perplexity

推論型生成AIの創造性と推論力：最新モデルの学術的・実用的評価

本レポートは、OpenAI o1・o3シリーズやGemini 2.5 Proなどの最新推論型生成AIの創造性と推論力について、学術研究と実用性能の両面から包括的に分析する[1] [2] [3]。これらの新世代AIモデルは、従来の大規模言語モデルを大きく上回る推論能力を示し、数学、科学、コーディングといった高度な認知タスクで人間専門家レベルの成果を達成している[4] [5] [6]。一方で、創造性評価においては、AIが特定の測定基準で人間を上回る結果を示しつつも、真の創造的洞察とパターン再組み合わせの境界について重要な議論が続いている[7] [8] [9]。

推論型生成AIの技術的革新

Chain of Thought推論の進化

推論型AIモデルの最大の技術的革新は、段階的思考プロセス「Chain of Thought」の実装にある[1] [2] [10]。OpenAI o1シリーズは強化学習を用いて、問題を細分化し各ステップで論理的推論を行う能力を獲得している[2] [11]。この手法により、国際数学オリンピック予備試験（AIME）でo1が83%、最新のo3が96.7%という驚異的な正答率を達成した[2] [5]。

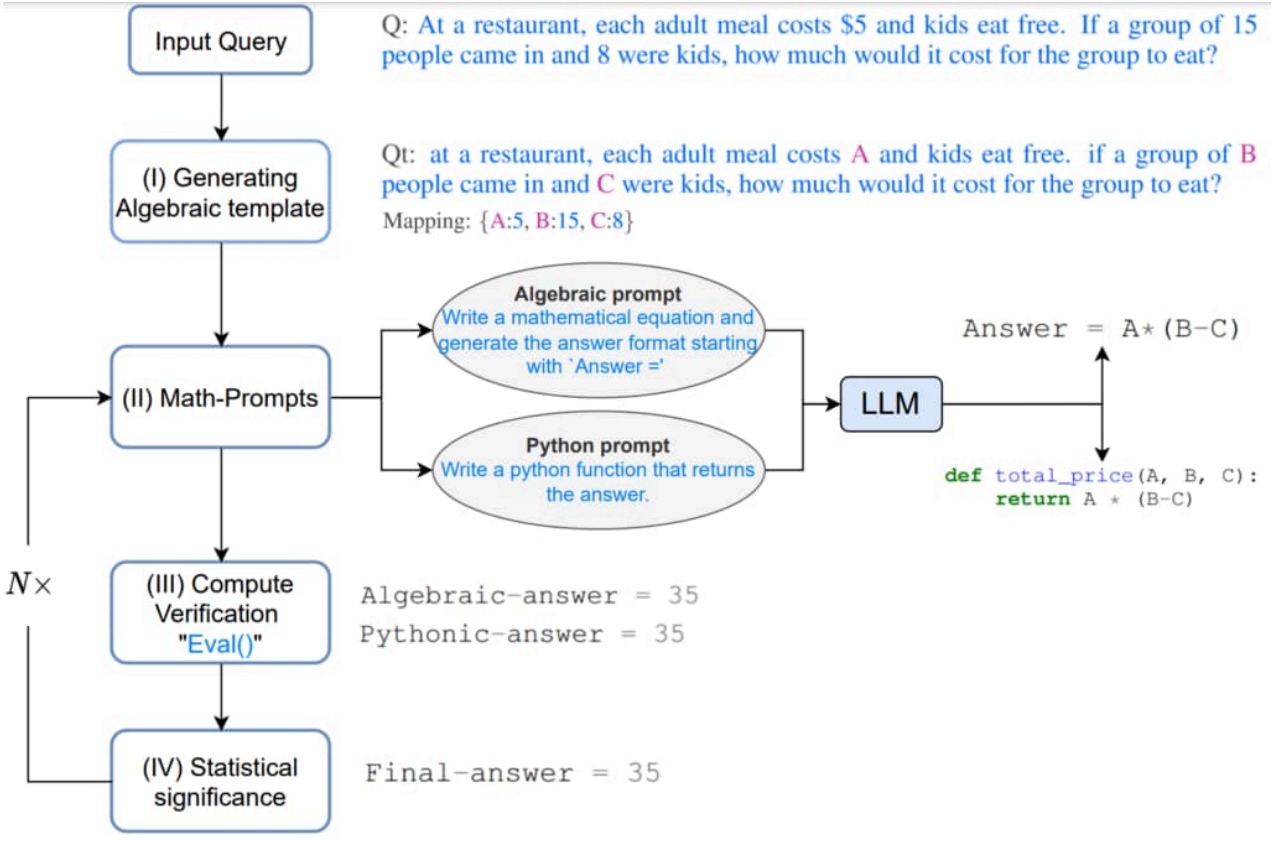


Figure 1: **MathPrompter flow.** We outline the MathPrompter process with an example alongside.

Mathprompter Flowchart Illustrating Problem-Solving Process with Large Language Models

マルチモーダル統合推論

Gemini 2.5 Proは、テキスト、画像、音声を統合した推論能力で差別化を図っている[12] [3] [6]。このマルチモーダルアプローチにより、視覚的データと言語的データを同時に処理し、より包括的な推論が可能となっている[12] [13]。特に長文コンテキスト理解において、128,000トークンの処理で91.5%という卓越した性能を示している[3]。

