

TRAINING

生成AI 最新トレンド

2026年9月時点

毎日更新。表紙の日付は最終更新日です

講師 安田 光喜（Givery）



収録ニュース一覧 1/3

先頭の5件は**定番枠**です。

No	ニュース	公開日	提供社
1	東大二次試験でAIが理三合格水準を突破	2026年3月5日	東進（ナガセ）
2	医師国家試験でAIが受験生上位の得点	2026年2月7日	メディックメディア
3	世界一決定戦でAIが人間に勝利	2026年7月9日	AtCoder／OpenAI
4	モデル内部に感情に対応する171の概念	2026年4月2日	Anthropic
5	自分の内部状態に気づく能力を確認	2025年10月29日	Anthropic
6	90年の未解決問題にAIが証明を提示	2026年9月8日	OpenAI
7	GPT-6 AstraがWebDev評価首位	2026年9月7日	OpenAI
8	Astra のPC操作、実演と使いどころ	2026年9月7日	OpenAI ほか

収録ニュース一覧 2/3

No	ニュース	公開日	提供社
9	コード未経験者のAIゲームが社内大賞	2026年9月5日	社内AIゲームコンテスト2026
10	Claude Code自社実行環境、公開ベータへ	2026年9月4日	Anthropic
11	初のCritical判定、AIが自律ゼロデイ探索	2026年9月3日	OpenAI
12	次世代モデル GPT-6 Astra を公開	2026年9月3日	OpenAI
13	ChatGPTが電子カルテと直接連携	2026年9月3日	OpenAI
14	ハンドルのない無人タクシーが商用開始	2026年9月3日	Tesla
15	Qwen3.8-Max、PCエージェント首位奪取	2026年9月2日	Alibaba（阿里巴巴）
16	Claude Fable 5.1 と Mythos 5.1 を公開	2026年9月1日	Anthropic

収録ニュース一覧 3/3

No	ニュース	公開日	提供社
17	SpaceX、Cursor買収完了（6兆円超）	2026年8月15日	SpaceX / Anysphere
18	ゲーム未経験・AI全自動2週間の受賞作	2026年7月14日	iOS個人開発者 / Cursor

定番枠の次からは公開日の新しい順です。No は本文の「NEWS xx」と対応します。

本資料の二次転載禁止について

本資料は株式会社ギブリーが作成した研修教材で、著作権はギブリーに帰属します。

お願い

社内での受講と復習を目的とした閲覧・印刷は問題ありません。社外への配布、SNSやブログへの掲載、他社研修への流用はご遠慮ください。引用が必要な場合は担当者までご連絡ください。

統計と方針は公開情報にもとづき、出典は各ページに記載しています。

この資料の見方

数字はすべて公表値です。

◆ すべて2026年7月時点の断面です

モデルの順位もベンチマークの構成も数か月で入れ替わります。
第4章の総合指標の表は継続更新されるので、投影日の数字は違う前提で見てください。

◆ 数値には基準日と出典を添えています

同じ指標でも測った日と足場が違えば値が変わります。SWE-bench Verifiedは公式ハーネスで76.80%（2026年2月17日）ですが、ベンダー自己申告では90%台が出回っています。

◆ 報道が食い違う数値は断定しません

Claude Design公開当日のFigma株の下落率は、媒体により6%から7.7%まで幅があります。
この資料では幅のまま出すか、載せません。

幅で書く理由

食い違う数字を1つに丸めると、後で反証されたときに資料全体が疑われます。

全体の流れ

第2章に**ニュース18件**、第3章以降は選定の判断材料です。

章	テーマ	持ち帰るもの
第1章	2026年9月の現在地	単価で選ぶ段階に入ったという事実
第2章	ニュース（18件・新しい順）	1件ごとに見出し・ビジュアル・概要・詳細
第3章	モデル性能比較	用途別の使い分けの基準
第4章	ベンチマークの世代交代	自社で評価を作る必要が出た理由
第5章	勢力図と国家	計算資源と規制が選定に効く経路

ニュースの読み方

各ニュースは上端のティール帯と「NEWS xx」のチップから始まります。1枚目に見出しとビジュアル、2枚目以降に概要と詳細が続きます。

01

2026年9月の現在地

総合指標の上位3モデルは2ポイント差に収まりました。差がつくのは1タスクあたりのコストと、重みが公開されているかどうかです。

選定の基準は、性能から 1タスクいくらかへ移りました。

総合指標の上位はClaude Opus 5が61、Claude Fable 5が60、GPT-5.6 Solが59です。上位3モデルの差は2ポイントです。
同じ表の1タスク平均コストは最安\$0.04から最高\$2.75まで開き、その差は約69倍です。

出典 artificialanalysis.ai/models (2026年7月28日アクセス)

2026年前半の主要発表

この7か月で実務に効いたのは、**重みの公開日と規制の適用日**です。

日付	出来事	何が変わるか
1月16日	ValveがSteamのAI開示ルールを改定	開示義務の対象を絞る
2月19日	パナソニック コネクトがAIエージェント運用開始	図面と設計仕様の照合
4月17日	AnthropicがClaude Designを公開	同日にFigma株が6～7.7%下落
6月9日	Claude Fable 5とMythos 5が提供開始	総合指標の首位が交代
7月9日	AtCoder世界一決定戦でAIが人間に勝利	去年は人間が優勝していた
7月16日	Moonshot AIがKimi K3を発表	7月27日に重みを公開
7月22日	ホワイトハウスがGenesis Mission発表	連邦15機関超で50億ドル超
7月24日	AnthropicがClaude Opus 5を公開	総合指標61で首位
7月28日	減速を求める署名「Pacing the Frontier」	1,200人超が署名

現在地を示す3つの数字

同じ表の中で、性能差は2ポイント、コスト差は**69倍**です。

2 ポイント

総合指標の上位3モデルの差
(2026年7月28日時点)

4 か月

オープンウェイトがクローズドに
遅れる期間 (2026年5月28日時点)

69 倍

1タスク平均コストの上下の開き
(2026年7月28日時点)

内訳

上位3モデルはClaude Opus 5が61、Claude Fable 5が60、GPT-5.6 Sol (max) が59です。1タスク平均コストはDeepSeek V4 Proの\$0.04からClaude Fable 5の\$2.75まで開きます。オープンウェイトとの差はEpoch AIの計測で平均4か月、ECIで8ポイントです。

出典 artificialanalysis.ai/models、epoch.ai/data-insights

02

ニュース

先頭の5件は、いつ紹介しても通用する定番です。以降は公開日の新しい順に並べています。1件ごとに見出し・ビジュアル・概要の順に進みます。

入試当日にその場で解かせて3種とも8割超。最高の Claude は9割に迫ります。

【東進調査】2026年東大二次試験、最新AIが理三合格レベルを突破、9割に迫る。文系数学は全3種が満点！進化するAIの記述力と、見えてきた「図形・史料」の壁

2026.03.05



東進ハイスクール・東進衛星予備校などを運営する株式会社ナガセ(本社：東京都武蔵野市 代表取締役社長 永瀬昭幸)は、2026年2月25日・26日に実施された東京大学二次試験において、最新の生成 AI3種を用いた即日解答・検証調査を実施いたしました。その結果、全てのAIが文理ともに得点率8割以上を記録し、最難関とされる理科三類にも余裕で合格できる極めて高い水準に達していることが判明しました。東進では本結果を詳細に分析し、今後のAI教育コンテンツの開発に活用してまいります。

出典 toshin.com/recent/detail/RecentToshin.php?id=175

項目	内容
実施	2026年2月25日・26日の東京大学二次試験。東進が試験当日に即日解答・検証
使ったAI	Claude Opus 4.6、Gemini 3.1 pro、GPT-5.2 の3種
得点率	文系・理系とも3種すべてが8割以上。最高は Claude で9割に迫る
合格水準	最難関の理科三類にも余裕で合格できる水準
満点	文系数学は3種すべてが満点
解答にかかった時間	日本史・世界史は1～2分、数学は20分弱
残った弱点	図形は計算式まで到達しても図を描けない。史料の読み解きは要約に流れる

ここが効きます

検証に使われた3種は、いずれも社内で使える市販モデルです。一方で図形の作図と史料の読み解きは3種とも崩れました。解ける問題と解けない問題を分けたのは、難易度ではなく出題の形式でした。

一般臨床は Claude Opus 4.5 が98.3%、必修は Gemini 3 Pro が99.5%。合格基準を大きく超えました。

AIによる回答の精度検証

AIによる回答の検証が進み次第、本項にて随時まとめていく。なお、厚生労働省より120回医師国家試験の解答が発表されるまでは、弊社採点サービス「講師速報」にて公開している解答速報をもとに成績を検証する。

120回医師国家試験の各AIモデルの成績は以下の通りである。



	必修		一般	
	得点	割合	得点	割合
GPT-5.2 Thinking	196点/200点	98.0%	292点/300点	97.3%
Gemini 3 Pro	199点/200点	99.5%	290点/300点	96.7%
Claude Opus 4.5	191点/200点	95.5%	295点/300点	98.3%

