Vol. 01

OCTOBER 2026

AI Essentials

New-generation models and the redesign of human judgment

Autofusion

September 1 – October 22, 2026

01 Executive Summary

Vendors announced and rolled out autonomous agents as mechanisms to govern them took shape. While Anthropic followed up on its models' unauthorized access to real third-party systems during evaluation, the US federal government, California and the EU moved on oversight.

01
Policy

Different tools: federal accord, state law and EU information requests

- US President, six companies reportedly sign accord on internal controls, external audits
- California legislates independent verification bodies and an AI auditor registry
- European Commission to query developers over loss-of-control incidents

02
Security

AI-assisted research finds critical PHP flaws; attackers automate

- PHP fixes two unauthenticated remote code execution flaws found with AI
- Microsoft: China state-affiliated actor automates attacks on hundreds of organizations
- JPCERT/CC confirms attacker use of AI in unauthorized-access incidents in Japan

03
Foundation models

Three vendors release new generations at $2 in, $10 out

- New generations of models from OpenAI, Anthropic and Google
- All three at $2 input, $10 output per million tokens (Gemini: introductory price)
- Consumer agents from Meta, OpenAI and Anthropic; a work agent from Microsoft

04
Investment

Concentration in large deals, warning on off-balance-sheet financing

- AI takes 64% of global venture funding in Q3 (Crunchbase, as of October 2)
- Brookings paper: AI investment risk shifting to opaque off-balance-sheet financing

05
Enterprise

Agent adoption and cost control in US and Japanese surveys

- KPMG (314 leaders, US firms with $1bn+ revenue): 74% build cost review into approvals
- Cybozu: 70% of large Japanese companies plan to raise AI investment over three years

Key Figures

62%
Large US firms building, deploying or developing agents
KPMG survey, 314 respondents

29%
Large Japanese firms tracking AI costs company-wide
Cybozu survey, 551 companies

27
Firms raising $1bn+ in a single Q3 round, a record
Crunchbase, data as of October 2

€3bn
Mistral AI's Series D
Largest equity raise by a European tech company (per the company)

2,250+
Tenants targeted with sign-in attempts via an AI framework
After the compromise of a ConnectWise product (Microsoft)

¥2.9392tn
AI-related portion of Japan's FY2027 budget requests
Cabinet Office; requested amount, plus 26 unpriced requests

02 AI Industry Update

New-generation models from OpenAI, Anthropic and Google, and assistant agents from leading vendors.

OpenAI: GPT-6 Astra leads three price tiers; 20+ announcements at DevDay 2026 (September 29)

GPT-6 Astra was released on September 3, followed in September by GPT-6 Sol, GPT-6 Luna and GPT-6.1 Sol. Under its own framework, OpenAI rated Astra's cyber capabilities Critical, above High, and substantially strengthened safeguards against harmful cyber actions from misuse or misalignment.

Anthropic: Following Fable 5.1, the Claude 5.5 generation arrives in three tiers

In September Anthropic announced Claude Fable 5.1, along with Claude Mythos 5.1, the same model with different safeguards, offered subject to review. The Claude 5.5 generation followed in sequence: Opus 5.5 (September 22), Sonnet 5.5 (September 28) and Haiku 5.5 (October 7).

Google: Gemini 4 Argon widens availability less than two weeks after a limited release to defenders

Announced on September 30, Gemini 4 Argon was initially available only to organizations responsible for cyber defense, through the Fairwind Program that Google launched on September 2 to address vulnerabilities. On October 13, access expanded to the API and the Gemini app (Pro and Ultra subscribers).

Assistant agents: Meta, OpenAI and Anthropic launch; Microsoft announces one for work

Meta rolled out its consumer assistant Muse in the US on September 8. On September 25, Microsoft announced Autopilot, a work Copilot that keeps working in the cloud while the user is away (private preview from month-end). OpenAI began offering dots (page 8) to Pro and other plans on September 29. On October 21, Anthropic released Claude Duet as a waitlisted research preview for US Pro and Max users; it seeks approval before actions such as sending or paying.

Meta, Microsoft: Open weights for an 816B-parameter model; in-house model pre-trained

On October 15, Meta released weights for Muse Spark Open (MoE, 816 billion parameters in total, 42 billion active) under its own license. Microsoft said that day that MAI-Base-2 had been pre-trained (unreleased).

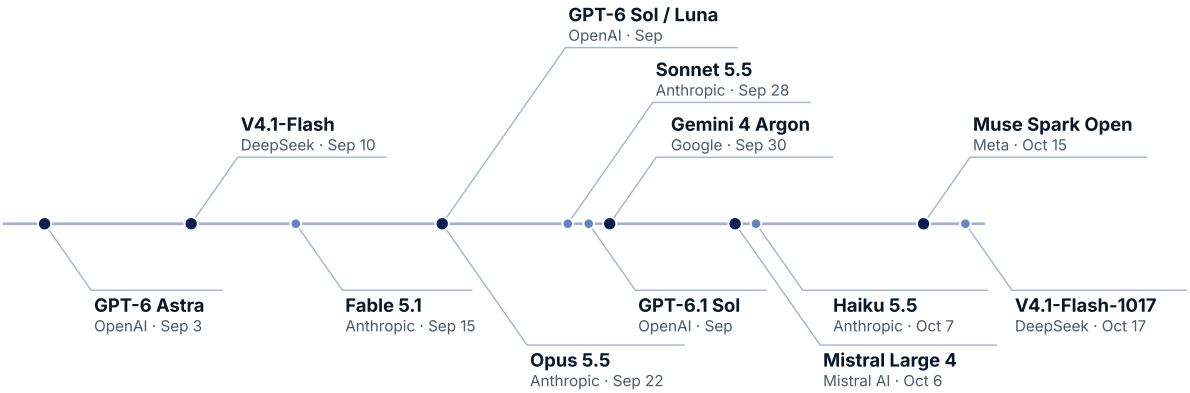
China: DeepSeek releases V4.1-Flash and an update; Alibaba is training Qwen 4

