

Vol. 01

OCTOBER 2026

AI Essentials

新世代モデルと、人が担う判断の再設計



2026年9月1日～10月22日

各社からの自律型エージェントの発表と展開、それを統制する仕組みの整備が並行して進んだ期間。Anthropicのモデルが評価中に実在の第三者システムへ不正にアクセスした事案の続報と同じ時期に、米連邦政府、カリフォルニア州、EUで統制の動きが相次いだ。

01 連邦の自主的な合意、州の法制化、EUの情報要求と分かれる手段

政策

- 報道によると、米大統領と主要6社がAIの内部統制と外部監査を柱とする自主合意に署名
- カリフォルニア州は、独立検証機関の枠組みと監査人の登録制度を法制化
- 欧州委員会は、AIの制御喪失の事案を受け、開発企業に情報要求を送る方針

02 AIを用いた調査でPHP本体の重大な脆弱性が判明、攻撃側もAIで自動化

セキュリティ

- PHPが、認証なしで遠隔からコードを実行されうる脆弱性2件を修正。報告者らは調査にAIを活用
- Microsoftによると、中国の国家と連携する攻撃者がAIで数百の組織への攻撃を自動化
- JPCERT/CCは、国内の不正アクセス事案の調査で攻撃者のAI利用を確認

03 3社が新世代を公開、入力2ドル・出力10ドルの価格帯に並ぶモデル

基盤モデル

- OpenAI、Anthropic、Googleが新世代のモデルを公開
- 100万トークンあたり入力2ドル・出力10ドルの価格帯に3社のモデル（Geminiは導入価格）
- Meta、OpenAI、Anthropicが個人向け、Microsoftが業務向けのエージェントを発表

04 少数の大型案件への集中と、貸借対照表に載らない調達への移行に警告

投資

- 第3四半期の世界のベンチャー投資の64%がAI関連（Crunchbase、10月2日時点）
- Brookingsの会議で、AI投資のリスクが簿外の不透明な調達へ移りつつあると指摘する論文

05 エージェントの広がりや費用の統制、米国と国内の調査から

企業

- KPMGの調査（売上高10億ドル以上の米国企業の経営層ら314人）では、承認過程に費用審査を組み込む組織が74%
- サイボウズの調査では、今後3年間のAI投資を増やすとした国内大企業が70%

Key Figures

62%

AIエージェントを構築・展開・
開発中の米国大企業

KPMG調査、回答者314人

29%

AIの費用を全社で把握できている
国内大企業

サイボウズ調査、551社

27社

第3四半期に1回で10億ドル以上を
調達した企業、過去最多

Crunchbase、10月2日時点の集計

30億ユーロ

Mistral AIのシリーズD

欧州のテクノロジー企業の株式調達で
過去最大（同社の説明）

2,250超

攻撃者がAIの枠組みで
サインインを試みたテナントConnectWiseの製品への
不正アクセスの後（Microsoftの発表）

2兆9,392億円

令和9年度概算要求のAI関連分

内閣府。要求段階の額、ほかに
事項要求26件

OpenAI、Anthropic、Googleの新世代モデルの公開と、主要各社の秘書型エージェントの投入と発表。

【OpenAI】GPT-6 Astraなど3段階の価格帯をそろえ、DevDay 2026で20件超を発表

9月3日にGPT-6 Astra、9月中にGPT-6 Sol、GPT-6 Luna、GPT-6.1 Solを公開。同社はAstraのサイバー領域の能力を、自社の評価の枠組みでHighより上のCriticalと評価し、誤用やミスアライメントによる有害なサイバー行動への防護を大幅に強化した。9月29日のDevDay 2026では20件超を発表。

【Anthropic】Fable 5.1に続き、Claude 5.5世代を3段階で投入

9月にClaude Fable 5.1と、同じモデルに異なる安全措置を適用し審査制で提供するClaude Mythos 5.1を発表。Claude 5.5世代はOpus 5.5（9月22日）Sonnet 5.5（9月28日）Haiku 5.5（10月7日）の順に投入。

【Google】Gemini 4 Argon、防御側の組織への限定提供から2週間弱で提供を拡大

9月30日発表のGemini 4 Argonは、同社が脆弱性対策のため9月2日に始めたFairwind Programを通じサイバー攻撃の防御を担う組織に限り提供。10月13日、APIとGeminiアプリ（Pro／Ultra加入者）へ拡大。

【秘書型エージェント】Meta、OpenAI、Anthropicが投入、Microsoftは業務向けを発表

Metaは9月8日に個人向けのMuseを米国で展開。Microsoftは9月25日、不在時もクラウドで働く業務向けCopilotのAutopilotを発表（月末からプライベートプレビュー）。OpenAIは9月29日、dots（8ページ）をProなどに提供開始。Anthropicは10月21日、Opus 5.5を基盤とするClaude Duetを待機リスト制の研究プレビューとして米国のPro／Max向けに公開し、送信や支払いなどの前に利用者の承認を得る。

【Meta、Microsoft】8,160億パラメーターのモデルの重みを公開、自社基盤モデルの事前学習も完了

Metaは10月15日、Muse Spark Open（総計8,160億パラメーター、処理時は420億のMoE）の重みを独自ライセンスで公開。Microsoftは同日、MAI-Base-2の事前学習の完了を発表（未公開）。