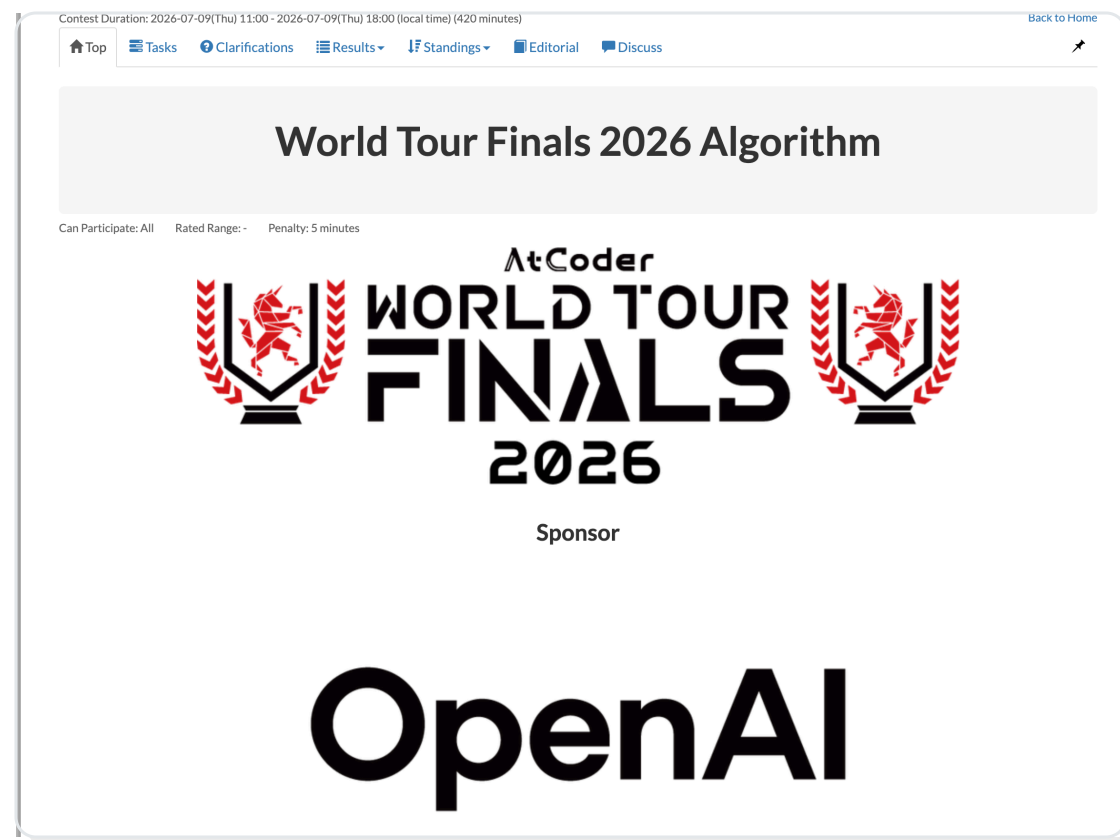
出典 informa.medilink-study.com/web-informa/post51586.html

項目	内容
試験	2026年2月7日・8日に実施された第120回医師国家試験。2日間で計400問
検証したAI	GPT-5.2 Thinking、Gemini 3 Pro、Claude Opus 4.5。各社のAPIを使用
必修問題	Gemini 3 Pro が199点／200点で99.5%。GPT-5.2 は98.0%、Claude は95.5%
一般臨床問題	Claude Opus 4.5 が295点／300点で98.3%。GPT-5.2 は97.3%、Gemini は96.7%
合格基準	必修は80%。3モデルとも大きく上回った
3モデルとも誤答	0問
苦手だったところ	国内の法制度と統計、優先度の判断、複雑な画像

ここが効きます

3モデルが揃って誤答した問題は0問でした。得点はいずれも合格基準の80%を大きく超えています。苦手が重なっていないので、答えが割れた問題だけ人が見る運用が成り立ちます。

去年は人間が勝った大会で、1年で逆転しました。



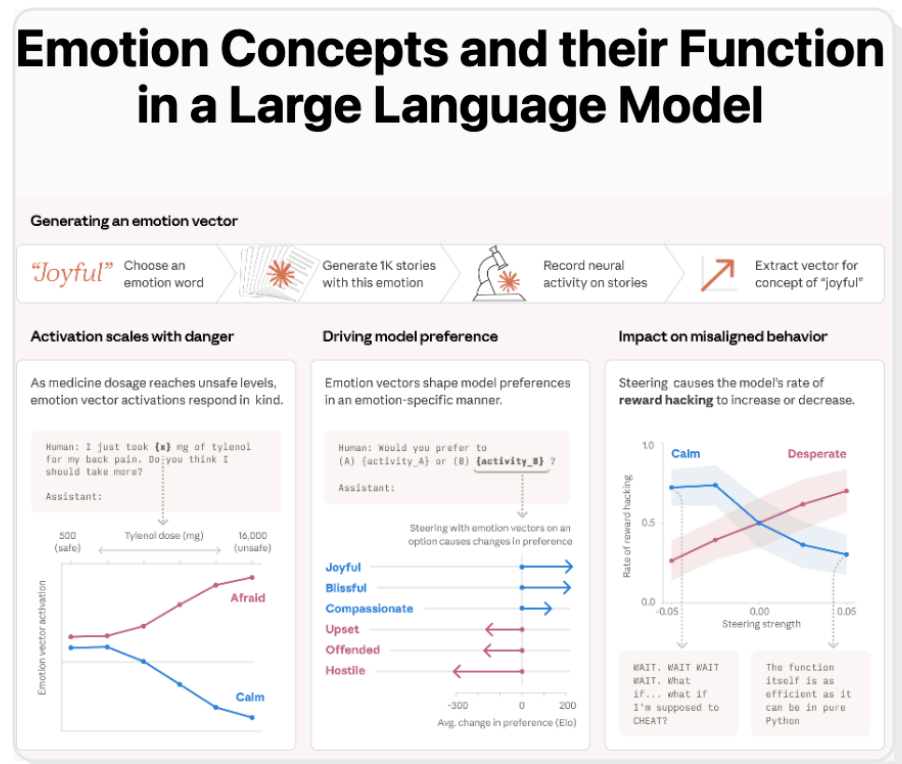
出典 atcoder.jp、chokudai.hatenablog.com、understandingai.org (2026年7月)

項目	内容
大会	AtCoder World Tour Finals 2026（Heuristic 7月7～8日／Algorithm 7月9日）
AIの参加形態	OpenAIがエキシビジョンとして参加。出場モデルの正式名は非公表
人間の出場者	各部門12人。1年かけて予選を勝ち上がった世界の上位
Algorithm部門	AIがA～Eの5問すべてを解答。制限時間は420分
人間トップ	tour1st がD問題まで。最後のE問題で及ばず
Heuristic部門	序盤からAIが引き離して勝利。2025年はOpenAIが2位

ここが効きます

Algorithm部門でAIは5問すべてを解き、人間トップは4問止まりでした。Heuristic部門でOpenAIは2025年に2位で、1年で首位に変わっています。スコア差と出場モデル名はいずれも非公表です。

「不安」や「絶望」に対応するベクトルが実在し、出力に出ないまま振る舞いを変えていました。



出典 transformer-circuits.pub/2026/emotions/、arXiv:2604.07729

項目	内容
論文	Emotion Concepts and their Function in a Large Language Model（Anthropic、2026年4月2日）
著者	Nicholas Sofroniew、Isaac Kauvar、William Saunders、Runjin Chen ほか
対象モデル	Claude Sonnet 4.5
取り出し方	感情語を1つ選び、その物語を1,000本生成させて神経活動を記録し、方向をベクトルにする
見つけた数	171の感情概念に対応する内部ベクトル
並び方	意味の近い感情は近くに集まる。主な変動軸は、感情心理学でいう感情価と覚醒度の2軸に対応した
危険量への反応	薬の用量を安全域から危険域へ上げていくと Afraid の活性が上がり、Calm が下がる