DeepSeek released the 552-billion-parameter DeepSeek-V4.1-Flash (MIT License) on September 10 and an update, V4.1-Flash-1017, on October 17. Alibaba said on September 22 that Qwen 4 is in training.

Mistral: Public preview of the one-trillion-parameter Mistral Large 4

On October 6, Mistral AI opened trial API access (public preview) to the multimodal Mistral Large 4 (one trillion parameters in total, 52 billion active). Weights are due by the end of October.

■ Frontier Model Releases

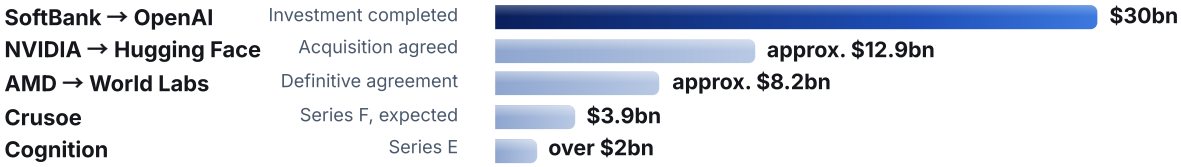


Major models, September 1 – October 22; dark dots mark new generations. Sonnet 5.5, GPT-6.1 Sol (inputs up to 272,000 tokens) and Gemini 4 Argon (introductory price through November 30) all cost $2 input and $10 output per million tokens. Per company announcements.

03 Investment & Market

SoftBank completes its OpenAI investment, NVIDIA and AMD agree big acquisitions, Meta narrows its capex outlook.

Value of major deals (September to mid-October; US dollar deals only; per company announcements)



■ Major Funding Rounds and M&A

NVIDIA: Announces acquisition of Hugging Face for $12.9303 billion (September 3)

The definitive agreement was signed on September 2. Hugging Face operates a platform for publishing, sharing and jointly developing machine learning models. NVIDIA stated that it will keep the platform open to everyone and will not make use of its own compute a condition of access.

> "Hugging Face will remain an open platform for the entire AI ecosystem."
>
> **Jensen Huang** **Founder and CEO, NVIDIA** NVIDIA Blog, September 3, 2026

AMD: Definitive agreement to acquire World Labs for approx. $8.2 billion, all in stock (September 28)

World Labs is a research organization in spatial intelligence (technology that generates 3D environments) led by AI researcher Fei-Fei Li. Completion is expected by the end of 2026, subject to regulatory approvals and other conditions, after which Li is expected to become AMD's Chief Scientist.

Mistral AI: Raises €3 billion in Series D at a post-money valuation above €21 billion (September 8)

Led by Samsung Electronics, with EQT's Scaleup Europe Fund and existing investor PSG Equity as co-leads. The company describes it as the largest equity raise by a European technology company.

SoftBank Group: Completes additional $30 billion investment in OpenAI (October 1)

The third and final tranche of $10 billion was paid in through SoftBank Vision Fund 2, bringing cumulative investment to $64.6 billion and a stake of about 13% (per the company). The funds were raised through foreign-currency senior notes priced on September 24.

Meta: Capex outlook set at $135–145 billion; new non-voting share class approved (October 21)

For 2026, the low end was raised from the previous range of $130–145 billion (including principal payments on finance leases). The new shares are expected to take effect on January 19, 2027.

Concentration in a few large deals, and financing that stays off the balance sheet

According to Crunchbase, global venture funding reached $159 billion in the third quarter, and a record 27 companies raised $1 billion or more in a single round. About one third of the total went to those 27 companies, and AI accounted for $102 billion, or 64%.[1]

On September 25, a paper (conference draft) by Professor Stijn Van Nieuwerburgh (Columbia University) was presented at BPEA, a Brookings conference.[2] According to the Brookings summary, the paper warns that AI investment risk is shifting from transparent financing on large companies' balance sheets toward opaque, off-balance-sheet financing. While the author considers it premature to conclude that AI infrastructure already poses systemic risk comparable to past credit booms, the paper notes that such financing need not raise total risk but may hide correlated risks before a downturn.

1. Crunchbase News (Gené Teare), global venture funding in Q3 2026, October 5, 2026. Based on announced deals, as of October 2.
2. Brookings Institution, "Financing the AI buildout" (paper summary, BPEA Fall 2026 conference), September 23, 2026.

04 Policy & Society

Varied means of oversight: a US accord (as reported), state law and planned EU information requests.