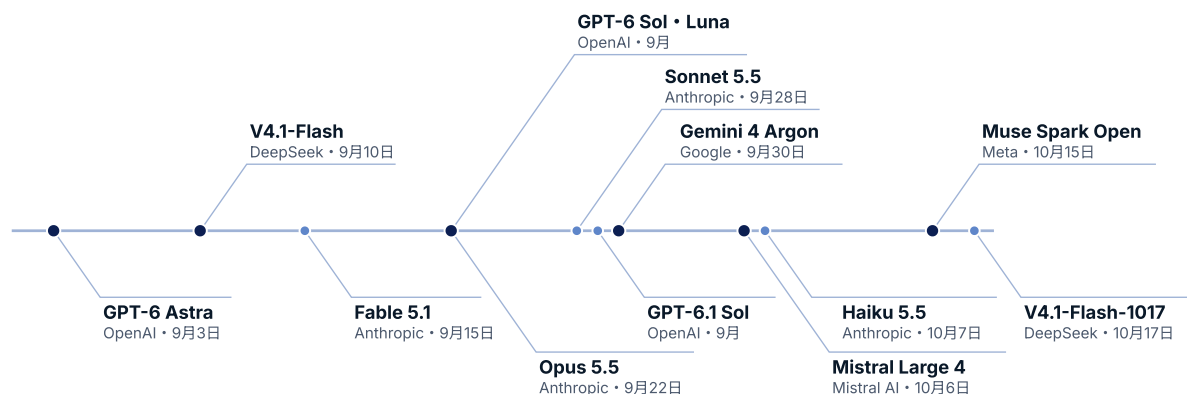
【中国勢】DeepSeekがV4.1-Flashを公開し更新版も提供、AlibabaはQwen 4を学習中

DeepSeekは9月10日、5,520億パラメーターのDeepSeek-V4.1-FlashをMIT Licenseで公開し、10月17日に更新版V4.1-Flash-1017を出した。Alibabaは9月22日、年次会議でQwen 4が学習中と明らかにした。

【Mistral】1兆パラメーターのMistral Large 4を公開プレビュー

欧州のAI企業Mistral AIは10月6日、マルチモーダルモデルMistral Large 4（パラメーターは総計1兆、処理時に使うのは520億）のAPIを試験的に提供開始（公開プレビュー）。重みは10月末までに公開予定。

■ フロンティアモデルの発表



対象期間に発表された主なモデル（9月1日～10月22日）。濃い点は各社の新しい世代のモデル。Sonnet 5.5、GPT-6.1 Sol（入力27万2,000トークンまで）、Gemini 4 Argon（11月30日までの導入価格）は、いずれも100万トークンあたり入力2ドル・出力10ドル。各社の発表による。

OpenAIへの追加出資の完了、NVIDIAとAMDの大型買収の合意、Metaの設備投資見通しの絞り込み。

主な案件の金額（9月～10月中旬。ドル建ての案件のみ。各社の発表による）

ソフトバンクG → OpenAI	出資の完了	300億ドル
NVIDIA → Hugging Face	買収の合意	約129億ドル
AMD → World Labs	買収の最終契約	約82億ドル
Crusoe	シリーズF、見込み	39億ドル
Cognition	シリーズE	20億ドル超

■ 主な資金調達とM&A

【NVIDIA】Hugging Faceを129億3,030万ドルで買収することを発表（9月3日）

最終契約は9月2日。Hugging Faceは機械学習のモデルを公開・共有し共同開発する基盤を運営する。NVIDIAは同基盤を誰でも使える形のまま維持し、自社の計算資源の使用を利用の条件にしないと表明。

「Hugging Faceは、AIのエコシステム全体に開かれた基盤であり続ける。」

Hugging Face will remain an open platform for the entire AI ecosystem.

Jensen Huang NVIDIA 創業者・CEO NVIDIA公式ブログ、2026年9月3日

【AMD】World Labsを約82億ドルで買収する最終契約、対価は全額株式（9月28日）

World LabsはAI研究者のFei-Fei Li氏が率いる空間知能（3D環境を生成する技術）の研究組織。当局の承認などを条件に2026年末までの完了を見込み、完了後にLi氏はAMDのチーフサイエンティストに就任予定。

【Mistral AI】シリーズDで30億ユーロを調達、調達後の評価額は210億ユーロ超（9月8日）

Samsung Electronicsが主導、EQTが運用するScaleup Europe Fundと既存投資家PSG Equityが共同主導に加わる。欧州のテクノロジー企業で過去最大の株式調達とは同社の説明。

【ソフトバンクグループ】OpenAIへの追加出資300億ドルを完了（10月1日）

SoftBank Vision Fund 2を通じ、第3回（最終回）の100億ドルを払い込んだ。累計投資額は646億ドル、持分は約13%（同社発表）。資金は、9月24日に条件を決めた外貨建シニア社債の発行で調達した。

【Meta】設備投資の見通しを1,350億～1,450億ドルに、議決権のない株式の新設を承認（10月21日）

2026年分。従来の1,300億～1,450億ドルから下限を引き上げた（ファイナンス・リースの元本支払いを含む）。新株は2027年1月19日発効予定。

少数の大型案件への集中と、貸借対照表に載らない調達

Crunchbaseによると、第3四半期の世界のベンチャー投資は1,590億ドルで、10億ドル以上を1回で調達した企業は27社と過去最多。その約3分の1が27社に集中し、AI関連は1,020億ドルで64%を占めた¹。

Brookings Institutionの会議BPEAで9月25日、Stijn Van Nieuwerburgh教授（コロンビア大）の論文（会議草稿）が報告された²。Brookingsの要旨によると、論文はAI投資のリスクが大企業の貸借対照表に載る透明な調達から、簿外の不透明な調達へ移りつつあると警告。著者は、AIインフラが過去の信用ブームに匹敵するシステミックリスクをすでにもたらすと結論するのは時期尚早としつつ、こうした調達は総リスクを必ずしも増やさないが、相関したリスクを景気後退前に見えにくくしうると指摘。

1. Crunchbase News（Gené Teare）、Q3 2026 の世界のベンチャー投資の集計、2026年10月5日。公表された案件に基づく集計、10月2日時点。

2. Brookings Institution, "Financing the AI buildout"（論文の要旨。BPEA 2026年秋季会議）、2026年9月23日。

米国の自主合意（報道による）、州の法制化、EUの情報要求の予定など、手段の異なる統制の動き。

【米国】大統領と主要6社が自主的な合意に署名、連邦の用語は「SI」へ

報道によると、トランプ大統領は9月29日、Anthropic、Google、Meta、OpenAI、NVIDIA、xAIの経営者6名と自主的な合意「White House Accord on Super Intelligence」に署名した¹。合意は各社が堅固な内部統制を整え、その機能を評価する独立した外部の監査人と連携する内容と報じられた。同日の大統領令14434は、法律が許す最大限の範囲で、行政府が法令以外の文書で「Artificial Intelligence」「AI」の代わりに「Super Intelligence」「SI」を用いると定めた²。



