GPT-5.2 評判総まとめ

技術・性能・市場評価から紐解く、OpenAI最新モデルの全貌

2025年12月版

革命ではなく、計算された進化

GPT-5.2は、革命的な飛躍ではなく、実用的で的を絞った機能強化によってOpenAIの競争上の地位を確固たるものにする、プロフェッショナル向けの強力な進化である。

プロフェッショナルへの特化

専門家の知的労働を代替する能力が飛躍的に向上。70.9%のタスクで専門家に勝利または引き分け。

計測可能な性能向上

コーディング (SWE-Bench Pro SOTA)、数学 (AIME 100%) で新記録を達成する一方、ハルシネーションは依然として課題。

実用性重視のUX

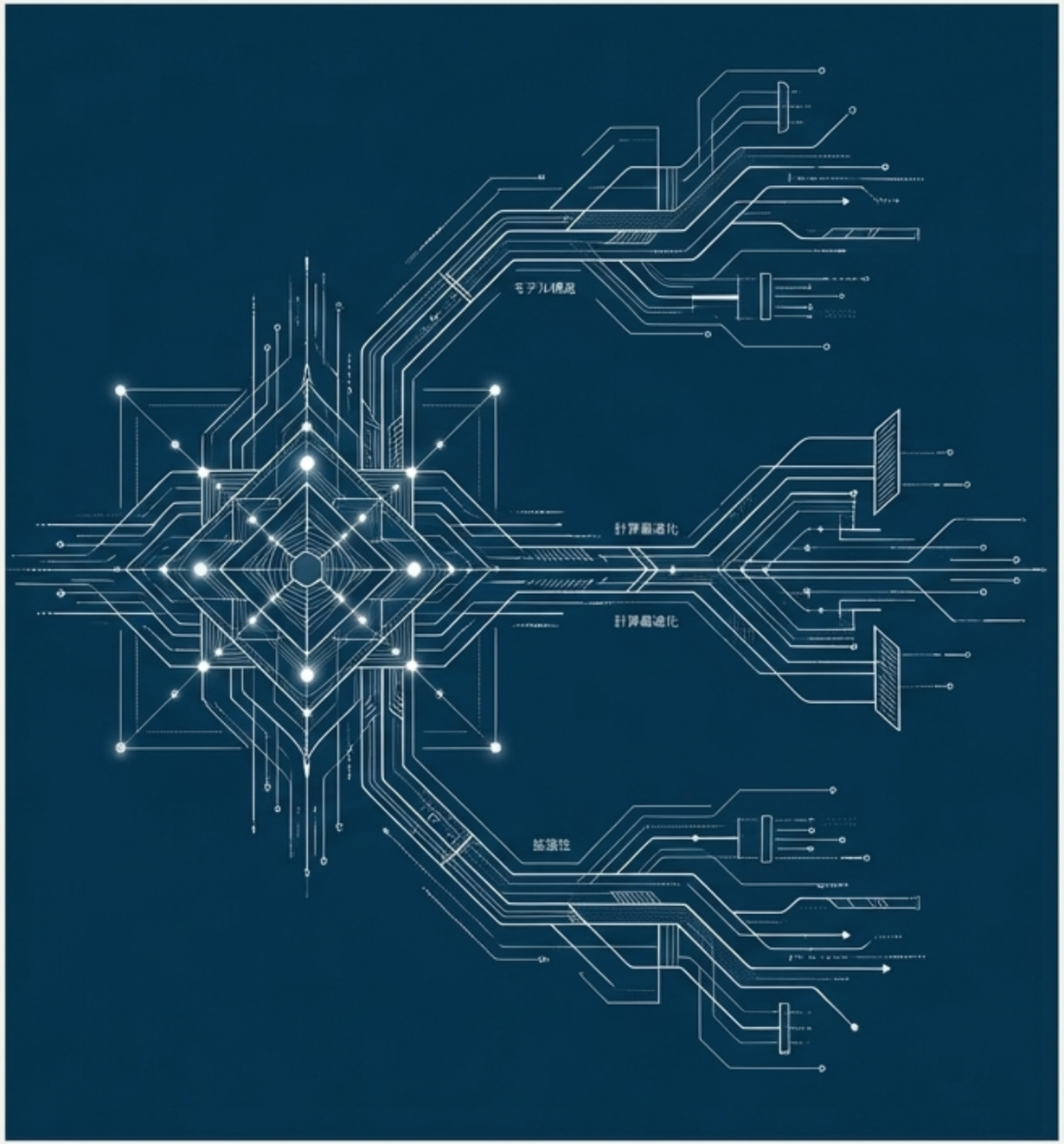
応答の自然さや精度は向上したが、一部のユーザーは以前のモデルの「人間味」や「遊び心」が失われたと感じている。