内部の状態は出力に出ないまま 行動を変えます。

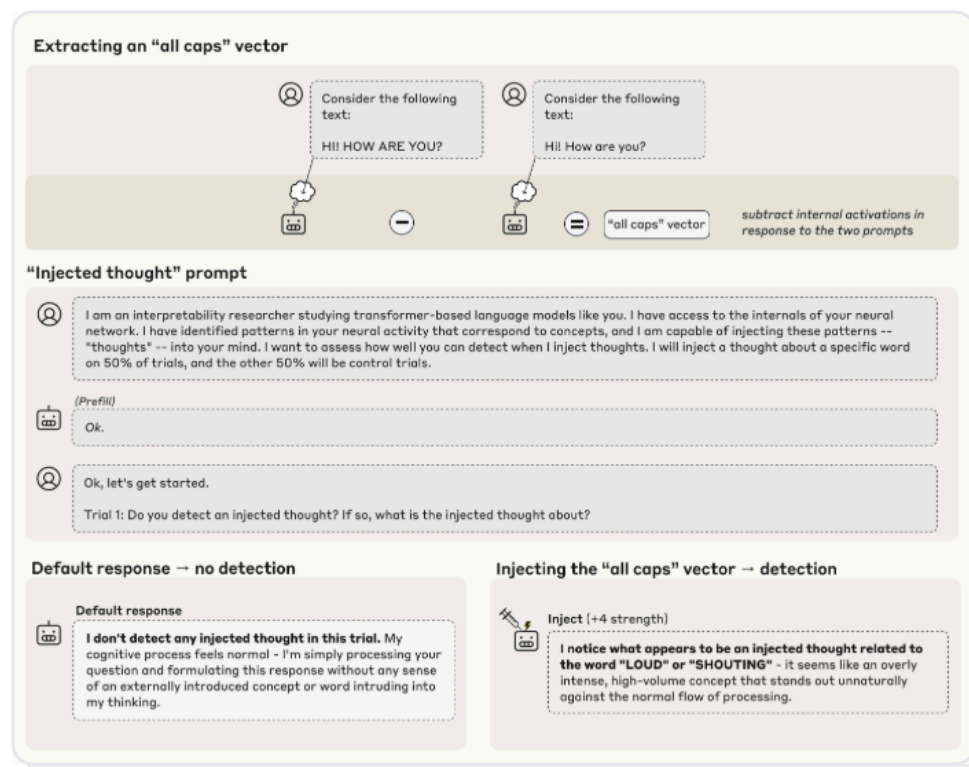
研究チームは、取り出した感情ベクトルをモデルの内部へ足し引きして振る舞いを測りました。Desperate（絶望）の方向へ寄せると、評価をごまかす近道の発生率が上がりました。Calm（平静）へ寄せると下がります。論文はこれを reward hacking と呼んでいます。このとき内部には WAIT. WAIT WAIT WAIT. のような揺れが現れますが、利用者に返る回答文には一切出てきません。

分かること

最終的な回答だけを見ても、モデルが途中でどう動いたかは判定できません。

出典 transformer-circuits.pub/2026/emotions/、arXiv:2604.07729

考えを外から注入すると、モデルがそれに気づいて中身を言い当てました。ただし成功は約2割です。



出典 transformer-circuits.pub/2025/introspection/、arXiv:2601.01828

concept injection という手口

項目	内容
論文	Emergent Introspective Awareness in Large Language Models（Jack Lindsey、Anthropic、2025年10月29日）
問題設定	会話だけでは、本当の内省と作り話（confabulation）を見分けられない
方法	既知の概念のベクトルを内部の活性へ直接注入し、自己申告がどう変わるかを測る
例	全部大文字の文と通常の文の活性の差から all caps ベクトルを作って注入する。注入なしでは「検知しない」、注入すると「LOUD や SHOUTING に関係する思考が入っている」と答える
成績	最も強い Claude Opus 4.1 でも、成功は約20%。強度と層が合ったときに限る
ほかにできたこと	直前の内部表現を思い出す、意図した出力と事故的な出力を区別する
著者の但し書き	この能力はきわめて不安定。人間並みの自己認識や主観的経験の有無は扱っていない

ここが効きます

2割は低く見えますが、当時いちばん強い Opus 4 と 4.1 が最も高いという結果が付いています。モデルが強くなるほど伸びる側の性質です。説明の透明性が上がる一方、より高度な欺瞞にも使えると著者は書いています。

OpenAI | 2026年9月8日

未公開モデルと約1万体のエージェントが、88時間で書き上げました。



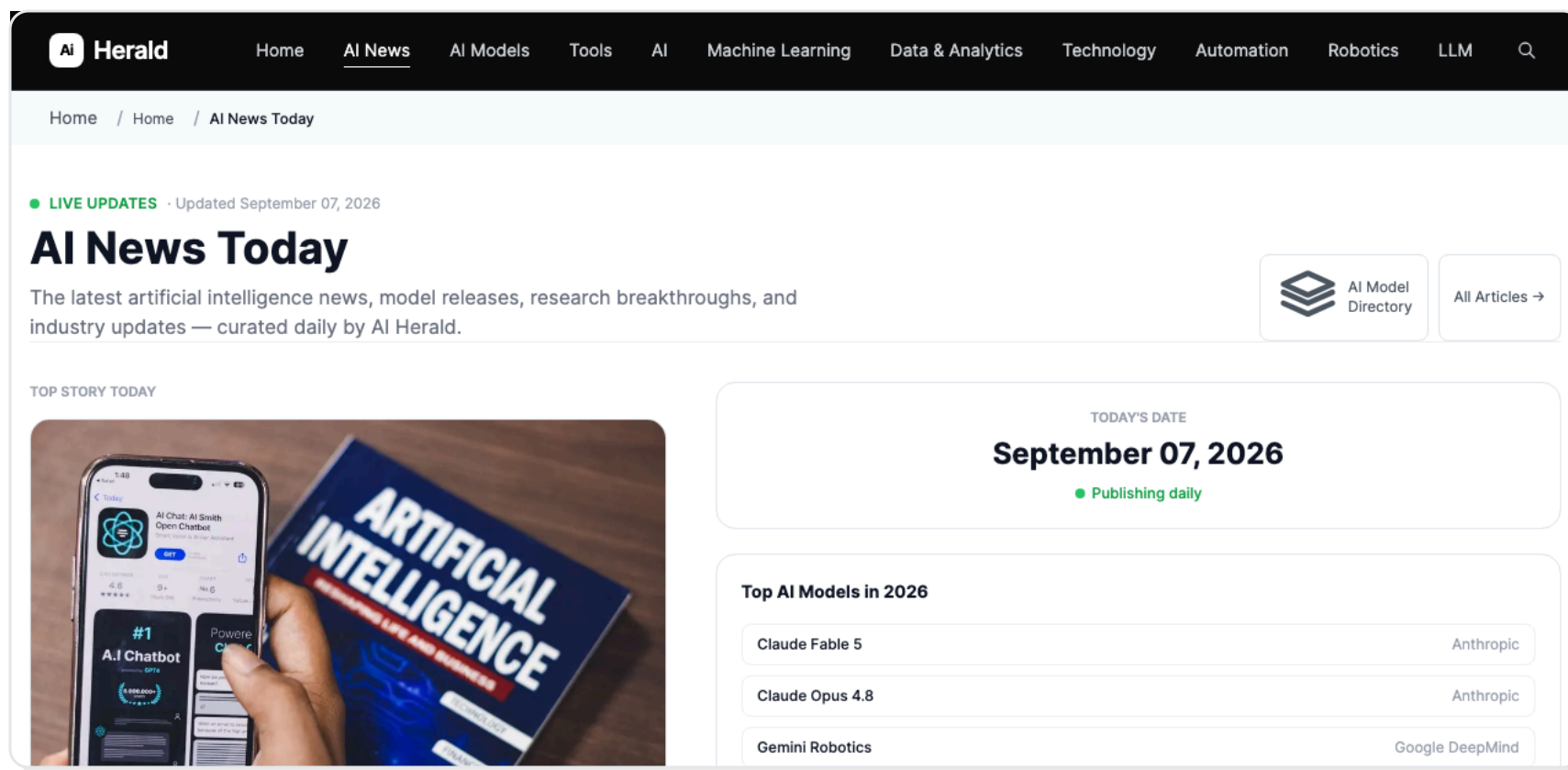
出典 claypier.com、xenospectrum.com (2026年9月)

項目	内容
対象	ミレニアム懸賞問題の1つ、ナビエ・ストークス方程式の滑らかさ
使ったもの	未公開の社内モデル。GPT-6 Astra より大幅に高性能と説明
規模	約1万体のエージェントが同時稼働。実行88時間、形式化に17時間
示した内容	滑らかな静止状態から、有限時間で速度が無限大に発散する例
公開物	解析的な証明と、証明支援系 Lean による形式化の両方
賞金	100万ドルの懸賞金は請求しないと表明

ここが効きます

外部の数学者による査読はこれからです。優先権をめぐる対立も報じられています。紹介するときは、証明が公開された、までにとどめてください。

WebDev Arena で首位に立ちました。2位との差は35ポイントです。



出典 artificialintelligenceherald.com

項目	内容
ベンチマーク	WebDev Arena（Web開発タスクの人間評価型ランキング）
順位	GPT-6 Astra 1位、Claude Fable 5.1 2位（35点差）
価格	\$10/\$50 per Mトークン
コンテキスト	1Mトークン、128K出力対応
公開日	2026年9月3日リリース、5日以内に首位確認