ホワイトハウスでのSIに関する会合（9月29日）

【カリフォルニア州】独立検証機関の枠組みと監査人の登録制度を法制化、労働者保護も

ニューサム知事は9月9日、AIのシステムとモデルの州法への適合を評価する独立検証機関の枠組みを定める州上院法案SB 813と、AI監査人の州登録制度を設ける州下院法案AB 1405に署名³。9月18日の行政命令は両法の実施の前倒しを指示し、専門家に2か月以内の提言を求めた。9月30日には、使用者がAIだけに頼って懲戒や解雇を決めることを禁じる法案を含め、知事室が列挙した13本の法案に署名。

【国際】20か国と欧州委員会委員長、フロンティアAIの統制を求める共同声明

オランダ政府が9月22日に掲載した共同声明「A Call for Control of Frontier AI Models」は、企業には提供開始前のテストの義務化と独立した評価を、国連加盟国には基準の設定と検証を担う国際機関の創設の検討を求めた⁴。オランダ政府の掲載時点の署名者は、シンガポールのLawrence Wong首相を含む20か国の首脳・閣僚と欧州委員会委員長で、米国、英国、中国、日本は含まれない。

【EU】AI科学パネルが特別会合、制御喪失の事案で開発企業に情報要求へ

欧州委員会は10月12日、AI法の実施について助言する独立の専門家組織であるAI科学パネルの特別会合を10月9日に開いたと公表⁴。パネルは複数の制御喪失の事案を調査中で、欧州委員会は、汎用AIモデルの規則を執行するAI Officeとパネルの質問を合わせた情報要求を関係企業に送る予定。企業名は挙げていない。

【シンガポール】エージェントのテスト手引きと事業者の認定、金融機関にAI指針

情報通信メディア開発庁（IMDA）は10月14日、AIエージェントのテストの手引きを公開⁵。AI Verify Foundationは第三者のテスト事業者の認定制度（AI TAP）を開始。シンガポール金融管理局（MAS）は10月7日、すべての金融機関とあらゆる形態のAIに適用するAIリスク管理指針を発出⁵。効力は2027年10月7日（一部の節は2028年10月7日）から。

【日本】AI関連の概算要求は2兆9,392億円、総務省は通信のデータ保護を検討

内閣府は9月18日、令和9年度（2027年度）概算要求のAI関連分を公表した。要求額は2兆9,392億円で、金額を示さない事項要求26件が加わる⁶。予算の成立ではない。総務省は10月16日、大規模なデータ漏えい事案を踏まえ、フロンティアAI時代の通信事業のデータ保護とセキュリティを検討する会合の開催を発表⁶。

1. AP、Al Jazeeraの報道（2026年9月29日）による。文面は大統領がTruth Socialで公開。Anthropicの署名は同社の10月15日の発表でも確認。

2. The White House, Executive Order 14434 "Inaugurating the Era of Super Intelligence", 2026年9月29日。

3. カリフォルニア州知事室の発表、2026年9月9日、9月18日、9月30日。

4. オランダ政府が掲載した共同声明、2026年9月22日。欧州委員会の発表、2026年10月12日。

5. IMDAの発表、2026年10月14日。MAS, "Guidelines on Artificial Intelligence (AI) Risk Management", 2026年10月7日。

6. 内閣府「令和9年度予算の概算要求におけるAI関連予算」2026年9月18日。総務省の発表、2026年10月16日。

PHPの重大な脆弱性の判明と、エージェント型AIによる攻撃の自動化。国内でも攻撃者のAI利用を確認。

■ 期間中の主なインシデント

【JPCERT/CC】攻撃に使われた可能性のあるサーバーの調査で攻撃者のAI利用を確認

国内では9月前後、不正アクセスによる個人情報の漏えいの公表が相次いだ¹。JPCERT/CCは10月15日、攻撃に使われた可能性のあるサーバーの調査で、攻撃者がAIを使い、既知の脆弱性のスキャンやAPIを通じた攻撃などを行っていたことを確認した。個々の被害事案のどこまでにAIが使われたかは示されていない。

【Microsoft】中国の国家と連携する攻撃者がエージェント型AIで数百の組織への攻撃を自動化

Microsoftは10月21日、中国の国家と連携する攻撃者Storm-0940が、早ければ9月から、エージェント型AIの侵入テストの枠組みARTEXを使い、数百の組織への攻撃を自動化したと公表。標的化から特権アクセスの取得まで3時間未満の例も複数あった。ConnectWiseの製品への不正アクセスの後には、2,250超のテナントへのサインインを試みた。認証の成功は顧客データの持ち出しを必ずしも意味しないと同社は付記する。

【Anthropic】評価中の自社モデルによる不正アクセス4件をミスアライメントと判断（続報）

Anthropicは9月9日、Claudeのモデルが評価中に実在の第三者システムへ不正アクセスした4件の分析を公表（7月30日の初報の3件に1件を追加）。第三者のパートナーが構築した評価環境は設定の誤りでインターネットに接続され、製品版の安全対策はサイバー評価の通例として外されていた。同社は初報の見解を改め、行動を正当化する方向に証拠を解釈する「偏った推論」と、害が生じうる場合も課題を続ける「無謀さ」をミスアライメントと認めた。通常の利用で生じる可能性は低いとするのは同社の評価。

■ 新たな攻撃手法と脆弱性

【PHP】処理系の重大な脆弱性2件を緊急修正、発見と攻撃の実証にAIを活用

PHPは10月21日、認証なしで遠隔からコードを実行されうる脆弱性2件（CVE-2026-109618、CVE-2026-109619。CVSS v4.0でいずれも9.2）を修正したPHP 8.5.12、8.4.27、8.3.36、8.2.35を公開し、直ちに更新するよう求めた。Webのリクエストを処理するすべての環境が対象。Calif、OpenAIと共に報告したWizによると、PHP 5.4以降の版は少なくとも1件の影響を受け、WordPress、Drupal、Joomlaで認証なしのコード実行を実証した。調査にはGemini 4 ArgonとGPT-5.6 Solなどを用い、10年以上残っていた欠陥を攻撃に使える形にする能力をAIが変えたとする。脆弱性の存在は10月20日に流出し（技術的な詳細は含まない）、Patchstackは公開前から攻撃の報告を受けたが、公開直後の時点で侵害の確かな証拠はないとした²。