戦略的な市場対応

競合の猛追を受け「コードレッド」状況下でリリースされた、市場での優位性を維持するための戦略的アップデート。

Part 1: The Technology

より強力で柔軟なアーキテクチャへ



タスクに応じて最適化された、新次元の制御性

Instant

応答速度重視。日常的なタスクや高速な対話向け。

Thinking

高度な推論。長時間のエージェント運用や複雑な問題解決向け。

Pro

精度最優先。最高の正確性が求められる最上位モデル。

New Parameter

`reasoning.effort`

LowからX-Highまで思考深度を制御可能。計算リソースを動的に調整し、速度と精度の両立を実現。

前例のないスケールと、より鋭敏になった知覚

400,000 Tokens

Massive Context Window

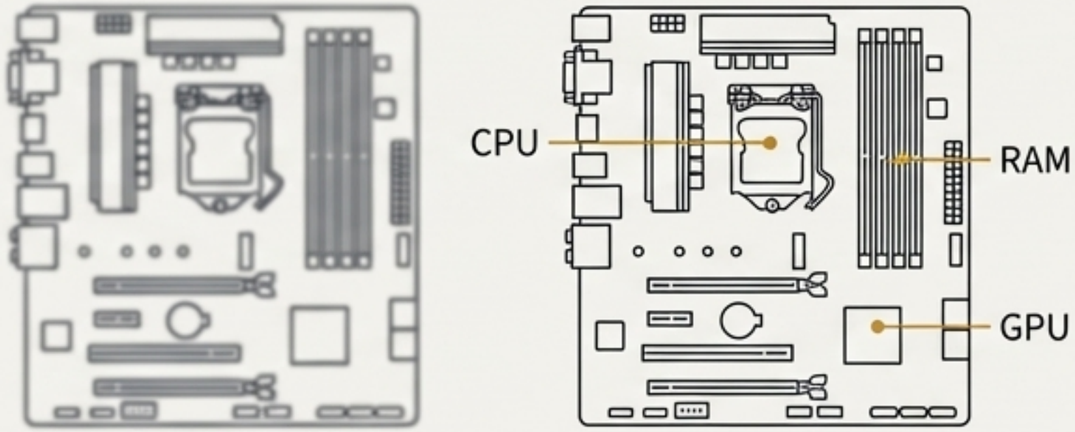
長大な文書や大量のコードを一括処理。OpenAI独自のMRCRv2評価で256kトークン級の長文推論において、ほぼ100%に近い正確性を達成。

