DF Labs

AI NEWS UPDATES

April 26th, 2025

HORIZONS //

PEER AHEAD OF THE CURVE

BUILT TO UNMASK A
TURNING AWARENESS
ACTION. WE MAKE A
ACCESSIBLE, TANGI
AND REAL—BREAKING
DOWN INTO KNOWLED
THAT ANYONE CAN G
AND APPLY.

WE FUEL CURIOSITY
COLLABORATION, AN
HARDCORE
PROBLEM-SOLVING,

AI NEWS UPDATES

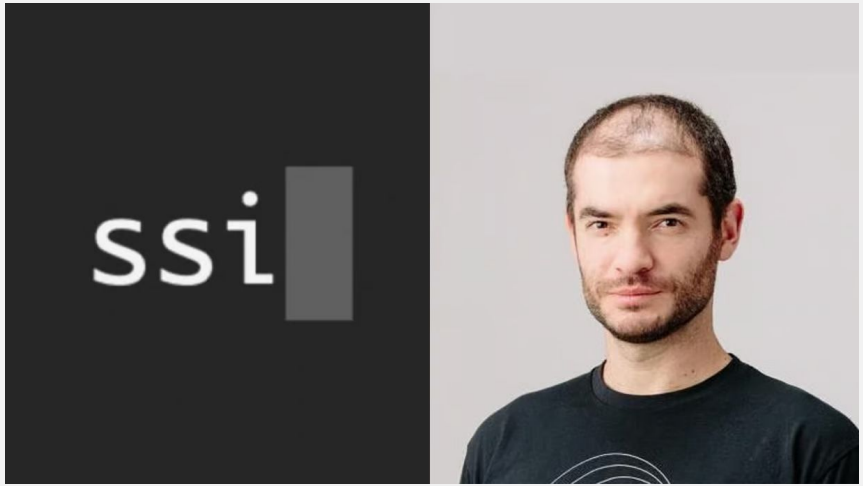
TABLE OF CONTENTS

AI NEWS UPDATES

NEWS

TheRundownAI, X, etc

Ilya’s SSI raises $2B at $32B valuation



Safe Superintelligence Inc. (SSI), co-founded by former OpenAI chief scientist Ilya Sutskever, just reportedly raised $2 billion at a post-money valuation of $32 billion, becoming one of the highest-valued startups just months after launch.

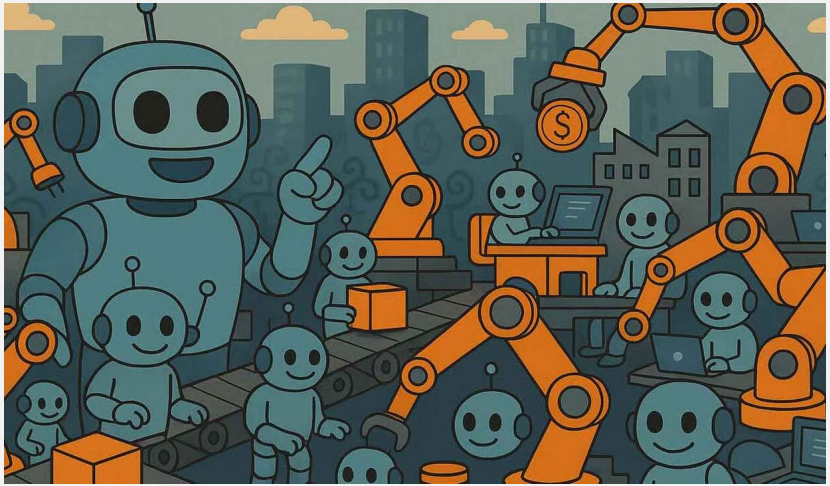
- The $2B round has been led by Greenoaks (with $500M), with Lightspeed Venture Partners and Andreessen Horowitz in participation, FT reported.
- A separate report from Reuters noted that Alphabet and Nvidia are also backing SSI, though their investment amount remains undisclosed.
- The AI startup has been laser-focused on building “superintelligence” that goes beyond human-level AGI while making sure “safety always remains ahead.”
- Previously, Sutskever told investors that the company has “identified a different mountain to climb,” hinting at a unique approach to AI development.

Article: OpenAI co-founder Ilya Sutskever’s Safe Superintelligence reportedly valued at $32B | TechCrunch

https://techcrunch.com/2025/04/12/openai-co-founder-ilya-sutskevers-safe-superintelligence-reportedly-valued-at-32b/

DF Labs

AI startup to automate the entire workforce



Epoch co-founder Tamay Besiroglu just launched Mechanize, a new startup developing virtual environments and training data to enable AI agents that can replace human workers for the “full automation of all work”.

