2025 年 6 月 24 日　全 6 頁

DeepSeek ショックから半年、米国優位は続くか

オープン型 AI モデルがもたらした変化

経済調査部 主任研究員　田邉 美穂

[要約]

- 2025 年初頭の「DeepSeek ショック」以降、生成 AI モデルの開発は加速し、オープン型（技術情報が公開されている）AI モデルの性能向上と軽量化および低コスト化が進展。限られた計算資源でも高い推論能力等を発揮するモデルが複数登場し、従来クローズ型（技術情報が非公開）AI モデルが保持していた技術的優位性は急速に揺らぎつつある。

- 軽量版モデルの性能向上により、ローカル運用や商用利用が現実的となり、企業の導入選択肢が拡大しつつある。これに対抗し、クローズ型 AI モデル各社も価格引き下げや軽量版モデルの投入を進め、競争は激化している。米国企業は性能やクラウド連携の面等で優位を維持しているが、今後は価格や柔軟性、透明性も競争力の鍵となり、従来の優位性だけでは安泰とは言えない状況になりつつある。

- また、高性能な計算資源である GPU の需要は引き続き拡大している。軽量モデルの普及により、企業が自社環境やデータ処理を端末内で行うことが可能なエッジデバイスで AI を運用するケースも増え、GPU の需要は各所に広がっている。さらに生成 AI の進化に伴い、GPU に求められる能力も多様化している。GPU の需要は数量だけでなく用途や性能面でも広がりを見せており、今後も減少する可能性は低いとみられる。

- 今後は、企業による生成 AI の導大拡大が見込まれることから、各国では利用ポリシーや法制度の整備が急がれる。さらに今後は、フィジカル AI 等の「行動する AI」の普及も注目される。生成 AI は社会や産業の基盤を支える技術として新たなフェーズに突入しつつあり、新興企業が既存の技術覇権を揺るがす展開も十分に想定される。

1. はじめに

2025 年 1 月に中国のスタートアップ企業 DeepSeek が新 AI モデル、DeepSeek-R1(以下、R1 モデル)を発表したことを契機に、NVIDIA をはじめとする AI 関連企業の株価が大幅下落するという出来事が発生した。いわゆる「DeepSeek ショック」と呼ばれるこの出来事から、早くも半年が経過しようとしている。

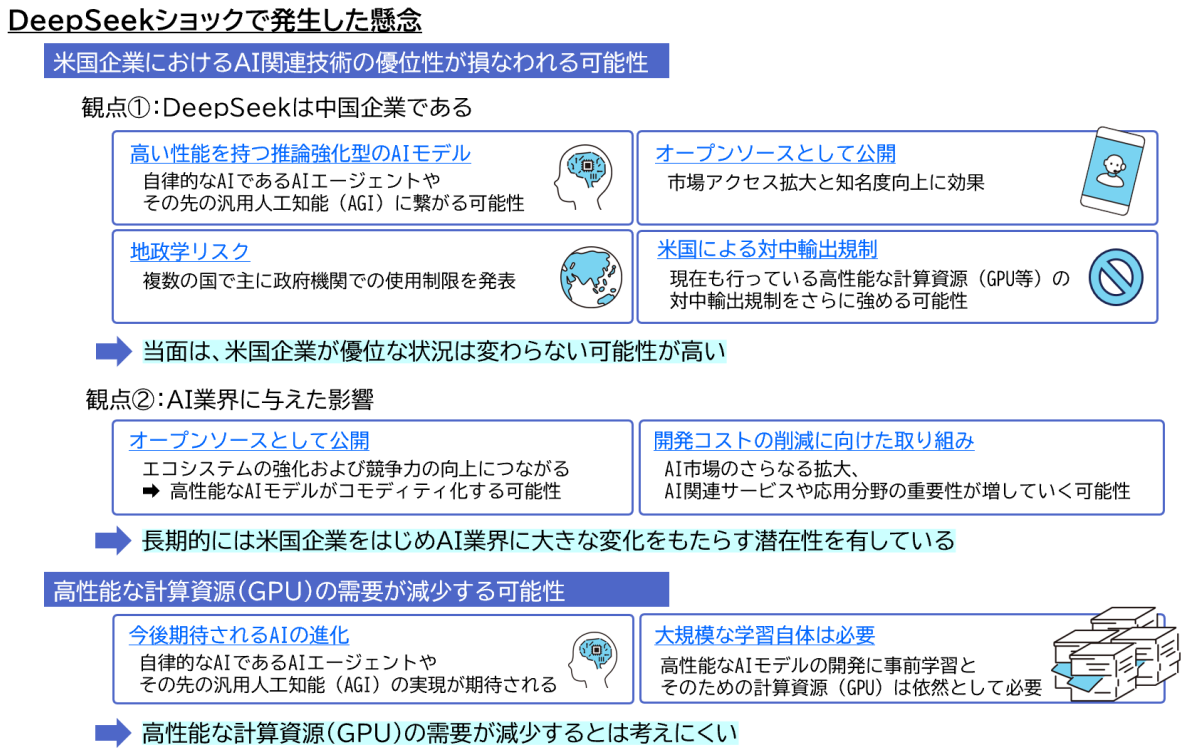
筆者が当時公表したレポート[1]では、世間一般が抱く DeepSeek ショックにまつわる 2 つの懸念点、①米国企業における AI 関連技術の優位性が損なわれる可能性、②高性能な計算資源（GPU）の需要が減少する可能性について整理を行った（**図表 1**）。