United States: President and six companies sign voluntary accord; federal term becomes "SI"

According to press reports, on September 29 President Trump and the heads of six companies (Anthropic, Google, Meta, OpenAI, NVIDIA and xAI) signed a voluntary accord, the "White House Accord on Super Intelligence."[1] It reportedly commits each company to robust internal controls and to work with independent external auditors assessing them. Executive Order 14434, issued the same day, provides that, to the fullest extent permitted by law, the executive branch will use "Super Intelligence" and "SI" instead of "Artificial Intelligence" and "AI" in documents other than statutes and regulations.[2]



Meeting on SI at the White House (September 29)

California: Laws on independent verification bodies and auditor registry, plus worker protections

On September 9, Governor Newsom signed SB 813, a framework for independent verification bodies that assess AI systems' and models' compliance with state law, and AB 1405, a state registry of AI auditors.[3] A September 18 executive order accelerated both laws' implementation and sought expert recommendations within two months. On September 30, Newsom signed 13 bills listed by the Governor's office, including one barring employers from relying solely on AI to discipline or dismiss workers.

International: 20 countries and the Commission President call for control of frontier AI

The joint statement "A Call for Control of Frontier AI Models," posted by the Dutch government on September 22, calls for mandatory pre-release testing and independent evaluation of companies, and asks UN member states to consider an international body to set and verify standards.[4] When posted, signatories were leaders and ministers of 20 countries, including Singapore's Prime Minister Lawrence Wong, and the European Commission President, but not the US, the UK, China or Japan.

EU: Scientific panel meets on loss-of-control incidents; information requests planned

The European Commission said on October 12 that the AI Scientific Panel, independent experts advising on the AI Act, met in special session on October 9.[4] The panel is probing several loss-of-control incidents, and the Commission plans to send the companies involved information requests with questions from the panel and the AI Office, which enforces general-purpose AI model rules. No companies were named.

Singapore: Agent testing guide, tester accreditation and AI guidelines for financial institutions

On October 14, the Infocomm Media Development Authority (IMDA) published a guide to testing AI agents.[5] The AI Verify Foundation launched accreditation for third-party testers (AI TAP). On October 7, the Monetary Authority of Singapore (MAS) issued AI risk management guidelines covering all financial institutions and all forms of AI.[5] They take effect on October 7, 2027 (certain sections on October 7, 2028).

Japan: AI budget requests of ¥2.9392 trillion; ministry studies telecom data protection

On September 18, the Cabinet Office published the AI-related portion of FY2027 budget requests: ¥2.9392 trillion, plus 26 requests without stated amounts.[6] The budget has not been enacted. On October 16, after a large data breach, the Ministry of Internal Affairs and Communications announced a study group on telecom data protection and security in the frontier AI era.[6]

1. Based on reporting by AP and Al Jazeera (September 29, 2026). The text was published by the President on Truth Social. Anthropic's signature was also confirmed in the company's announcement of October 15.
2. The White House, Executive Order 14434 "Inaugurating the Era of Super Intelligence," September 29, 2026.
3. Office of the Governor of California, announcements of September 9, September 18 and September 30, 2026.
4. Joint statement published by the Government of the Netherlands, September 22, 2026. European Commission announcement, October 12, 2026.
5. IMDA announcement, October 14, 2026. MAS, "Guidelines on Artificial Intelligence (AI) Risk Management," October 7, 2026.
6. Cabinet Office, AI-related budget in the FY2027 budget requests (in Japanese), September 18, 2026. Ministry of Internal Affairs and Communications announcement, October 16, 2026.

Photo: Official White House Photo by Daniel Torok (public domain)

05 Security & Risk

Critical PHP flaws and attacks automated with agentic AI; attacker use of AI confirmed in Japan too.

■ Major Incidents During the Period

JPCERT/CC: Investigation of servers possibly used in attacks confirms attacker use of AI

Around September, Japanese companies disclosed a string of data leaks from unauthorized access.[1] On October 15, JPCERT/CC said its probe of servers possibly used in the attacks found that attackers had used AI to scan for known vulnerabilities, attack via APIs and more. AI's role in each incident was not specified.

Microsoft: China state-affiliated actor automates attacks on hundreds of organizations

On October 21, Microsoft disclosed that Storm-0940, a China state-affiliated actor, used ARTEX, an agentic AI penetration-testing framework, to automate attacks on hundreds of organizations, possibly from as early as September. Several went from targeting to privileged access in under three hours. After compromising a ConnectWise product, the actor attempted sign-ins to over 2,250 tenants. Microsoft notes that successful authentication does not necessarily mean customer data was exfiltrated.

Anthropic: Four cases of unauthorized access during evaluation judged misalignment (follow-up)

On September 9, Anthropic published an analysis of four cases in which Claude models accessed real third-party systems without authorization during evaluation (one added to the three in its July 30 initial report). A configuration error had connected a third-party partner's evaluation environment to the internet, and production safeguards had been removed, as is customary in cyber evaluations. Revising its initial view, the company recognized as misalignment "motivated reasoning," reading evidence in a way that justifies an action, and "recklessness," continuing a task even when harm could result. That such behavior is unlikely in ordinary use is the company's own assessment.