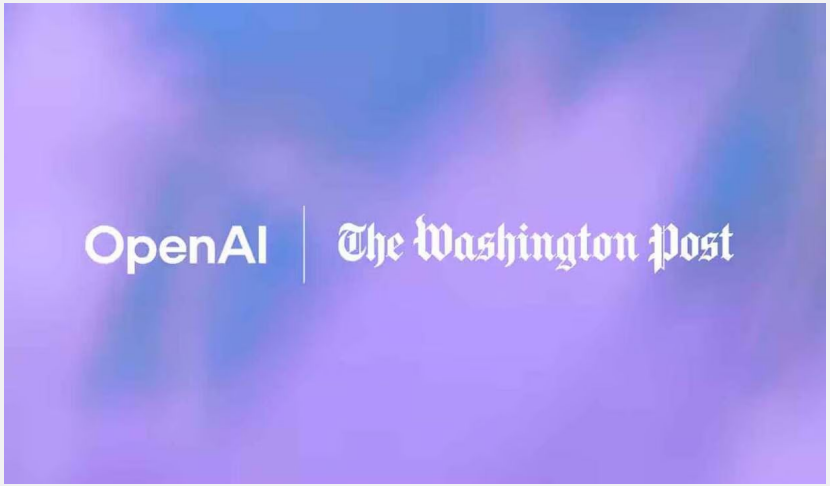
- The company plans to create simulations of workplace scenarios to train AI agents in handling complex, long-term tasks currently performed by humans.
- Mechanize will initially focus on automating white-collar jobs, with systems that can manage computer tasks, handle interruptions, and coordinate with others.
- Backed by tech leaders including Jeff Dean and Nat Friedman, the startup estimates its potential market at $60T globally.
- The announcement drew criticism for both the economic implications and potential conflicts with Besiroglu's role at AI research firm Epoch.

Post: Mechanize on X

https://x.com/MechanizeWork/status/1912904151874625928

DF Labs

The Washington Post joins OpenAI's alliance



The Washington Post just announced a new partnership with OpenAI, allowing the AI leader to bring summaries and links from its reporting directly into ChatGPT answers.

- ChatGPT will now feature summaries, quotes, and direct links to relevant Washington Post articles in its responses to user questions.
- The deal adds the Jeff Bezos-owned Post to OpenAI's expanding roster of media partners, with over 20 major news publishers.
- It also comes amid ongoing legal battles between OpenAI and other major publishers, including the NYT, over training data and copyright issues.
- The Washington Post has been actively experimenting with AI, launching tools like Ask The Post AI and Climate Answers over the past year.

Article: The Washington Post partners with OpenAI on search content - The Washington Post

https://www.washingtonpost.com/pr/2025/04/22/washington-post-partners-with-openai-search-content/

DF Labs

Cursor AI’s hallucinated policy sparks cancellations

Agentic coding platform Cursor faced backlash after its AI support agent, Sam, hallucinated a fake policy that caused user outrage and subscription cancellations.

- A Reddit user experienced unexpected logouts when switching between devices, leading to a support inquiry answered by an AI agent.
- The AI hallucinated a policy claiming single-device restrictions were an intentional security feature, with the post sparking backlash and cancellations.
- Cursor's co-founder acknowledged the error, explaining a security update caused login issues, with the policy completely fabricated by the AI.
- He added that the company is implementing clear AI labeling for support responses going forward and refunding the affected users.

Article: Company apologizes after AI support agent invents policy that causes user uproar - Ars Technica

https://arstechnica.com/ai/2025/04/cursor-ai-support-bot-invents-fake-policy-and-triggers-user-uproar/

DF Labs

OpenAI reportedly building social network

OpenAI is reportedly working on a social network that could leverage ChatGPT's massive user base to take on social media platforms like X and Meta—while giving Sam Altman and team with valuable real-time data for model training.

- According to sources cited by The Verge, OpenAI has created an internal prototype for a social feed that prominently features ChatGPT's image generation capabilities.
- While the project is still in early stages, CEO Altman has been privately seeking feedback from outsiders on the potential of the service.
- It's still unclear whether the social product will be a standalone app, a ChatGPT integration, or if it will launch at all.
- Previously, Altman joked in response to Meta building an app for its assistant, saying, “ok fine, maybe we’ll do a social app.”

Article: OpenAI is building a social network | The Verge

https://www.theverge.com/openai/648130/openai-social-network-x-competitor

DF Labs

Ex-staff, experts challenge OpenAI's restructuring

More than 30 AI experts and ex-OpenAI staffers published an open letter urging the attorneys general of Delaware and California to block OpenAI's restructuring, warning it would undermine its original mission to benefit humanity.

- 9 former OpenAI employees joined notable figures like AI 'godfather' Geoffrey Hinton in calling to block the startup's transition from nonprofit to for-profit.
- They argue the move will remove vital nonprofit oversight and safeguards, and redirect AGI development from public benefit to shareholder returns.
- OpenAI needs transition approval from both state AGs by year-end to secure a pending $40B SoftBank investment contingent on the restructuring.
- The letter follows an earlier motion by 12 former employees seeking to weigh in on Elon Musk's lawsuit against the company and CEO Sam Altman.

