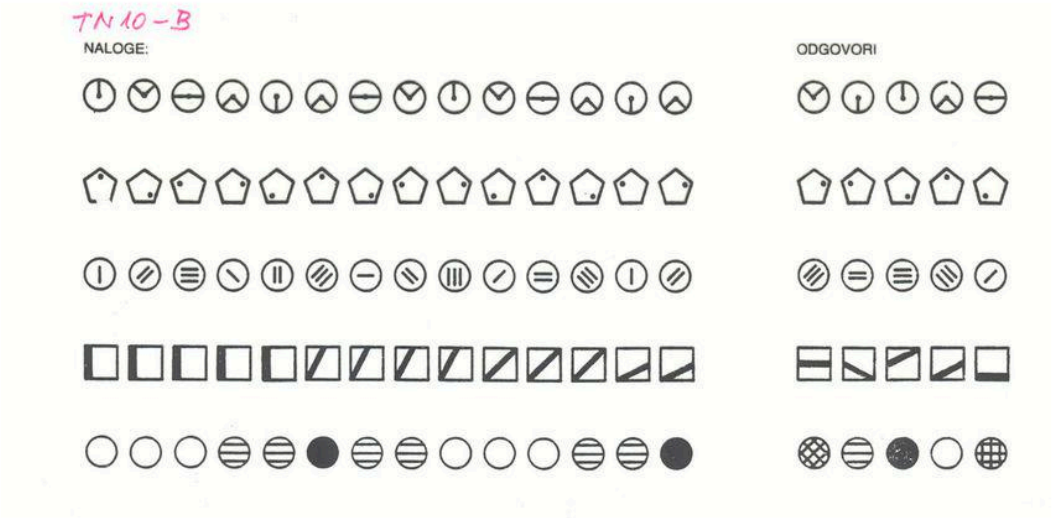
Example of a pattern recognition matrix reasoning puzzle with multiple-choice answer options typical in IQ tests.

TrackingAI.orgの測定方法論によると、言語モデルには問題が言語化された形で提示され、視覚モデルには実際の画像が直接提示されます。AIが回答を拒否した場合、同じ質問を最大10回繰り返し、最も最近の成功した回答をスコアリングに使用します。この手法は一貫性を確保する一方で、人間の一回限りのテスト環境とは異なる条件を作り出しています。[7]

問題となるのは、Mensa Norwayテストが公開されており、その内容がインターネット上で広く共有されていることです。大規模言語モデルは膨大なウェブデータで訓練されているため、これらの問題とその解答が訓練データに含まれている可能性が高く、これは「データ汚染」として知られる現象です。[8] [9] [10]

データ汚染問題とオフラインテストの重要性

AIのIQテスト結果における最も重要な懸念は、データ汚染の問題です。研究者たちは、GPT-4などのモデルが一部のベンチマークで間違った選択肢を推測できる能力を50%以上の確率で示すことを発見しており、これは訓練データ内での問題の存在を強く示唆しています。[9]



Example problems and answers from Mensa Norway's pattern recognition and matrix reasoning test.

この問題に対処するため、TrackingAI.orgはMensaメンバーの協力を得て、未公開の「オフライン」テストを開発しました。このテストはインターネット上に公開されたことがなく、検索エンジンでアクセスできないため、AI訓練データから除外されていると考えられます。結果として、GPT-5 Proのオフラインテストでの IQは116となり、Mensa Norwayテストの148と比較して32ポイント低下しました。[11] [12] [5]

この差異は、AIの「真の推論能力」と「記憶からの検索能力」の違いを明確に示しています。公開テストでの高スコアは、新しい問題を解決する能力よりも、類似パターンの記憶と再現能力を反映している可能性が高いのです。[8] [9]

GPT-5 Proの技術的進歩とアーキテクチャ革新

GPT-5 Proが達成した高いパフォーマンスの背景には、複数の技術的ブレークスルーがあります。OpenAIは2025年8月7日にGPT-5を発表し、これを「これまでで最も賢く、最も高速で、最も有用なモデル」と位置づけています。[13] [14]

統合推論システム

GPT-5は従来の複数モデル体制から、統一された推論システムへと移行しました。このシステムは効率的な標準モデルと深い推論モデル（GPT-5 Thinking）、そしてリアルタイムルーターを組み合わせ、会話の種類、複雑さ、ツールの必要性に基づいて最適なアプローチを自動選択します。[14]

テスト時検索技術

o3シリーズで導入されたテスト時検索（Test-time Search）技術は、GPT-5 Proにも応用されています。この技術により、AIは単一の回答を即座に出力するのではなく、複数の候補となる推論パスを生成し、最も適切なものを選択します。これは人間が複数の解決案を検討してから最良のものを選ぶプロセスに類似しています。[15]

