ChatGPT

DeepSeek-V3.1総合レポート

モデルの性能

言語理解と推論能力の特徴と改善点

DeepSeek-V3.1は**言語理解**と**推論**の両面で大きく強化されています。特に**ハイブリッド推論構造**（後述）により、一つのモデルで通常対話と深い推論を切り替え可能になりました。このアップデートによってモデルはより効率的に長い推論を行え、DeepSeek-R1（2025年5月版）のような推論専用モデルと同等の回答品質を保ちつつ応答時間が短縮されています。実際、**MMLU**（学術・常識知識の総合評価）ではV3.1の成績は**約85–93%**（Exact Match）に達し、前版V3（2025年3月版）の約81%を上回りました [1] [2]。これはOpenAIのGPT-4と同等の水準であり [3]、大規模評価では**Chatbot Arena Elo**スコアで1382を記録し、全モデル中トップクラス（オープンモデルでは最高位）となっています [4]。また**GSM8K**（小学生算数テスト）のような推論課題でも、DeepSeek-V3.1は約85%前後の高い正答率を示し、GPT-4に匹敵する水準に達していると報告されています [3]。これらの結果から、V3.1は知識問題・数学的推論において従来のオープンソースモデルを凌駕し、閉源の最先端モデルにも迫る性能を獲得しています [5]。

加えて、**長大なコンテキスト**処理能力の向上も性能面で重要です。V3.1では**文脈ウィンドウが128,000トークン**に拡張され、約300ページの書類に相当する情報を一度に保持できます [6]。これは従来モデルの4倍にあたり、大規模な文書要約・分析や長編対話での一貫性維持が飛躍的に向上しました [7]。長文文脈への対応強化のため、**マルチヘッド潜在アテンション（MLA）**による深層の注意機構や**マルチトークン予測（MTP）**による並列生成などの新技術も導入されています [8]。これらにより、一度に複数トークンを予測して生成速度を向上させつつ、長文に対する内部メモリ管理を改善しています [9]。さらに**FP8精度**（8-bit浮動小数点、UE8M0フォーマット）の採用により、高速化とメモリ効率が図られています。こうした工夫のおかげで**トレーニングコスト**は$5.6M相当に抑えられ（約278万GPU時間） [10]、同規模の従来モデルに比べ非常に効率的です。

コード生成能力とベンチマークスコア

コード生成性能はDeepSeek-V3.1の大きな強みであり、前バージョンからさらに改善されています。特に**HumanEval**（プログラミング問題集）でのPass@1スコアは**82.6%**に達し、GPT-4の80.5%を上回りました [11] [12]。2024年末時点でも既にDeepSeekモデルはHumanEvalで80%以上を記録していましたが [13]、V3.1で遂に閉源モデルを凌駕する水準に到達したことになります。さらに**Codeforces**（競技プログラミング）ではDeepSeek-V3.1が**51.6**という高いパーセンタイルを示し、これはGPT-4の23.6を大幅に上回っています [14]。また**Aider**ベンチマーク（多言語コードアシスタント評価）でも71.6%の精度を達成し、Claude 4（Anthropic社の最新モデル）やDeepSeek-R1を僅かに上回りました [15] [3]。こうした結果から、**実用的なコーディング・デバッグ能力**においてV3.1はClaudeやGPT-4と比肩し、一部では優位性を示しています。特にアルゴリズム問題や多言語コード生成など**開発者ワークフロー直結の課題**で強みが顕著です [16]。

高いコード性能を支える要因として、前述のMTPによる効率的生成に加え、**大規模コードデータの継続学習**や、以前は分離していた「DeepSeek-Coder」モデル系列の知見統合が挙げられます [17]。実際、2024年にはDeepSeekはチャットモデルとコーダーモデルを統合してV2.5をリリースし、HumanEvalで89%という当時最高クラスのスコアを達成していました [18]。V3.1ではその路線を押し進めつつ、**フロントエンド開発**に特化した最適化も行われています。公式チームによれば、V3.1は「より美麗で実行可能なコード」を生成するようチューニングされており、WebページやゲームのUIコード出力の品質が向上しています [19]。総じて、