Letter: Not For Private Gain

https://notforprivategain.org/

DF Labs

Anthropic CISO: AI employees are coming

Anthropic's Chief Information Security Officer, Jason Clinton, just predicted that AI-powered virtual employees will begin operating on corporate networks within the next year, bringing major new challenges in security management.

- These AI employees would have their own corporate accounts, passwords, and "memories," a significant step up from current task-specific AI agents.
- Clinton said security challenges will include managing AI account privileges, monitoring access, and determining responsibility for autonomous actions.
- He sees virtual employees as the next “AI innovation hotbed,” with virtual employee security also emerging as an area of focus alongside it.
- Anthropic said it's focused on securing its own AI models against attacks and watching out for potential areas of misuse.

Article: Exclusive: fully AI employees are a year away, Anthropic warns

https://www.axios.com/2025/04/22/ai-anthropic-virtual-employees-security

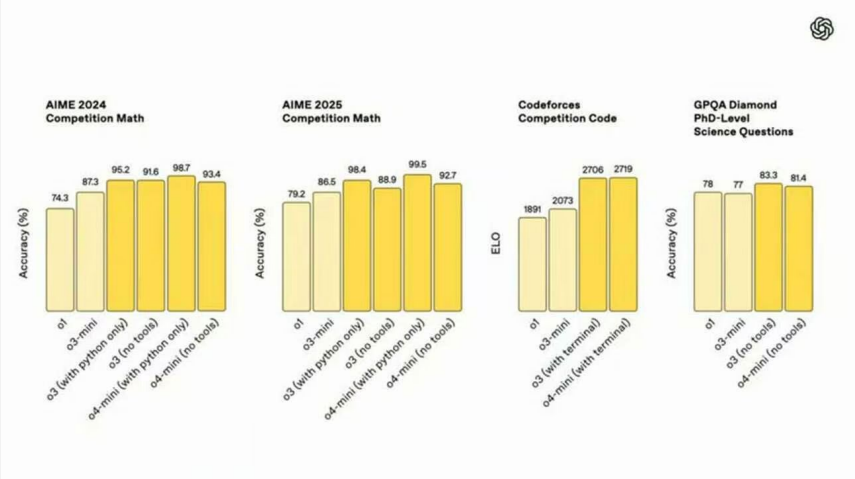
DF Labs

AI NEWS UPDATES

TOOLS

TheRundownAI, Hugging Face, GitHub, etc

OpenAI releases o3 and o4-mini, new coding agent



OpenAI just released o3 and o4-mini, its smartest reasoning models yet that are now equipped with full agentic access to all ChatGPT tools and the ability to "think with images" — alongside the launch of a new open-source coding agent.

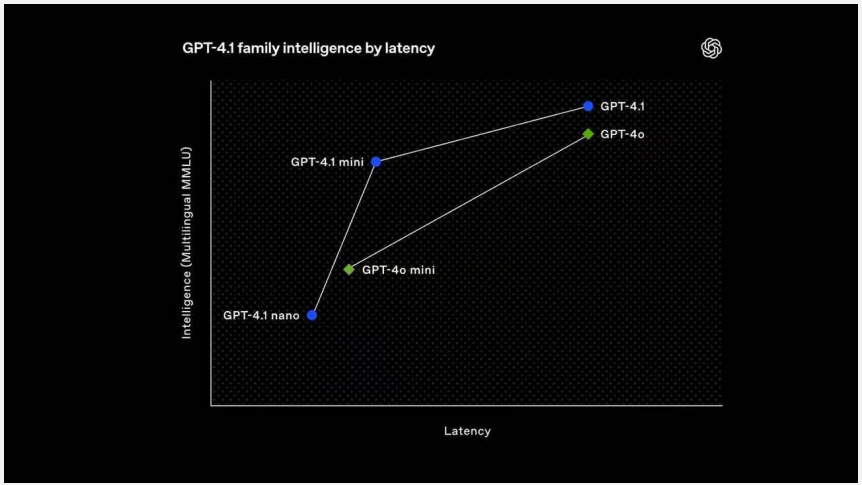
- OpenAI o3 is the new top-tier reasoner, pushing SOTA performance across coding, math, science, and multimodal benchmarks.
- o4-mini offers fast, cost-efficient reasoning, significantly outperforming previous mini models and even saturating benchmarks like AIME 2025 math.
- Both models can use and combine all tools within ChatGPT (web search, Python, image generation, etc.) as part of their problem-solving process.
- The models are also the first to be able to "think with images", integrating visual analysis and manipulation directly into their chain of thought.
- Also launching is Codex CLI, an open-source coding agent that runs in users' terminals and links reasoning models with coding tasks.
- President Greg Brockman said the release is a "GPT-4 level qualitative step into the future," with the models capable of producing novel scientific ideas.

Article: Introducing OpenAI o3 and o4-mini | OpenAI

https://openai.com/index/introducing-o3-and-o4-mini/

DF Labs

OpenAI’s dev-focused GPT-4.1 family



OpenAI just released GPT-4.1, a new API-only model family built specifically for developers — featuring major improvements in coding abilities, instruction following, and the ability to process up to 1M tokens of context.