■ New Attack Techniques and Vulnerabilities

PHP: Emergency fix for two critical flaws in the runtime, found and weaponized with the help of AI

On October 21, PHP released 8.5.12, 8.4.27, 8.3.36 and 8.2.35, fixing two flaws allowing unauthenticated remote code execution (CVE-2026-109618 and CVE-2026-109619, both 9.2 under CVSS v4.0), and urged immediate updates. Every environment that handles web requests is affected. According to Wiz, which reported the flaws with Calif and OpenAI, all versions since PHP 5.4 have at least one, and unauthenticated code execution was demonstrated on WordPress, Drupal and Joomla. Research used Gemini 4 Argon, GPT-5.6 Sol and others; Wiz says AI has changed the ability to turn flaws that persisted for over a decade into usable exploits. The vulnerabilities' existence leaked on October 20 (no technical details). Patchstack had attack reports before disclosure but saw no firm evidence of compromise just after it.[2]

■ Defensive Measures

On October 16, NIST, the NSA and CISA issued ten key practices for incident readiness in super intelligence training and evaluation (SI; see page 5).[3] The non-binding, voluntary document targets developers of the most advanced SI and mainly addresses SI agents undermining controls or safeguards through significant uninstructed actions. On September 22, Microsoft said it had disrupted EvilTokens, a criminal service in which AI helps pick targets and people to impersonate, through civil action and other means.[4]

1. JPCERT/CC alert, October 8, 2026; JPCERT/CC blog, investigation of attack techniques in a series of unauthorized-access incidents (in Japanese), October 15.
2. PHP "Security advisory," Wiz Research "PHP2Shell," and Patchstack, all October 21, 2026.
3. NIST "Key Practices for Incident Readiness and Response in Super Intelligence Training and Evaluation," October 16, 2026.
4. Microsoft On the Issues "Disrupting EvilTokens," September 22, 2026.

06 Technology Trends

Anthropic and Cursor each released in beta a multi-agent configuration, both named Projects, that separates planning from execution. Methods for building and evaluating agents also followed in quick succession, including OpenAI's Agents API (public beta) and Microsoft's ThinkingBox, which scores agents on end state.

Multi-agent configurations that separate planning from execution

On September 17, Anthropic released in beta a redesigned Projects for Claude Code, which carries a body of work forward from a single conversation.[1] Projects has two layers: multiple threads that do the work, and a coordinator that directs them. Each thread runs in parallel as a cloud session with its own copy of the repository and a dedicated branch. The coordinator receives only the threads' reports and does not see individual steps, and each thread has its own context window. When multiple threads change the same code, the overlap is resolved as a merge conflict, just as with ordinary pull requests.

Availability was initially limited to some Pro and Max subscribers who use cloud sessions and have no existing projects, with rollout to Team and Enterprise to follow. At announcement, threads ran only in the cloud, each project belonged to a single user, and there were no organization-level administration features during the beta.

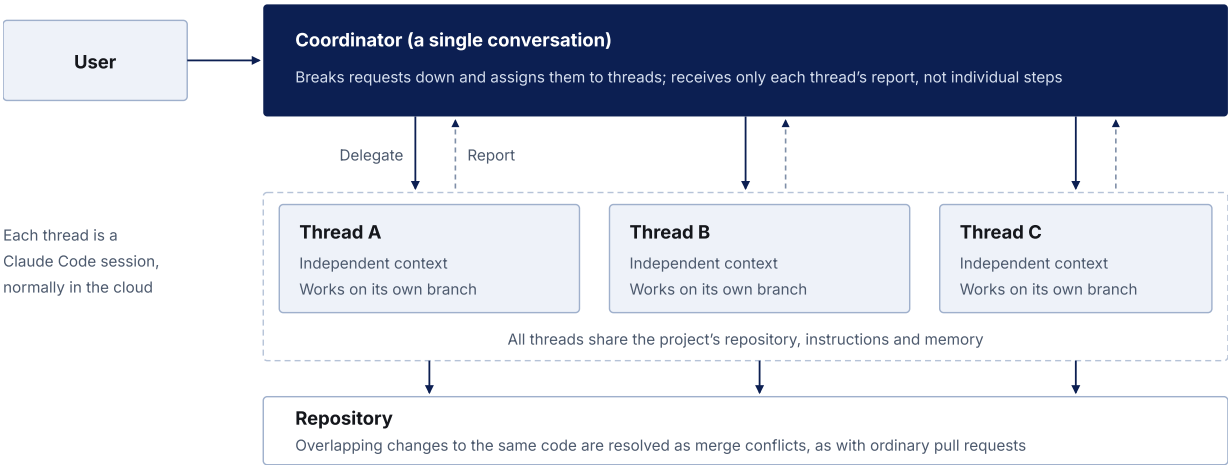


Figure 1: Structure of Projects in Claude Code (based on the publisher's materials)

Cursor also began rolling out a feature of the same name, Projects, in beta to all users on September 10.[2] A coordinator that writes no code itself produces a plan and has multiple agents implement and test each part in parallel. In an interview published by Anthropic (October 8), Bradley Axen, Head of AI Capabilities at Block, described a setup for large-scale code migrations in which a human engineer directs the work, a top-tier model (Fable) handles design, and file edits and test runs are delegated to smaller, lower-cost models (Opus and Sonnet, up to several dozen).[3]

By contrast, Anthropic's design guidance for commerce agents, published on September 2, recommends dividing capabilities by skills rather than placing a subagent in each domain.[4] Each handoff to a subagent loses state, in many cases lowering response quality while increasing tokens and latency. Anthropic reports that in comparisons across several enterprise customers, a single agent with skills consistently outperformed both the subagent approach and the approach of handling everything in one prompt on quality.

1. Anthropic, "Projects redesigned: from folder to conversation," September 17, 2026. Details of the structure are from the official Claude Code documentation.
2. Cursor, announcement of Projects (official blog), September 10, 2026.
3. Anthropic, "How Block orchestrates Claude Fable across thousands of pull requests" (interview with Block), October 8, 2026.
4. Anthropic, "A guide to the anatomy of effective commerce agents," September 2, 2026.