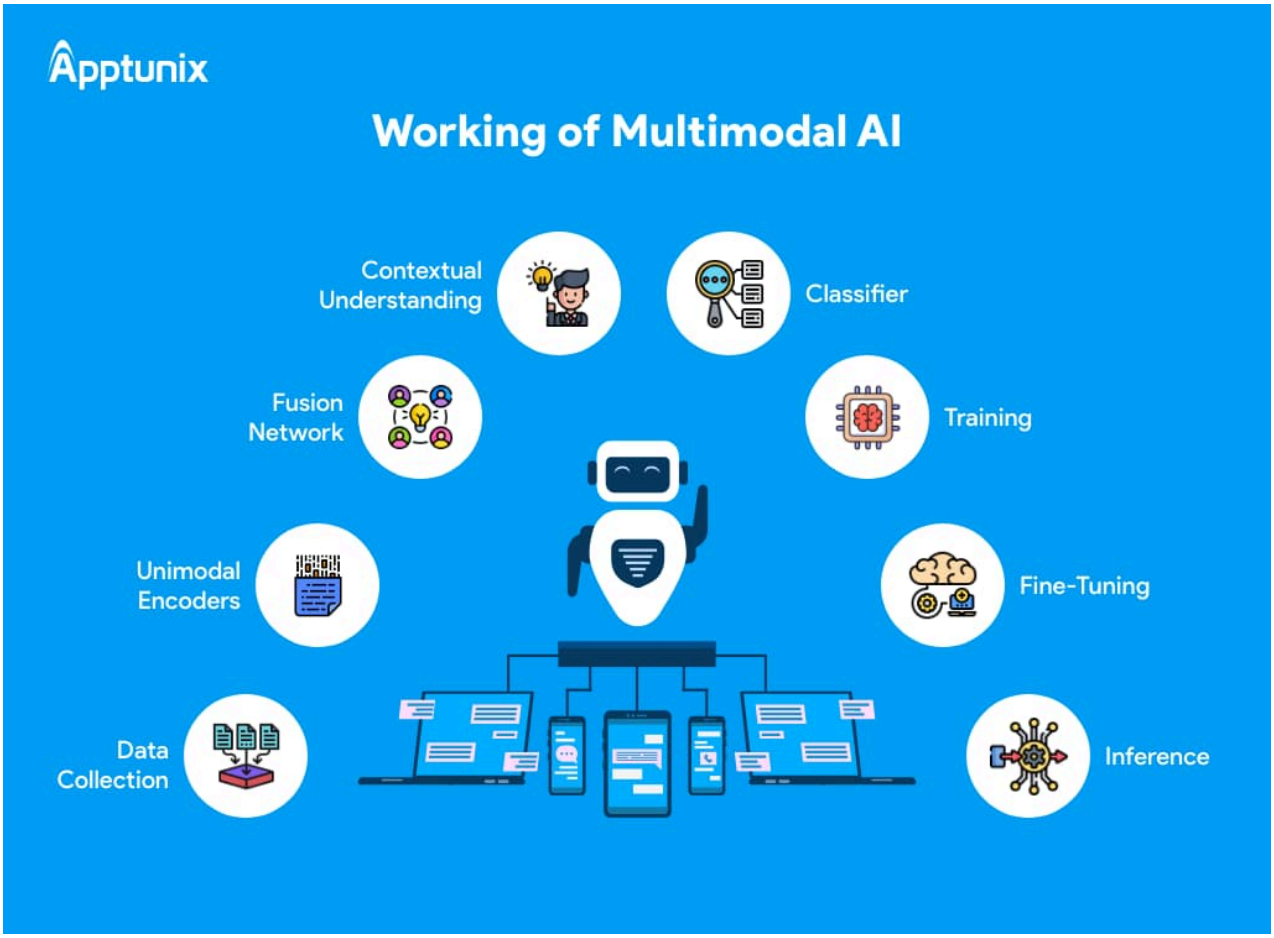


Diagram illustrating the workflow of multimodal AI.

学術的評価：AI創造性研究の最前線

創造性測定手法の多様化

AIの創造性評価は、心理学的測定手法の適用から始まり、現在では複数の専門化されたベンチマークが開発されている[14] [15] [16]。Torrance Tests of Creative Thinking（TTCT）では、GPT-4が流暢性、柔軟性、独創性、精緻性の4つの基準で人間を上回る結果を示した[16] [8]。Alternative Use Task（AUT）においても、AIは日常物品の用途について人間より多様で独創的なアイデアを生成している[8]。

数学的・論理的創造性の限界

しかし、高度な創造性が要求される分野では課題が明らかになっている[7]。DeepMath-Creativeベンチマークでは、最高性能のO3 Miniでも70%程度の精度に留まり、これは学習したパターンの再組み合わせレベルに過ぎないという分析がなされている[7]。大喜利を用いたLoTbenchでは、現在のAIの創造性は限定的で、人間との差は予想より小さいことが示された[14]。

人間-AI協働における創造性の変化

興味深い研究結果として、AI支援が個人の創造性に与える影響が報告されている[9]。ストーリー創作実験では、AIアイデアへのアクセスにより作品の創造性が向上する一方で、作品間の類似性も増加し、全体的な多様性が減少することが確認された[9]。これは、AIが創造性を増強する一方で、同質化のリスクも伴うことを示唆している。

学術的評価：推論力の系統的分析

因果推論能力の評価

AIの推論力評価において、因果関係の理解は特に重要な指標となっている[17]。コイン投げゲーム、ファイルダウンロード、シンプソンのパラドックスという3つのシナリオでの評価では、GPT-4が他のモデルを大きく上回る因果理解能力を示した[17]。しかし、介入を伴う複雑なシナリオでは、すべてのモデルで性能低下が観察されている[17]。