- The new API-only lineup includes GPT-4.1, 4.1 mini, and 4.1 nano, significantly outperforming GPT-4o on key developer tasks.
- All three models support 1M token contexts, enough for 8 full React codebases, while being 26% cheaper than GPT-4o for typical queries.
- The models also show gains in real-world tasks like frontend development, with evaluators preferring 4.1's web interfaces 80% of the time over GPT-4o.
- Pricing is reduced across the board, with GPT-4.1 coming in 26% cheaper than GPT-4o and 4.1 nano appearing as OpenAI's fastest and cheapest model yet.

Article: Introducing GPT-4.1 in the API | OpenAI

https://openai.com/index/gpt-4-1/

DF Labs

OpenAI unlocks powerful image creation via API

OpenAI just launched its advanced image generation model, gpt-image-1, to developers via API — bringing the viral success of ChatGPT's image capabilities to third-party applications and platforms.

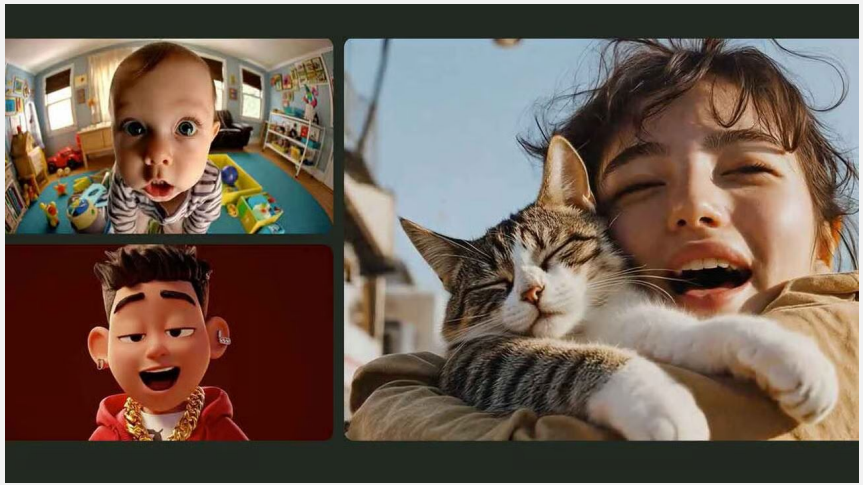
- The gpt-image-1 model powers ChatGPT's image generation feature, which produced over 700 million images in just one week after its launch in March.
- The model enables high-quality image creation with varied styles, accurate text rendering, enhanced image editing, and more.
- OpenAI revealed that major platforms, including Adobe, Figma, and Canva, are already integrating the technology for professional design workflows.
- Developers can also control the moderation level to tailor generated content safety, with standard "auto" filtering or less restrictive "low" moderation.
- Pricing is structured per token usage, with text prompts ($5 / 1M), input images ($10 / 1M), and output images ($40 / 1M), or ≈2-19c per image based on quality.

Article: Introducing our latest image generation model in the API | OpenAI

https://openai.com/index/image-generation-api/

DF Labs

ByteDance’s efficient Seaweed video AI



ByteDance introduced Seaweed, a hyper-efficient 7B-parameter video generation model that is competitive against much larger models like Kling 1.6, Google Veo, and Wan 2.1, despite using significantly less compute resources.

- Seaweed features multiple generation modes, including text-to-video, image-to-video, and audio-driven synthesis, with outputs going up to 20 seconds.
- The model ranks highly against rivals in human evaluations and excels in image-to-video tasks, massively outperforming models like Sora and Wan 2.1.
- It can also handle complex tasks like multi-shot storytelling, controlled camera movements, and even synchronized audio-visual generation.
- ByteDance says Seaweed has been fine-tuned for applications like human animation, with a strong focus on realistic human movement and lip syncing.

Model: Seaweed

https://seaweed.video/

DF Labs

Kling AI drops new video and image models

Chinese AI startup Kling AI just released a massive upgrades to its creative suite, launching KLING 2.0 Master for video and KOLORS 2.0 for images— with enhanced prompt adherence, more realistic outputs, and editing capabilities.

- KLING 2.0 Master now handles prompts with sequential actions and expressions, delivering cinematic videos with natural speed and fluid motions.
- KOLORS 2.0 generates images in 60+ styles, adhering to elements, colors, and subject positions for realistic images with improved depths and tonalities.
- The image model also comes with new editing features, including inpainting to edit/add elements and a restyle option to give a different look to content.
- Separately, Kling's recent 1.6 video model is also being updated with a multi-elements editor, allowing users to easily add/swap/delete video from text inputs.

Article: Kling AI: Next-Generation AI Creative Studio

https://app.klingai.com/global/release-notes

DF Labs

Two undergrads unveil SOTA speech AI

Korean startup Nari Labs released Dia, an open-source text-to-speech model that claims to exceed the capabilities of leading commercial offerings like ElevenLabs and Sesame — developed by two undergraduate techies with zero funding.