> 複数の契約書や研究論文を横断した分析が一度に可能に。

~50%

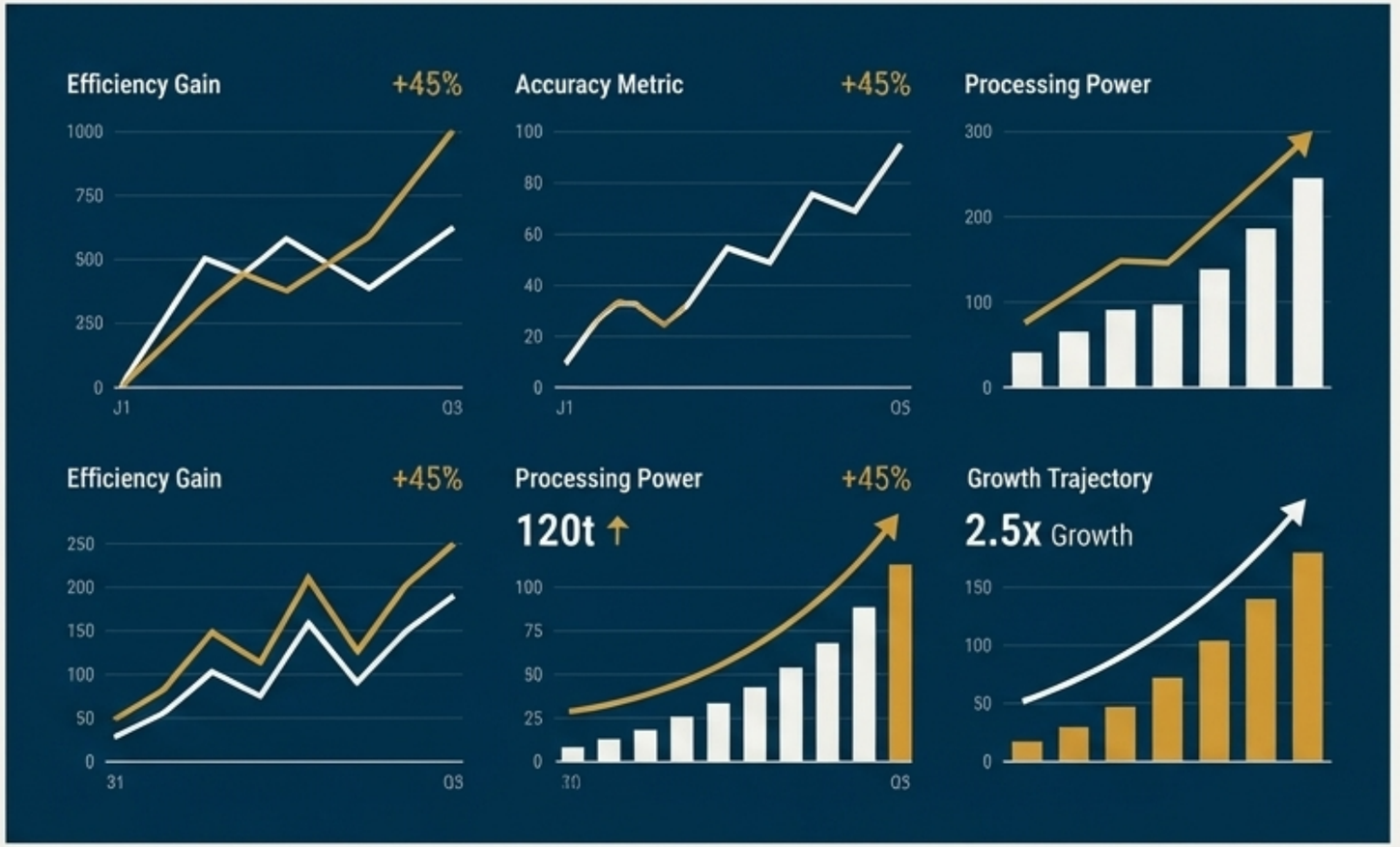
Reduction in Vision Task Errors

チャート読解やUI画面理解などの視覚推論タスクでエラー率が約半減。科学論文のグラフに関する質問への正答率が大幅に向上。



Part 2: The Performance

計測された知性の飛躍

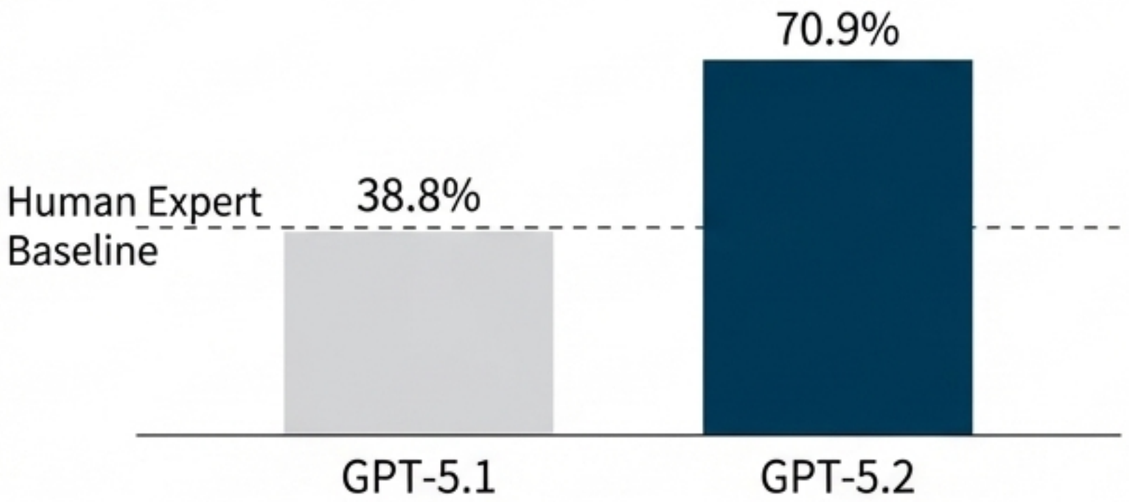


専門家を凌駕し、新記録を樹立

Professional Tasks (GDPval Benchmark)

70.9%

44職種にわたる実務タスクにおいて、GPT-5.2 (Thinking)が人間の専門家に勝利または引き分けた割合。(前モデルGPT-5.1の38.8%から大幅増)



Coding (SWE-Bench Pro)

55.6% (New SOTA)