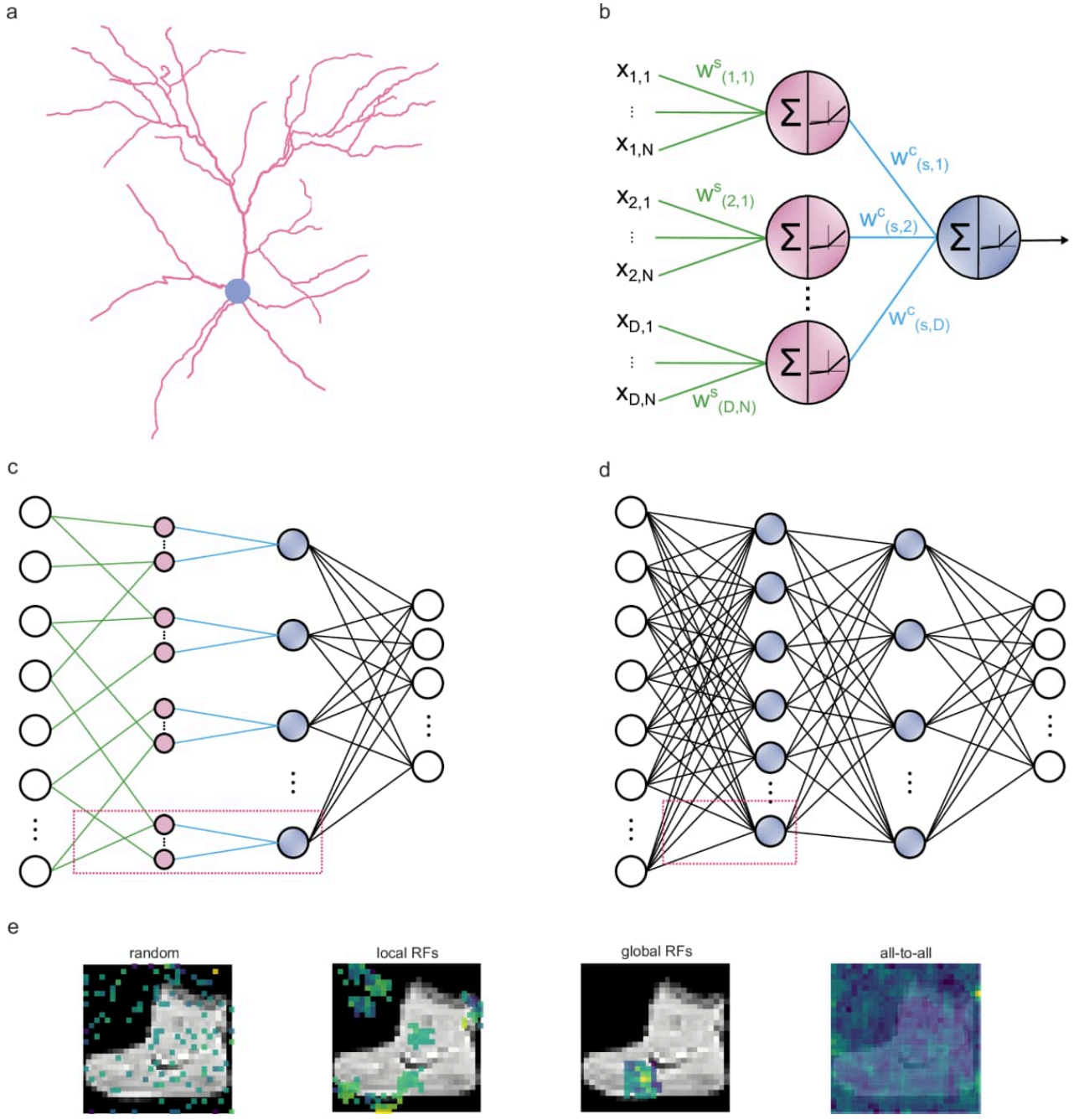


Diagram of artificial neural networks with varying connectivity patterns and receptive fields.

推論パターンの系統的分析

OpenAI o1の推論パターンについて、数学、コーディング、常識推論の3領域での比較研究が行われている[11]。研究では、o1が6つの特徴的な推論パターンを示すことが確認され、特に段階的問題分解と自己検証機能が高い精度に寄与していることが明らかになった[11] [18]。