DeepSeek-V3.1はプログラミングタスクで一貫して高水準の結果を出しており、開発支援AIとして非常に評価が高いモデルとなっています [20] [21]。

ハイブリッド推論モードとエージェント機能

DeepSeek-V3.1の最大の特徴の一つが、**ハイブリッド推論アーキテクチャ**の導入です。単一モデル内に**"Non-Think"（通常チャット）モード**と**"Think"（深掘り推論）モード**の両方を持ち、プロンプトの形式を変えるだけで切り替えられます。例えばユーザが難問を入力した場合、V3.1は自動的に深い推論が必要と判断して内部でThinkモードを発動し、チェーン・オブ・ソート（思考の連鎖）による段階的推論を行います [22]。これにより、従来は手動で推論専用モデルに切り替えていた手間が省かれ、**常時推論モードオンの場合より40%応答時間を短縮**しつつ複雑タスクでの精度を維持できたと報告されています [23]。実際、公式チャットUIでも「DeepThinkボタン」でThinkモードをワンタッチ切替でき、ユーザ体験が向上しました。このHybridモードの実現には、新たに特殊トークン`<think>`や`<|search_begin|>`が導入されており、モデル内部で**推論プロセスの開始**や**外部ツール検索**を明示する仕組みになっています [24]。これらのトークンにより、モデルは必要に応じ自発的にインターネット検索や計算などの**ツール使用**へ移行でき、回答精度を高めるエージェント的振る舞いが可能です [24]。

ツール使用や複雑なマルチステップ課題への対応力も、V3.1では大幅に強化されました。開発チームの**ポストトレーニング最適化**により、ツールAPIの呼び出しフォーマット遵守や連続対話型のエージェントタスク処理が改善され、例えば**コードエージェント評価**のSWE-Benchで従来比+20%以上の正答率向上が達成されています [25] [26]。具体的には、SWE-bench Verified（コード問題の自動解決率）で**66.0**、マルチリンガル版で**54.5**、ターミナル操作ベンチマークで**31.3**と、いずれも前モデルを大きく上回るスコアを記録しました [25]。これはDeepSeek-R1（2025年5月版）の各スコア（例えばVerified 44.6）と比べても飛躍的な改善です [27]。加えて、**外部ブラウザ検索付きの難問**に挑む内部評価「BrowseComp」でも、V3.1-Thinkは英語版で30.0、中文版で49.2の達成率を示し、R1（それぞれ8.9、35.7）から大幅に向上しています [28]。これらの結果は、DeepSeek-V3.1が**ツール接続型エージェント**としても非常に高い能力を備えており、複雑な情報検索・推論タスクで従来モデルを凌駕することを示しています。

使用者の評価・評判

DeepSeek-V3.1の公開直後、研究者や開発者コミュニティからは概ね好意的な反応が寄せられました。モデルのオープンソース公開により、Hugging Face上では即座に数百の「いいね」が付き、ダウンロード数も急増しています [29]。Twitter(X)上でも「DeepSeek v3.1は既に絶賛のレビューを受けている」と紹介され [30]、実際リリース当日にコミュニティ主導のベンチマークサイト（ChatLLMやLivebenchなど）へ実装されるなど注目の高さがうかがえます。オープンソースAIの支持者からは「GPT-4やClaudeに匹敵する性能を**数分の一のコスト**で提供する画期的モデル」と評価されており [31] [11]、特に**開発者向け**には「コード生成が鋭敏で実務的」「月額費用を劇的に抑えられる」と好評です [20] [32]。事実、あるガイドでは「DeepSeek V3.1は最強のオープンソースモデルであり、ChatGPT代替として生産環境に投入できる水準だ」とまで謳われています [33] [34]。無料で使える**公開デモやAPIの太っ腹なフリーティア**も提供されており、一般ユーザーからも「試しやすい」「他社サービス障害時のバックアップとして頼もしい」といった声が上がっています [35] [36]。

一方で、一部ユーザーからは**懸念や批判的なフィードバック**も見られます。例えばRedditのコミュニティでは、「V3.1は**クリエイティブな文章生成が苦手**で、前版(V3-0324)より会話のニュアンス理解が劣る」との指摘がありました [37]。この投稿者はV3.1の出力を「ありきたりなクリシェが増えた」と感じており、創作的応答に関しては旧モデルの方が優れていたという評価です。一方、中国のネット小説愛好家コミュニティからは逆の報告が出ており、「**中国語小説の文体生成はR1より改善している**」「創作用途でもR1を超えた出来」といった声もあります [38]。実際、DeepSeekチームはV3以降で中国語文章のスタイル向上にも注力したと述べており [39]、その効果を評価するユーザーもいるようです。総合すると、**技術系タスク（コード・数学）への評価は非常に高く**、一部に「推論力はR1と大差ない」との意見があるものの [40]、多くのユーザーが性能向

