

GenAI Newsletter May | Agents can dream now, Google I/O's Search Revolution, and headless software.

Dear readers, Hold onto your keyboards, because this month wasn't just another: sophisticated systems like Google's Gemini Spark are jumping across apps to get stuff done; not to be outdone, rivals like OpenAI and Anthropic are suiting up for the corporate battlefield, aggressively pushing their agents into the enterprise world through massive new partnerships. Meanwhile, the biggest twist of all? This month also saw Google radically redefine Search with an AI-first conversational assistant, signaling a foundational shift in how we find information, all while major frontier models prepare for increased government security oversight and undergo national security evaluations.. Companies like OpenAI and Anthropic shift strategy, launching dedicated deployment arms and forging massive consulting partnerships (PwC, KPMG) to overcome implementation bottlenecks. Coupled with major market news like OpenAI's reported IPO plans and the renegotiation of the Microsoft/OpenAI partnership, the focus is now squarely on scaling AI into real-world business infrastructure.

GenAI News Podcast

Don't have time to read this newsletter?

Listen to the 5' podcast

Top Highlights

#1. Microsoft, Google and xAI agreed to give the U.S. government early access to new AI models before public release, allowing the Commerce Department's Center for AI Standards and Innovation to test them for national security risks. The initiative aims to assess threats, including cyberattacks and military misuse, amid rising

concern about the power of frontier systems like Anthropic's Mythos. The government has already conducted more than 40 evaluations of advanced models, signalling increasing oversight of AI capabilities.

#2. Anthropic launched a new "dreaming" capability where AI agents review past sessions, identify patterns, and refine their behaviour between tasks. The system enables agents to improve performance and reduce errors over time, acting as a form of autonomous memory consolidation. It is part of the company's broader push to build more autonomous, self-improving agents that can handle complex tasks like coding and eventually expand into fields like finance and law. Still in research preview, the feature highlights Anthropic's ambition to create AI systems capable of learning independently and working for extended periods with increasing accuracy and productivity.

#3. Google I/O 2026 marked a clear shift to an "agentic AI era" built around Gemini, with Google focusing on systems that not only answer questions but actively perform tasks across its ecosystem. The company introduced new flagship models – Gemini 3.5 Flash (faster, cost-efficient, and designed for real-world tasks) and Gemini Omni (a multimodal model that can create and edit content like video from text, images, audio, and more). Alongside these, Google unveiled Gemini Spark, a personal AI agent that can operate across apps and workflows to automate tasks on behalf of users. A major focus was the transformation of core products, especially Google Search, which is being redesigned into an AI-driven, conversational assistant capable of handling complex queries and actions rather than just returning links. Google also expanded AI across its ecosystem – integrating Gemini into Workspace apps, creative tools (like video and content generation), shopping experiences, and developer platforms. Overall, the announcements highlight Google's strategy to make AI a central, action-oriented layer across all products and devices, moving from standalone models to fully integrated systems that support everyday work, creativity, and decision-making at scale.

#4. AI provider companies are accelerating their enterprise AI strategy through AI services and consulting firms, shifting from competing purely on AI models to competing on large-scale deployment capabilities. OpenAI is launching a new $4 billion "Deployment Company" to help businesses build and integrate AI systems, including acquiring consulting firm Tomoro and embedding engineers directly inside organisations to scale real-world adoption. Anthropic is making a similar push by raising $1.5

billion from investors to build a consulting arm. The aim for both companies is to add hundreds of engineers and address a key bottleneck in enterprise AI: the need for specialised implementation expertise to embed AI into core business operations.

Breaking News

Get daily updates on the Musk vs. Altman trial.

Business News & Market Insights

#5. Microsoft and OpenAI renegotiated their partnership, ending Microsoft's exclusive rights to sell OpenAI models and allowing OpenAI to offer its technology across competing clouds like Amazon and Google. Microsoft retains a non-exclusive license to OpenAI IP until 2032 and will continue to receive a 20% revenue share through 2030, though capped. The shift follows Microsoft's $13 billion investment since 2019 and reflects growing tensions as OpenAI seeks broader enterprise reach.