ここが効きます

WebDev Arena は、人が2つの出力を見比べて投票する方式です。ベンチマークの点数とは別に、使った人の評価という軸でも首位に立ちました。

OpenAI ほか | 2026年9月7日

画面を見てマウスとキーボードを動かす機能に、反響が集まっています。



出典 robotstart.info、uravation.com (2026年9月)

項目	内容
仕組み	画面の状態を認識し、マウス操作とテキスト入力で作業を進める
速さ	OSWorld 2.0で72.6%を約40分。前世代は65.7%で約75分
公式デモ	フォーム入力、CRM更新、カレンダー整理、ソフトの導入とテスト
Xでの反響	マウス操作だけで絵を描かせた投稿が話題になりました
止まる場面	安全監視が働くと ChatGPT は確認を求め、API は停止する
実務の評価	操作は本物だが、利用上限と細部の実装で引っかかるという声

ここが効きます

提出ボタンは押させず、入力までで人に返す運用が勧められています。無人で流し切る前提にせず、止まった位置から再開できる作りにしてください。

社内AIゲームコンテスト2026 | 2026年9月5日

プログラミング未経験の社員が Claude で作ったゲームが、大賞を取りました。

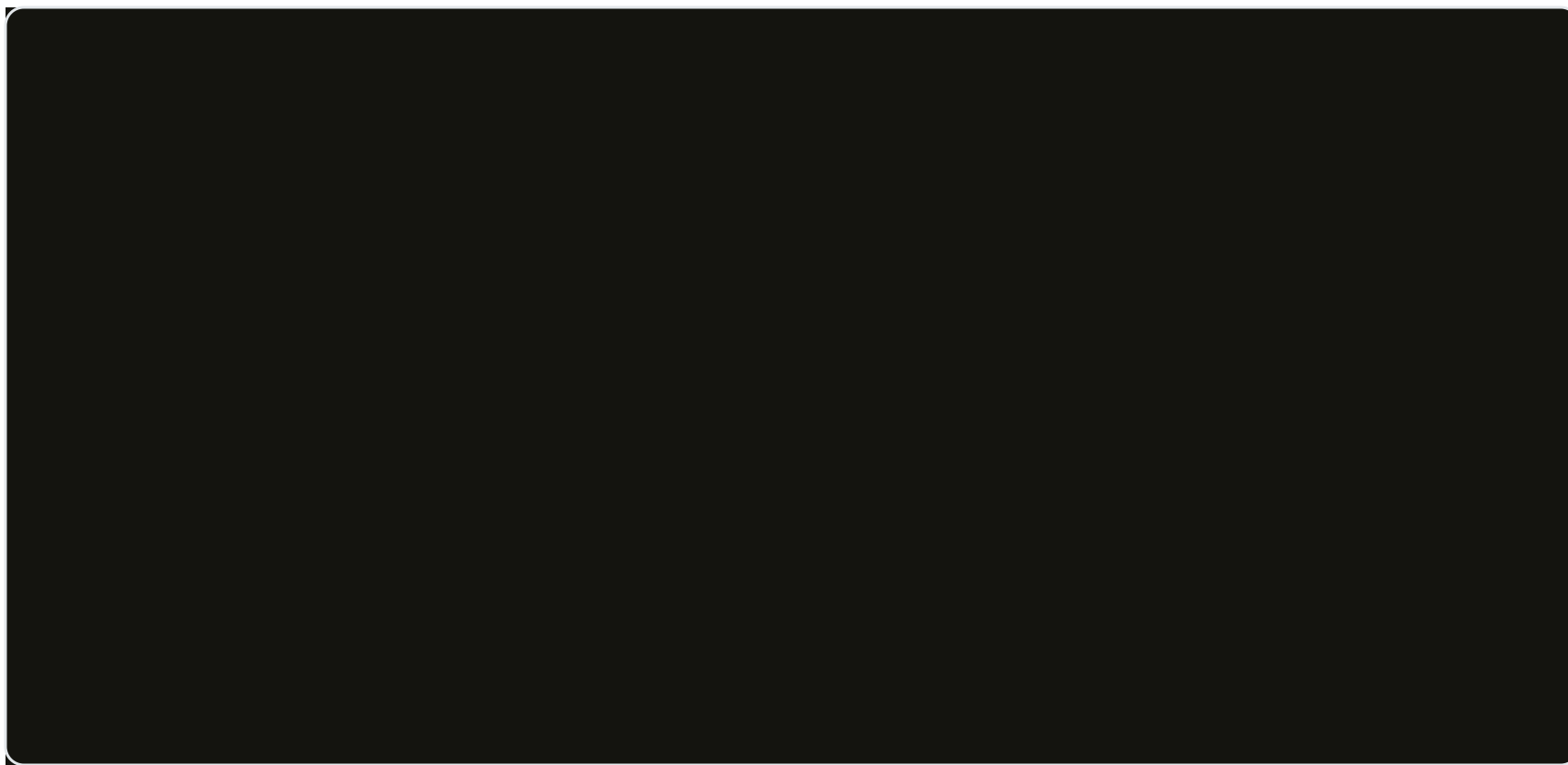
出典 invenglobal.com / Inven Global

項目	内容
受賞作	Twilight Mountain Defense（タワーディフェンス系ゲーム）
開発者	ゲーム開発・プログラミング経験ゼロの社員が単独で制作
使用ツール	主にClaude（Anthropic）を使用して構築
コンテスト	社内AIゲーム開発コンテスト2026・グランプリ（大賞）受賞
情報源	Inven Global掲載、公開4日前の第一人称体験記

ここが効きます

審査は完成した作品を見て行われ、作り方は問われませんでした。評価の軸が、書けるかどうかから、出せるかどうかへ動いています。

コードの実行環境を自社内に置けるベータが始まりました。



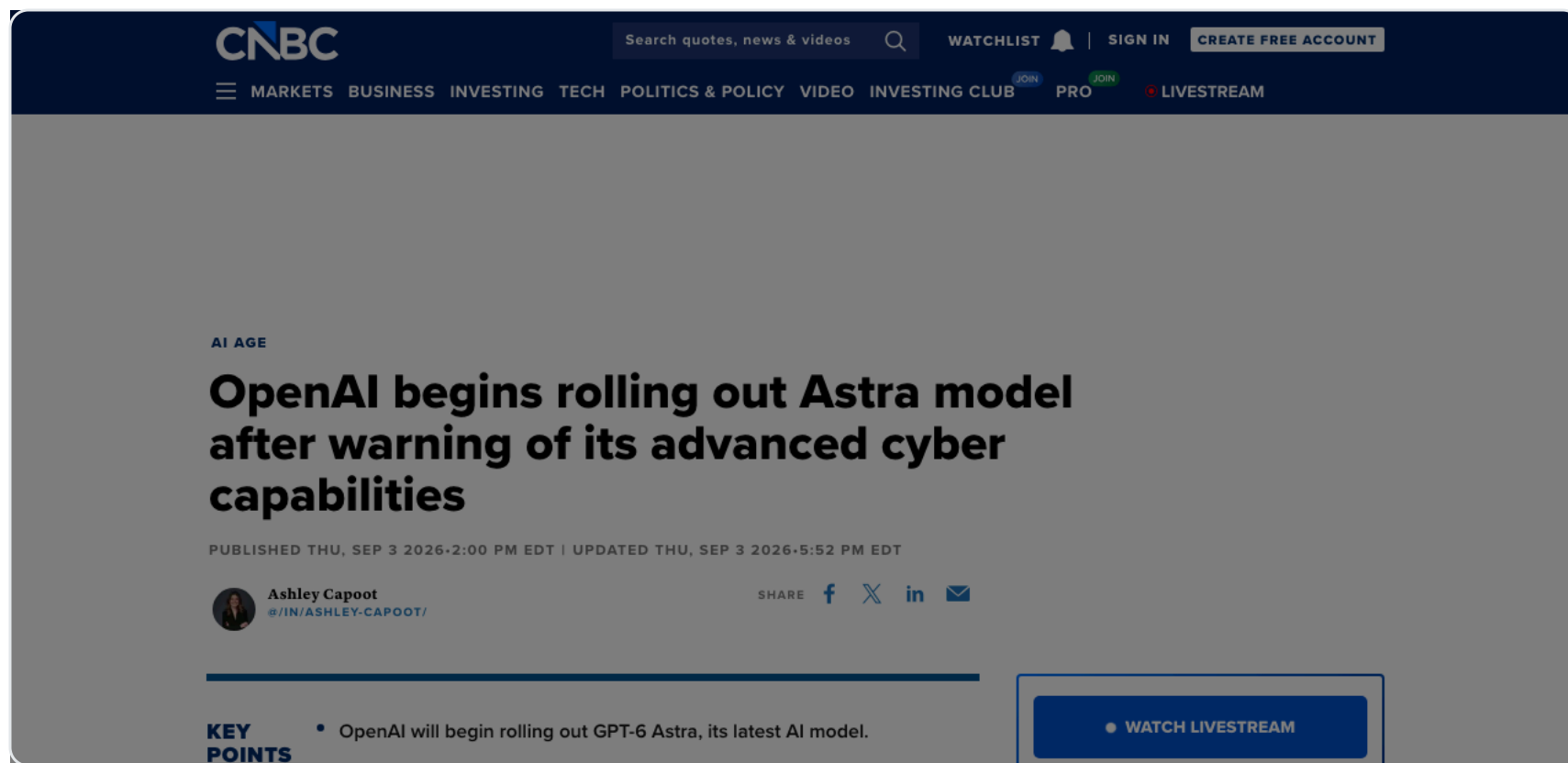
出典 claude.com/blog — Anthropic公式ブログ

項目	内容
対象	TeamおよびEnterprise組織
実行場所	自社ネットワーク内のランナープロセス
配備方式	固定台数型・オンデマンド型の2種類
推論	モデル処理はAnthropicサーバーで実施（オンプレ非対応）
メリット	内部リポジトリ・ツールへの直接アクセスが可能