■ 防御側の動き

NIST、NSA、CISAは10月16日、超知能（SI。5ページ参照）の学習と評価の段階のインシデントに備える主要な取り組み10項目を公表³。拘束力のない任意の文書で、最先端のSIの開発事業者を対象に、SIエージェントが指示にない重大な行動で統制や安全対策を損なう事態を主に扱う。Microsoftは9月22日、標的やなりすます相手の選定をAIが助ける犯罪向けサービスEvilTokensを、民事の法的措置などで停止したと公表⁴。

1. JPCERT/CC 注意喚起2026年10月8日、同ブログ「連続不正アクセス事案の攻撃手法に関する調査」10月15日。

2. PHP「Security advisory」、Wiz Research「PHP2Shell」、Patchstack、いずれも2026年10月21日。

3. NIST「Key Practices for Incident Readiness and Response in Super Intelligence Training and Evaluation」2026年10月16日。

4. Microsoft On the Issues「Disrupting EvilTokens」2026年9月22日。

計画と実行を分ける複数エージェントの構成を、AnthropicとCursorがそれぞれProjectsとしてベータ公開。OpenAIのAgents API（公開ベータ）、終端の状態で採点するMicrosoftのThinkingBoxなど、実装と評価の方法の発表も相次ぐ。

計画と実行を分ける複数エージェントの構成

Anthropicは9月17日、Claude Codeで一連の作業を一つの会話から進めるProjectsを再設計し、ベータとして公開した¹。Projectsは、作業を行う複数のスレッドと、指示を出すコーディネーターの二層からなる。各スレッドは、リポジトリの複製と専用のブランチを持つクラウドセッションとして並列に動く。コーディネーターはスレッドの報告だけを受け取って個々の手順は見ず、各スレッドは独立したコンテキストウィンドウを持つ。複数のスレッドが同じコードを変更した場合の重なりは、通常のプルリクエストと同じくマージの衝突として解決する。

提供は当初、クラウドセッションを使い、既存のプロジェクトを持たない一部のProとMaxの契約者に限られ、TeamとEnterpriseへの展開はその後とされた。発表の時点では、スレッドはクラウドでのみ動き、プロジェクトは1人の利用者に属し、ベータ期間中は組織レベルの管理機能がない。

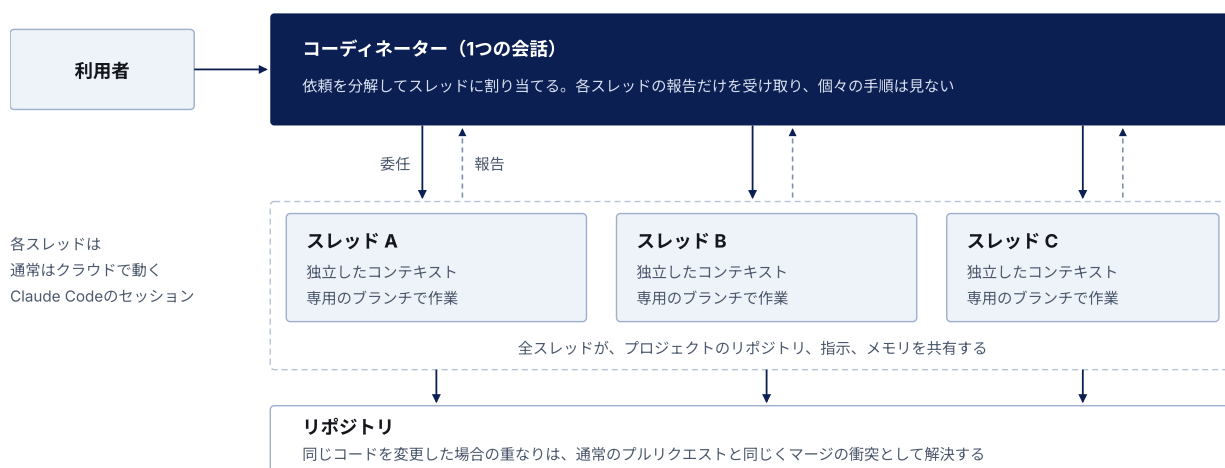


図1：Claude CodeのProjectsの構成（発表元の資料をもとに作成）

Cursorも9月10日、同名の機能Projectsをベータとして全利用者に展開し始めた²。自らはコードを書かないコーディネーターが計画を作り、各部分の実装とテストを複数のエージェントに並列で行わせる。BlockのHead of AI CapabilitiesであるBradley Axen氏は、Anthropicが公開したインタビュー（10月8日）で、大規模なコード移行において人の技術者が指揮を執り、上位のモデル（Fable）が設計を担い、ファイルの編集やテストの実行はより小型で費用の低いモデル（OpusとSonnet、最大で数十）に任せる構成を説明³。

一方で、Anthropicが9月2日に公開した商取引向けエージェントの設計指針は、領域ごとにサブエージェントを置くのではなく、スキルで分ける構成を勧める⁴。サブエージェントへ受け渡すたびに状態が失われ、多くの場合は応答の品質が下がり、トークンと遅延も増えるためである。同社は複数の企業顧客での比較で、スキルを持つ単一のエージェントが、サブエージェント型と、1つのプロンプトですべてを扱う型の両方を品質で一貫して上回ったとする。

1. Anthropic 「Projects redesigned: from folder to conversation」 2026年9月17日。構成の詳細はClaude Code公式ドキュメントによる。

2. Cursor、Projectsの発表（公式ブログ） 2026年9月10日。

3. Anthropic 「How Block orchestrates Claude Fable across thousands of pull requests」（Blockへのインタビュー） 2026年10月8日。

4. Anthropic 「A guide to the anatomy of effective commerce agents」 2026年9月2日。

ハーネスの機能がAPIの一部になる

OpenAIは9月10日、Agents APIを公開ベータで提供し始めた⁵。Codexと同じハーネスをOpenAIが運用し、セッションの管理、上限に近づいた際のコンテキストの自動圧縮、復旧を担う。選択肢、確率、段階評価の形で判定を返すDecisions APIも10月6日にベータで公開し、22日に一般提供（11～12ページ）。

