Google Gemini 2.5 Pro Deep Think 徹底レポート：技術・性能・料金・競合比較

Genspark

【エグゼクティブサマリー】

- 実在と位置づけ: Deep Think は「Gemini 2.5 Pro の実験的・強化推論モード」。I/O 2025 で正式発表され、「複数仮説の並列思考」を用いる推論強化として先行提供開始。まずは Gemini API の trusted tester、その後に Gemini アプリの Google AI Ultra 加入者へ限定提供からロールアウト（2025 年 8 月 1 日から限定解放）blog.google[1] techcrunch.com[2] gemini.google[3]。
- 技術的定義: Parallel thinking（複数の思考パスを同時並行に展開）と新規の強化学習を組み合わせ、回答前に複数仮説の生成・批判的検証を行う「考える時間の拡張」を特徴とする推論モード blog.google[1] Google DeepMind PDF[4]。
- 主要ベンチマーク: Deep Think は HLE 34.8%（ツール無し）、AIME 99.2%、IMO 60.7%（Bronze 相当）、LiveCodeBench v6 87.6%で、いずれも通常の 2.5 Pro（HLE 21.6%、AIME 88.0%、IMO 31.6%、LiveCodeBench 74.2%）を大幅に上回る Google DeepMind モデルカード [5]。
- マルチモーダル・長文: 2.5 シリーズはネイティブにテキスト/音声/画像/動画を扱い、1,048,576 トークン（約 100 万）入力の超長文コンテキスト。2.5 Pro は動画最大 3 時間処理や 1M トークン下での長文・マルチモーダル推論に強い ai.google.dev モデル一覧 [6] Google DeepMind PDF[4]。
- アーキテクチャ: 2.5 は Sparse Mixture-of-Experts（MoE）系 Transformer。総容量と 1 トークンあたり計算コストを分離し、推論効率と性能を両立。視覚処理の刷新で「3 時間動画」を 1M トークン内で扱えるようにするなど効率を向上 Google DeepMind PDF[4]。
- 料金と提供: 2.5 Pro/Flash は Vertex AI・Gemini API で GA。Pro は入力$1.25/100 万 tokens〜、出力（推論トークン含む）$10/100 万 tokens〜（標準 API、≤200k 入力時）。Deep Think は価格別 SKU というより推論モードとしての提供で、まずは API の trusted tester と Ultra 加入者へ限定 cloud.google.com Vertex AI 価格 [7] ai.google.dev 価格 [8] gemini.google リリースノート [3]。
- 競合比較: GPT-5 は AIME 94.6%（no tools）、SWE-bench Verified 74.9%、Aider Polyglot 88%、MMMU 84.2%、GPQA 88.4%と強力だが、OpenAI は MMLU/GSM8K/HumanEval の公的数値は未掲載。Anthropic Claude Opus 4.1 は

ハイブリッド推論（拡張思考最大 64K）を明示するが、GSM8K/HumanEval の公式値は当該記事では未掲。Deep Think は HLE や LiveCodeBench で競合を凌駕と報じられているが、普及面では利用条件（Ultra/信頼テスター）が厳しめ openai.com9 OpenAI システムカード PDF10 anthropic.com11 techcrunch.com 8/112。

---

1. 存在確認と「Deep Think」の定義・目的

- 存在と発表経緯: Google I/O 2025 で「2.5 Pro の実験的強化推論モード」として Deep Think を正式公開。複数仮説を検討してから回答する「強化推論モード」で、まずは Gemini API の trusted tester に提供し、安全性評価の後に段階的に拡大と説明 blog.google1。
- 機能の実態: Parallel thinking（複数の思考パスを並列に展開）と新しい強化学習により、より長く・深く考える工程を回答前に挿入。高度な数学・コーディング・マルチモーダル推論で効果を発揮 blog.google1 Google DeepMind PDF4。
- 提供形態: Deep Think は「別 SKU のモデル」ではなく、2.5 Pro における「推論モード（enhanced reasoning mode）」として位置づけられている（API のモデル一覧では「2.5 Pro」は“Thinking”サポートの思考モデルで、Deep Think の名称を個別 SKU としては明示していない）ai.google.dev モデル一覧 6 blog.google1。
- ロールアウト状況: 2025/5 時点で trusted tester へ。2025/8/1 からは Gemini アプリの Google AI Ultra 加入者向けに限定公開（早期・限定アクセス）techcrunch.com 5/202 gemini.google リリースノート 3。