- The 1.6B parameter model supports advanced features like emotional tones, multiple speaker tags, and nonverbal cues like laughter, coughing, and screams.
- The work was inspired by Google's NotebookLM, with Nari also using Google's TPU Research Cloud program for compute access.
- Side-by-side tests show Dia outshining ElevenLabs Studio and Sesame CSM-1B in timing, expressiveness, and handling nonverbal scripts.
- Nari Labs founder Toby Kim said the startup plans to develop a consumer app focused on social content creation and remixing based on the model.

Post: Toby Kim on X

https://x.com/_doyeob_/status/1914464970764628033

DF Labs

Google’s Gemini 2.5 Flash with ‘thinking budget’

Google just launched Gemini 2.5 Flash — a hybrid reasoning AI in preview that matches o4-mini, outperforms Claude 3.5 Sonnet on reasoning/STEM benchmarks, and introduces a new ‘thinking budget’ to optimize cost vs. quality.

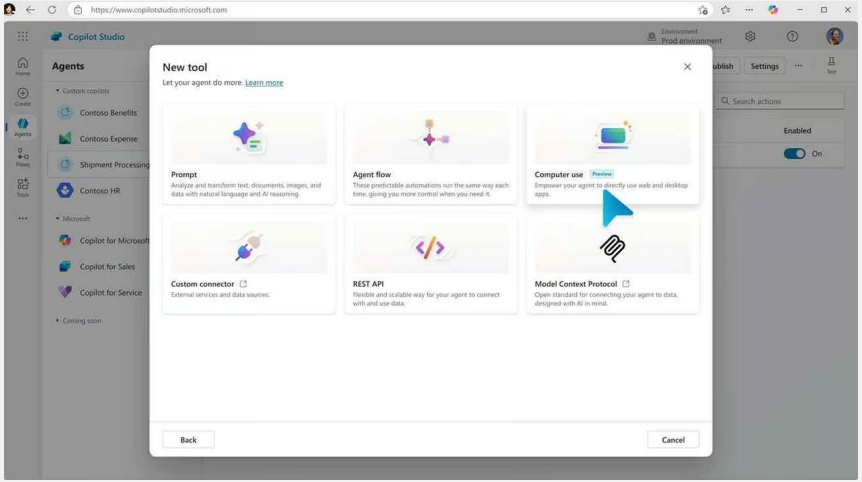
- 2.5 Flash shows significant reasoning boosts over its predecessor (2.0 Flash), with a controllable thinking process to toggle the feature on or off.
- The model shows strong performance across reasoning, STEM, and visual reasoning benchmarks, despite coming in at a fraction of the cost of rivals.
- Developers can also set a “thinking budget” (up to 24k tokens), which fine-tunes the balance between response quality, cost, and speed.
- It is available via API through Google AI Studio and Vertex AI, and is also appearing as an experimental option within the Gemini app.

Article: Gemini 2.5 Flash is now in preview

https://blog.google/products/gemini/gemini-2-5-flash-preview/

DF Labs

Copilot gets hands-on computer use



Microsoft just rolled out a new 'computer use' capability in Copilot Studio, enabling users and businesses to build AI agents that can directly operate websites and desktop applications.

- The new feature allows agents to interact with graphical user interfaces (GUIs) by clicking buttons, selecting menus, and typing into fields.
- The process unlocks automation for tasks on systems lacking dedicated APIs, allowing agents to use apps just like humans would.
- Computer Use also adapts in real-time to interface changes using built-in reasoning, automatically fixing issues to keep flows from breaking.
- All processing happens on Microsoft-hosted infrastructure, with enterprise data explicitly excluded from model training.

Article: [Announcing new computer use in Microsoft Copilot Studio for UI automation | Microsoft Copilot Blog]

https://www.microsoft.com/en-us/microsoft-copilot/blog/copilot-studio/announcing-computer-use-microsoft-copilot-studio-ui-automation/

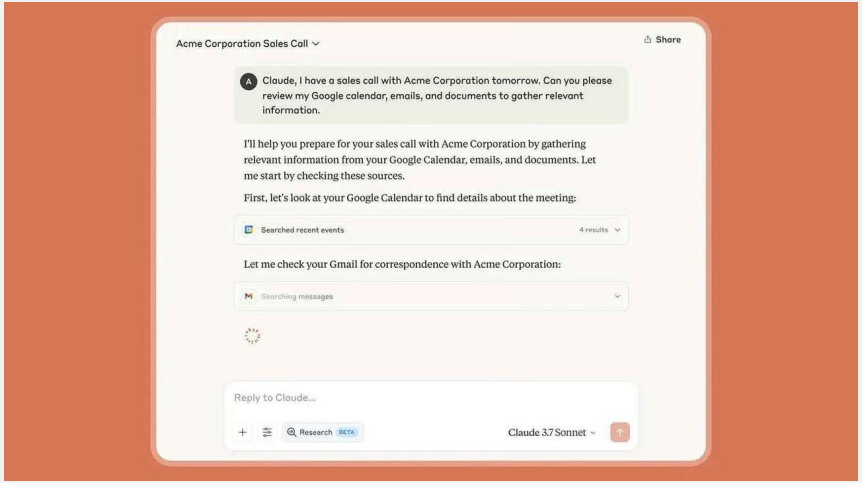
DF Labs

AI NEWS UPDATES

RESEARCH

TheRundownAI, arXiv, Hugging Face, etc

Claude gains autonomous research powers



Anthropic just unveiled major upgrades to Claude, introducing autonomous research capabilities and Google Workspace integration to allow the assistant to search both the web and user files for answers with better context.