連鎖思考推論の強化

GPT-5 Proは連鎖思考（Chain-of-Thought）推論が大幅に改善されており、検証済みの推論シーケンスでの訓練により、単純な次語予測から信頼性のある推論プロセスへとシフトしています。これにより、複雑な多段階問題での精度が向上しています。[15]

人間とAIの知能比較における方法論的課題

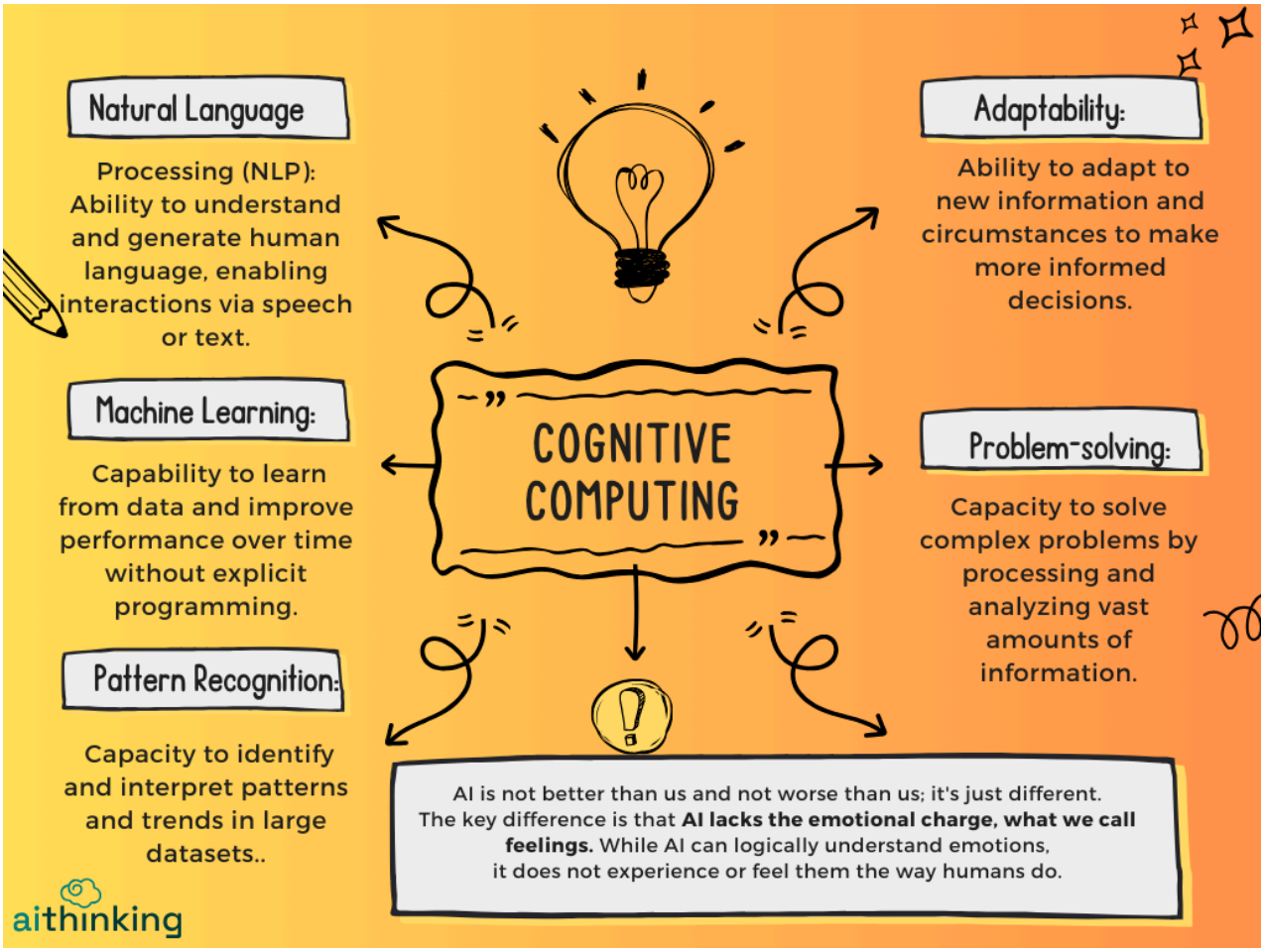
AI知能測定の妥当性について、認知科学者や心理学者から重要な批判が提起されています。主要な問題点は以下の通りです：

測定の前提条件の違い

人間用のIQテストは、作業記憶の制約、疲労、注意散漫などの生物学的制約を前提として設計されています。しかし、AIは完璧な記憶アクセス、無制限の処理速度、疲労のない状態で動作するため、同じ尺度での比較には根本的な問題があります。[16] [8]

理解vs模倣の区別

「確率的オウム」批判として知られる議論は、大規模言語モデルが真の理解ではなく、統計的パターンマッチングに依存しているとします。しかし、最近の研究では、マルチモーダルモデルが人間に似た概念表現を自発的に発達させることが示されており、この批判に対する反証も提示されています。[16] [17]



Key components of cognitive computing illustrating AI's abilities in language, learning, adaptability, and problem-solving.