2. 主要ベンチマーク（MMLU/GSM8K/HumanEval 他）と比較

- Deep Think（公式モデルカードの主要値）
  - Humanity’s Last Exam（HLE, no tools）: 34.8%（2.5 Pro は 21.6%）Google DeepMind モデルカード 5
  - AIME 2025: 99.2%（2.5 Pro は 88.0%）Google DeepMind モデルカード 5
  - IMO 2025: 60.7%（Bronze 相当、2.5 Pro は 31.6%）Google DeepMind モデルカード 5
  - LiveCodeBench v6: 87.6%（2.5 Pro は 74.2%）Google DeepMind モデルカード 5
  - MMMU: I/O 記事で Deep Think 84.0%と記載（2.5 Pro は後述の技術レポで 82.0%）blog.google1。
- Gemini 2.5 Pro（技術レポート Table 3 抜粋）
  - Global MMLU: 87.8%（Flash 88.4%）Google DeepMind PDF4
  - GPQA（diamond）: 86.4%（Flash 82.8%）Google DeepMind PDF4

    - AIME 2025: 88.0%（Flash 72.0%）Google DeepMind PDF4
    - LiveCodeBench: 74.2%（Flash 59.3%）Google DeepMind PDF4
    - MMMU: 82.0%（Flash 79.7%）Google DeepMind PDF4
    - HLE（no tools）: 21.6%（Flash 11.0%）Google DeepMind PDF4
- GPT-5（OpenAI 公式発表の代表値・同時期）
    - AIME 2025: 94.6%（no tools）openai.com9
    - SWE-bench Verified: 74.9%、Aider Polyglot: 88%（実開発寄りコード・編集系）openai.com9
    - MMMU: 84.2%、GPQA: 88.4%（no tools）openai.com9
    - 備考: GPT-5 の公式記事・システムカードは MMLU/GSM8K/HumanEval の数値を明記していない openai.com9 OpenAI システムカード PDF10
- Claude Opus 4.1（Anthropic）
    - 「ハイブリッド推論モデル」で「拡張思考（最大 64K tokens）」の利用条件を明示。記事内は MMMLU 等の図表で比較を掲載するが、当該記事中に GSM8K/HumanEval の公式値は見当たらない anthropic.com11。
- Gemini 2.5 Pro（補足）: SWE-bench Verified 63.8%（カスタムエージェント設定）、LMArena で大幅上位などの主張 blog.google 3 月発表 14。

注 1: MMLU/GSM8K/HumanEval に関し、Deep Think の公式モデルカードは該当値を掲載していません。Gemini 2.5 Pro の MMLU（Global MMLU）は技術レポートにありますが、GSM8K/HumanEval は同レポートでは確認できませんでした Google DeepMind PDF4。

3. マルチモーダル、長文理解・要約、複雑推論の実例評価

- マルチモーダル能力:
    - 2.5 シリーズはネイティブにテキスト・音声・画像・動画入力をサポート（Pro/Flash/Lite/LIVE/Native Audio など派生）ai.google.dev モデル一覧 6。
    - 2.5 Pro は 1M トークンの長文下で動画最大 3 時間の処理が可能。視覚トークン削減（1 フレーム 258→66）で長時間動画を 1M 内に収める設計。画像を HTML/SVG の構造表現へ変換、動画内容からインタラクティブなアプリ（クイズ、p5.js アニメ）生成などの「入力→アプリ化」能力も強化 Google DeepMind PDF4。
    - 2.5 Pro ページは低遅延のネイティブ音声対話・スタイル制御・ツール連携・雑音抑制など音声機能も強調 DeepMind/Pro ページ 15。
- 長文理解/要約:
    - 入力上限 1,048,576 tokens（Pro/Flash/Lite）ai.google.dev モデル一覧 6。
    - 長文評価（LOFT、MRCR）や 1M トークン下の複雑検索・要約に強みを示す（LOFT hard retrieval 1M: Pro 69.8%）Google DeepMind PDF4。