1 つ目の米国企業における AI 関連技術の優位性が損なわれる可能性については、DeepSeek が中国企業であることから、短期的には米国企業が優位な状況は変わらないとした。その一方で、オープンソースとして公開されたことにより、高性能な AI モデルがコモディティ化していくことも予想され、長期的には米国企業をはじめ AI 業界に大きな変化をもたらす潜在性を有していると分析した。

2 つ目の高性能な計算資源（GPU）の需要が減少する可能性については、今後期待される AI の進化や、DeepSeek の新モデルにおいても事前学習として大規模な学習を要していたことから、今後も需要が減少するとは考えにくいと分析した。

半年が経過した現在、これらの懸念点はどのように変化したのか。本レポートでは、AI 業界の現状を評価するとともに、今後の方向性について分析を行う。

図表 1 前回のレポートにおける議論内容



（注）地政学リスクは、モデル自体をダウンロードし利用環境を整えることで、中国へのデータ流出のリスクを抑えることは可能
（出所）大和総研作成（イラストはソコスト（https://soco-st.com/））

[1] 田邉美穂[2025]「DeepSeek は何が衝撃的なのか：低コストで開発された高性能な AI モデルが世界にもたらす影響」大和総研レポート（2025 年 2 月 26 日）

2. 生成 AI モデルのオープン化の加速と軽量化・低コスト化の進展

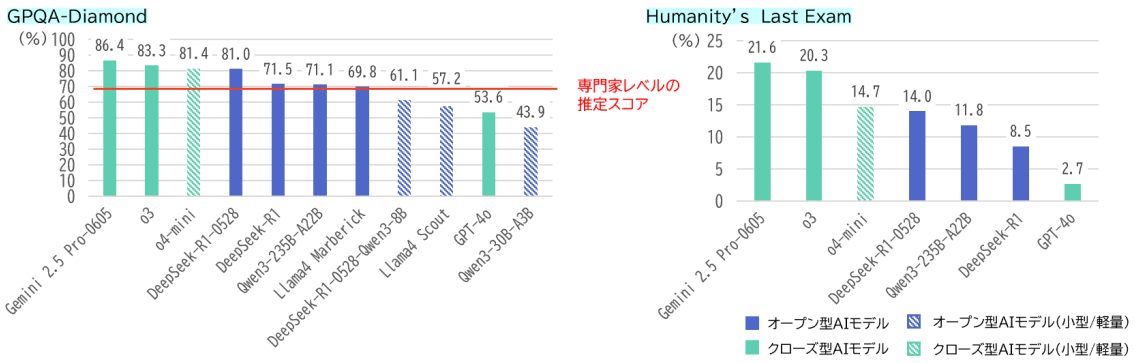
米国企業における AI 関連技術の優位性が損なわれる可能性を検討するにあたり、まずは DeepSeek ショック以降の生成 AI モデル開発の動向を整理する（**図表 2 上・下**）。