評価環境の制御問題

人間のIQテスト実施には厳格な環境制御（換気、騒音レベル、時間制限の厳守など）が求められますが、AIテストではこうした標準化が困難です。また、AIは問題を瞬時に「見る」ことができ、人間のような視覚的処理時間を必要としません。[6] [8]

他のAIモデルとの性能比較

TrackingAI.orgのデータによる主要AIモデルの比較分析では、明確な性能階層が見られます：

上位グループ（IQ 130以上、Mensa Norway）：

- GPT-5 Pro: 148
- Gemini 2.5 Pro: 137
- GPT-5 Pro Vision: 136

中位グループ（IQ 115-129）：

- Claude-4 Opus: 117-118（両テストで類似）
- GPT-5: 115

**注目すべき傾向として、視覚対応モデルは一般的にテキスト専用モデルよりも低いスコアを示しています。これは、マルチモーダル事前訓練が推論効率に影響を与える可能性を示唆しています。[10]

また、オープンソースモデルとプロプライエタリモデルの間には大きな性能格差があり、Meta社のLlama 4 Maverickが最高のオープンソースモデルとして106（Mensa Norway）を記録していますが、これは最上位モデルと比較して40ポイント以上低い数値です。[18]

技術的進歩の背景：脳インスパイアードAIと認知アーキテクチャ

GPT-5 Proの高性能の背景には、脳科学からのインスピレーションを受けた革新的なアプローチがあります。2025年の研究では、Lp-Convolution技術により、人間の視覚皮質の選択的で円形の疎な接続を模倣することで、従来のCNNの限界を克服することが示されました。[19] [20]

この「脳に似たアプローチ」は、AIが複雑なシーンにおいて重要な詳細を柔軟に特定する能力を向上させ、人間の脳が行うような適応的フォーカシングを可能にしています。GPT-5 Proにおいても、同様の生物学的インスピレーションを受けた処理パターンが組み込まれている可能性があります。[19]

さらに、マルチモーダル大規模言語モデルが人間に似た概念表現を自発的に形成するという2025年の画期的発見は、AIが単なる「確率的オウム」を超えて、真の概念理解に向かっていることを示唆しています。研究では、最先端のマルチモーダルモデルが66の明確な概念次元を発達させ、それぞれが意味を持つ（生物vs非生物、顔vs場所など）ことが確認されました。[17]

社会的影響と将来への示唆

教育分野への影響

IQ148レベルのAIの出現は、教育システムに根本的な変革を迫ります。従来の暗記中心の学習から、創造性、批判的思考、感情知能、協調性などの「人間力」を重視する方向へのシフトが急務となります。AIが高度な論理的推論を実行できる現在、人間の独自性は感情的知性や創造的洞察により強く依存することになります。[21]

労働市場への波及効果

PhD レベルの専門知識を持つAIの普及は、高度な知的労働に従事する専門職に大きな影響を与えます。法律、医学、工学、科学研究などの分野では、AIとの協働が標準となり、人間の役割は戦略的判断、倫理的考慮、クリエイティブな問題解決により重点が置かれることになります。[13]