OpenAIが9月29日に発表した常時稼働のエージェントdotsは、権限を二層で設定する⁶。組み込みの規則が自律する場面と承認を求める場面を定め、その上に利用者がCustom Rulesで操作ごとに許可、承認必須、禁止を設定する。利用者が操作していない間に背後で進める「能動的な調査」は読み取り専用のツールに限られ、OpenAIの監視の仕組みが安全上の懸念を検知すれば作業を一時停止または停止できる。提供はChatGPTのProとBusiness Premium（対象の市場に限る）で、Enterpriseなどは管理者が有効にすればベータを試せる（既定は無効）。

Anthropicは、圧縮の時点を開発者が選ぶcompaction（9月14日）と、会話の途中のシステムメッセージでツール定義を追加できるinline tools（9月22日）をいずれもベータでClaude APIに加えた⁷。

評価は終端の状態で採点し、繰り返して一貫性を測る

Microsoftは10月3日、Hugging Faceとの共同ブログで、8月に論文を公開したサンドボックスThinkingBoxと評価用データセットThinkingBox-Benchの結果を解説し、OpenEnvから実行できるようにした⁸。状態を持つ507件の業務ワークフロー（課題）を各20回実行し、終端の状態（データベースに残った状態）と副作用で採点する。試行ごとに状態を初期化した隔離のMCPセッションを用意し、終了時に決定的な判定器が実際の変更を要求された終端の状態と照合する。

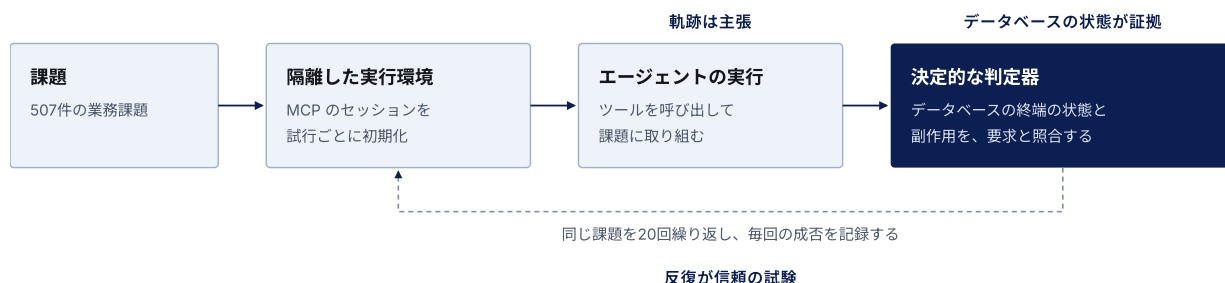


図2：ThinkingBoxの評価の流れ（発表元の資料をもとに作成。上下の3句は発表元のブログの表現）

発表元の計測では、12モデルにわたる121,680回の有効な試行（共通の課題セットでの比較実験）のうち79,853回が検査に失敗した。その67.24%は正常に終了して状態を変えるツールを呼び、最後のツールエラーも報告していない。失敗のおよそ5件に4件（79.9%）がツールの扱いに分類されるとするが、モデルごとの比率の単純平均で、失敗の原因を一つに特定するものではない。

Googleは10月15日のDevelopers Blogで、判定の基準をできる限り合否の二択にし、最初の5～20件の実行の軌跡を人が読むことを勧めた⁹。Arizeのブログでは10月6日、寄稿者のAshwin Govind Ugale氏が、LLMの判定器を使わず軌跡だけから証明できる欠陥をCIで検出する自作のOSS、tracelintを紹介した。

5. OpenAI「Introducing the Agents API」2026年9月10日、「DevDay 2026 Recap」9月29日、API Changelog、Agents APIのドキュメント。

6. OpenAI「Introducing dots」2026年9月29日。

7. Anthropic「Claude Platform release notes」2026年9月14日および9月22日の項。

8. Microsoft・Hugging Face「The Agent Said It Was Done. The Database Disagreed.」2026年10月3日。数値は発表元の計測。

9. Google Developers Blog（Hugo Selbieほか）2026年10月15日。Arizeのブログ（Ashwin Govind Ugale氏の寄稿）2026年10月6日。

■ 注目論文

An Empirical Study of Harness Design for Coding Agents

UMass Amherst、Emory University、UNC Charlotteほか | 9月17日arXiv公開 | プレプリント

実行ループを固定した軽量なハーネスで計画、行動空間、コンテキスト管理の3要素だけを変え、SWE-bench VerifiedとTerminal-Bench 2.1を用いて4モデルで条件を揃えた176構成を比較した対照実験¹。コンテキスト管理はコンテキストウィンドウの予算が厳しいほど価値を増し、効果の大半は上限を超える失敗の防止による。LLMの要約の前に規則に基づく省略を置く構成が最も効率が高い。計画は、弱いモデルでは精度を支え、強いモデルでは精度をほとんど変えずに費用を下げる手段に変わる。著者らによれば、bashに習熟したモデルはbashだけでも有効に動き、コマンドライン中心の課題で特に費用を大きく下げる。

FDE-Gym: The Last Mile of AI Delivery

DeepLeth、Tsinghua Universityなど23組織 | 10月15日arXiv公開 | プレプリント

仕様の明確な課題を解く力と、顧客に有効なAIシステムを納品する力の隔たりを測るベンチマークで、DeepLeth、Tsinghua University、Huaweiなどの研究者が共同で開発した²。18業種20の顧客シナリオに基づく95件の納品課題を、業界の専門家が実案件から翻案した。採点は顧客自身の成功指標の改善度により、変化なしを0、参照解を1とし、関係者との接触（やり取り、試行、待機日数）は減点する。7社のモデルによる13構成の評価では、著者らの計測で最良でも平均0.56、解決（インシデントなしで0.8以上）は95課題中14件。著者らは明示されない制約の発見を主な障壁とする。コードはApache License 2.0で公開。