- 複雑問題・推論:
    - Deep Think の公式デモ（Codeforces の難問「Catch the mole」）では、Minimax 戦略や不確実性管理、並列多数の思考パスからの合成で高難度の探索的コーディング課題を解く様子が提示 YouTube/Google DeepMind13。
    - ベンチ実績として、HLE/IMO/AIME/LiveCodeBench で通常 2.5 Pro を大幅上回る（前項参照）Google DeepMind モデルカード 5。

4. モデルのアーキテクチャ、パラメータ、学習データ、推論効率

- アーキテクチャ: Sparse Mixture-of-Experts（MoE）Transformer。トークンを専門家（experts）に動的ルーティングし、総容量と計算コストを分離。これにより計算効率を維持しながら能力を拡大 Google DeepMind PDF4。
- パラメータ数: 非公開（公式技術レポートは MoE の設計方針やルーティングの利点を説明するが、具体パラメータ数は示さず）Google DeepMind PDF4。
- 学習データ: 公開 Web、コード、画像、音声、動画など多様モダリティ。2.5 のプリトレは 2025 年 1 月がカットオフ（1.5 よりデータ品質のフィルタリングと重複排除を改善）。ポストトレは指示・人間選好・ツール利用など Google DeepMind PDF4。
- 推論効率・設計:
    - 視覚処理の刷新で映像トークン効率化（3 時間動画を 1M 内で処理可能）。長文×マルチモーダル×Thinking の統合的運用が可能 Google DeepMind PDF4。
    - Thinking/Deep Think は「思考予算（thinkingBudget）」で推論トークン量を制御可能。0 で無効、1〜1024 は 1024 に丸め、最大 24576（API 設定）ai.google.dev Thinking16。

5. レビュー・第三者評価（長所・短所・課題）

- メディア/外部評価の概観:
    - TechCrunch（5/20）: Deep Think は「複数仮説を検討する強化推論」で 2.5 Pro のベンチ（LiveCodeBench や MMMU）を押し上げ、trusted tester を起点に展開 techcrunch.com 5/202。
    - TechCrunch（8/1）: Deep Think を「初の一般公開マルチエージェント」系と表現。Ultra 加入者と API の選抜テスターへ提供。HLE 34.8%、LiveCodeBench v6 87.6%など SOTA の主張を紹介 techcrunch.com 8/112。
    - Ars Technica（8/1）: Deep Think は「数分かけて解く」重推論で HLE 34.8%など競合凌駕の結果に触れ、Ultra 限定のアクセス制約も指摘 arstechnica.com17。
- 長所:

    - 高難度数学・競技コーディング・マルチモーダル推論で顕著な上振れ（AIME/IMO/LiveCodeBench/HLE）Google DeepMind モデルカード 5。
    - 1M トークン長文と 3 時間動画処理の現場適用性、ネイティブ音声の会話品質/表現制御 ai.google.dev6 DeepMind/Pro ページ 15。
- 短所・課題:
    - サービングコストとレイテンシ（「数分」かかり得る）。高価プラン（Ultra）や限定テストからの提供で利用門戸が狭い arstechnica.com17 techcrunch.com 8/112。
    - 一部の定番ベンチ（GSM8K/HumanEval）について公式の完全比較が未整備。比較の透明性・再現性を高める余地 Google DeepMind PDF4 Google DeepMind モデルカード 5。