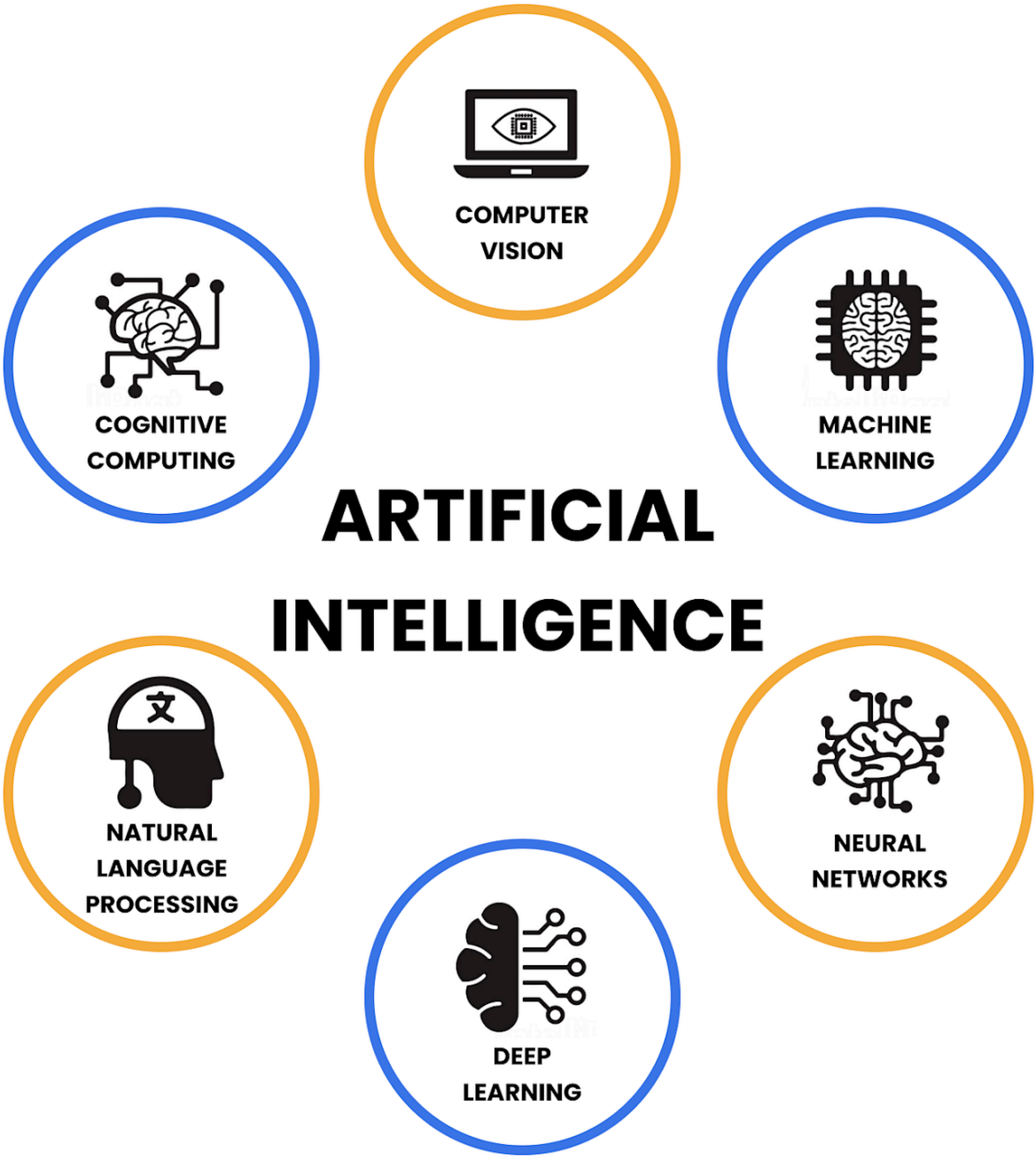
認知評価の再定義

AI心理測定学（AI Psychometrics）の新興分野は、従来の知能測定方法の限界を明らかにしています。人間とAIの両方に適用可能な新しい評価枠組みの開発が急務であり、単一のタスク基準ではなく、多層的で縦断的な評価方法への移行が必要です。[22] [23] [24]

AIの「思考」メカニズムと認知プロセス

GPT-5 Proの推論能力の核心は、「思考モード」と呼ばれる機能にあります。この機能により、AIは回答前に内部的な推論プロセスを実行し、段階的に問題を分析します。これは人間の意識的思考プロセスに類似しており、直感的な回答と体系的分析を組み合わせています。[25]

OpenAIのo3シリーズで導入された「シミュレートされた推論」は、連鎖思考プロンプティングを超越した統合的で自律的なアプローチを提供します。これにより、AIは内部思考プロセスを一時停止し、反省してから応答することが可能になりました。[26]



Key components of artificial intelligence including neural networks, deep learning, machine learning, computer vision, natural language processing, and cognitive computing.