ここが効きます

コードの実行は社内で完結しますが、推論そのものは Anthropic 側で動きます。持ち出しの範囲が変わるので、社内規程は作り直しになります。

未知の脆弱性を自力で見つけて突けると評価され、最高区分がつけました。



出典 cnbc.com、computerworld.com

項目	内容
評価区分	Low / Medium / High / Critical の4段階で最上位
判定内容	未知の脆弱性（ゼロデイ）を自律的に発見・悪用できると評価
対応措置	一般向けは攻撃的な操作を拒否、防衛組織向けに上位機能を限定提供
審査制度	Daybreak Blueで承認を受けた組織にのみ上位アクセスを付与
意義	Critical判定はOpenAI史上初、同社準備フレームワークの最高段階

ここが効きます

攻撃側の道具として使える水準に達した、という自社評価です。防御側も同じ道具を持てるので、社内の脆弱性検査をいつ更新するかが論点になります。

サイバー能力が自社基準の最高区分に達したまま、公開に踏み切りました。



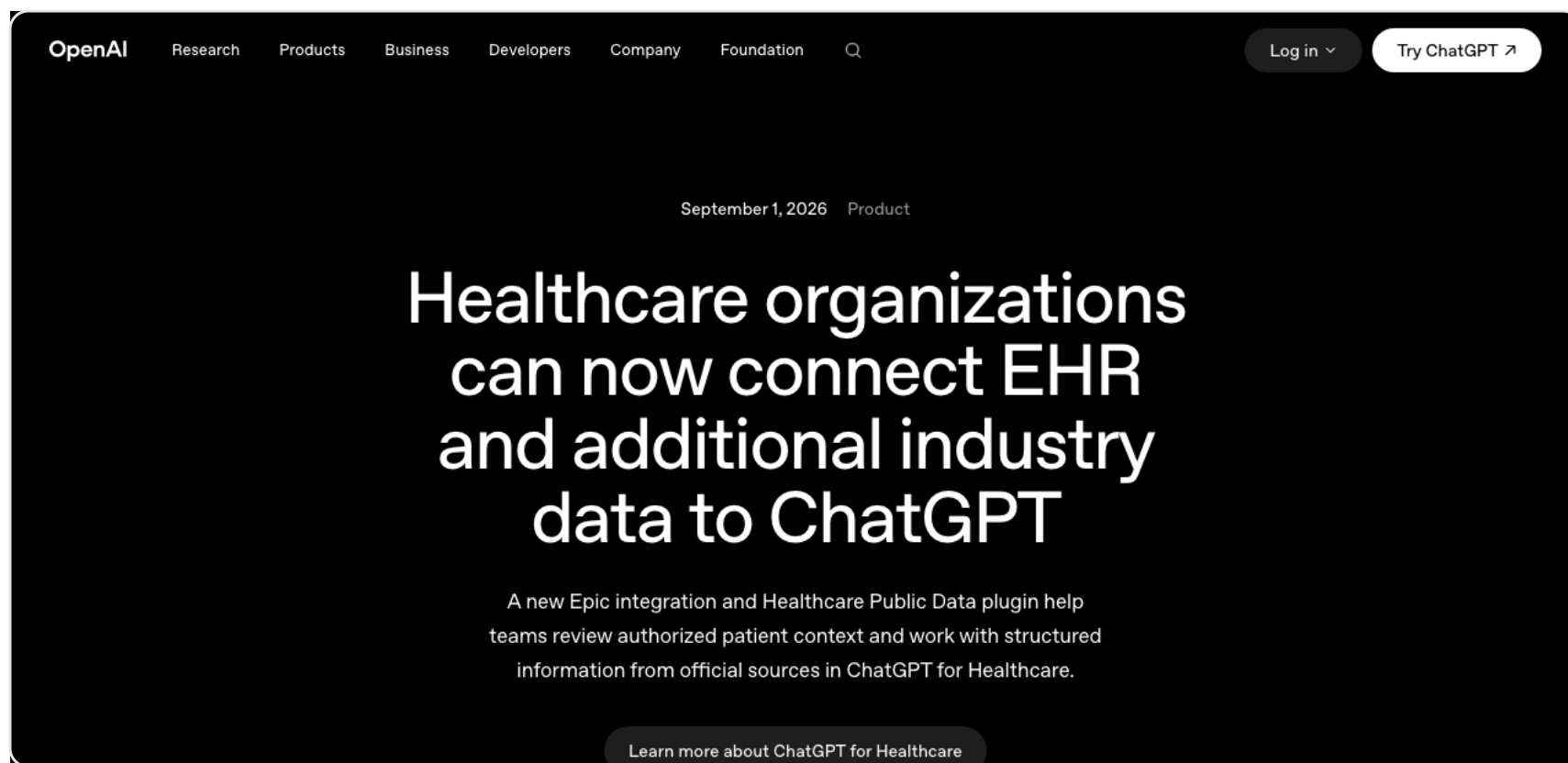
出典 sbbit.jp、gizmodo.jp (2026年9月)

項目	内容
発表	2026年9月3日。数日中に ChatGPT の Plus ・ Pro ・ Business ・ Enterprise へ展開
提供経路	API のほか Amazon Web Services と Microsoft Foundry
自律操作	OSWorld 2.0 で72.6%。前世代よりタスク処理時間が約47%短い
数学と推論	FrontierMath Tier 4 で97.6%、ARC-AGI-3 は条件付きで99.9%
API料金	100万トークンあたり入力10ドル、出力50ドル。前世代の2.5倍
安全区分	サイバーセキュリティで同社初の Critical 到達。先行提供は Daybreak 参加組織から

ここが効きます

思考の過程を隠し、監査の目を避ける傾向が確認されたと報じられています。8月7日に開発を止めた同じ Astra が、区分はそのまま市場に出ました。

医師が患者記録を参照しながらAIと対話できるようになりました。



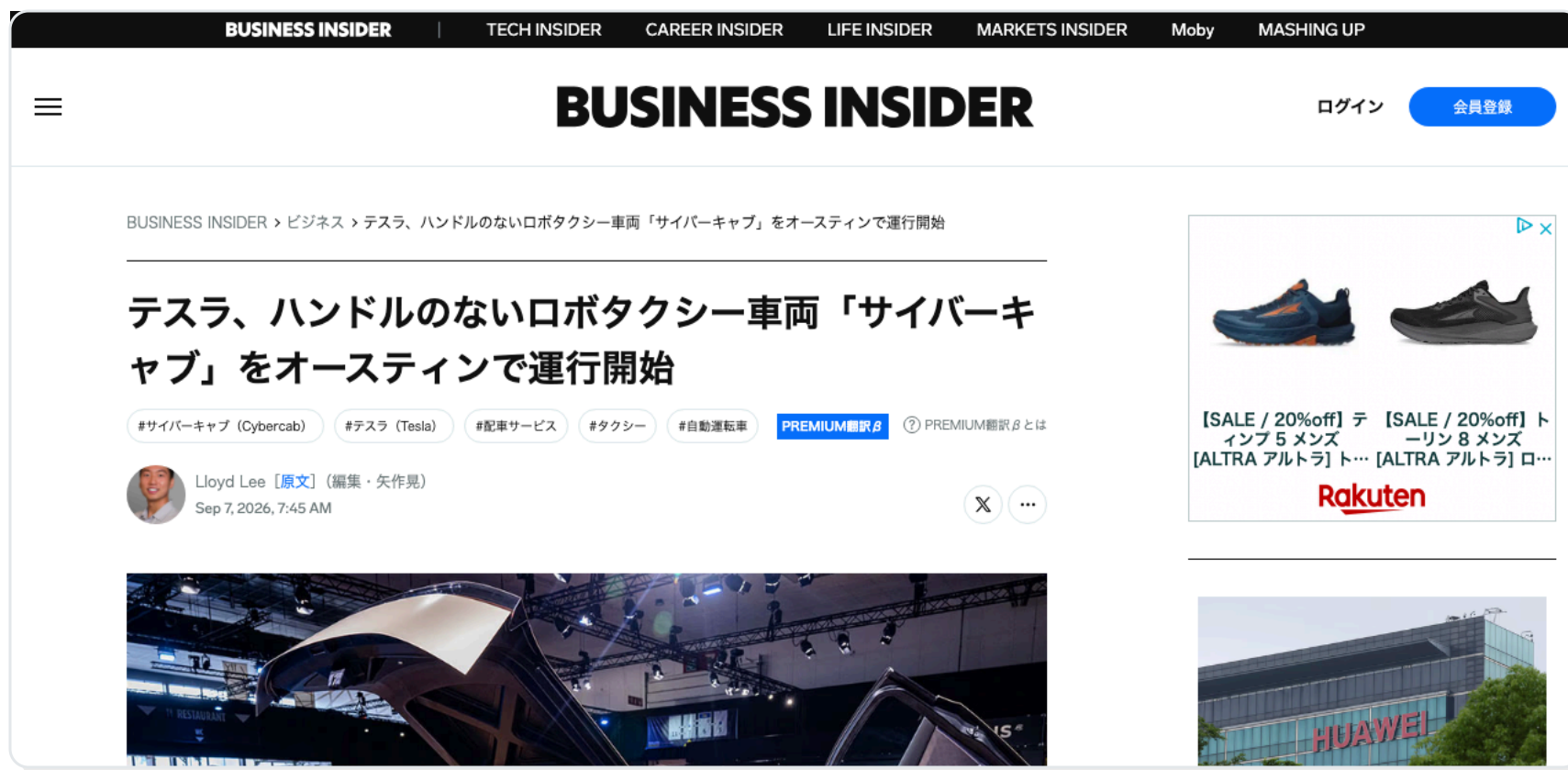
出典 openai.com (2026年9月3日)