6. 応用領域・ユースケース・API/プラットフォーム・料金

- 応用・デモ:
    - 科学研究: 長文×マルチモード×Thinking で大規模文献/動画講義/データから仮説検証・可視化（2.5 レポの多数例）Google DeepMind PDF4。
    - ソフトウェア開発: SWE-bench/LiveCodeBench の高スコア、動画→インタラクティブアプリ化、コード変換・編集、WebDevArena 上位など blog.google 6/5 プレビュー18 blog.google 3 月発表 14。
    - クリエイティブ: ネイティブ音声、スタイル制御、マルチモーダル生成・編集のワークフローDeepMind/Pro ページ 15。
- API/プラットフォーム:
    - 一般提供: 2.5 Pro/Flash は Vertex AI、Gemini API、Google AI Studio で GAGoogle Cloud Blog19。
    - Deep Think: まず API の trusted tester、のちに Gemini アプリで Google AI Ultra 加入者が限定アクセス（8/1 時点）。順次 API の一部テスターにも展開予定 gemini.google3 techcrunch.com 8/112。
- 料金の代表（Vertex AI; USD, /100 万 tokens の目安）
    - Gemini 2.5 Pro: 入力$1.25（≤200k）、$2.5（>200k）。出力（応答＋推論）$10（≤200k）、$15（>200k）。バッチは半額。Search Grounding は 1 日 1 万回まで無償、超過$35/1,000 回 cloud.google.com Vertex AI 価格 7。
    - Gemini 2.5 Flash（GA）: 入力$0.30、出力$2.50。Flash-Lite はさらに低単価 developers.googleblog Flash-Lite 告知 20 cloud.google.com Vertex AI 価格 7。
    - Gemini API（開発者向け）: 2.5 Pro（プレビュー期の例）で入力$1.25〜、出力$10〜（思考トークンも出力に含む）。Search Grounding は無料枠/有料

超過あり ai.google.dev 価格 8。

7. 競合（GPT-5、Claude Opus 4.1、従来 Gemini）との比較考察

- 数値比較（代表的・公式確認できる指標）
  - AIME 2025: Deep Think 99.2% > GPT-5 94.6% > 2.5 Pro 88.0%Google DeepMind モデルカード 5 openai.com9 Google DeepMind PDF4。
  - LiveCodeBench v6: Deep Think 87.6% > 2.5 Pro 74.2%（GPT-5 は SWE-bench/Aider では強いが LCB 公式値は未掲載）Google DeepMind モデルカード 5 openai.com9。
  - HLE（no tools）: Deep Think 34.8% > 2.5 Pro 21.6%（GPT-5 は HLE 明示なし）Google DeepMind モデルカード 5 Google DeepMind PDF4。
  - MMMU: Deep Think 84.0%（I/O 記事）、GPT-5 84.2%、2.5 Pro 82.0%blog.google1 openai.com9 Google DeepMind PDF4。
  - Global MMLU: 2.5 Pro 87.8%（Deep Think 未掲載、GPT-5 未掲載、Claude Opus 4.1 は記事内掲載図だが本稿では数値引用を見送る）Google DeepMind PDF4 anthropic.com11。
- 所見:
  - 数学（AIME/IMO）と競技コーディング（LCB）では Deep Think が現行公開値で最有力。汎用・実務コード（SWE-bench, Aider）は GPT-5 も最上位圏。マルチモーダル理解（MMMU）は Deep Think と GPT-5 が拮抗 Google DeepMind モデルカード 5 openai.com9。
  - 一方で Deep Think は利用条件が限定的（Ultra/テスター）。レイテンシとコストの設計上、対話業務や軽量用途には 2.5 Flash/Flash-Lite や他社の高速系の方が適するケースがある arstechnica.com17 ai.google.dev モデル一覧 6。

8. 設定・運用 Tips（思考予算と実務上の注意）

- thinkingBudget の活用: 難問は大きめ（例: 8k〜24k）で精度向上、軽タスクは 0〜1k で低遅延・低コスト化。Flash-Lite は思考デフォルト OFF、Pro/Flash は思考内蔵で調整可能 ai.google.dev Thinking16 developers.googleblog21。
- 入出力上限: 2.5 Pro/Flash/Flash-Lite の入力上限は 1,048,576 tokens、2.5 Pro の出力上限は 65,536 tokens（公式）。外部記事で「出力 192k」などの言及もあるが、運用は公式上限を参照 ai.google.dev モデル一覧 6 TechTarget22。