上を実感しています。また「出力が冗長すぎる」との声もありましたが [41]、これはThinkモード時にチェーン・オブ・ソートが可視化される仕様上、推論過程が文章に表れるためと考えられます。もっともUI上の切替表示の問題であり、**回答自体の有用性には大きな影響はない**と見る向きが一般的です。

モデルの**安定性**についてもおおむね良好な評価です。大型MoEモデルで懸念される暴走や未収束はV3.1では報告されておらず、「学習過程で致命的なLossスパイクもロールバックも一度も無かった」と技術報告で述べられています [42]。そのため、**研究用途や商用検証**に安心して使えるとの声もあります [31]。ただ、一部には「R2（次世代推論モデル）の開発停滞をV3.1で穴埋めしているだけでは」との業界観測もありました [43] [44]。実際、DeepSeek社が期待されていたR2モデルをまだリリースしていない点については議論があり、V3.1公開時にチャットUIからR1表記が消えたことも相まって「路線転換か？」との憶測を呼んでいます [45] [46]。しかしユーザーから見れば、まずは**現行モデルとして使いやすさと性能を両立したV3.1**への評価が優先しており、「オープンソース界隈でClaudeやGPT-4に迫る選択肢が出てきた」と歓迎する声が大勢を占めています [31] [47]。

他の主要なLLMとの比較

GPT-4との比較（OpenAI）

DeepSeek-V3.1は度々**「ChatGPT（GPT-4）の対抗馬」**として言及されますが、その比較ポイントは**性能面**と**利用面**の両方にわたります。性能に関しては、**MMLUやHumanEval、数学問題**といった定量ベンチマークではV3.1がGPT-4に匹敵、あるいは凌駕する結果を示しています。前述の通りHumanEvalではV3.1が82.6%でGPT-4の80.5%を上回り [11]、MMLUでも87.5% vs 86.4%と僅差ながらV3.1が上回る項目が見られます [3]。一方で**数学競技系**（例えばMATHデータセット）ではGPT-4が若干高精度（76.6% vs 74.2%）という評価もあり [3]、純粋な数学問題ではGPT-4の底力がわずかに勝るケースもあります。ただ総じて、DeepSeek-V3.1は**従来「GPT-4の牙城」と思われたコード生成や難問推論の領域で肩を並べる**までになっており [12] [3]、これはオープンモデルとして驚異的な成果といえます。加えて、GPT-4が不得意とされる**長文文脈**（標準では8Kトークン、拡張版でも32K）の処理において、DeepSeekは標準128Kという圧倒的な強みを持ちます [7]。例えば小説全文や大規模コードベースの解析では、GPT-4では複数回の分割が必要なところをDeepSeek-V3.1は一度にこなせるため、**長文タスクの使い勝手**はDeepSeekが上回ります。

利用体験・コスト面では、**DeepSeekはオープンソースかつ低コスト**、GPT-4はクローズドかつ高コストという対照的な差があります。DeepSeek-V3.1のAPI利用料金は入力1Mトークンあたり約$0.07〜$0.27、出力でも$1.10/M程度とされ、GPT-4 API（入力$2.50、出力$10 per M tokens）の**約1/9以下**に抑えられています [48]。ある比較記事では「DeepSeekはGPT-4の同等性能を**90%安いコスト**で提供する」とされ [49] [50]、費用対効果の面でDeepSeekを選ぶ企業も出始めています。さらにDeepSeekは**モデル重みがMITライセンスで公開**されており、自社サーバでのホスティングやモデル改変が可能です [51]。対するGPT-4はOpenAIのクラウド経由でしか利用できず、サービス規約や使用制限も厳格です。機能面では、GPT-4は**画像入力などマルチモーダル対応**やプラグイン連携が可能という利点がありますが [52]、DeepSeekも**関数呼び出し（Function Calling）API**やAnthropic形式のAPIリクエストに対応し（互換APIを提供）、企業システムへの統合を容易にしています。総合すると、**GPT-4は汎用性と完成度、マルチモーダル対応で勝り**、**DeepSeek-V3.1はコスト効率や長文・専門タスク適性で優れる**と言えます。実運用では両者を組み合わせ、コストのかかる処理をDeepSeekに任せつつGPT-4をクリエイティブ用途に使うといったハイブリッド活用も提案されています [53]。

