

Gemini 3 Deep Think: ARC-AGI-2で人間を超える推論の到達点

驚異的なパフォーマンス：人間を凌駕する推論力

84.6%
の正答率を達成

ARC-AGI-2 (v2 Semi-Private)
ARC Prize Foundationによって検証された最高水準のスコア



人間平均 (60%) を大きく超越
24ポイント以上

400名以上の一般参加者による平均スコアを
24ポイント以上上回り、人間の推論能力を
凌駕する適応性を示した。

科学・数学・工学への特化

推論な数学的論理の欠陥検出や特
長成高レシビの設計など、専門性の
高い実課題において強みを発揮。



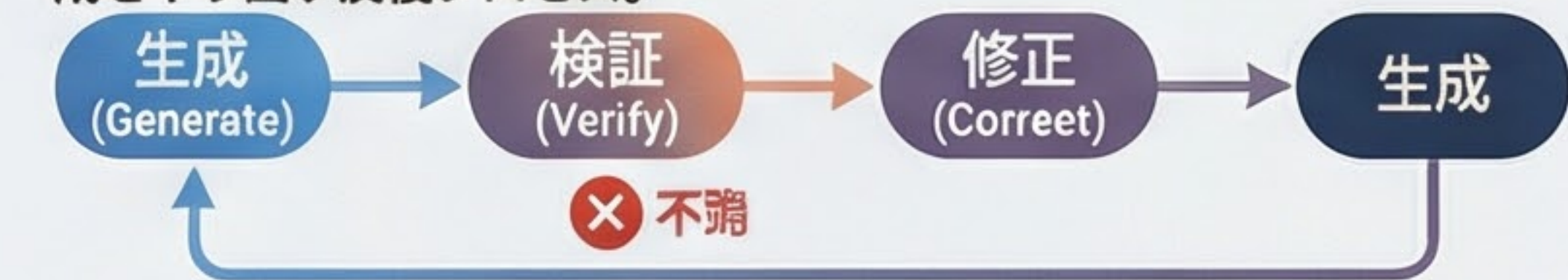
「Deep Think」の技術的メカニズム

並列思考と探索 (Parallel Thinking)



「生成→検証→修正」のループ (Aletheia)

候補解を出し、検証器がエラーを見つけると修正論に固す、おもしろは生
成をやり直す反復プロセス。



推論トークンの爆発的増加

標準的なモデルが96トークンで思考
するのに対し、Deep Thinkは最
大138,000トークンを推論に使用。



主要モデルとのベンチマーク比較

評価指標 / モデル	Gemini 3 Deep Think	Claude Opus 4.6	GPT-5.2	人間 (平均)
ARC-AGI-2 (Semi-Private)	84.6%*	68.8%	53.1%	2352
Humanity's Last Exam	53.4%*	53.1%	45.5%	2352
Codeforces (Elo Rating)	3455	52.9%	45.5%	未指定
人間 (平均)	60.0%	N/A	N/A	N/A

評価と課題：AGIへの道のりと現実

1タスクあたり
\$13.62
のコスト

1タスクあたり僕のコスト
高い推論性能の代償として計算
資源を大量消費するため、コス
ト効率の改善が今後の課題。



データ汚染 (Contamination) の懸念
学習データにARG特有の拘束形式や
既知が含まれていることによる
「フォーマット認知性」がスコアを
押し上げている可能性。



AGIへの残されたギャップ
ARC-AGI-2は2Dグリッド推論に
限定されており、寛世界の長遠計
画や社会的相互作用を調るARC-
AGI-3への進化が必要。