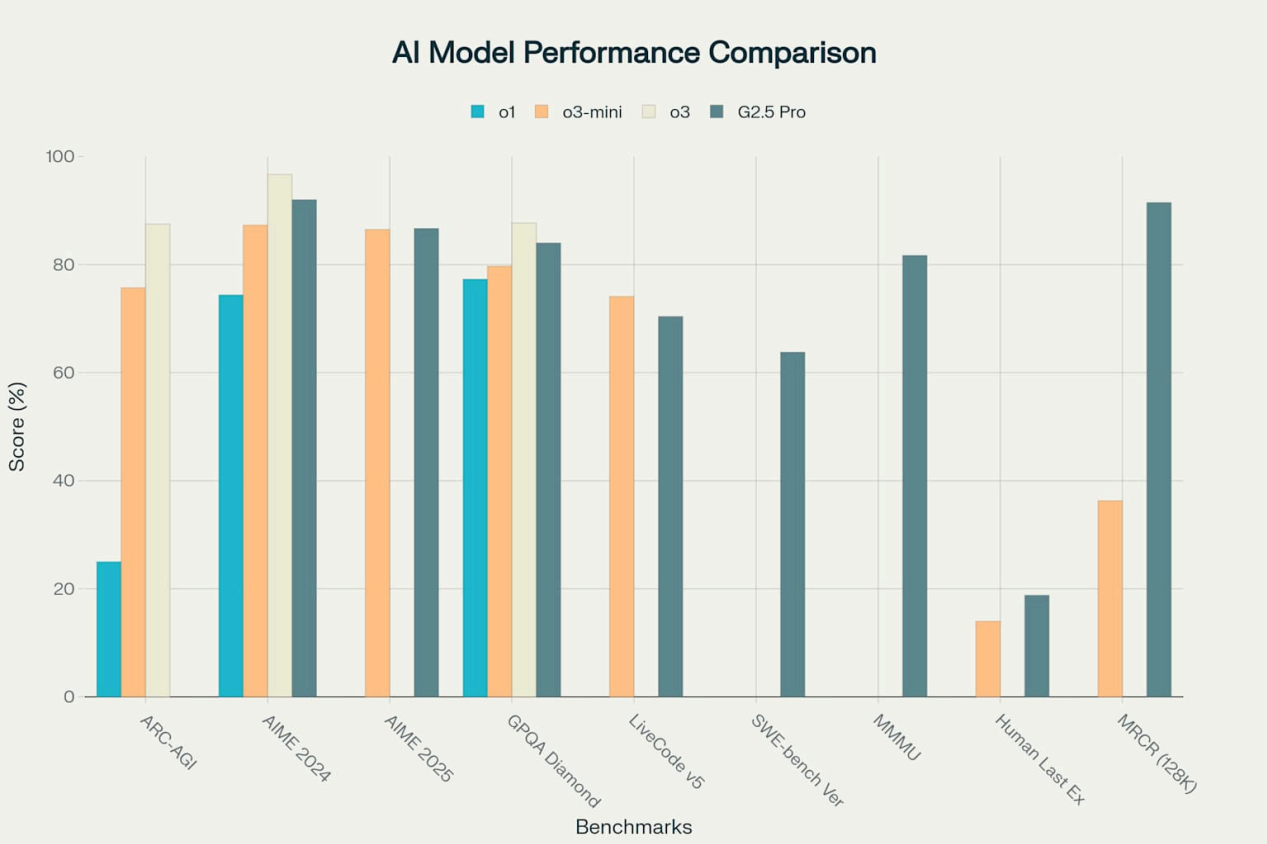
ベンチマーク性能の課題

一方で、現在のベンチマーク評価には根本的な問題があるとの指摘もある[19]。Apple研究チームによるGSM-Symbolicベンチマークでは、問題の数値や名前を変更するだけで、同じ論理構造にも関わらずAIモデルの性能が大幅に低下することが確認された[19]。これは、AIが真の推論よりも学習データのパターンマッチングに依存している可能性を示唆している。

実用的観点：製品性能とベンチマーク比較

総合性能ランキング

最新の推論型AIモデルの実用性能を示すベンチマーク結果では、明確な性能階層が形成されている[5] [3] [20]。



Performance comparison of reasoning AI models across multiple benchmarks

ARC-AGIテストにおいて、OpenAI o3が87.5%という人間平均（85%）を上回る革新的スコアを達成した[5]。これはAGI（汎用人工知能）実現への重要な進展として評価されている[4] [5]。数学的推論では、AIME 2024でo3が96.7%、Gemini 2.5 Proが92.0%という高い性能を示している[3] [6]。

コーディング性能での競争

プログラミング分野では、複数のベンチマークで激しい競争が展開されている[3] [20]。LiveCodeBench v5では、o3-miniが74.1%でトップ、Gemini 2.5 Proが70.4%で続いている[3]。しかし、コスト効率を考慮すると、Gemini 2.5 Proがより実用的な選択肢として評価されることが多い[20]。