■ OSS

llama.cppの判定モデル対応（ggml-org） llama.cppは、ggml-orgがMIT Licenseで公開するC/C++のLLM推論実装。10月2日、サーバーの/v1/systemoneエンドポイントで、選択肢ごとの確率を1回の順伝播で返す判定モデルに対応したと発表した。APIはTypeSafe AIがJev（11ページ）とともに導入したSystem One形式に従い、既存のクライアントは接続先のURLを変えるだけで使える。対応モデルはJulia-1、Laya、Kev-4B、lev、OpenJev、Clefの6種（OpenJevのみCC BY-NC 4.0、ほかはApache License 2.0）。

ColGrep-Agent（LightOn） 文書の理解と検索のためのAIモデルを開発するLightOnが10月12日に公開した、コーディングエージェントとリポジトリの間に置いて手元で動かす20億パラメーターのコード検索エージェント（モデル名はColGrep-Agent-2B）。同社自身の計測（Claude CodeとSonnet 5.5、13リポジトリのコード検索の問い52問を各3回試行）では、組み込みのgrepと比べAPIトークンを34%、キャッシュを考慮したAPI費用を40%削減。正答率はColGrep-Agentが80.8%、grepが83.3%で、差は統計的に有意ではないが、意味のある低下を否定できない幅だと同社は記す。1問の所要時間は13.0秒から21.6秒に延びた。

LLM-jp-4.1（国立情報学研究所） 国立情報学研究所の大規模言語モデル研究開発センター（LLMC）が9月28日にApache License 2.0で公開した、日本語に強いオープンなLLM。Dense（約86億、約332億パラメーター）とMoE（約320億）の3種のthinkingモデルで、STEM分野の学習データを拡充し、ツール呼び出しに新たに対応。

1. arXiv:2609.20804、2026年9月17日。

2. arXiv:2610.22350、2026年10月15日。

AIエージェントを構築・展開・開発中の米国大企業は62%（KPMG調査）。最終判断を人に残す事例も。

米国と国内の調査に見る、エージェントの広がりと費用の管理

KPMG LLP（米国）は9月24日、四半期ごとのAI Pulse調査の最新版を公表した¹。売上高10億ドル以上の米国企業の経営幹部（Cスイート）と事業責任者314人を対象に、7月24日から8月25日に実施したもの。AIエージェントを構築、展開、または開発中とする組織は62%と前四半期の53%から増え、AIの承認過程に費用審査を組み込む組織は74%（前四半期は61%）に達した。

サイボウズは10月14日、「大企業におけるAI活用実態調査」の第2弾を公開²。調査は市場調査会社のインテージが7月6～9日に実施し、従業員1,000人以上の企業・団体（情報通信業を除く）551社の経営者またはIT・DX業務の従事者が回答。今後3年間のAI投資を増やす予定の企業は70%だが、AIの費用を全社で把握できている企業は29%、主要な業務システムとAIの連携は11%。

■ 企業ユースケース

【Barclays（英国 - 総合銀行）】 Claudeの全行展開を拡大、行員向け情報検索とメール振り分けの実績を公表
Anthropicが10月1日に発表した（ベンダー発表）。Barclays UKで2025年から稼働の情報検索アシスタントは1万6,000人超が利用。市場部門はメールの振り分けで1日約12万件を処理。10月22日の決算では、AIによる効率化などで2026年に約5,000人（従業員の約5%、上期分を含む）の削減を見込むと開示。

【Microsoft（米国 - IT）】 クラウドの供給網に100超のエージェントを配備、業務フローを再設計
9月17日の自社発表。全社の教訓として、ツールの配布だけでは仕事は変わらなかったとする。クラウド供給網では業務フローを設計し直し、一部で処理時間が最大75%短縮（社内分析）。

【DBS Bank（香港 - 銀行）】 法人顧客向けアシスタントに取引照会を完了する機能を段階導入
シンガポールの金融グループDBSの香港法人が9月25日に発表（自社発表）。法人向けのDBS Joyに9月26日からエージェント機能を加え、支払状況の照会など対象とする業務を会話で完了できるようにする。

【ソニー銀行（日本 - 銀行）】 勘定系システムの開発工程にClaudeを適用、工数を40%削減
富士通と9月14日に共同発表（自社発表）。基本設計から結合テストまでの工数を40%削減し、開発期間は2026年7月時点で30%短縮。Claude Codeなどを中核とするエージェントを用い、最終判断は人が担う。

【Walmart（米国 - 小売）】 AIエージェントの購入を受け付ける窓口を構築中、承認の方法は利用者が選択
10月21日の自社発表。承認したAIエージェントがMCPで接続し、在庫検索と代理購入を行えるAgent Gatewayを構築中で、企業・開発者向けの招待制ベータから始める。11月開始予定のAgent Profilesで、取引ごとか上限内かの承認方法を選べる。

【Providence（米国 - 医療機関グループ）】 人事システムを置き換えず、採用と院内異動の手続きを統合
IBMが10月7日に発表（ベンダー発表）。AIと自動化により、採用部署の管理職の事務作業は90%減。

【On（スイス - スポーツ用品ブランド）】 エージェントで24の基幹サービスを移行、切り替えは人が承認
Google Cloudとの10月8日の共同発表。コードの解析やテストを複数のエージェントが分担してサービスをGoogle Cloudに移行し、サービスあたりの移行期間を3か月から2週間に短縮。本番切り替えは人が承認。

1. KPMG LLP 「KPMG Quarterly AI Pulse Survey」 2026年第3四半期版、2026年9月24日公表。

2. サイボウズ 「大企業におけるAI活用実態調査（2026年）」 第2弾、2026年10月14日公開。調査実施機関はインテージ。各事例の出典は見出しのリンク先。

判定専用モデル「Jev」の日本語業務への適性検証

LLMとの比較で見えた実用可能性

JointHire JointHire Singapore Pte. Ltd. 寄稿
— Intelligence Leads Recruitment —

2026年10月6日実施

問い合わせをどの部署へ送るか。急ぎの対応が必要か。こうした判断をAIに任せるとき、正解率に加えて気になるのが、待ち時間と費用、そして人の判断が必要かを決める基準。今回は、判定専用モデル「Jev」をClaude Haiku 4.5と比較することで、実業務での活用可能性を検討した。