図表 2 DeepSeek ショック後の主なオープン型 AI モデル（上）、ベンチマークの比較（下）

主なオープンソースモデル

	DeepSeek(中国)	Alibaba(中国)	Meta(米国)
モデル名	DeepSeek-R1-0528	Qwen3-235B-A22B	Llama4 Maverick
公開時期	2025年5月	2025年4月	2025年4月
ライセンス	MITライセンス 商用利用、改変、再配布等が可能	Apache License 2.0 商用利用、改変、再配布等が可能	Metaライセンス 商用利用も可能(Metaの規約に準拠)
特徴	・推論能力が向上 ・MoEアーキテクチャを採用 ・蒸留による既存小型モデルの高性能化 (DeepSeek-R1-0528-Qwen3-8B等)	・ThinkingモードとNon-Thinkingモードの2つの動作モードをもつ ・MoEアーキテクチャを採用 ・軽量で効率的なモデルもリリース (Qwen3-30B-A3B)	・推論や画像理解など高性能・多機能 ・MoEアーキテクチャを採用 ・軽量で高速なモデルもリリース (Llama4 Scout)

推論能力等の評価に用いられるベンチマークの一例



（注）ベンチマークは脚注 2 を参照
（出所）各種資料より、大和総研作成

生成 AI モデルのオープン化の加速

最近のオープン型 AI（技術情報が公開されている）モデルを見ると、Mixture of Experts（以下、MoE）アーキテクチャと呼ばれる計算資源を効率的に使用する仕組みを利用し、高い推論能力を持つモデルが次々と公開されている。一般に、生成 AI モデルの性能評価は、さまざまな観点から複数のベンチマークを用いて行うため、単一の指標だけで優劣を判断することはできない。しかし、推論能力等を測るベンチマーク[2]の一例である GPQA-Diamond を見ると、OpenAI の

[2] **図表 2 下**左の GPQA（A Graduate-Level Google-Proof Q&A Benchmark）は、理系分野を中心に高度な専門知識や推論力（博士課程レベル）等を測ることが出来るベンチマークであり、Diamond はその中でも特に問題の難易度が高いものを指す。**図表 2 下**右の Humanity's Last Exam は、人間の専門知識のフロンティアで AI 知識の限界をテストするために設計されたベンチマークであり、こちらの指標ではまだ差があることがわかる。
Rein, D., B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman[2023] "GPQA: A Graduate-Level Google-Proof Q&A Benchmark" https://arxiv.org/abs/2311.12022
Phan, L., A. Gatti, Z. Han, N. Li, J. Hu, H. Zhang, C. B. C. Zhang, M. Shaaban, J. Ling, S. Shi, M. Choi, A. Agrawal, A. Chopra, A. Khoja, R. Kim, R. Ren, J. Hausenloy, O. Zhang, M. Mazeika, S. Yue, A. Wang, and D. Hendrycks[2025] "Humanity's Last Exam" https://arxiv.org/abs/2501.14249

「o3」やGoogleの「Gemini2.5 Pro」といった最新のクローズ型（技術情報が非公開）AIモデルが依然として先行しているものの、それらのスコアに迫るオープン型AIモデルも現れ始めている。

商業利用可能な軽量版モデルがビジネス活用の新たな選択肢に

これらのオープン型AIモデルは小型/軽量版モデルも同時に公開されており、大規模モデルと比較すると性能は劣るものの、推論コストの削減や応答速度の高速化が期待できる。また、オープン型AIモデルは、無料または低価格で利用可能な場合が多く、かつ商用利用も可能[3]なことから、ビジネス用途での活用も期待される。

導入においては、環境構築や運用に一定の専門知識が前提となるが、モデルサイズによっては、クラウドに依存せずローカル環境[4]での実行も可能だ。また、ローカル環境や国内のクラウドサービス等、最適な利用環境を選択することで、ビジネス活用で課題となる地政学リスクの回避にもつながる[5]。