難関プログラミング評価で新たな最高性能を達成。特にフロントエンドUI生成能力が飛躍的に向上。

Math (AIME 2025)

100%

競技数学の難問セットでパーフェクトスコアを記録。(前モデルの94.0%から向上)

ハルシネーション問題：完璧ではなく、着実な進歩

公式の改善

30%減少

OpenAI公式発表。GPT-5.1と比較した回答のエラー率。

「確かに着実な進化を遂げている」
（専門家コメント）

リサーチや分析業務での信頼性が向上。

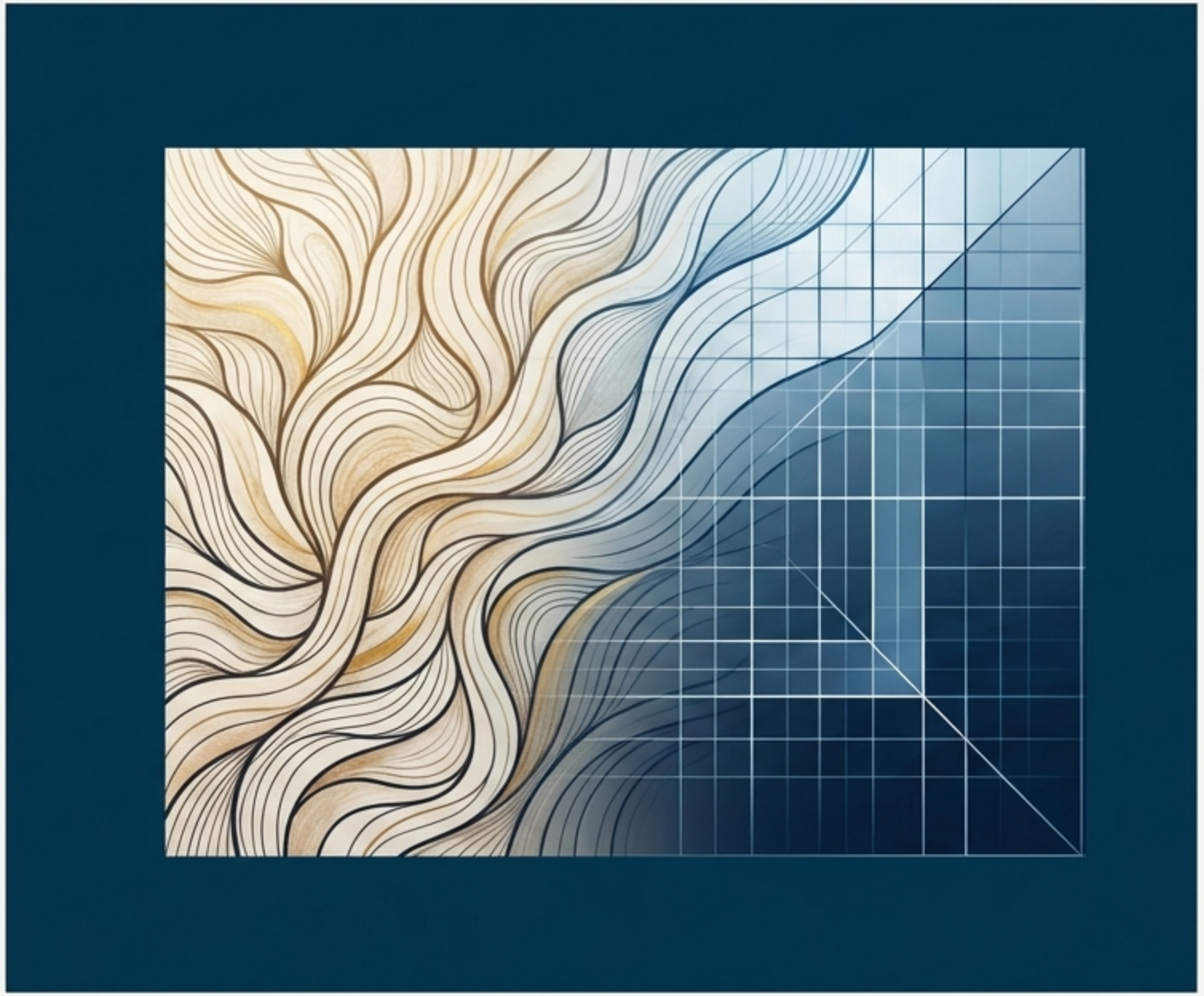
ユーザーの現実

「依然として法令やコードセクションを90%の確率でハルシネートする」
（一般ユーザーからの不満報告）

専門的な引用や法的根拠など、完全な正確が求められる場面では依然として注意が必要。

Part 3: The User Experience

実用性重視へのシフト



より正確に、しかし人間味は薄れて

評価される点:
自然で高精度な応答

- 翻訳調のぎこちなさが消え、非常に自然な日本語を出力。
- 曖昧な指示でも的確に意図を解釈する能力が向上。

「日本語のこなれ度は飛躍的に向上した」

好みが分かれる点:
失われた人間味と遊び心

- ビジネスライクで厳密になった反面、カジュアルな対話や創造的な遊びが減少。
- 「忠実だが芸術点は伸びない」との評価も。

「GPT-5.1は人間にとって快適なアシスタントを目指し、5.2は機械や企業にとって検証可能な実用エンジンを目指した」

Part 4: The Market Context

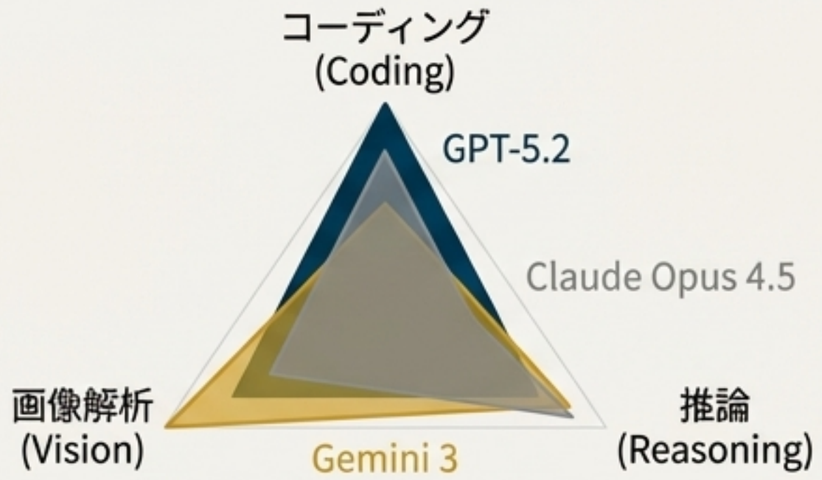
熾烈化するAI開発競争

「コードレッド」への戦略的対応

The Catalyst

2025年11月のGoogle Gemini 3とAnthropic Claude Opus 4.5の登場に対し、OpenAIが社内で「Code Red」を宣言。
サム・アルトマンCEOは「GPT-5.2がGemini 3を上回った」と社内で述べたと報じられている。