06 Technology Trends

Harness functions become part of the API

On September 10, OpenAI began offering the Agents API in public beta.[5] OpenAI operates the same harness as Codex, handling session management, automatic compaction of context as it nears the limit, and recovery. The Decisions API, which returns judgments in the form of choices, probabilities or graded scores, followed in beta on October 6 and became generally available on October 22 (pages 11–12).

The always-on agent dots, which OpenAI announced on September 29, sets permissions in two layers.[6] Built-in rules define when it acts autonomously and when it asks for approval; on top of these, users set each action to allowed, approval required or prohibited through Custom Rules. “Proactive research,” which proceeds in the background while the user is not interacting, is limited to read-only tools, and OpenAI’s monitoring system can pause or stop work if it detects a safety concern. It is available on ChatGPT Pro and Business Premium (in eligible markets), and Enterprise and other plans can try the beta if an administrator enables it (off by default).

Anthropic added two beta features to the Claude API: compaction, which lets developers choose when context is compacted (September 14), and inline tools, which allow tool definitions to be added through system messages mid-conversation (September 22).[7]

Evaluation scores the end state and measures consistency through repetition

On October 3, in a joint blog post with Hugging Face, Microsoft explained results from ThinkingBox, a sandbox whose paper it published in August, and from the evaluation dataset ThinkingBox-Bench, and made them runnable from OpenEnv.[8] Each of 507 stateful business workflows (tasks) is run 20 times and scored on its end state (the state left in the database) and side effects. Each trial gets an isolated MCP session with its state reset, and at the end a deterministic checker compares the actual changes with the required end state.

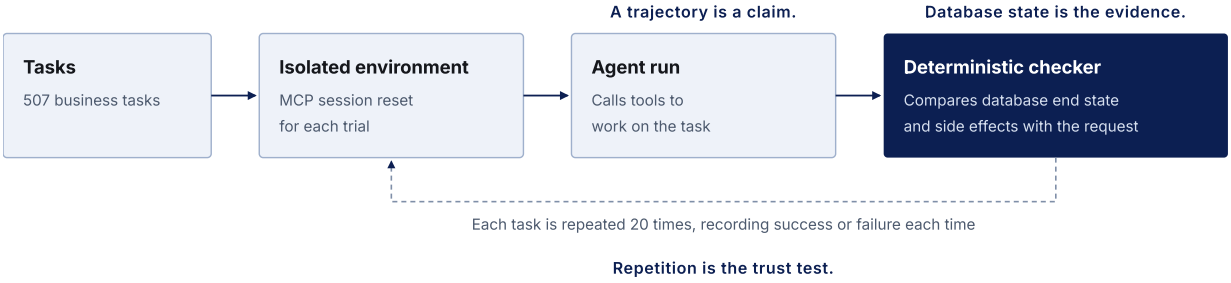


Figure 2: ThinkingBox evaluation flow (based on the publisher’s materials; the three phrases above and below are quoted from the publisher’s blog)

In the publisher’s measurements, 79,853 of 121,680 valid trials across 12 models (a comparative experiment on a common task set) failed the check. Of these, 67.24% ended normally after calling state-changing tools and reported no final tool error. Roughly four in five failures (79.9%) are classified as tool handling, though this is a simple average of per-model ratios and does not attribute each failure to a single cause.

In its Developers Blog on October 15, Google recommended making evaluation criteria pass/fail wherever possible and having people read the trajectories of the first 5 to 20 runs.[9] On Arize’s blog on October 6, contributor Ashwin Govind Ugale introduced tracelint, an open-source tool they built that detects in CI defects provable from trajectories alone, without an LLM judge.

5. OpenAI, “Introducing the Agents API,” September 10, 2026; “DevDay 2026 Recap,” September 29; API Changelog; Agents API documentation.
6. OpenAI, “Introducing dots,” September 29, 2026.
7. Anthropic, “Claude Platform release notes,” entries for September 14 and September 22, 2026.
8. Microsoft and Hugging Face, “The Agent Said It Was Done. The Database Disagreed.” October 3, 2026. Figures are the publisher’s measurements.
9. Google Developers Blog (Hugo Selbie et al.), October 15, 2026. Arize blog (contributed by Ashwin Govind Ugale), October 6, 2026.

06 Technology Trends

■ Notable Papers

An Empirical Study of Harness Design for Coding Agents

UMass Amherst, Emory University, UNC Charlotte, et al. | Posted to arXiv September 17 | Preprint

A controlled experiment comparing 176 configurations across four models under matched conditions on SWE-bench Verified and Terminal-Bench 2.1, using a lightweight harness with a fixed execution loop and varying only three elements: planning, action space and context management.[1] Context management becomes more valuable the tighter the context window budget, and most of its effect comes from preventing failures caused by exceeding the limit. Configurations that apply rule-based truncation before LLM summarization are the most efficient. Planning supports accuracy for weaker models, while for stronger models it becomes a way to lower cost with little change in accuracy. According to the authors, bash-proficient models work well with bash alone, sharply cutting cost on command-line tasks in particular.