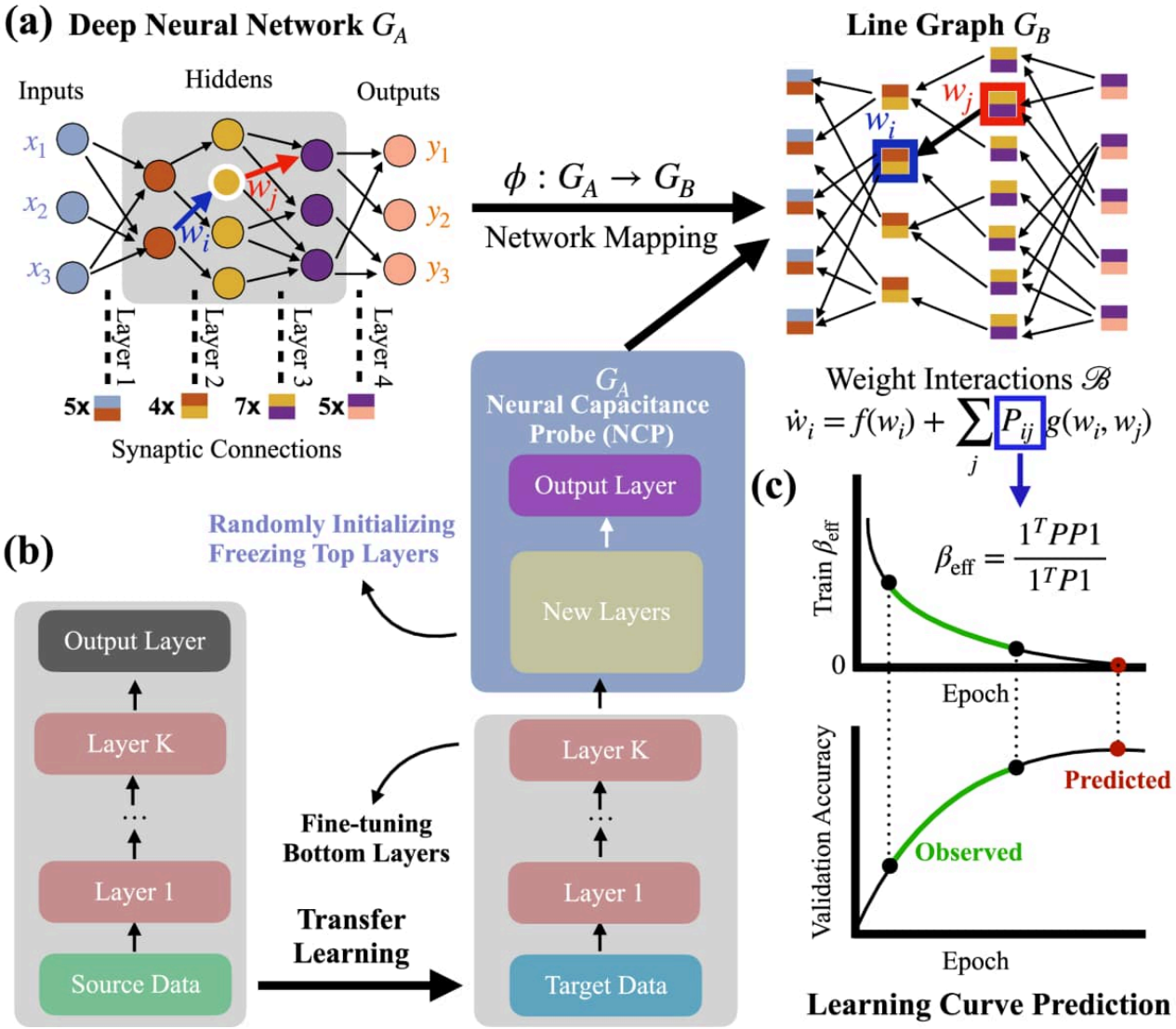
マルチモーダル理解の優位性

Gemini 2.5 Proは、マルチモーダル理解（MMMU）で81.7%の最高スコアを記録し、この分野での明確な優位性を示している[3]。特に長文コンテキスト処理（MRCR 128K）では91.5%という圧倒的性能で、他モデルを大きく引き離している[3] [21]。

マルチモーダル推論の技術的進展

統合的認知処理の実現

2025年のマルチモーダル推論AIは、単一モダリティの限界を超えた統合的認知処理を実現している[22] [13]。これまでのLLMが抱えていた現実世界への接地の問題を、視覚、聴覚、言語情報の統合により解決している[13]。OpenAI o3、Microsoft Magma、Google Gemini 2.5などの最新モデルは、シーンの分析、推論、行動決定を統合的に行う能力を持つ[13]。



Schematic diagram of a deep neural network analysis framework using network mapping, transfer learning, and learning curve prediction.

