Gemini Embedding 2, Defense AI Procurement, and the New Compute Race

AI High Signal Digest

2026-03-11

Gemini Embedding 2, Defense AI Procurement, and the New Compute Race

By AI High Signal Digest • March 11, 2026

Google pushed multimodal retrieval and Workspace AI further into product, while healthcare studies, defense procurement shifts, and large infrastructure commitments showed where AI competition is becoming most concrete. This brief covers the main research, product, industry, and policy developments.

Top Stories

Why it matters: This cycle brought three concrete shifts: multimodal retrieval became an API product, healthcare AI produced measurable screening and clinical results, and both compute procurement and government procurement became strategic battlegrounds [1, 2, 3, 4].

1) Gemini Embedding 2 makes multimodal retrieval a platform feature

Google released Gemini Embedding 2, its first fully multimodal embedding model, in public preview via the Gemini API and Vertex AI [1, 5]. The model places text, images, video, audio, and PDFs in a single embedding space, supports 100+ languages and 8,192-token text inputs, offers native audio embeddings, flexible 3,072 / 1,536 / 768 output sizes via MRL, and accepts up to 6 images, 120-second video, and 6-page PDFs per request [5]. Release notes and ecosystem writeups positioned it for simpler RAG, semantic search, clustering, and other cross-modal retrieval tasks [6, 7].

Impact: One model can now cover retrieval across five modalities, reducing the need for separate embedding systems for each content type [6].

Useful links: docs [8] · blog [8]

2) Compute spending keeps scaling up

Thinking Machines said it is partnering with NVIDIA to power frontier model training and customizable AI, bring up 1GW or more of compute starting with Vera Rubin, and co-design systems and architectures; NVIDIA also made a significant investment in the company [9, 3]. Separately, Nscale raised a $2 billion Series C at a $14.6 billion valuation to expand regional capacity, grow engineering and operations, and strengthen the platform layer for training and inference at scale [10].

Impact: The cycle's infrastructure news points to the same conclusion: access to large-scale compute remains a primary competitive lever for frontier AI [3, 10].

3) U.S. government AI procurement is splitting vendors

DeepLearningAI said OpenAI signed a contract to provide AI systems for processing classified U.S. military data after Anthropic refused terms allowing less restrictive military and intelligence use of its models [4]. The same post said the deal followed a White House move barring Anthropic from government contracts, while separate posts citing Axios said the Trump administration was preparing an order to remove Anthropic AI from federal operations [4, 11, 12]. Microsoft later filed an amicus brief supporting Anthropic's complaint against the administration [13].

Impact: Choices about surveillance, warfare, and national-security use are now directly shaping contracts, vendor access, and inter-company alliances [4, 13].

4) Google reports measurable breast-cancer screening gains

Google Research said two *Nature Cancer* studies with Imperial College and NHS UK found its experimental AI screening system identified 25% more interval cancers while reducing screening workloads by an estimated 40% [14, 2]. Google framed the papers as a turning point in screening technology and early detection efforts [14].

Impact: This is a concrete clinical result tied to a real workflow, with both detection and workload outcomes reported [2].

Research & Innovation

Why it matters: The research picture this cycle was less about abstract benchmark gains and more about grounded reasoning, clinical evaluation, tool creation, and compact multimodal performance [15, 16, 17, 18].

AMIE posts prospective clinical results

Google said it ran a prospective clinical study of its AMIE medical chatbot at Beth Israel Deaconess Medical Center urgent care, using it for history taking and to present potential diagnoses for patient-provider discussion [16]. In blinded assessment, AMIE and primary care providers showed similar overall quality on differential diagnosis and management plans, with no significant differences reported for diagnosis, management appropriateness, or safety; primary care providers still outperformed AMIE on management practicality and cost-effectiveness [16]. Paper: https://arxiv.org/abs/2603.08448 [19]

Enterprise evals are getting more grounded

Databricks' OfficeQA Pro benchmark measures end-to-end enterprise reasoning: finding the right documents, extracting the right values, and performing analyses. Frontier agents still score below 50% [15]. AI21 made a similar point from the retrieval side, arguing that standard RAG breaks on aggregative questions across large corpora; its Structured-RAG approach induces a schema at ingestion, maps documents to SQL records, and translates queries to SQL at inference [20]. AI21 also released two new aggregative QA benchmarks with the paper [20].