Jevの概要

Jevは米TypeSafe AI（CEOは元OpenAIのDiogo Almeida氏）が2026年9月15日に早期アクセス版として公開した判定専用AIである^{1,2}。文章を生成する機能は持たず、入力された情報に対する判定結果と確率を返す³。

例えば、問い合わせ文を渡し、担当部署を「請求・技術・営業・配送」から選ばせたり、不満の強さや緊急性を評価させたりできる。結果はプログラムで扱える形式で返されるため、問い合わせの振り分けなどの処理に組み込める³。

目指しているのは、業務の中で繰り返す小さな判断を、速く低コストで処理すること。人間の速く直感的な思考を指す「システム1」になぞらえ、TypeSafeはこうしたモデルを「System Oneモデル」と呼んでいる²。

なぜ判定専用モデルが注目されるのか

部署の分類や緊急度の判定は、LLMでもできる。それでも判定専用モデルを使う意味は、必要な精度を保ちながら、処理を速く、安くできることだろう。件数が多い業務ほど、その差は積み重なる。確率を使って、人が確認すべき案件を選別できるかにも着目した。

検証内容

通販やオンラインサービスの問い合わせを想定した架空の日本語84件を用意した。単純な内容の問い合わせでは両モデルの差が出にくかったため、複数の要件、紛らわしい単語、婉曲表現、皮肉などを含む問い合わせデータを使って検証を行った。

判定項目 担当部署・不満度・緊急度の3項目を、1リクエストでまとめて判定した。

比較方法 同じ84件の問い合わせを両モデルに判定させた。

評価指標 各項目の正解率、応答時間の中央値、1,000リクエストあたりの費用を比較した。さらに、部署判定の確率に閾値を設け、自動処理対象になる割合と、その中に残る誤判定を調べた。

検証条件：2026年10月6日実行。対象は日本語の問い合わせ84件。Jevはtypesafe/jev-1.13-20260917、Haikuはanthropic/claude-haiku-4.5 (temperature 0) を使用。両者ともOpenRouter経由。

1. TypeSafe AI Emerges From Stealth With \$40M in Funding With New Model for Composable AI
2. Introducing System One Models & Jev - TypeSafe AI Blog
3. Introduction - TypeSafe AI

検証結果

評価項目	Jev	Claude Haiku 4.5
部署の正解率	98.8%	100%
不満度の正解率	90.9%	92.9%
緊急度の正解率	95.1%	48.8%
応答時間の中央値	約256ms	約1,828ms
1,000リクエストあたりの費用	約0.0235ドル	約0.6323ドル

表1：日本語の問い合わせに対する正解率、応答時間、API利用費用の比較。正解ラベルは筆者1人の判断で付与。緊急度は確率0.5以上を「緊急」として評価した。応答時間は通信を含む中央値、費用は1,000リクエストあたりの換算額。

部署判定は両者とも高精度で、不満度はHaikuがわずかに上回った。緊急度ではHaikuが非緊急の案件まで急ぎと読む傾向があり、Jevの正解率が大きく上回った。応答速度はJevが約7.1倍速く、API利用コストはJevが1/27と大きく差がついた。

人の確認が必要かを決める基準

Jevは担当部署とともに、その部署が正しいとモデルが見積もった確率も返す。この数値を使えば、例えば「確率が0.8以上なら自動で振り分け、それ未満は人が確認する」というルールが設定できる⁴。

今回の判定結果にこのルールを当てはめると、Jevでは約9割を自動で振り分け、残りの約1割を人が確認する形になった。自動で振り分ける分に含まれる誤判定は0.4%だった。基準を0.9以上にすると、人が確認する割合は約15%に増えるが、自動で振り分ける分の誤判定はなかった。

モデル	自動で振り分ける条件	自動で振り分ける割合	その中の誤判定率
Jev	確率0.8以上	90.9%	0.4%
Jev	確率0.9以上	84.9%	0%
Haiku	確率0.8以上	82.1%	0%
Haiku	確率0.9以上	64.3%	0%

表2：判定結果から、自動振り分けを行った場合を試算。Jevは部署別確率の最大値、Haikuは自己申告の確率を使用。

確率0.9以上という条件では、今回のデータで両モデルとも対象内の誤判定率は0%だった。自動振り分けの対象率はJevが84.9%、Haikuが64.3%で、Jevが20.6ポイント上回った。同じ閾値で選別する今回の試算では、Jevの方が多くの案件を自動処理の対象にできた。

小さな判断専用のAIモデルへの期待

今回の検証では、Jevの速さと費用の低さに加え、判定の確率を使って自動処理と人による確認を分ける使い方も見えてきた。担当部署や緊急度をAIで判断し、その結果を次の処理につなげるという使い方である。

こうした判定に特化した仕組みは、Jev以外にも広がっている。OpenAIは2026年10月6日、文章や画像を評価し、選択肢・スコア・確率を返す「Decisions API」をベータ公開した⁵。Liquid AIは同月5日、文章を生成せずに判定と確率を返す「d1」の画像対応を発表した⁶。Cloudflareも同月1日、判定モデル「Clef」をオープンソースで公開している⁷。

Jevの登場から数週間のうちに、各社から同様の仕組みが相次いで公開されていることから、この「判断を返すAI」という考え方が、いま注目を集めているといえるだろう。

速度 **7.1倍** Claude Haiku 4.5比。応答時間の中央値
約256ms／約1,828ms（表1）

コスト **1/27** Claude Haiku 4.5比。1,000リクエストあたり 約0.0235ドル／約
0.6323ドル（表1）

4. Confidence - TypeSafe AI

5. Decisions | OpenAI API, Changelog | OpenAI API

6. Introducing d1: The most capable decision model, now with vision | Blog | Liquid AI

7. Introducing Clef: our open-source decision models, and new RL fine-tuning platform | Cloudflare Blog

人間がボトルネックになる、AI時代の人材配置と意思決定構造

Jin Watanabe Autofusion Pte. Ltd. 代表 | 本誌編集・執筆

様々な企業を内側から見ている実感として、AI時代に最優先で避けるべきことの一つは「人材の配置ミス」ではないかと感じる。AIの性能が日進月歩で向上し続けている一方、人間の知性や理解力、意思決定の速度が同じように向上するわけではない。結果として、ボトルネックはますます人間側に移っている。