Claude 2/3との比較（Anthropic）

Anthropic社のClaude（特にClaude 2や最新版Claude 4 Opusなど）は、**長大な連続対話**や**安全性**に定評のあるLLMです。DeepSeek-V3.1との性能比較では、**総合的な推論力や知識問題**でほぼ同等、場合によってClaudeが僅かに上、といった評価が見られます。例えばMMLUにおいてClaude 3.5 (Sonnet)は88.2%、DeepSeek-V3.1は87.5%とごく僅かにClaude側が高く [3]、逆にコード生成系ではDeepSeekがClaudeを上

回るケースが多いです。前述の通り、HumanEvalはClaude 3.5が79.2%に対しDeepSeek 82.6% [3]、CodeforcesもClaude ~31相当に対しDeepSeek 51.6と大差でした [3]。AnthropicのClaude 2以降は**100K以上の長文コンテキスト**と**拡張推論モード**を持ち、DeepSeekも128Kコンテキストで匹敵しますが、最新版Claude 4では**20万トークン超**に拡大しており、この点ではややClaudeが先行しています [54] [55]。もっともDeepSeekも128Kで実用上不足は少なく、両者とも**超長文での安定した応答**が可能です。

利用面では、Anthropic ClaudeはAPI提供や商用利用でOpenAIに近い立ち位置ですが、**料金はGPT-4より割安**に設定されており、DeepSeekとの差も小さくなります。とはいえDeepSeekのオープンソース無料提供というインパクトには及ばず、「DeepSeekはClaude等に対し90%のコスト削減を実現したディスラプターだ」と評する向きもあります [56]。機能比較では、Claudeは現時点で**画像や音声入力には非対応**で、モダリティはテキスト中心でDeepSeekと同様です。ただしClaudeは人権や倫理に配慮した**ハラスメント対策や政治的中立**など安全面の調整が特徴であり、**出力の一貫性・無害性**では定評があります。一方、DeepSeekはオープンで調整可能な分、使い手のプロンプト設定によっては回答がぶれる可能性も指摘されています（もっともR1世代以降、DeepSeekも安全性や一貫性の向上に努めています）。またAnthropicはClaudeの**拡張思考モード**（Extended Thinking Mode）を搭載しており、非常に長い推論チェーンもこなすとされます [57]。DeepSeek-V3.1も同様のチェーン推論を自動化していますが、ユーザーからは「R1と比べて劇的には賢くなっていない」との指摘もあり [40]、**超高度な常識推論や倫理判断**ではClaudeシリーズに一日の長がある可能性があります。もっとも、コード処理や数学的問題解決といった**実務領域ではDeepSeekが優位**との声が強く [16]、用途によって評価が分かれるところです。総じて、Claude 2/3は「安定志向の大規模対話AI」、DeepSeek-V3.1は「開発者フレンドリーでコスパに優れた新鋭モデル」として棲み分けが進んでいます。

Gemini 1.5との比較（Google DeepMind）

GeminiはGoogle DeepMindが開発する次世代LLMで、マルチモーダル対応や超長文コンテキストを武器としています。2025年8月時点でGeminiの最新版は「Gemini 2.5 Pro」とされ、ここでは質問に挙げられた**Gemini 1.5**（おそらく初期のGeminiシリーズ）との比較を行います。Gemini 1.5世代は既に**テキスト・画像・音声を統合**した入出力が可能で、DeepSeek-V3.1がテキスト専用であるのに比べ**モダリティの幅広さ**で優れています [58] [59]。またGeminiはGoogle検索や各種クラウドサービスとネイティブに連携でき、リアルタイム情報へのアクセスやGmail/Docsとの統合など**エコシステム統合**が強みです [58]。一方、DeepSeekは現時点でテキスト応答に特化していますが、その分**専門領域の深掘り**に注力しており、特に数学的推論では「**State-of-artの推論性能（MATHで97.3%正解）**」と評価されています [49]。Geminiも総合知能は非常に高く、2025年時点の業界ベンチマークで**コーディング能力や汎用知識**においてトップクラスとされていますが [60]、DeepSeekのV3/R1モデルも**ほぼ同等の性能を低コストで実現**している点が注目されています [56]。実際、ある企業向け比較では「Geminiは包括的機能と安定基盤、DeepSeekは90%低コストで高度推論能力を提供」とまとめられています [56]。