Tool creation remains a bottleneck for autonomous agents

Tool-Genesis evaluates whether LLMs can infer interfaces, generate schemas, and implement reusable tools directly from natural-language descriptions [17]. The authors highlight a central limitation: current models often create plausible-looking interfaces that break downstream, which makes autonomous tool creation a weak point for self-evolving agents [17]. A strong finding from the benchmark is that closed-loop repair with execution feedback helps substantially, but the gain is scale-dependent and smaller models benefit less [17]. Paper: https://arxiv.org/abs/2603.05578 [17]

Compact multimodal models keep improving

Microsoft released Phi-4-reasoning-vision-15B, a compact open-weight multimodal model that reportedly rivals much larger models on math, science, and computer-use tasks while using a fraction of the training compute [18]. More: https://www.microsoft.com/en-us/research/blog/phi-4-reasoning-vision-and-the-lessons-of-training-a-multimodal-reasoning-model/ [18]

Google explores Bayesian-style reasoning

A Google research blog described fine-tuning LLMs on Bayesian model outputs so they learn to reason like optimal Bayesian agents, reporting stronger probabilistic belief-updating across domains [21].

Products & Launches

Why it matters: Product work is moving beyond chat interfaces toward source-grounded office workflows, visual learning, and developer tooling that can run and schedule agents [22, 23, 24].

Gemini expands across Workspace

Google said new Gemini features are rolling out in beta to AI Ultra and Pro subscribers: Docs can draft from contextual sources and help match document format; Slides can generate layouts and editable diagrams; Sheets can build and edit entire spreadsheets; and Drive's Ask Gemini can surface AI Overviews and answer questions across documents, email, calendar, and the web [22, 25, 26, 27, 28]. Google also said the rollout starts today, globally in English for Docs, Sheets, and Slides, and in the U.S. for Drive [22]. Sundar Pichai added that users can choose grounding sources for Doc drafts, build complex Sheets 9X faster, and get summarized answers directly in Drive search results [29].

More: https://goo.gle/4uAEKn8 [22]

ChatGPT adds interactive visual explanations for learning

OpenAI rolled out dynamic visual explanations for more than 70 core math and science concepts across all ChatGPT plans starting today [23]. Users can manipulate variables and formulas and see graphs and relationships update in real time [30]. OpenAI also said 140 million people already use ChatGPT weekly to understand math and science concepts, and Nick Turley said a Codex workflow helps convert common questions into visual learning blocks [31, 30].

More: https://openai.com/index/new-ways-to-learn-math-and-science-in-chatgpt/ [23]

Developer tooling keeps getting more agent-native

- Ollama can now run prompts on a schedule in Claude Code for recurring work such as PR checks, research tasks, bug triage, and reminders [24, 32, 33, 34, 35].
- LangGraph added single-command deployment to LangSmith via `langgraph deploy` [36].
- Together introduced an official MCP server so coding agents can build AI apps, fine-tune models, or spin up clusters faster [37, 38].

Moondream updates segmentation

Moondream said its segmentation model now delivers better masks, new SOTA benchmarks, and a 40% speedup [39]. The update is already live on Moondream Cloud, with a local model and technical whitepaper coming later this week [39]. More: https://moondream.ai/blog/segmenting-update-2026-03-10 [39]

Industry Moves

Why it matters: Corporate strategy this cycle centered on agent distribution, inference infrastructure, and folding more AI functionality into existing platforms [40, 41, 42].

Meta buys Moltbook

Axios reported that Meta acquired Moltbook, a social network for AI agents [43]. Follow-on posts said Moltbook's founders are joining Meta Superintelligence Labs and that the deal gives Meta early technology and expertise for building platforms where millions of AI assistants can interact and transact across Facebook, WhatsApp, and Instagram [44, 40].

NVIDIA deepens its vLLM bet through Inferact

Inferact said NVIDIA is now its latest investor, extending a collaboration around vLLM [41]. The companies pointed to an uptick in NVIDIA pull requests to the vLLM repo and closer integration with NVIDIA Dynamo, ModelOpt, and Nemotron products [41]. Inferact also said it is using successive NVIDIA architectures from Ampere to Hopper to Blackwell to improve inference performance [45].

OpenAI reportedly plans to add Sora video generation to ChatGPT

A report shared by *The Information* said OpenAI is adding Sora video-generation capabilities to ChatGPT, while continuing to operate the standalone Sora app for now [42]. The report said the move could increase both ChatGPT usage and cost [42]. Source: https://www.theinformation.com/articles/openai-plans-launch-sora-video-ai-chatgpt-strategy-shift [42]

Anthropic expands in Asia-Pacific

Anthropic said it is expanding to Australia and New Zealand and will soon open an office in Sydney, its fourth Asia-Pacific office after Tokyo, Bengaluru, and Seoul [46].