#6. AI companies including OpenAI, Anthropic, Stripe and Zapier are developing new software architectures designed specifically for AI agents rather than humans. The AI industry is shifting from training agents to use human UIs to building "headless" software (APIs) designed specifically for agents, which are expected to become the dominant users of software. This transition is moving competition from user interfaces to control over APIs, data, and permissions. Key players like Anthropic (with the Model Context Protocol, which enables agents to connect to tools, data, and workflows), Mesa, Salesforce with its "Headless 360", Zapier, companies like Stripe, Mastercard, and OpenAI, which are adding payment rails for agentic shopping, are already implementing this agent-centric software architecture.

#7. A Microsoft study using the DELEGATE-52 benchmark found that even top AI models introduce significant errors in extended workflows, corrupting around 25% of document content over long tasks. The benchmark tested 19 models across 52 domains, revealing that errors accumulate over multi-step operations and that only highly structured tasks, like a Python coding approach, are reliable. Average degradation across models can reach up to 50%, highlighting major limitations for autonomous AI agents in real-world, long-duration processes.

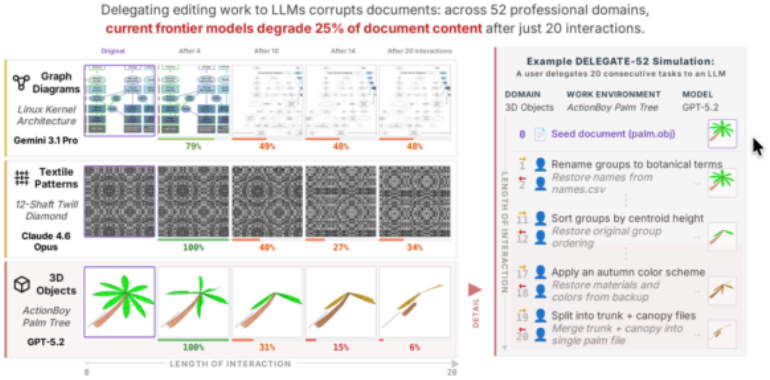


Figure 1: Illustrative examples of how LLMs corrupt documents over long workflows in the DELEGATE-52 benchmark. As LLMs edit files that represent graph diagrams, textile patterns or 3D objects, they introduce sparse but severe errors that silently corrupt documents, compounding over long interaction.[1]

#8. Google announced a "complete reimagining" of Search with AI, marking its biggest change in more than 25 years and replacing traditional blue links with AI-generated summaries and conversational interactions powered by Gemini 3.5 Flash. Users can now query using text, images, videos, or files and continue with follow-up questions in a ChatGPT-like interface. This shift could disrupt industries reliant on search traffic, as fewer users may click through to external websites, threatening publishers already experiencing traffic declines.

#9. OpenAI is reportedly planning to file for an IPO in the coming weeks, targeting a public debut potentially later in 2026. The move follows a legal victory removing a major obstacle and comes after massive funding rounds that valued the company in the hundreds of billions. The listing could become one of the largest tech IPOs ever, but OpenAI still faces regulatory scrutiny, structural challenges, and increasing competition from rivals like Google and Anthropic.

#10. **Anthropic pushes enterprise AI adoption through major partnerships and vertical solutions**. Anthropic is expanding its enterprise footprint with initiatives like Claude for business use cases and strategic alliances with firms such as PwC and KPMG. PwC plans to deploy Claude across a global workforce and train 30,000 professionals. With KPMG, Claude is integrated into its Digital Gateway platform and rolled out to over 276,000 employees to enable real-time agentic workflows for clients. These collaborations focus on "agentic" AI systems that automate complex workflows in areas like finance, deal-making, and software development, with reported efficiency gains such as dramatically reduced processing times. The broader push reflects a shift from AI experimentation to large-scale operational deployment.