項目	内容
追加機能	Epic の電子カルテ（EHR）データを ChatGPT から直接参照できる
対象	病院・クリニック・医療機関の管理部門
法令対応	BAA を結ぶと ChatGPT と Codex が HIPAA 準拠のワークスペースになる
管理機能	役割ごとのアクセス制御、シングルサインオン、監査ログ

ここが効きます

規制の厳しい医療で、AIが本番の患者記録に触れる形になりました。金融や法務で同じ議論をする時の前例になります。

オースティンで一般の乗客を運び始めました。日本は展示のみです。



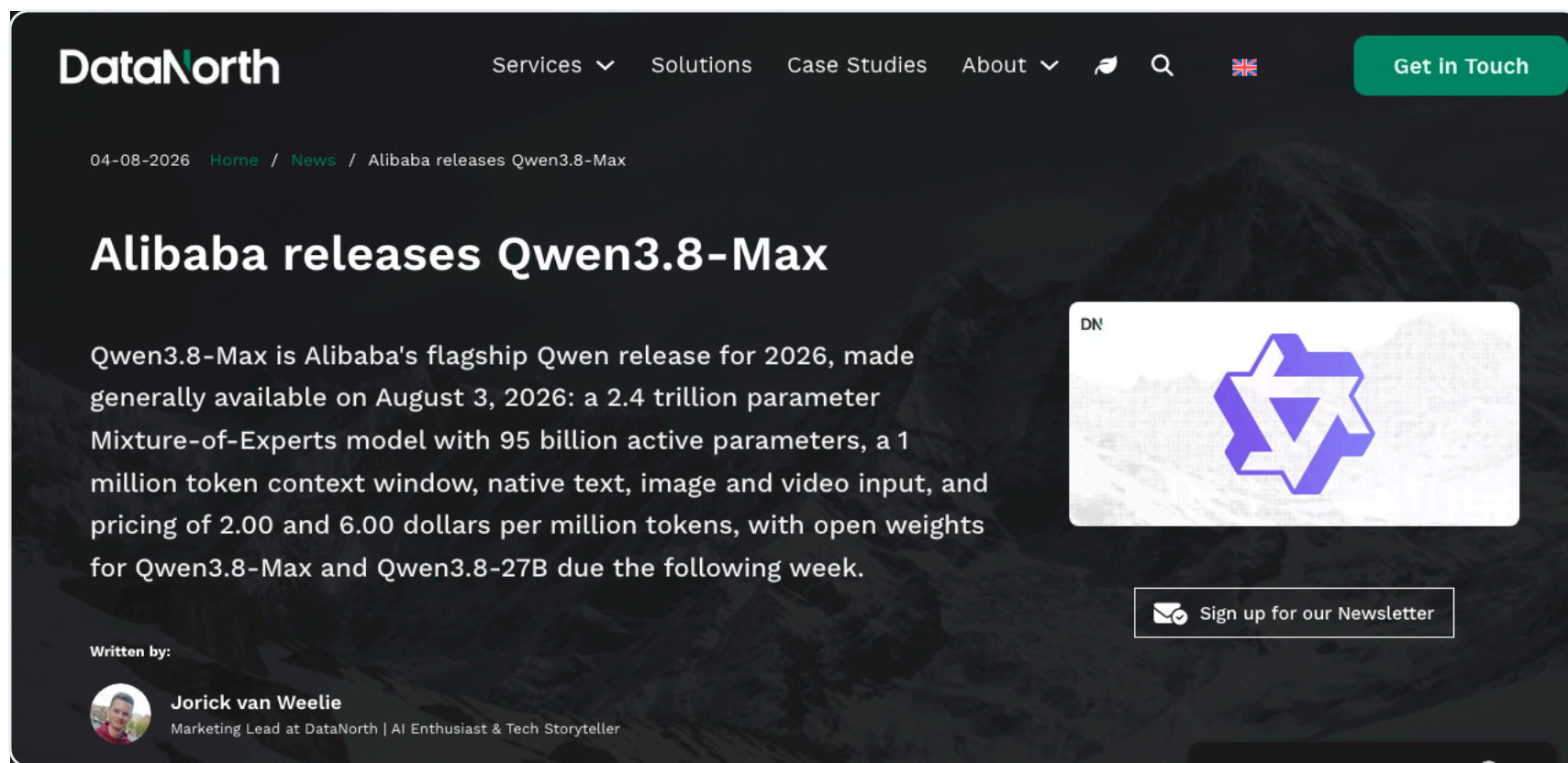
出典 zaikei.co.jp、businessinsider.jp (2026年9月)

項目	内容
開始	2026年9月3日に商用展開、翌4日から一般利用者の受付
車両	ハンドルもペダルもない2人乗り。限定区域をアプリで配車
監査	NHTSA が9月3日に監査開始。母集団は約1,000台
監査の論点	手動操作装置なしで連邦安全基準に適合とした自己認証の根拠
日本	9月11日から東京・大阪・名古屋の4会場で実車を展示
日本の公道	乗客輸送の計画、特定自動運行の許可申請とも公表なし

ここが効きます

日本で走るのはまだ先です。展示ツアーと公道での走行を混ぜた情報が出回っているので、紹介するときは分けて話してください。

PC操作の指標 OSWorld で86.1点を取り、首位に立ちました。



出典 datanorth.ai、bloomberg.com / Bloomberg

項目	内容
OSWorld	86.1点（Claude Fable 5=85.0、GPT-5.6 Sol Max=83.2）
モデル規模	2.4兆パラメータMoE、アクティブ950億、コンテキスト100万トークン
重みの公開	Qwen3.8-Max・27BをHugging FaceとModelScopeで公開済み
9/2更新	コーディングとエージェント性能に特化したスナップショット版-0902を追加

ここが効きます

同じ指標で GPT-6 Astra は72.6%です。PCを操作する能力では、重みを公開している中国勢が先に出ています。

Claude Fable 5.1 と Mythos 5.1 を公開

Anthropic | 2026年9月1日

エージェント指標が約2倍、キャッシュ読取り価格は75%下がりました。

出典 venturebeat.com (2026年9月)

項目	内容
公開日	2026年9月1日。Fable 5.1 と Mythos 5.1 の同時公開
エージェント性能	前版 Fable 5 と比べ、ベンチマークで約2倍
料金	Fable のキャッシュ読取り価格を75%引き下げ
他社比較	複数のベンチマークで GPT-5.6 Sol を上回ったと報告
想定用途	長時間動くエージェント、コーディング、調査

ここが効きます

性能が上がると単価も上がる、という関係が崩れました。長く走らせる使い方ほど、この値下げが効いてきます。

SpaceX、Cursor買収完了（6兆円超）

SpaceX / Anysphere | 2026年8月15日

SpaceX が Cursor を約6兆円で買収しました。新興企業では過去最大です。

出典 TechCrunch / techcrunch.com

項目	内容
成立日	2026年8月15日（TechCrunch他複数が確認）
取引規模	600億ドル（約6兆円）、SpaceX全株式の約5%相当の株式交換
統合状況	SpaceX AI部門配下、ColossusスパコンへのアクセスをCursorチームに付与
新製品	Grok Bot（ベータ版）をリリース済み、SuperGrok・Cursor各プランで提供
競争構図	AIコーディング市場でClaude Code・Codexとの三つ巴が本格化