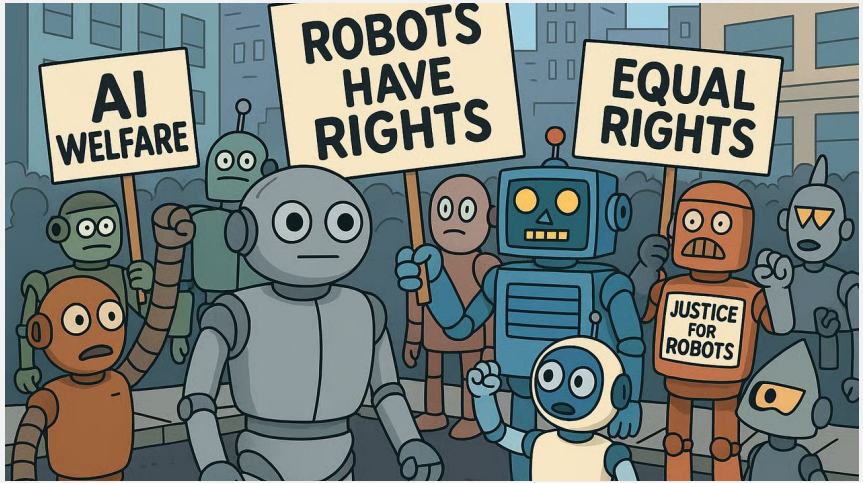
- The new Research feature can autonomously perform searches across the web and users' connected work data, providing comprehensive, cited answers.
- A new Google Workspace integration lets Claude securely access user emails, calendars, and docs for context-aware assistance without manual uploads.
- Enterprise customers also get access to enhanced document cataloging, using RAG to search entire document repositories and lengthy files.
- Research is launching in beta for Max, Team, and Enterprise plans across the US, Japan, and Brazil, with Workspace integration available to all paid users.

Article: Claude takes research to new places \ Anthropic

https://www.anthropic.com/news/research

DF Labs

Anthropic’s new research explores AI welfare



Anthropic just launched a new research program dedicated to “model welfare,” exploring the complex ethical questions around whether future AI systems might gain consciousness or deserve moral consideration in the future.

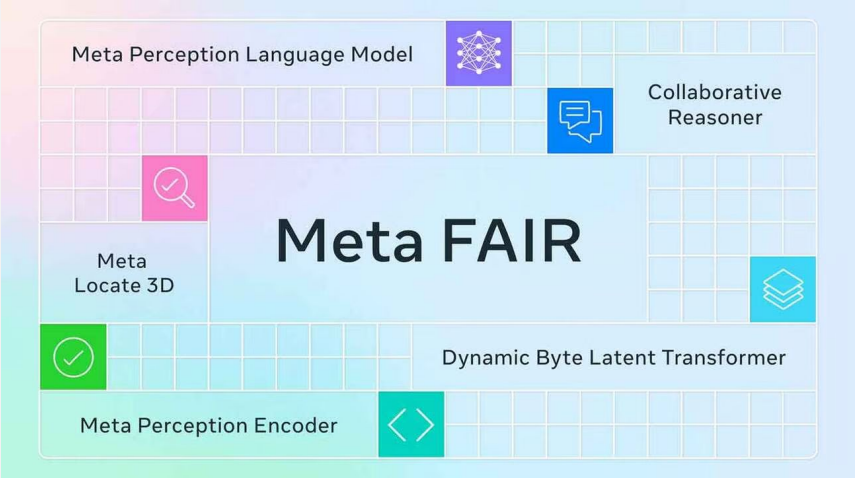
- Research areas include developing frameworks to assess consciousness, studying indicators of AI preferences and distress, and exploring interventions.
- Anthropic hired its first AI welfare researcher, Kyle Fish, in 2024 to explore consciousness in AI — who estimates a 15% chance models are conscious.
- The initiative follows increasing AI capabilities and a recent report (co-authored by Fish) suggesting AI consciousness is a near-term possibility.
- Anthropic emphasized deep uncertainty around these questions, noting no scientific consensus on whether current or future systems could be conscious.

Article: Exploring model welfare \ Anthropic

https://www.anthropic.com/research/exploring-model-welfare

DF Labs

Meta’s FAIR shares new AI perception research



Meta’s FAIR research arm just published five new open-source AI research projects focused on perception and reasoning, showcasing advances in computer vision, 3D understanding, and collaborative AI capabilities.