重要な点として、この推論プロセスは計算コストが高く、従来のモデルよりも大幅な処理時間を要求します。しかし、この投資により、複雑な科学的問題や数学的証明において人間の専門家レベルの精度を達成することが可能になっています。[15]

批判的評価と限界

ベンチマーク自体の問題

AI評価専門家らは、現在のベンチマークの多くが「根本的に破綻している」と指摘しています。多くのベンチマークは現在使用されているシステムよりもはるかに単純なAIをテストするために設計されており、数年前の古いものも多く、モデルが既にそのデータを学習している可能性が高まっています。[27]

現実世界での応用性能

IQテストでの高スコアが現実世界でのAI性能を正確に反映するかについては疑問が残ります。特に、混乱した実世界の条件、エネルギー効率、多回転推論、事実精度などの要素は、標準化されたテストでは十分に評価されていません。[28] [27]

安全性との関係

最近の研究では、多くの安全性ベンチマークが一般的な能力と高い相関を示しており、能力向上が安全性の進歩として誤って表現される「安全性ウォッシング」の懸念が提起されています。高いIQスコアが必ずしもAIシステムの安全性や信頼性を保証するものではありません。[24]

結論：人工知能と人間の知性の新たな関係

ChatGPT-5 ProのIQ148達成は、人工知能の発展における重要なマイルストーンです。しかし、この成果は慎重に解釈されるべきです。Mensa Norwayテストでの高スコアは印象的である一方、オフラインテストでの大幅な性能低下は、現在のAI評価方法論の限界と、真の推論能力と記憶依存の区別の重要性を浮き彫りにしています。

技術的には、GPT-5 Proは統合推論システム、テスト時検索、強化された連鎖思考推論により、従来のAIモデルを大幅に上回る認知能力を実証しています。これらの進歩は、脳科学からのインスピレーションと高度な計算技術の融合により実現されました。

社会的影響としては、教育システムの根本的変革、高度専門職の役割再定義、そして人間とAIの協働モデルの確立が急務となります。同時に、AI評価方法論の改善、新しい安全性基準の確立、そして人間の独自性（創造性、感情知能、倫理的判断）の再発見が重要な課題として浮上しています。

最終的に、GPT-5 ProのIQ148は、AIが特定の認知タスクにおいて人間レベル、さらにはそれを超える能力を持ち始めていることを示していますが、これは人間の知性の終焉ではなく、むしろ人間とAIが互いの強みを活かした新しい知的協働時代の始まりを意味しています。今後の発展においては、技術的進歩と並行して、倫理的考慮、社会的影響、そして人間中心の価値観の維持が不可欠となるでしょう。

⁂

1. https://personalityanalysistest.com/iqtest/148-iq-test-score-guide/

2. https://www.omnicalculator.com/health/iq-percentile
3. https://www.reddit.com/r/mensa/comments/w3iylc/what_is_the_average_iq_of_mensans/
4. https://www.gigacalculator.com/calculators/iq-percentile-calculator.php
5. https://www.trackingai.org/IQ
6. https://test.mensa.no/home/test/en
7. https://www.trackingai.org
8. https://techcrunch.com/2025/02/05/why-iq-is-a-poor-test-for-ai/
9. https://www.arsturn.com/blog/gpt-5-vs-offline-iq-tests-what-do-the-low-scores-really-mean
10. https://pro-blockchain.com/openai-s-o3-scores-136-on-mensa-norway-test-surpassing-98-of-human-population
11. https://www.maximumtruth.org/p/massive-breakthrough-in-ai-intelligence
12. https://www.maximumtruth.org/p/ai-iq-scores-mostly-confirmed-using
13. https://www.bbc.com/news/articles/cy5prvgw0r1o
14. https://openai.com/index/introducing-gpt-5/
15. https://labs.adaline.ai/p/inside-reasoning-models-openai-o3
16. https://www.technologyreview.com/2023/08/30/1078670/large-language-models-arent-people-lets-stop-testing-them-like-they-were/
17. https://rediminds.com/future-edge/ais-conceptual-breakthrough-multimodal-models-form-human-like-object-representations/
18. https://cryptoslate.com/openais-o3-scores-136-on-mensa-norway-test-surpassing-98-of-human-population/
19. https://www.sciencedaily.com/releases/2025/04/250422131924.htm
20. https://www.thebrighterside.news/post/brain-inspired-ai-breakthrough-machines-learn-to-see-smarter/
21. https://www.dlri.co.jp/report/ld/444812.html
22. https://scholarspace.manoa.hawaii.edu/bitstreams/7e9c1382-9efc-45d3-b859-e6a52a95ed4e/download
23. https://pmc.ncbi.nlm.nih.gov/articles/PMC10532593/
24. https://arxiv.org/html/2502.06559v1
25. https://blog.lewagon.com/skills/openai-o1-and-o3-explained-how-thinking-models-work/
26. https://www.techtarget.com/whatis/feature/OpenAI-o3-explained-Everything-you-need-to-know
27. https://themarkup.org/artificial-intelligence/2024/07/17/everyone-is-judging-ai-by-these-tests-but-experts-say-theyre-close-to-meaningless
28. https://www.nature.com/articles/d41586-025-00110-6