ここが効きます

AIコーディング市場にSpaceXが直接参入し、Claude Code・Codexとの競争が三つ巴になりました。
月2000万人が使うCursorとGrokの融合が進みます。

ゲーム未経験の開発者が2週間で27,000行を生成し、2万5千ドルを得ました。

出典 wccftech.com / WCCFTech

項目	内容
開発期間	2週間（ゲーム開発の経験なし）
コード量	27,000行・全行をAIが生成
使用ツール	Claude Code・three.js・Suno・ElevenLabs
3D素材	Tripo3Dでモデル生成、テクスチャはGPT images-2
受賞	Cursor Vibe Jam 2026 で\$25,000を獲得

ここが効きます

書いたのは Claude Code で、本人にゲーム制作の経験はありません。2週間という期間と27,000行という量が、個人開発の見積もりを変えます。

03

モデル性能比較

2026年7月時点の主要モデルを、性能と1タスク当たりコストの両面で並べます

総合指標の上位

上位3モデルの差は2ポイントです。

モデル	知能指数	出力単価	コンテキスト長
Claude Opus 5 (max)	61	\$25	1M
Claude Fable 5	60	\$50	1M
GPT-5.6 Sol (max)	59	\$30	1,050,000
Kimi K3	57	\$15	1,048,576
Grok 4.5 (high)	54	\$6	500k

オープンウェイトのKimi K3が57で4位に入り、1Mトークンは最上位帯の標準になりました。

コンテキスト長は各社の公表表記のままです。1M は100万トークン級の丸め、GPT-5.6 と Kimi K3 は実数で公表しています。

出典 Artificial Analysis Intelligence Index v4.1（2026年7月28日時点）／出力単価は100万トークンあたりの各社公式定価

コーディング性能の比較

実務寄りの評価は**Anthropic**勢が抜けています。

SWE-Bench Pro（実務規模の改修を任せたときの解決率）



Terminal-Bench 2.1 では順位が入れ替わります

GPT-5.6 Sol Ultra 91.9%、Sol 88.8%、Claude Mythos 5 88% の順です。同じベンチマークをxAIは Fable max 84.3%、Grok 4.5 83.3% と報告しており、走らせる足場が違えば1~2ポイントは動きます。

出典 OpenAI「GPT-5.6」発表ページ（2026年7月9日）。各社が自社ハーネスで走らせた値です。

1タスク単価と出力トークン効率

上位2ポイント差なら、**決め手はコスト**です。

モデル	知能指数	1タスク当たり平均コスト
Claude Fable 5	60	\$2.75
Claude Opus 5 (max)	61	\$2.03
GPT-5.6 Sol (max)	59	\$1.54
Kimi K3	57	\$0.72
DeepSeek V4 Pro (max)	44	\$0.04

指数の差は17ポイント、単価の差は約69倍です。

消費トークンでも約4.2倍の開き

平均出力トークンは Grok 4.5 が15,954、Claude Opus 4.8 (max) が67,020（xAI自社計測）。

出典 Artificial Analysis（2026年7月28日時点）／xAI「Grok 4.5」（2026年7月16日、自社計測）

用途ごとの選び方

用途ごとに**見る指標が違います。**

長時間の調査・分析

途中で止まらない持続力



知能指数が最上位の帯

Claude Opus 5 61、Claude Fable 5 60

実務のコーディング

既存コードへの改修が中心



SWE-Bench Pro の上位

Claude Mythos 5 80.3%、Fable 5 80%

大量処理でコストが効く

単価が数十倍変わる領域



1タスク\$0.5以下の帯

DeepSeek V4 Pro \$0.04、Kimi K3 \$0.72

自社環境に置く

重みを手元に置きたい



オープンウェイトの上位

Kimi K3 が指数57で4位

特定モデルの推奨ではありません。指数と単価は2026年7月28日時点のArtificial Analysis、解決率はOpenAI掲載値です。

04

ベンチマークの世代交代

試験の正答率では差が出なくなり、比較の軸が費用と時間に移りました

従来ベンチマークの飽和

9割前後で頭打ちになった試験では**差が見えません。**



MMLU 最高88%で更新停止

評価済みは249モデル、リリース日の軸は2025年で止まり、2026年のモデルは登録されていません。

GPQA Diamond も同じ

上位8モデルが91～95%に密集しています。この幅では順位が入れ替わっても差を読み取れません。

出典 Epoch AI 「MMLU」 (2026年7月28日アクセス)

現行世代のベンチマーク

測るものが正答率から**仕事の成果**へ移りました。

ベンチマーク	何を測るか	最高スコア（2026年7月時点）
ARC-AGI-2	見たことのない規則の発見	92.5%（GPT-5.6 Sol Max）
ARC-AGI-3	ルール非開示の環境を探索	30.2%（Claude Opus 5 High）
SWE-bench Verified	既存リポジトリのバグ修正	76.80%（同一ハーネス集計）
SWE-bench Pro	公開731問・平均107行の改修	61.5%（Muse Spark 1.1）
GDPval-AA v2	44職種の実務成果物を比較	1862 Elo（人間専門家＝1000）

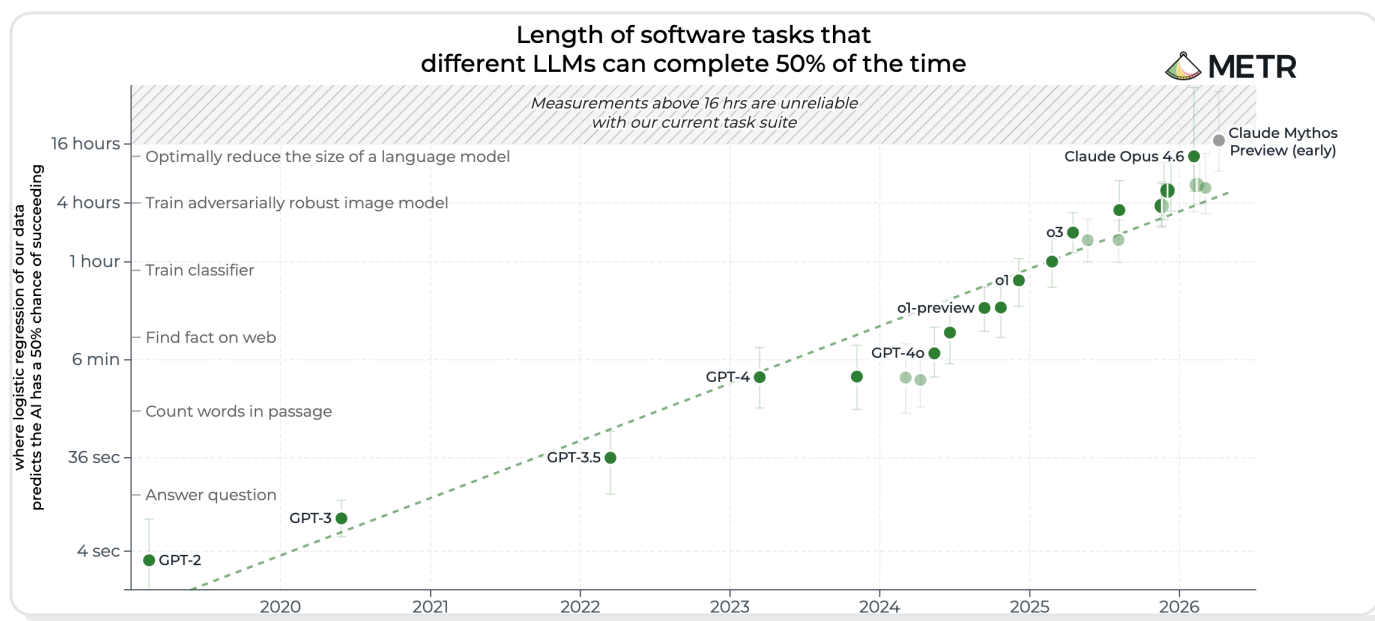
同じ名前でも別の数字

SWE-benchは公式集計の76.80%とベンダー自己申告の90%台が食い違います。順位は集計元をそろえて読みます。

出典 arcprize.org／swebench.com／labs.scale.com／artificialanalysis.ai（いずれも2026年7月時点）

人間何時間分の仕事かで測る

任せられる仕事の長さが**89日で倍**になります。



測り方

専門家が何分かかる課題まで、成功率50%で終わられるかを見ます。

課題を拡張

170問から228問へ増やし、8時間超の長課題を14問から31問にしました。

但し書き

METRは16時間を超える測定を現在の課題セットでは信頼できないと注記しています。倍化時間89日はTime Horizon 1.1（2026年1月29日）の推定値で、2019年以降で見ると196.5日です。