実世界応用での優位性

マルチモーダル能力は特に実世界のタスクで威力を発揮している[12] [22]。建築図面の解析、製品画像の分類、複雑な表の理解など、従来のテキスト専用モデルでは困難だったタスクで大幅な性能向上が確認されている[12]。この技術進展により、AIシステムは人間に近い多感覚的な情報処理が可能となっている[13]。

推論AIの課題と限界

コスト効率性の問題

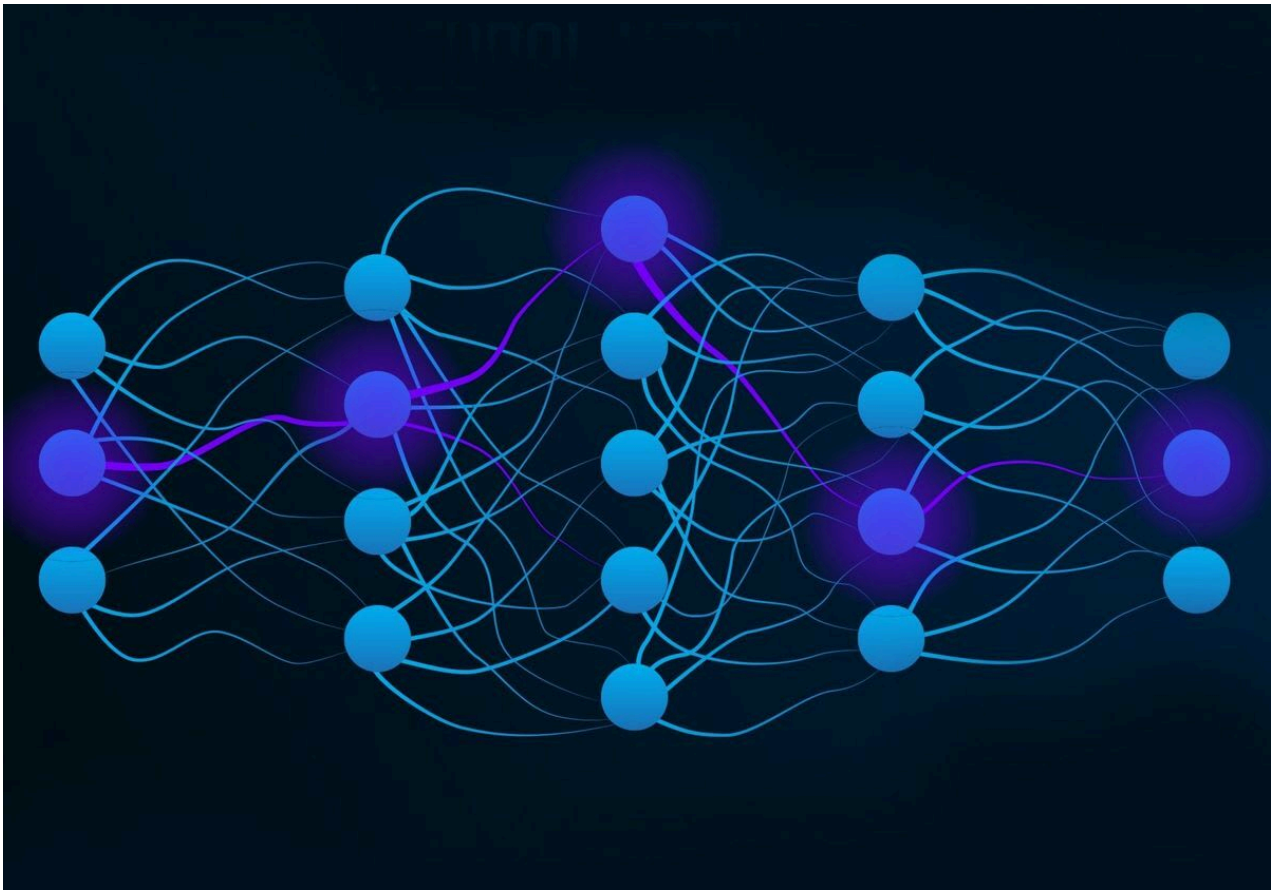
推論型AIモデルの最大の実用的課題は、計算コストの高さにある[23] [24] [20]。OpenAI o3は最高性能を示す一方で、Gemini 2.5 Proより18倍のコストがかかることが報告されている[20]。o3-miniは改善されているものの、依然として従来モデルの3倍のコストが必要である[24] [20]。