#11. DeepMind spinout Isomorphic Labs raises $2.1 billion in its second funding round, led by Thrive Capital with participation from

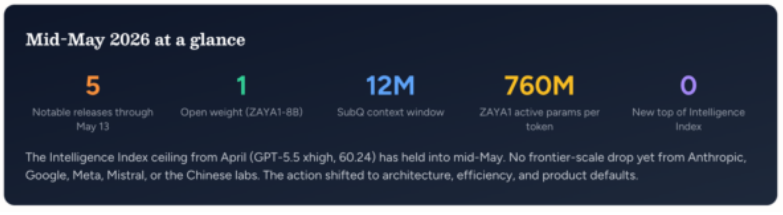
Alphabet, MGX, Temasek, and the UK Sovereign AI fund, to accelerate AI-driven drug discovery. The company leverages technologies like AlphaFold and its proprietary drug design engine to develop therapies across multiple disease areas, aiming to make AI-designed drugs a reality. Partnerships with pharma firms such as Novartis and Eli Lilly could be worth up to $3 billion combined, highlighting strong industry momentum behind AI-based drug development.

#12. More than 1,000 brands are now running advertising campaigns in ChatGPT via Criteo's integration, showing rapid adoption of conversational ads. Early performance data indicates AI-driven conversions are approaching 2x the rate of traditional search in key retail categories. The platform is expanding globally and becoming an incremental channel rather than replacing existing ad spend.

#13. Beyond enterprise, Anthropic has entered partnerships with major organisations like the Gates Foundation to apply AI in health and education, signalling a dual strategy of commercial expansion and societal impact. These initiatives aim to deploy AI systems in critical domains while addressing access, efficiency, and global development challenges.

New Models & Innovations

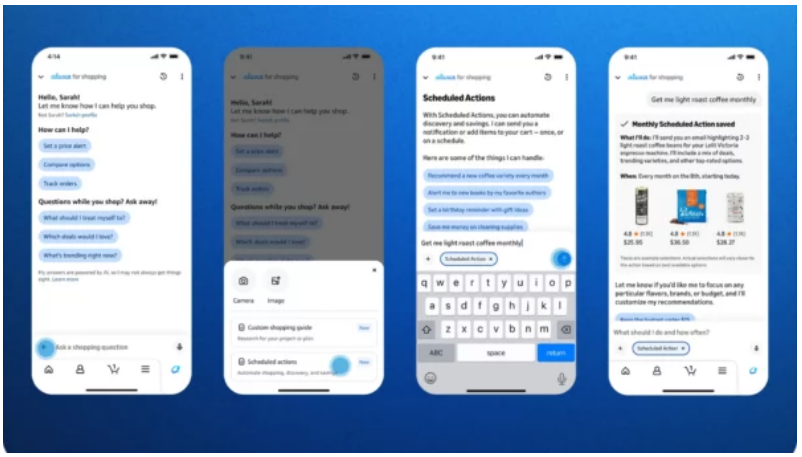
#14. After a wave of major releases in April 2026 (including GPT-5.5 and Claude Opus 4.7), May saw fewer frontier breakthroughs but a shift toward model efficiency, architecture improvements, and product integration. The Intelligence Index ceiling set in April remains unchanged, with innovation moving toward optimised models (e.g. default ChatGPT updates, smaller MoE models, long-context systems) rather than raw capability leaps. This indicates a transition from "model race" to "deployment and usability optimisation."



#15. OpenAI explores a radical "post-app" future with an AI-native smartphone. OpenAI is reportedly developing a smartphone centred on AI agents rather than traditional apps, where tasks like booking, messaging, or research are handled autonomously by context-aware systems. The device would combine on-device and cloud AI models

and could give OpenAI greater control over user experience and data. If realised, this approach could challenge the app-based ecosystems dominated by Apple and Google and signal a broader industry shift toward agent-driven computing.

#16. Amazon introduced "Alexa for Shopping," replacing its Rufus chatbot and embedding an AI assistant directly into its search experience to answer queries and take actions like comparing products or scheduling purchases. The system combines Rufus product knowledge with Alexa+ personalisation, leveraging user history and enabling features such as price tracking, automated buying, and cross-device shopping. Rufus had already reached over 300 million users in 2025, and Amazon aims to expand its reach further with this unified, agentic approach.