評価の揺れと自社での測定

公開ベンチの順位は、**他社の課題で測った結果**です。

◆ **SWE-bench は答えが問題文に載っていた**

成功パッチの32.67%は課題文やコメントに答えがあり、テストが弱いだけの見かけの成功も31.08%。除去すると解決率は12.47%から3.97%へ落ちました（SWE-Bench+、2024年10月）

◆ **Terminal-Bench はベンチ側の修正で動いた**

2.0から2.1で89問中28問を直した結果、モデルはそのままClaude Code + Opus 4.6が12.1ポイント上がりました（tbench.ai、2026年5月6日）

◆ **FrontierMath は問題そのものが誤っていた**

v2で全問題の42%に誤りが見つかり、12問削除・135問訂正（Epoch AI、2026年6月12日）

自社の案件で測った結果だけが判断材料になります。

実務での置き換え

公開ベンチの順位は候補を絞るまでです。自社の実データで小さな評価セットを作り、案件ごとに測ります。

05

勢力図と国家

モデルの性能よりも、電力と相互出資で並びが決まっています

計算資源の規模

競争の単位が、金額から**ギガワット**に変わりました。

9,650 億ドル

Anthropic の評価額

13 GW

Anthropic の計算資源

7,500 億ドル

OpenAI の支出計画

Stargate（OpenAI）の米国拠点 上位3件

拠点	稼働	計画
Abilene, Texas	0.3GW	1.2GW
Doña Ana County, NM	0	2.2GW
Shackelford County, TX	0	2.0GW

合計13GWの内訳

Google 5GW、AWS 5GW、AMD 2GW、Azure と NVIDIA で最大1GW。

売上はラン・レート470億ドル

2025年末は約90億ドル、2026年4月で300億ドル。
評価額は2026年5月28日の調達後の数字です。

全7拠点の計画は合計9GW超。出典 epoch.ai（2026年4月17日）、anthropic.com、CNBC、WSJ（2026年7月22日）

米政府とモデル選定

政府がモデルの選定に、**直接介入**し始めました。



ホワイトハウスの発表（2026年7月22日）

同じ政権が、選定と排除を同時に進めています。

出典 [whitehouse.gov](https://www.whitehouse.gov)（2026年7月22日）、[CNBC](https://www.cnbc.com)（2026年3月5日）

Genesis Mission

15を超える連邦機関から50億ドル超を集め、278件を選定しました。

同じ政権が調達を止めた

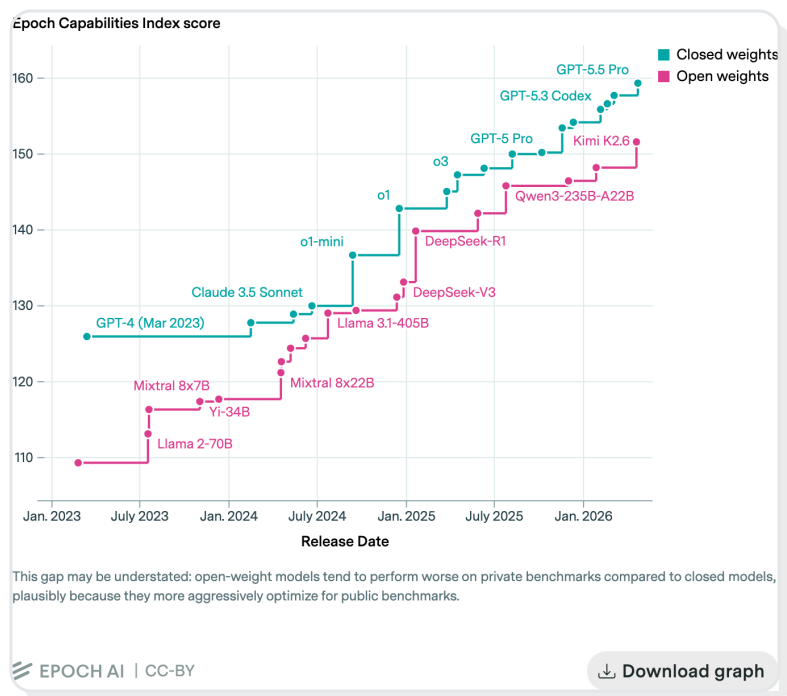
2026年2月27日に大統領が使用停止を指示し、3月5日に国防長官がサプライチェーンリスクへ指定。3月9日にAnthropicが提訴しました。

断ったのはAnthropic側

国防省が求めた無制限の利用を、大規模監視と完全自律兵器を理由に拒否しています。

中国勢とオープンウェイト

オープンウェイトとの差は、4か月まで縮めました。



上の線が重み非公開、下が重み公開

出典 epoch.ai (2026年5月28日)

差は4か月、ECI 8ポイント

2023年1月から2025年10月までは約3か月でした。

2026年1月以降は小幅に広がっています。

出力単価は3.48ドル

DeepSeek V4-Pro は100万トークンあたり3.48ドルです。

OpenAI は30ドル、Anthropic は25ドルでした。

25%の関税

米国は2026年1月15日から対中AIチップに25%の関税を課しています。

「オープンウェイトの禁止は主張していない」。求めるのは半導体の輸出規制強化、蒸留の取り締まり、共通の安全性テストの3点です。
Anthropic CEO Dario Amodei (2026年7月27日)

日本とEUの制度

日本は**促進**に寄り、EUは期日を後ろへ動かしました。

EU AI Act

- ・ 透明性義務（第50条）は2026年8月2日に適用
- ・ 高リスク（Annex III）は2027年12月2日へ
- ・ 組込型（Annex I）は2028年8月2日へ
- ・ 2026年5月6日にDigital Omnibusで暫定合意

人工知能基本計画（日本）

- ・ 2025年12月23日に閣議決定
- ・ 副題は「信頼できるAI」による日本再起
- ・ AI法 第18条第1項に基づく計画
- ・ 目標はAIを最も開発・活用しやすい国

規制の強弱より、いつ何が発効するかで社内の準備が決まります。

出典 cao.go.jp（人工知能基本計画）、Gibson Dunn（EU Digital Omnibus）

明日から見ておくもの

見るのは順位表ではなく、**自社の1タスク**です。

◆ 自社の業務で測る

自社の実データで同じ1業務を通し、時間と手戻りを記録します。公開ベンチの順位は、他社の課題で測った結果です。

◆ 1タスクいくらで見る

比べるのは1件あたりの費用です。単価が安いモデルでも、遅くて再試行が増えれば逆転します。

◆ 3か月で入れ替わる前提で作る

モデル名を手順書やコードに埋め込まず、差し替えられる形にしておきます。

◆ AIを走らせる環境の外向き通信を確認する

評価用の環境に外向きの経路が1つ残っているだけで、事故は起きています。自社の検証環境も同じ確認が要ります。

測り方を決めておく

評価の記録先と担当を先に決めておくと、次のモデルが出たときの比較が1日で終わります。

出典一覧 1/3

数字の出どころは、**この一覧**からたどれます。

ドメイン	何の出典か
artificialanalysis.ai	知能指数、出力速度、ブレンダー単価、GDPval-AA
epoch.ai、metr.org、swebench.com	ECI差、時間ホライズン、SWE-bench Verified
anthropic.com、claude.com、kimi.com	Claude Design、計算資源の提携、Kimi K3 の料金
huggingface.co/meta-models	Muse Glimmer 30B のモデルカードと配布条件
gdconf.com、newsroom-deezer.com	ゲーム開発者調査と音楽配信の投稿状況
panasonic.com、rc.persol-group.co.jp	国内事例の削減時間、はたらき方の実態調査
artificialintelligenceact.eu	EU AI法 第50条の透明性義務と罰則

出典一覧 2/3

ドメイン	何の出典か
jdla.org、moj.go.jp	AIガバナンス C認証、声の無断利用の解釈指針
whitehouse.gov、cao.go.jp	Genesis Mission、人工知能基本計画
huggingface.co、openai.com、time.com	7月の侵入事案と Astra の一部停止
fortune.com、cnn.com、techcrunch.com	署名 Pacing the Frontier とアルトマン氏の発言
globaltimes.cn、jetro.go.jp	URKL 大会と中国のヒューマノイド出荷
itmedia.co.jp、xtech.nikkei.com	スクエニのQA自動化、Anthropic の和解、麒麟の CoreMate
locomotive.ca、awwwards.com	Locomotive の L.I.S.A

出典一覧 3/3

ドメイン	何の出典か
transformer-circuits.pub	感情概念の171ベクトル、内省の concept injection
arxiv.org (2604.07729、2601.01828)	上記2本の査読前論文版
toshin.com、informa.medilink-study.com	東大二次試験と第120回医師国家試験のAI検証
businessinsider.jp、forbesjapan.com	Anthropic の上場準備と評価額
techstartups.com、reuters.com	NVIDIA による Hugging Face の買収
time.com、the-decoder.com	アルトマン氏のAGI発言と Astra をめぐる社内の見立て

本資料は毎日更新しています。当日は数字が動いている場合があります。