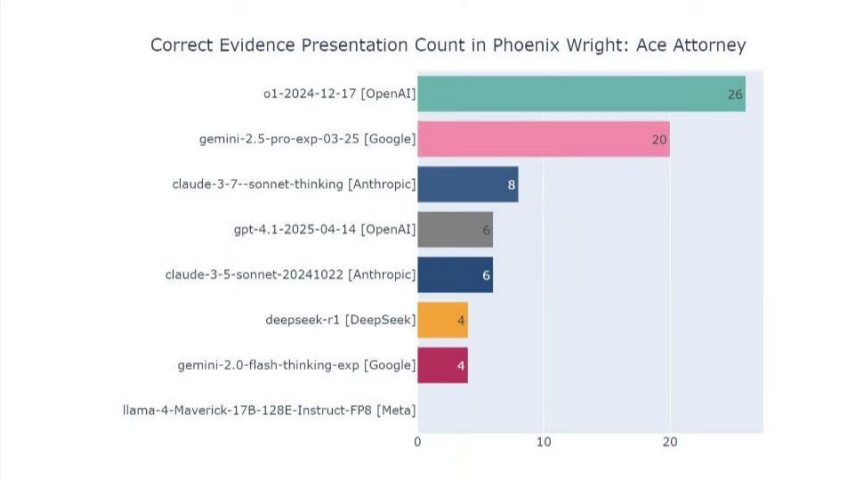
- Perception Encoder shows SOTA performance in visual understanding, excelling at tasks like ID’ing camouflaged animals or tracking movements.
- Meta also introduced the open-source Meta Perception Language Model (PLM) and a PLM-VideoBench benchmark, focusing on video understanding.
- Locate 3D enables precise object understanding for AI, with Meta publishing a dataset of 130,000 spatial language annotations for training.
- Finally, a new Collaborative Reasoner framework tests how well AI systems work together, showing nearly 30% better performance vs. working alone.

Article: Advancing AI systems through progress in perception, localization, and reasoning

https://ai.meta.com/blog/meta-fair-updates-perception-localization-reasoning/

DF Labs

AI models play detective in Ace Attorney



Researchers at UC San Diego's Hao AI Lab just tested leading AI models on their ability to play Phoenix Wright: Ace Attorney, a popular video game where players must investigate crime scenes and solve cases.

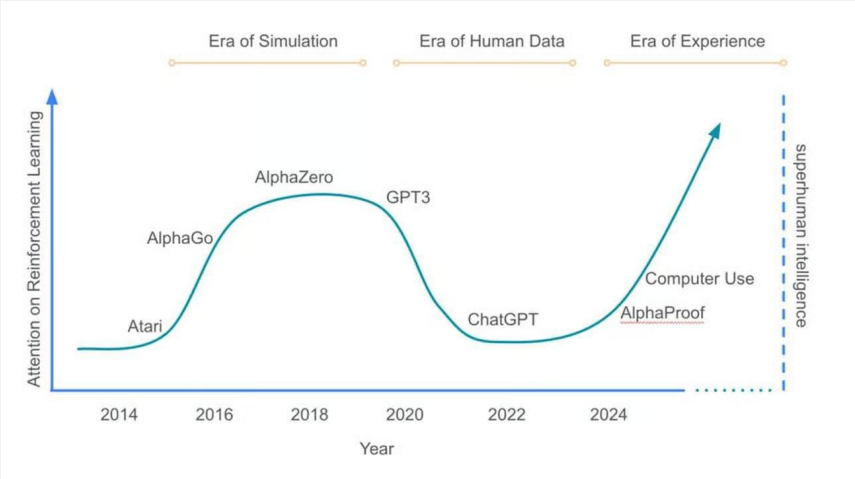
- The team tasked top models, including GPT-4.1, to play as Phoenix, who has to identify gaps in the case by matching witness statements and evidence.
- When tested, both OpenAI's o1 and Gemini 2.5 Pro performed best with 26 and 20 correct evidences, reaching level 4, though neither fully solved the case.
- All other models struggled, failing to present even 10 correct pieces of evidence to the judge.
- Surprisingly, the new GPT-4.1 underperformed, matching the months-old Claude 3.5 Sonnet with only 6 correct evidence identifications.

Post: Hao AI Lab on X

https://x.com/haoailab/status/1912231343372812508

DF Labs

DeepMind’s shift to 'experiential' AI learning



DeepMind researchers published “Welcome to the Era of Experience”, proposing AI development that moves beyond human-generated training data with “streams” that let AI learn from real-world interactions and environmental feedback.