コンテキスト長では、Gemini 1.5が既に数十万〜100万トークン級（Pro版で最大100万）のウィンドウを持つのに対し、DeepSeek-V3.1は128Kと**相対的に短め**です [61] [62]。そのため超長文ドキュメントやマルチメディア入力が絡むタスクではGeminiが適しています。逆に、DeepSeekは**オープンソース（MITライセンス）**であることから、研究者や開発者が内部挙動を分析・改良しやすく、コミュニティ主導での拡張が可能です [63] [64]。GeminiはGoogle独自モデルでブラックボックス的要素が多いため、組み込みやチューニングの自由度ではDeepSeekが勝ります。また**コスト面**では、GeminiのAPIは従量課金制ながらGoogleの高性能インフラ上で提供されるため安価とは言えず、DeepSeekは**オフピーク時75%割引**など攻めた価格設定で大幅なコスト優位を打ち出しています [65] [66]。現にDeepSeekは「競合比で90%のコスト削減」を掲げており [49]、予算重視のプロジェクトには有力な選択肢です。

まとめると、**Geminiは企業向けの信頼性や多機能性**（マルチモーダル＆Google連携）が強みであり、**DeepSeek-V3.1はコスト効率と高度なテキスト推論能力**で勝負するモデルです [56]。大規模インフラやマルチモーダル処理を必要とする場合はGemini、純粋なテキスト知能を安価に活用したい場合はDeepSeekと、用途に応じて使い分けるのが望ましいでしょう [67]。Gemini 2.x世代ではさらに性能向上が報じられています

が、DeepSeekもV3およびR1シリーズで閉源モデルに肉薄する成果を上げており、それをオープンかつ低コストで提供している点でAIエコシステムにおける存在感を高めています [68] [60] 。

---

[1] [13] [18] [25] [26] [39] Change Log | DeepSeek API Docs
https://api-docs.deepseek.com/updates

[2] [27] [28] deepseek-ai/DeepSeek-V3.1 · Hugging Face
https://huggingface.co/deepseek-ai/DeepSeek-V3.1

[3] [7] [11] [12] [14] [15] [16] [19] [22] [23] [32] [33] [34] [35] [36] DeepSeek V3.1 API Key Complete Guide: Access the Most Powerful Open-Source Model (August 2025) - Cursor IDE 博客
https://www.cursor-ide.com/blog/deepseek-v3-api-key-guide-2025

[4] [54] [55] [57] [61] AI by AI: Top 5 Large Language Models (July 2025) – Champaign Magazine
https://champaignmagazine.com/2025/07/11/ai-by-ai-top-5-large-language-models-july-2025/

[5] [42] arxiv.org
https://arxiv.org/pdf/2412.19437

[6] [43] [44] [45] [46] DeepSeek’s V3.1 update and missing R1 label spark speculation over fate of R2 AI model | South China Morning Post
https://www.scmp.com/tech/big-tech/article/3322481/deepseeks-v31-update-and-missing-r1-label-spark-speculation-over-fate-r2-ai-model

[8] [9] [10] [17] [20] [21] [24] [29] [31] [40] [47] [51] DeepSeek V3.1 Base : The ChatGPT killer is back | by Mehul Gupta | Data Science in Your Pocket | Aug, 2025 | Medium
https://medium.com/data-science-in-your-pocket/deepseek-v3-1-base-the-chatgpt-killer-is-back-1c0f05530677

[30] Bindu Reddy on X: "Deepseek v3.1 is out and is already getting rave ...
https://twitter.com/bindureddy/status/1957982737325043756

[37] [38] Deepseek V3.1 is bad at creative writing, way worse than 0324 : r/LocalLLaMA
https://www.reddit.com/r/LocalLLaMA/comments/1mv7zdl/deepseek_v31_is_bad_at_creative_writing_way_worse/

[41] DeepSeek v3.1 : r/LocalLLaMA
https://www.reddit.com/r/LocalLLaMA/comments/1muft1w/deepseek_v31/

[48] [52] [53] DeepSeek V3 vs GPT-4o: Which is Better?
https://www.analyticsvidhya.com/blog/2024/12/gpt-4o-vs-deepseek-v3/

[49] [50] [56] [58] [59] [60] [62] [63] [64] [65] [66] [67] [68] Google Gemini vs DeepSeek 2025: Complete LLM Comparison | Performance, Pricing, Features
https://aloa.co/ai/comparisons/llm-comparison/gemini-vs-deepseek