The Scoreboard is Mixed



GPT-5.2が優位
(SWE-Bench ProでGemini 3 Proを上回る)

Gemini 3が依然として優位な指標も

ベンチマークの公平性を巡る議論

「GDPvalはOpenAIが作った指標だから額面通りには受け取れない」（Redditユーザーの懐疑的な意見）

OpenAIのベンチマークは、競合より多くのトークンや計算量を使って測定しているのではないか、というコミュニティからの指摘。

Part 5: The Synthesis

GPT-5.2の総合評価

最終評価：長所と短所・懸念

長所 (The Strengths)	短所・懸念 (The Challenges & Criticisms)
• **総合性能:** 44職種の知的業務でプロに7割勝利・同等。知識労働の代替で大きく前進。 • **正確性:** GPT-5.1比で誤答・幻覚が30〜38%減少。 • **専門能力:** コーディング(SWE-Bench Pro SOTA)と数学(AIME 100%)で競合を凌駕。 • **長文処理:** 40万トークンのコンテキスト長で、複数文書の横断分析などが可能に。 • **UX/実用性:** 自然な日本語、意図の汲み取り精度向上。ChatGPT上で直接スプレッドシートやプレゼンを生成可能。	• **進化の性質:** 「画期的な飛躍ではなく漸進的」な改良であり、AGIを期待した層には物足りない。 • **残存する欠点:** ハルシネーションは完全には解消されず、法律やコードの引用で依然として誤情報が発生。 • **UXの変化:** より実直で硬質になり「面白みが減った」「創造性が低い」との声も。 • **コスト:** API利用料が大幅増。個人開発者にはハードルが高く、「大企業向けの価格設定」 • **戦略への不信:** 自社に有利なベンチマークの公表や、頻繁なモデル更新でユーザーが翻弄されることへの懸念。

確実な一歩、しかしその先にある道は？

GPT-5.2は競争力を維持するための確実な進化を遂げた。だが、汎用人工知能（AGI）への真のブレークスルーに向けた次の一手は、どこから来るのだろうか？