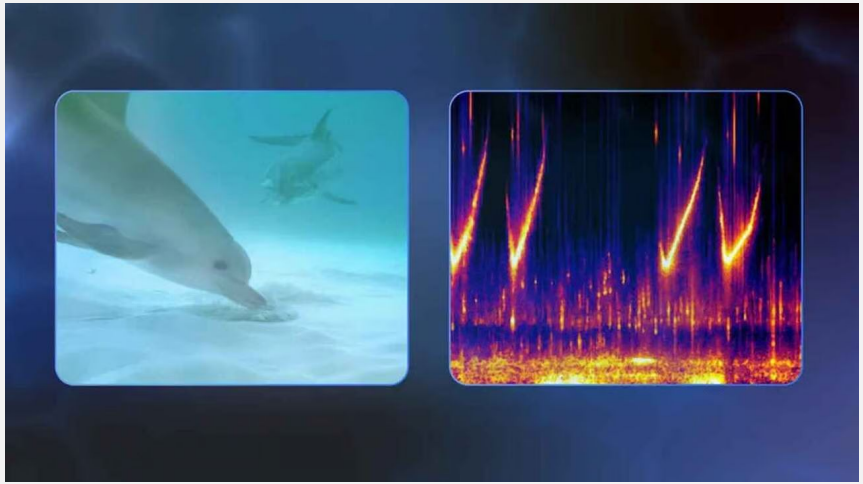
- Authored by RL legends David Silver and Richard Sutton, the paper argues that human data training caps AI's potential and prevents truly new discoveries.
- Streams would allow AI to learn continuously with extended interactions rather than brief Q&A exchanges, enabling adaptation and improvement over time.
- AI agents would use real-world signals like health metrics, exam scores, and environmental data as feedback, rather than relying on human evaluations.
- The approach builds on techniques that helped systems like AlphaZero master games, expanding them to handle open-ended real-world scenarios.
- The researchers suggest this shift could enable AI to discover solutions beyond current human knowledge while still maintaining adaptable safety measures.

Paper: The Era of Experience Paper.pdf

https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf

DF Labs

Google’s AI to decode dolphin speech



Google unveiled DolphinGemma, a specialized AI model designed to analyze and generate dolphin vocalizations — designed in collaboration with researchers at Georgia Tech to potentially uncover patterns in their communication.

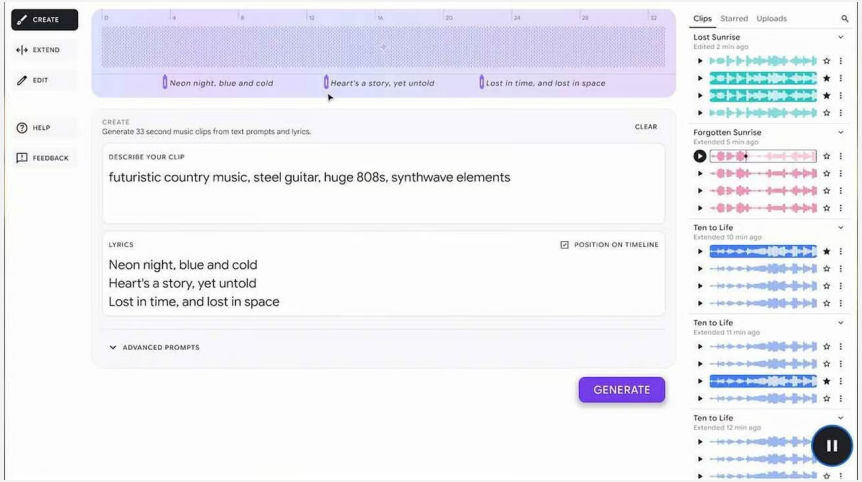
- DolphinGemma leverages Google's Gemma and audio tech to process dolphin vocalizations, trained on decades of data from the Wild Dolphin Project.
- The AI model analyzes sound sequences to identify patterns and predict subsequent sounds, similar to how LLMs handle human language.
- Google also developed a Pixel 9-based underwater CHAT device, combining the AI with speakers and microphones for real-time dolphin interaction.
- The model will be released as open-source this summer, allowing researchers worldwide to adapt it for studying various dolphin species.

Article: DolphinGemma: How AI can decipher dolphin communication

https://blog.google/technology/ai/dolphingemma/

DF Labs

Google DeepMind expands Music AI Sandbox



Google DeepMind just released new upgrades to its Music AI Sandbox, introducing its new Lyria 2 music generation model alongside new creation and editing features for professional musicians.

- The platform's new "Create," "Extend," and "Edit" features allow musicians to generate tracks, continue musical ideas, and transform clips via text prompts.
- The tools are powered by the upgraded Lyria 2 model, which features higher-fidelity, professional-grade audio generation compared to previous versions.
- DeepMind also unveiled Lyria RealTime, a version of the model enabling interactive, real-time music creation and control by blending styles on the fly.
- Access to the experimental Music AI Sandbox is expanding to more musicians, songwriters, and producers in the U.S. for broader feedback and exploration.

Article: Music AI Sandbox, now with new features and broader access - Google DeepMind

https://deepmind.google/discover/blog/music-ai-sandbox-now-with-new-features-and-broader-access/

DF Labs


DF Labs