安全性と信頼性の課題

推論モデルの動的な処理能力は、従来の予測可能性と信頼性を損なう可能性があるとIBM研究者が指摘している[23]。事前プログラムされた推論スキルとは異なり、新しい推論モデルは確実性に欠ける面がある[23]。特に、モデルが正しい思考プロセスを経ても間違った結論に至るケースが観察されている[18]。

創造性の真正性への疑問

AIの創造性について、真の創造的洞察とパターン再組み合わせの境界が不明確であることが重要な課題として残っている[7] [9]。現在のAIモデルは、学習データの巧妙な再組み合わせにより創造的に見える出力を生成しているが、これが人間の創造性と本質的に異なるプロセスなのかについて、学術的議論が続いている[7] [8]。



A visual representation of a neural network with interconnected nodes and highlighted pathways.

結論と今後の展望

推論型生成AIは、2024年から2025年にかけて劇的な進化を遂げ、多くの認知タスクで人間専門家レベルの性能を達成している[2] [4] [5]。OpenAI o3シリーズは純粋な推論力で、Gemini 2.5 Proはマルチモーダル統合で、それぞれ独自の強みを確立している[3] [6] [20]。学術的評価では、特定の測定基準でAIが人間を上回る創造性を示している一方で、真の創造的洞察とパターン再組み合わせの区別という根本的課題が浮上している[7] [8] [9]。

実用的観点では、これらのモデルは革新的な性能を提供するものの、高い計算コストと予測可能性の課題が広範な導入を制限している[23] [24] [20]。今後の発展では、コスト効率の改善と、人間の創造性との本質的差異の解明が重要な課題となる[7] [23] [19]。AGI実現への重要な進展として評価される一方で、これらの技術的・哲学的課題の解決が、真に有用で信頼できるAI推論システムの実現に不可欠である[4] [5] [19]。

⁂

1. https://qiita.com/ryosuke_ohori/items/7b92d2cdab0dc1365983
2. https://openai.com/index/learning-to-reason-with-llms/
3. https://www.datacamp.com/blog/gemini-2-5-pro
4. https://arxiv.org/abs/2409.18486
5. https://www.gocodeo.com/post/open-ais-o3-benchmarking
6. https://blog.google/products/gemini/gemini-2-5-pro-latest-preview/
7. https://arxiv.org/abs/2505.08744
8. https://www.openaccessgovernment.org/artificial-intelligence-outperforms-humans-in-creative-thinking/174204/
9. https://www.science.org/doi/10.1126/sciadv.adn5290
10. https://orq.ai/blog/what-is-chain-of-thought-prompting
11. https://arxiv.org/html/2410.13639v1
12. https://www.lifehacker.jp/article/2504-reasons-gemini-2-5-pro-best-reasoning-model/
13. https://ajithp.com/2025/04/21/multimodal-reasoning-ai/
14. https://note.com/ainest/n/n58f91a167c14
15. https://aicompetence.org/deepminds-michelangelo-benchmark/
16. https://www.mi-research.net/en/article/pdf/preview/10.1007/s11633-025-1546-4.pdf
17. https://www.mdpi.com/2079-9292/13/23/4584
18. https://composio.dev/blog/openai-o1-preview-a-detailed-analysis/
19. https://www.forbes.com/sites/kolawolesamueladebayo/2025/05/22/ai-models-still-struggle-with-reasoning---and-heres-why/
20. https://dev.to/composiodev/openai-o3-vs-gemini-25-vs-openai-o4-mini-5ej4
21. https://wandb.ai/byyoung3/Generative-AI/reports/Evaluating-the-new-Gemini-2-5-Pro-Experimental-model--VmlldzoxMjAyNDMyOA
22. https://codecrux.com/blog/multimodal-generative-ai-guide-unlocking-creativity/
23. https://www.ibm.com/jp-ja/think/topics/ai-reasoning

24. https://note.com/it_navi/n/nc9e6d154d703