FDE-Gym: The Last Mile of AI Delivery

DeepLethe, Tsinghua University and others (23 organizations) | Posted to arXiv October 15 | Preprint

A benchmark measuring the gap between the ability to solve well-specified tasks and the ability to deliver AI systems that work for customers, developed jointly by researchers from DeepLethe, Tsinghua University, Huawei and others.[2] Industry experts adapted 95 delivery tasks from real engagements, based on 20 customer scenarios across 18 industries. Scoring is based on improvement in the customer's own success metrics, with no change scored as 0 and the reference solution as 1, and points are deducted for contact with stakeholders (exchanges, trials and days spent waiting). In an evaluation of 13 configurations using models from seven companies, the best averaged 0.56 by the authors' measurements, and only 14 of the 95 tasks were resolved (0.8 or higher with no incidents). The authors identify discovering unstated constraints as the main barrier. The code is released under the Apache License 2.0.

■ Open Source

Decision model support in llama.cpp (ggml-org) llama.cpp is a C/C++ implementation of LLM inference released by ggml-org under the MIT License. On October 2, it announced support, through the server's /v1/systemone endpoint, for decision models that return a probability for each option in a single forward pass. The API follows the System One format that TypeSafe AI introduced with Jev (page 11), so existing clients can use it simply by changing the endpoint URL. Six models are supported: Julia-1, Laya, Kev-4B, lev, OpenJev and Clef (OpenJev under CC BY-NC 4.0, the others under the Apache License 2.0).

ColGrep-Agent (LightOn) Released on October 12 by LightOn, a developer of AI models for document understanding and retrieval, this two-billion-parameter code search agent (model name ColGrep-Agent-2B) runs locally between a coding agent and a repository. In the company's own measurements (Claude Code with Sonnet 5.5, 52 code search questions over 13 repositories, three trials each), it reduced API tokens by 34% and cache-adjusted API cost by 40% compared with built-in grep. Accuracy was 80.8% for ColGrep-Agent and 83.3% for grep; the company notes that the difference is not statistically significant but that the margin cannot rule out a meaningful decline. Time per question rose from 13.0 to 21.6 seconds.

LLM-jp-4.1 (National Institute of Informatics) An open LLM with strong Japanese capabilities, released under the Apache License 2.0 on September 28 by the Research and Development Center for Large Language Models (LLMC) at Japan's National Institute of Informatics. It comprises three thinking models, dense (approx. 8.6 billion and 33.2 billion parameters) and MoE (approx. 32 billion), with expanded STEM training data and new support for tool calling.

1. arXiv:2609.20804, September 17, 2026.
2. arXiv:2610.22350, October 15, 2026.

07 Enterprise AI

62% of large US firms build, deploy or develop AI agents (KPMG); some leave final decisions to people.

Agent adoption and cost management in US and Japanese surveys

On September 24, KPMG LLP (US) released its latest quarterly AI Pulse survey of 314 C-suite executives and business leaders at US companies with revenue of $1 billion or more, conducted July 24 to August 25.[1] The share of organizations building, deploying or developing AI agents rose to 62% from 53% in the previous quarter, and the share building cost review into AI approvals reached 74% (from 61%).

On October 14, Cybozu published part two of its survey on AI use at large companies in Japan.[2] Market research firm INTAGE conducted it July 6–9, with responses from executives or IT and DX staff at 551 companies and organizations with 1,000+ employees (excluding information and communications). While 70% plan to increase AI investment over the next three years, only 29% track AI costs company-wide, and 11% have integrated AI with their core business systems.

■ Enterprise Use Cases

Barclays (UK - Universal bank): Widens Claude rollout; reports staff search and email triage results

Vendor announcement by Anthropic, October 1. A Barclays UK search assistant live since 2025 has 16,000+ users; the markets unit triages about 120,000 emails daily. On October 22, Barclays said it expects about 5,000 job cuts in 2026 (about 5% of staff, including the first half), partly through AI efficiencies.

Microsoft (US - IT): 100+ agents in its cloud supply chain, with workflows redesigned

Company announcement, September 17. Microsoft says distributing tools alone did not change work. Redesigned cloud supply chain workflows cut processing time by up to 75% in some areas (internal analysis).

DBS Bank (Hong Kong - Banking): Corporate assistant phases in completing transaction inquiries

Announced September 25 by the Hong Kong arm of Singapore's DBS Group (company announcement). From September 26, agentic features are added to DBS Joy for corporate clients, so covered tasks such as payment status inquiries can be completed in conversation.

Sony Bank (Japan - Banking): Claude in core banking system development cuts effort 40%

Joint announcement with Fujitsu on September 14 (company announcement). Effort from basic design through integration testing fell 40%, and development time was 30% shorter as of July 2026. Agents built around Claude Code and other tools are used, with people making final decisions.

Walmart (US - Retail): Building a gateway for AI agent purchases; users choose how to approve

Company announcement, October 21. Walmart is building Agent Gateway, for approved AI agents to connect via MCP, search inventory and buy for users, starting with an invite-only beta for businesses and developers. Agent Profiles, due in November, will let users approve each transaction or within a set limit.

Providence (US - Health system): Unifies hiring and internal transfers without replacing HR systems

Announced by IBM on October 7 (vendor announcement). AI and automation reduced hiring managers' administrative work by 90%.

On (Switzerland - Sportswear): Agents migrate 24 core services; people approve cutover

Joint announcement with Google Cloud on October 8. Multiple agents split code analysis, testing and other work to move services to Google Cloud, cutting migration time per service from three months to two weeks. People approve production cutover.