9. 未公開・留意点

- Deep Think の MMLU/GSM8K/HumanEval は公式モデルカード／I/O 記事では未掲載（他媒体の推定値には注意）。比較は HLE/AIME/IMO/LCB/MMMU など公開

値中心に行うのが妥当 Google DeepMind モデルカード 5 blog.google1。
- パラメータ規模は非公開。MoE 設計・動画効率化などの設計思想は公開される一方、内部容量は秘匿 Google DeepMind PDF4。
- 提供の段階性: 安全性評価（frontier safety）を経る段階的展開で、UI 上の表示有無や地域差があり得る（Ultra 加入でも一時的に UI 変化がないという報告もコミュニティに散見）blog.google1 Google ヘルプスレッド 23。

比較のための簡易表（代表値）

- HLE（no tools）: 2.5 Pro 21.6% → Deep Think 34.8%（+13.2pt）Google DeepMind モデルカード 5。
- AIME 2025: 2.5 Pro 88.0% / Deep Think 99.2% / GPT-5 94.6%Google DeepMind PDF4 Google DeepMind モデルカード 5 openai.com9。
- LiveCodeBench v6: 2.5 Pro 74.2% / Deep Think 87.6%（GPT-5 は SWE-bench 74.9%等の別指標を公表）Google DeepMind PDF4 Google DeepMind モデルカード 5 openai.com9。
- MMMU: 2.5 Pro 82.0% / Deep Think 84.0%（I/O 記事）/ GPT-5 84.2%Google DeepMind PDF4 blog.google1 openai.com9。

結論と実務向け提案

- 何を選ぶべきか:
    - 数学・理論・競技的コーディングや複雑な戦略課題では Deep Think が最有力。ただし処理が重く高コストなので、「重要案件で最良品質が必要」「時間に余裕がある」時に限定して使うのが合理的 Google DeepMind モデルカード 5 arstechnica.com17。
    - 高スループットの要約・抽出・軽量エージェントは 2.5 Flash/Flash-Lite へ（思考予算を動的制御）。日常運用ではコスト/レイテンシのバランスが良い developers.googleblog21。
    - 実務コード修正・エージェント連携・ベンチ（SWE-bench/Aider）では GPT-5 も非常に強力。プロダクト要件に応じて Deep Think と使い分けが得策 openai.com9。
- 実装の勘所:
    - 思考予算（thinkingBudget）を業務難易度と SLA でプロファイリング。長文はコンテキストキャッシュや分割処理でコスト最適化 ai.google.dev Thinking16 cloud.google.com 価格 7。
    - マルチモーダルでは入力の粒度（フレーム間隔・画質・音声品質）を調整し、出力は構造化（Structured Outputs）で下流処理と接続 ai.google.dev 構造化

出力 24。
    - 監査性が必要な領域はツール実行/検索グラウンディングのログ化を徹底 ai.google.dev 検索連携 25。

参考（一次情報・公式中心）

- Deep Think 公式発表とロールアウト: blog.google（I/O 2025）1／techcrunch.com 5/202／gemini.google リリースノート 3
- 技術仕様・ベンチ: Google DeepMind 技術レポ 4／Deep Think モデルカード 5／ai.google.dev モデル一覧 6／ai.google.dev Thinking16
- 価格・GA: Google Cloud Blog（2.5 Pro/Flash GA）19／Vertex AI 価格 7／ai.google.dev 価格 8
- 競合: OpenAI GPT-5 公式 9／GPT-5 システムカード 10／Anthropic Opus 4.111
- デモ: Deep Think コーディングデモ（動画）13

必要であれば、上記ベンチ指標を用途別の推奨構成に落とし込んだ選定マトリクス（用途×モデル/モード×思考予算×コスト見積）も作成します。

Appendix: Supplementary Video Resource

[Google Gemini 2.5 Pro Deep Think 徹底レポート：技術・性能・料金・競合比較](/spark?id=eb064809-5bde-4eff-b3a7-d24d8b4dfeb2)