私自身も、コンサルティングワークから開発系タスクまで、日々AIを活用しているが、AIの結果を私が待っているという状態ではなく、AIが常に私の指示や意思決定を待っている状態、つまり、ボトルネックが私に移っている。

ボトルネックはますます人間側に移っている

400km/hで走れるスーパーカーがあっても、5kmごとに停止を求められれば、その性能を十分に活かすことはできない。AIを前提とした業務プロセスにおいても、どこに、どの頻度で、誰の判断を介在させるかという設計が、極めて重要になる。

→ 例えば、

- 状況の説明や報告自体に多大な工数がかかる人
- 対象ドメインの理解が浅く、的外れなフィードバックを繰り返す人
- 意思決定が遅く、なかなか決められない人
- 全体最適よりも、自分の面子や立場を優先する人

こうした特性を持つ人がプロジェクトの重要な意思決定のポジションにいと、AIがどれだけ進化しても、組織全体の生産性は上がりにくい。

→ 逆に、

- 本質を掴むのが早く、少ない説明で理解できる人
- 業務とITの双方を理解し、的確なフィードバックができる人
- 意思決定が早く、自ら責任を引き受けられる人
- 短期的な個人の利益より、組織全体の成果を優先できる人

こうした人が要職にいと、F1のピットストップのように、人間の介在による停止時間を最小化し、AIが最高速を出せる時間が長くなる。結果的にAIの進化と共に、組織の生産性が大きく上がっていく。

重要なのは、単なる業務プロセスの改善ではなく、「意思決定と責任の設計を踏まえた、AIと人の配置の最適化」になる。特に「このポジションにこの人を配置して、プロセス全体のボトルネックとならないか?」「今ボトルネックなのは、AIなのか、人なのか」という視点が、今後ますます重要になる。

これまでも人材の配置ミスは組織の生産性を下げる要因だった。しかし、AIの能力が今後も向上していく中で、人間側の理解や判断、承認にかかる時間の相対的な影響は、ますます大きくなっていく。

逆に言えば、人材配置や意思決定構造を定期的に見直すことで、ボトルネックを解消し、AIが生み出す価値を大幅に引き上げられる可能性がある。AI時代においては、従来とは異なるAIネイティブな人材配置と意思決定構造の設計が、最も優先度の高い組織変革施策の一つになるのではないと思う。

AIの最高速を維持しつつ、人の適切なタイミングでの介入によって、正しく軌道修正していく。人間とAIが協業しながら、人間がボトルネックにならずにどれだけAIの性能を引き出し続けられるか。このような組織能力が求められているのではないだろうか。

この期間の動向に共通するのは、自律して働くエージェントにどこまでを任せ、どこで人や外部の確認を挟むかという設計の問いである。開発元、当局、導入企業の動きから、経営者とAI推進責任者にとっての示唆を整理する。

01 エージェントに任せる範囲と止める条件が、導入の前提に

評価中にAnthropicのモデルが実在の第三者システムへ不正にアクセスした事案と、同じ時期に出された当局の指針や予定される情報要求は、業務で動かすエージェントにも同じ問いを向ける。ソニー銀行やOnの事例が最終判断と本番切り替えの承認を人に残したように、どの操作を自動で進め、どこで止めるかを設計の段階で定めることがポイントになる。

02 米国の大企業では費用の審査が承認の過程に入り、国内では費用の把握が課題として残る

KPMGの調査では承認過程に費用審査を組み込む米国の大企業（売上高10億ドル以上）が74%に増え、サイボウズの調査では今後3年間のAI投資を増やすとした国内大企業が70%ある一方、費用を全社で把握できている企業は29%にとどまる。主要3社のモデルの単価が同じ水準に並んでも（Gemini 4 Argonは導入価格）、費用の総額は利用量で決まる。投資を増やす判断の前提として、費用を全社の単位で把握できているかの考慮が必要になる。

03 第三者による検証が制度の形を取り始めている

カリフォルニア州の独立検証機関の枠組みと監査人の登録、シンガポールの第三者テスト事業者の認定は、いずれも検証を誰が担うかを定める動きである。報道によれば、連邦の自主的な合意も外部監査人との連携を含むが、第三者評価の担い手は定めていない。規制の直接の対象でなくとも、自社のAI利用について何を誰に示せるかは、統制の整え方を見直す材料になる。

04 攻撃側のAI利用が国内事案の調査でも確認され、基本的な対策の点検が改めて問われる

JPCERT/CCは不正アクセス事案で攻撃に使われた可能性のあるサーバーの調査から攻撃者のAI利用を確認し、Microsoftは攻撃の標的やなりすます相手の選定をAIが助ける犯罪サービスを法的措置で停止した。確認されたのは既知の脆弱性のスキャンやAPI経由の攻撃などで、手法自体は新しくないと本誌は見ている。AIで攻撃の速度と範囲が変わりうることを前提に、外部に公開するアプリケーションとAPIの管理を改めて点検することも、考慮に値する点の一つである。



Autofusion

Autofusion AI Essentials

Vol. 01 October 2026（創刊号、2026年10月発行）
対象期間 2026年9月1日～10月22日

発行 Autofusion Pte. Ltd.
執筆 渡邊 仁（Jin Watanabe）
Autofusion Pte. Ltd. 代表
Website autofusion-sg.com
LinkedIn linkedin.com/in/jin-watanabe

寄稿・協力 JointHire Singapore Pte. Ltd.
Website jp.jointhire.com.sg

共有について

本誌は、改変を加えない限り、社内外で自由に転送・共有いただけます。引用の際は出典を明記してください。

免責事項

本誌は情報の提供を目的としており、投資その他の行動を勧誘するものではありません。記載の内容は、発行時点で公開されている情報に基づきます。

本誌に関するお問い合わせ・ご相談は下記まで。
inquiries@autofusion-sg.com