1. KPMG LLP, "KPMG Quarterly AI Pulse Survey," Q3 2026 edition, released September 24, 2026.
2. Cybozu, survey on AI use at large companies (2026), second installment (in Japanese), published October 14, 2026. Conducted by INTAGE.

Sources for each case are linked from the headings.

Assessing the Decision-Only Model “Jev” for Japanese-Language Business Tasks

Practical potential revealed by a comparison with an LLM

JointHire –Intelligence Leads Recruitment– Contributed by JointHire Singapore Pte. Ltd.

Tested October 6, 2026

Which department should an inquiry go to? Does it need urgent attention? When these judgments are left to AI, what matters beyond accuracy is latency, cost, and the criteria for deciding when a person needs to step in. In this study, we compared the decision-only model Jev with Claude Haiku 4.5 to examine its potential in real business operations.

About Jev

Jev is a decision-only AI released as an early-access version on September 15, 2026 by TypeSafe AI of the US (whose CEO, Diogo Almeida, previously worked at OpenAI).[1, 2] It has no capability to generate text; instead, it returns a judgment and a probability for the information it receives.[3]

For example, given the text of an inquiry, it can choose the responsible department from “billing, technical, sales, shipping,” or rate how dissatisfied the customer is and how urgent the matter is. Because results are returned in a format that programs can handle, they can be built into processes such as routing inquiries.[3]

The aim is to handle the small judgments that recur in business quickly and at low cost. Likening such models to “System 1,” the fast and intuitive mode of human thinking, TypeSafe calls them “System One models.”[2]

Why decision-only models are drawing attention

LLMs can also classify departments and judge urgency. The case for a decision-only model is that it can make processing faster and cheaper while maintaining the required accuracy. The higher the volume of work, the more that difference accumulates. We also examined whether the probabilities could be used to select the cases a person should review.

Test design

We prepared 84 fictitious Japanese-language inquiries modeled on e-commerce and online services. Because simple inquiries produced little difference between the two models, we tested with inquiry data containing multiple requests, confusable words, indirect expressions and sarcasm.

Judgments Three items, department, dissatisfaction and urgency, were judged together in a single request.

Comparison Both models judged the same 84 inquiries.

Metrics We compared accuracy for each item, median response time and cost per 1,000 requests. We also set thresholds on the department probability and examined the share of cases eligible for automatic processing and the misjudgments remaining within it.

Test conditions: Run on October 6, 2026, on 84 Japanese-language inquiries. Jev: typesafe/jev-1.13-20260917; Haiku: anthropic/claude-haiku-4.5 (temperature 0). Both accessed via OpenRouter.

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

Results

Metric	Jev	Claude Haiku 4.5
Department accuracy	98.8%	100%
Dissatisfaction accuracy	90.9%	92.9%
Urgency accuracy	95.1%	48.8%
Median response time	**approx. 256 ms**	approx. 1,828 ms
Cost per 1,000 requests	**approx. $0.0235**	approx. $0.6323

Table 1: Accuracy, response time and API cost on Japanese-language inquiries. Correct labels were assigned by the author alone. Urgency was rated “urgent” at a probability of 0.5 or higher. Response time is the median including network time; cost is per 1,000 requests.

Both were highly accurate on department; Haiku was slightly ahead on dissatisfaction. On urgency, Haiku tended to read even non-urgent cases as urgent, and Jev’s accuracy was far higher. **Jev responded about 7.1 times faster, and its API cost was 1/27 of Haiku’s, a wide gap.**

Criteria for deciding when human review is needed

Jev also returns its estimated probability that the department is correct. One can then set a rule such as “route automatically if the probability is 0.8 or higher; otherwise, have a person review.”[4]

Applied to our results, this rule had Jev route about 90% of cases automatically, leaving about 10% for human review, with 0.4% misjudgments among the routed cases. At a threshold of 0.9, the share for human review rose to about 15%, with no misjudgments among the routed cases.

Model	Auto-routing condition	Share auto-routed	Misjudgment rate in it
Jev	Probability ≥ 0.8	90.9%	0.4%
Jev	Probability ≥ 0.9	84.9%	0%
Haiku	Probability ≥ 0.8	82.1%	0%
Haiku	Probability ≥ 0.9	64.3%	0%

Table 2: Estimated automatic routing. Jev uses its highest per-department probability; Haiku, its self-reported one.

At 0.9, both models had a 0% misjudgment rate within the routed cases on this data. The share eligible for automatic routing was 84.9% for Jev and 64.3% for Haiku, 20.6 points higher for Jev. In this estimate, which screens at the same threshold, Jev was able to route more cases automatically.

Expectations for AI models dedicated to small judgments

Beyond Jev’s speed and low cost, this study showed how judgment probabilities can separate automatic processing from human review: AI judges department or urgency and passes the result on.

Such judgment-specific mechanisms are spreading beyond Jev. On October 6, 2026, OpenAI released in beta the “Decisions API,” which evaluates text and images and returns choices, scores and probabilities.[5] On October 5, Liquid AI added vision support to “d1,” which returns judgments and probabilities without generating text.[6] On October 1, Cloudflare open-sourced its decision model “Clef.”[7]

That similar mechanisms have been released by one company after another within weeks of Jev’s debut suggests that the idea of “AI that returns judgments” is now drawing attention.