Policy & Regulation

Why it matters: Security standards, procurement rules, and consent features all appeared as active product and policy updates this cycle [47, 4, 48].

National-security rules are starting to alter vendor access

Posts this cycle described a White House move barring Anthropic from government contracts, a planned executive action to remove Anthropic AI from federal operations, and an OpenAI contract for classified military data processing after Anthropic refused looser military-use terms [4, 11]. One industry observer said even the threat was enough to get Anthropic dropped from some Fortune 100 vendor lists [49]. Microsoft's amicus brief shows the dispute is already drawing in other major vendors [13].

A frontier-model security standard is now public

The SL5 Task Force released the first public draft of the Security Level 5 standard, aimed at protecting frontier AI models against nation-state adversaries [47]. The v0.1 draft focuses on long lead-time interventions that need to start before SL5 is urgently required [47]. Draft: https://standard.sl5.org/ [47]

Compliance features are moving into day-to-day AI tools

Notion said AI Meeting Notes now supports automated consent notifications that individuals and enterprise admins can configure for recording and transcription workflows [50, 48]. This shows compliance controls being added directly to transcription features rather than handled only outside the product [48].

Quick Takes

Why it matters: These smaller items sharpen the picture on model use, eval quality, infrastructure, and where leading labs think AI is headed next [51, 52, 53].

- Google DeepMind marked AlphaGo's 10-year anniversary and tied its legacy to AlphaFold, AlphaProof + AlphaGeometry, Gemini Deep Think, and AlphaEvolve; Google said the combination of Gemini world models, AlphaGo-style search and planning, and specialized tools will be critical for AGI [51, 54, 55, 56, 57].
- Similarweb charts showed Claude daily active users rising sharply since the start of 2025 [53].
- FrontierMath and CritPt are showing nearly identical progress trends across models, suggesting shared capabilities behind math and physics research reasoning [58].
- Notion AI Meeting Notes says Japanese transcript and summary quality improved by just over 20%, and the system now transcribes tens of thousands of Japanese meeting hours per day [59].
- Hugging Face launched Storage Buckets [60].
- Hermes Agent reached #3 on GitHub's trending productivity repos; OpenClaw was #11 [61].
- Kalshi's use of LMSYS Arena results to settle real-money bets drew criticism over manipulation risk and whether arena scores should be used for consumer-facing markets at all [52, 62].
- Codex was reported back to stable after a reset, with rate limits restored [63, 64].

Sources

1. X post by @googleaidevs
2. X post by @ymatias
3. X post by @soumithchintala
4. X post by @DeepLearningAI
5. X post by @_philschmid
6. X post by @kimmonismus
7. X post by @qdrant_engine
8. X post by @_philschmid
9. X post by @thinkymachines
10. X post by @nscale_cloud
11. X post by @KobeissiLetter
12. X post by @scaling01
13. X post by @mirandanazzaro
14. X post by @GoogleResearch
15. X post by @DbrxMosaicAI
16. X post by @iScienceLuvr
17. X post by @dair_ai
18. X post by @dl_weekly
19. X post by @iScienceLuvr
20. X post by @AI21Labs
21. X post by @dl_weekly
22. X post by @Google
23. X post by @TheRealAdamG
24. X post by @ollama
25. X post by @Google
26. X post by @Google
27. X post by @Google
28. X post by @Google
29. X post by @sundarpichai
30. X post by @nickaturley
31. X post by @ChatGPTapp
32. X post by @ollama
33. X post by @ollama
34. X post by @ollama
35. X post by @ollama
36. X post by @LangChain
37. X post by @togethercompute
38. X post by @togethercompute
39. X post by @moondreamai
40. X post by @kimmonismus
41. X post by @inferact
42. X post by @steph_palazzolo
43. X post by @axios
44. X post by @TheRundownAI

45. X post by @woosuk_k
46. X post by @AnthropicAI
47. X post by @SL5TaskForce
48. X post by @zachtratar
49. X post by @TheEthanDing
50. X post by @zachtratar
51. X post by @Google
52. X post by @sarahookr
53. X post by @Similarweb
54. X post by @Google
55. X post by @Google
56. X post by @Google
57. X post by @Google
58. X post by @MinyangTian1
59. X post by @zachtratar
60. X post by @_akhaliq
61. X post by @Shaughnessy119
62. X post by @sarahookr
63. X post by @thsottiaux
64. X post by @reach_vb