クローズ型AIモデルも低価格へシフト

低コストで利用可能なオープン型AIモデルの進化を受け、クローズ型AIモデル各社においても価格戦略の見直しが進められている。例えば、Googleが2025年3月に大規模モデル「Gemini2.5 Pro」を発表し、同年4月にはその軽量版かつ低価格モデルとして「Gemini2.5 Flash」を発表。コスト効率と応答速度を両立したモデルとして話題となった。OpenAIは、同年4月に低価格モデル「GPT-4.1」を発表した一方で、同年2月に登場したばかりの高価格モデル「GPT-4.5」の提供を、同年7月をもって終了すると発表した。また直近では、「o3」の進化版として「o3-Pro」を発表すると同時に、「o3」の価格を従来比で80%引き下げる[6]と発表し、価格競争の激化を印象づけた。

性能面では米国企業が優位を保つも油断は許されない状況に

ここまでの議論を踏まえ、1つ目の懸念点である米国企業におけるAI関連技術の優位性が今後も維持されるのかを考察する。性能面では、OpenAIやGoogleといったクローズ型AIモデルを提供する米国企業が依然として優位を保っているものの、中国企業に代表されるオープン型AIモデルの進化は著しく、その差は急速に埋まりつつある。また、少ない計算資源かつ低コス

[3] モデルによっては、利用にあたり条件が設定されている場合もある。
[4] ただし、完全なローカル実行には一定の計算資源が必要であり、高性能なGPUを搭載したPCやワークステーション、あるいはデータ処理を端末内で行えるエッジデバイス向けに最適化された環境等が求められる。
[5] ここで言う地政学リスクとは、特定国のクラウドサービスやAPIに依存することで他国に情報が漏洩する等の国家安全保障に関するリスクを指す。モデルに内在する文化的・言語的バイアス等のリスクは含まれていない。
[6] OpenAIが提供するAPIサービスの価格を指す。OpenAI「API料金」より

トで利用可能な軽量型モデルにおいても一定の性能が担保されるようになったことで、企業が生成 AI を導入する際の選択肢が広がってきている。

クローズ型 AI モデルは、性能面だけではなく、Microsoft や Google といった大手クラウド基盤との連携により、導入の容易さやサポート体制の充実といった面で依然として強みを有している。さらに利用コストも低価格化にシフトしたことで、導入検討の場で優位に映る場面も多い。しかし、地政学的リスクやデータ規制への対応といった観点から、ローカル環境での運用が可能なオープン型 AI モデルへの関心も高まりつつある。

生成 AI 関連サービスの市場は今後も急速に拡大していくと見込まれており、企業の判断次第では、より柔軟でコスト効率の高いオープン型 AI モデルへの移行が進む可能性もある。米国企業が現在の優位性を維持するためには、性能だけでなく、価格、柔軟性、透明性といった多面的な価値提供が求められ、従来の優位性だけでは安泰とは言えない状況になりつつある。

3. 高性能な計算資源（GPU）に対する需要の拡大と多様化

高性能な計算資源（GPU）の需要は、前回のレポートで示した見通し通り、現時点においても減少しておらず、むしろ拡大している。NVIDIA をはじめとする主要企業の売上高の推移を見ても GPU 市場は依然として成長を続けており[7]、その背景には複数の要因がある。

まず、Microsoft、Amazon、Google 等の大手クラウド事業者が、AI 処理を支える大規模なデータセンターへの投資を継続していることが挙げられる。加えて、先述の軽量版モデルの流れから、企業が自社環境で小型モデルを運用するケースも増えてきており、これに伴ってエッジデバイスや PC 向けにも GPU の需要は拡大している。GPU の需要がクラウドに集中していた状況から、各所に分散され始めていると言える。