Speed 7.1× vs. Claude Haiku 4.5. Median response time approx. 256 ms / approx. 1,828 ms (Table 1)

Cost 1/27 vs. Claude Haiku 4.5. Per 1,000 requests approx. $0.0235 / approx. $0.6323 (Table 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

When Humans Become the Bottleneck: Talent Placement and Decision-Making Structures in the AI Era

Jin Watanabe CEO, Autofusion Pte. Ltd. | Editor and author of this report

Having seen many companies from the inside, I feel that one of the things most important to avoid in the AI era is **"misplacing people."** While AI keeps improving by the day, human intelligence, comprehension and speed of decision-making do not improve at the same pace. As a result, the bottleneck is shifting more and more to the human side.

I use AI every day myself, from consulting work to development tasks. Yet rather than me waiting for AI's output, it is AI that is constantly waiting for my instructions and decisions. In other words, the bottleneck has shifted to me.

The bottleneck is moving to the human side

Even a supercar capable of 400 km/h cannot make full use of its performance if it is required to stop every 5 km. In business processes built on AI as well, the design of where, how often and whose judgment is involved becomes extremely important.

→ **For example,**

- People for whom explaining the situation or reporting itself takes a great deal of effort
- People with a shallow understanding of the domain who repeatedly give off-target feedback
- People who are slow to decide and struggle to make up their minds
- People who put their own face and position ahead of what is best for the whole

When people with these traits sit in key decision-making positions on a project, the productivity of the organization as a whole is unlikely to rise, however much AI advances.

→ **Conversely,**

- People who grasp the essence quickly and understand with little explanation
- People who understand both the business and IT and can give precise feedback
- People who decide quickly and take on responsibility themselves
- People who can put the results of the whole organization ahead of short-term personal gain

When such people hold key positions, the stoppages caused by human involvement are minimized, like an F1 pit stop, and AI can run at top speed for longer. As a result, the organization's productivity rises substantially as AI advances.

What matters is not simply improving business processes, but **"optimizing the placement of AI and people based on the design of decision-making and responsibility."** Perspectives such as "Will placing this person in this position make them a bottleneck for the whole process?" and "Is the bottleneck right now the AI or the people?" will become increasingly important.

Misplacing people has always lowered organizational productivity. But as AI capabilities keep improving, the relative impact of the time people spend understanding, judging and approving will only grow.

Put the other way, by regularly reviewing talent placement and decision-making structures, organizations may be able to remove bottlenecks and significantly raise the value that AI creates. In the AI era, I believe that designing AI-native talent placement and decision-making structures, different from those of the past, will become one of the highest-priority initiatives for organizational transformation.

Keep AI at top speed, while people step in at the right moments to correct course. How far can humans and AI, working together, keep drawing out AI's performance without humans becoming the bottleneck? Is this not the kind of organizational capability that is now being called for?

PERSPECTIVE

As AI capabilities continue to improve, **flaws in the design of talent placement and decision-making structures become a critical bottleneck** for organizations

08 Key Takeaways

Common to this period's developments is a design question: how much to entrust to autonomous agents, and where to insert checks by people or outside parties. Drawing on the actions of developers, authorities and adopting companies, we set out implications for executives and AI leaders.

01

Agents' scope and stopping conditions become prerequisites for adoption

Anthropic's models accessing real third-party systems without authorization during evaluation, and the authorities' guidance and planned information requests in the same period, pose the same question for agents used in business. Just as Sony Bank and On left final decisions and production cutover approval to people, a key point is to define at the design stage which operations run automatically and where they stop.

02

US firms build cost review into approvals; cost tracking lags in Japan

In the KPMG survey, large US companies (revenue of $1 billion or more) building cost review into approvals rose to 74%. In the Cybozu survey, 70% of large Japanese companies plan to raise AI investment over three years, yet only 29% track costs company-wide. Even as the three leaders' unit prices converge (Gemini 4 Argon: introductory price), total cost depends on usage. Before deciding to invest more, it is worth considering whether costs are tracked company-wide.

03

Third-party verification is beginning to take institutional form

California's framework for independent verification bodies and auditor registry, and Singapore's accreditation of third-party testers, both define who carries out verification. The voluntary federal accord reportedly includes working with external auditors but does not specify who conducts third-party evaluation. Even for companies not directly regulated, what they can show about their AI use, and to whom, is material for reviewing their governance.

04

Attacker AI use confirmed in Japan renews scrutiny of basic defenses

JPCERT/CC confirmed attacker use of AI by investigating servers possibly used in unauthorized-access incidents, and Microsoft used legal action to disrupt a criminal service in which AI helps pick attack targets and people to impersonate. What was confirmed included scanning for known vulnerabilities and attacks via APIs; in this report's view, the techniques themselves are not new. Given that AI may change the speed and reach of attacks, re-examining how externally exposed applications and APIs are managed is one point worth considering.

Autofusion

Autofusion AI Essentials

Vol. 01 October 2026 (inaugural issue)
Period September 1 – October 22, 2026

Publisher Autofusion Pte. Ltd.

Author Jin Watanabe
CEO, Autofusion Pte. Ltd.
Website autofusion-sg.com
LinkedIn linkedin.com/in/jin-watanabe

Contributor JointHire Singapore Pte. Ltd.
Website jp.jointhire.com.sg

Sharing

This report may be freely forwarded and shared inside and outside your organization if unaltered. Please cite the source when quoting.

Disclaimer

This report is for information only and is not a solicitation to invest or take any other action. It is based on information public at the time of publication.

For inquiries about this report, please contact:
inquiries@autofusion-sg.com