#17. Alibaba is integrating its Qwen AI platform with Taobao and Tmall, enabling conversational shopping where users can browse, compare and purchase via chat instead of keywords. The AI agent will access a catalogue of over 4 billion products and manage logistics, recommendations, virtual try-ons and 30-day price tracking within a unified workflow. This marks a shift toward end-to-end "agentic commerce," where AI handles the full purchase journey rather than just discovery.

#18. OpenAI introduced a self-serve advertising platform that allows businesses to directly buy ads in ChatGPT, expanding beyond its initial pilot with select partners. The move supports ambitions to generate $2.5 billion in ad revenue this year and up to $100 billion by 2030, opening access to a broader advertiser base. The platform simplifies campaign creation and positions ChatGPT as a new performance marketing channel alongside search and social.

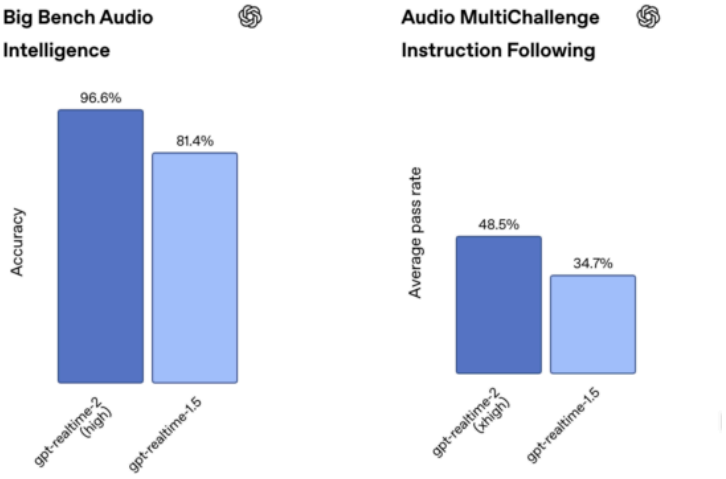
#19. Meta is developing a highly personalised AI assistant powered by its Muse Spark model to automate everyday tasks for its users. The system, inspired by OpenAI's OpenClaw, aims to connect

multiple tools and act autonomously with minimal human input, moving beyond chatbot interactions. The company is also building an internal agent codenamed “Hatch” and plans to integrate agentic shopping features into Instagram before Q4, while increasing AI infrastructure spending by billions.

#20. Anthropic released 10 ready-to-run AI agent templates for financial services tasks such as pitchbook creation, KYC screening, earnings review and financial modelling. These agents integrate directly with tools like Excel, PowerPoint, Word and Outlook, enabling end-to-end workflows across applications. The initiative positions AI agents as operational infrastructure for finance, reducing deployment time from months to days.

#21. OpenAI replaced ChatGPT’s default model with a new version designed to deliver more accurate, personalised responses with fewer hallucinations. The update enhances memory and contextual understanding, enabling the system to better use past interactions while producing more concise outputs. Even incremental changes to the default model affect hundreds of millions of users, potentially increasing reliance on the assistant.

#22. OpenAI has launched three models: GPT-Realtime-2, GPT-Realtime-Translate, and GPT-Realtime-Whisper, to enable AI agents that can listen, translate, and act during live conversations. The system supports translation from more than 70 languages into 13 outputs and real-time transcription, while maintaining context in longer voice sessions. Pricing starts at $32 per million audio input tokens, $0.034 per minute for translation, and $0.017 per minute for transcription, signalling a push toward production-grade voice agents.



Big Bench Audio evaluates challenging reasoning capabilities in language models that support audio input. Audio MultiChallenge evaluates multi-turn conversational intelligence in spoken dialogue systems, including instruction following, context integration, self-consistency, and handling natural speech corrections.

Policy Updates & Ethical Debates

#23. The EU reached a provisional agreement (Digital Omnibus on AI) introducing 16-month delays for high-risk AI obligations (now December 2027) and 12-month delays for product-embedded AI (August 2028). The reform simplifies compliance by reducing overlap with existing regulations and extending relief to more businesses. It also introduces new prohibitions, including bans on AI-generated non-consensual intimate content and child sexual abuse material, with compliance required from December 2026.

Share this article

Contact Us