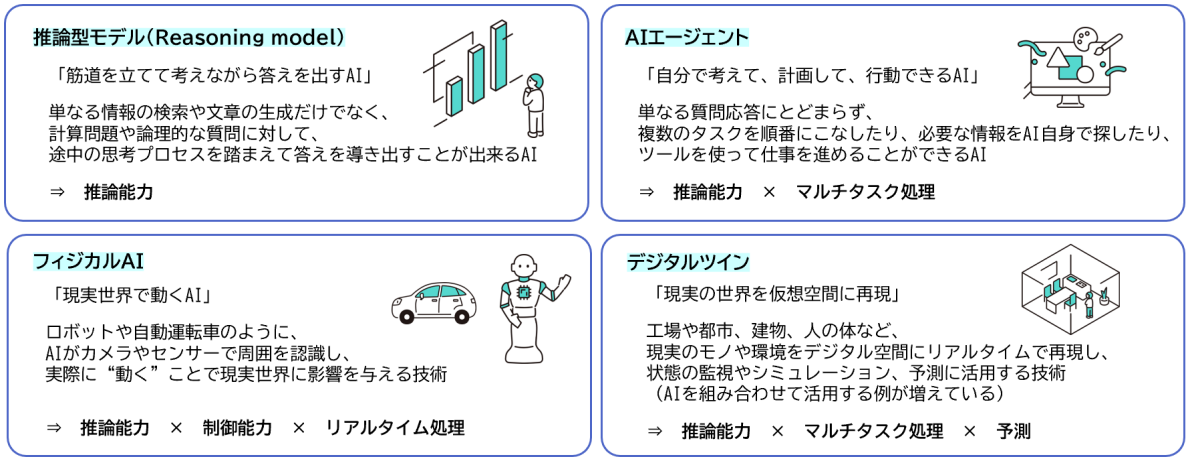
さらに、生成 AI モデルの進化に伴い、GPU に求められる能力も多様化している（**図表 3**）。OpenAI の「o3」や DeepSeek の「R1」等、現在主流となりつつある推論型モデル（Reasoning Model）は、より高度な推論能力を必要とするため、演算性能に優れた GPU が不可欠である。また、複数のタスクを自律的に計画・実行する高度な AI システムとして注目される AI エージェントにおいては、複雑な推論処理や複数のタスクを処理する能力が求められる。

また、今後を見据えると、フィジカル AI やデジタルツインといった現実世界とデジタル世界の融合を体現する技術が新たに注目されている。これらの技術の実現には、推論処理のみならず、リアルタイム制御や 3D シミュレーション等のより複雑で多様な処理にも対応できる GPU が求められる。このように、GPU に対する需要は単なる数量的な拡大にとどまらず、用途や性能要件の面でも多様化しており、生成 AI の進化とともに新たな技術領域への展開が進む中で、今後

[7] 日経クロステック「AI 半導体メーカーが躍進、2024 年の半導体売上高ランキング」（2025 年 2 月 12 日）、日経クロステック「NVIDIA、時価総額に加えて売上高でも世界 1 位の半導体メーカーに」（2025 年 4 月 15 日）上記記事によると、2024 年の半導体市場全体（AI 半導体以外も含む）の売上高は前年比 21.0%増だったとし、2025 年は前年比 13.7%増と予測し、AI 半導体の影響は大きいとしている。

もその需要が減少する可能性は低いと考えられる。

図表 3 GPU が必要とされる技術の例



（出所）大和総研作成（イラストはソコスト（https://soco-st.com/））

4. 生成 AI の進化による社会の変化と米国企業の立ち位置

前回のレポートにおいて、DeepSeek の与えた衝撃は、長期的には米国企業をはじめ AI 業界に大きな変化をもたらす潜在性を有していると指摘したが、この半年間でその兆候は一層明確になってきた。特に、オープン型 AI モデルの性能向上と軽量化の進展は、従来クローズ型モデルが保持していた技術的優位性を急速に侵食しつつある。性能面では米国企業が依然として先行しているものの、その差は着実に縮小している。

利用者にとっては、選択肢の拡大と利用コストの低下により、生成 AI の導入のハードルは下がってきている。企業における導入は、今後さらに加速する可能性が高い。一方で、企業における生成 AI の本格導入が現実味を帯びてきた今だからこそ、同技術は単なる技術革新ではなく、企業戦略や国家安全保障に直結する「戦略的資産」としての位置づけを強めている。こうした状況を踏まえると、利用形態も含めた各国の利用ポリシーや法規制の整備は急務であり、クローズ型モデルにおいても、これらの要求を満たす柔軟性や透明性の提供が不可欠となる。

さらに今後は、推論型モデルや AI エージェント、フィジカル AI といった「行動する AI」の普及が進むとみられ、それを支える GPU やインフラの設計思想も変化していくと考えられる。生成 AI は単なるツールから、社会や産業の構造を変える「基盤技術」へと新たなフェーズに突入しつつあり、新興企業が既存の技術覇権を揺るがす展開も十分に想定される。

以上