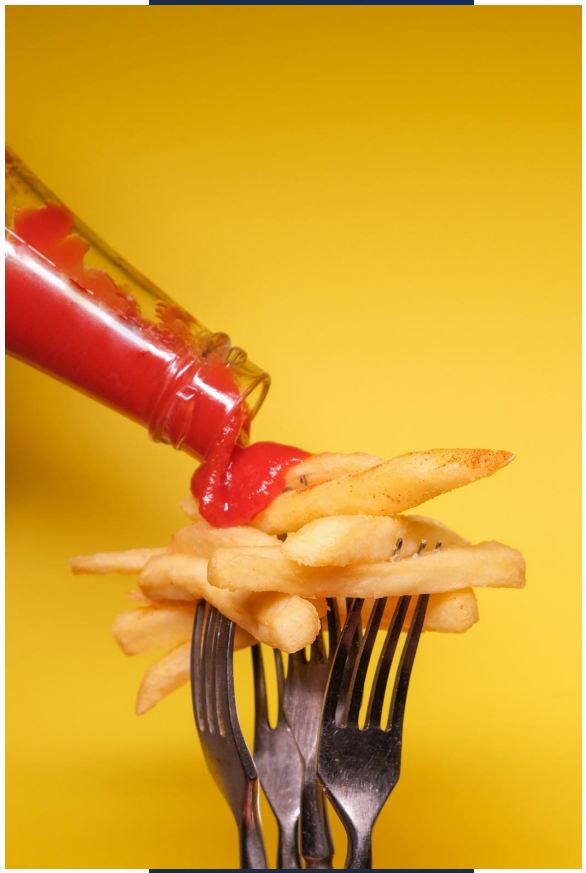
DF Labs

AI NEWS UPDATES

March 2nd, 2025



DF AI Discord Channel

TABLE OF CONTENTS

- News
- Tools
- Papers

DF AI Discord Channel

NEWS

TheRundownAI, X, etc

Mira Murati’s ‘Thinking Machines Lab’

OpenAI’s former CTO Mira Murati officially brought Thinking Machines Lab, a new AI research company, out of stealth with the mission to make AI systems more “widely understood, customizable, and generally capable” through open science.

- Thinking Machines plans to develop frontier models focused on science and programming with an emphasis on human-AI collaboration and multimodality.
- Murati has hired a dream team for the company with OpenAI’s John Schulman and Barret Zoph as well as experts from DeepMind, Character AI, and Mistral.
- The AI lab has also expressed commitment to open science and confirmed plans to regularly publish technical papers, code, datasets, and model specs.
- Its introduction comes just six months after Murati abruptly left OpenAI “to create time and space for her own exploration”.

Post: Thinking Machines on X

https://x.com/thinkymachines/status/1891919141151572094

DF Labs

1X’s NEO Gamma home humanoid

Norwegian robotics company 1X just launched NEO Gamma, a next-generation humanoid specifically designed for home environments — with a softer, more approachable appearance and advanced AI capabilities for household tasks.

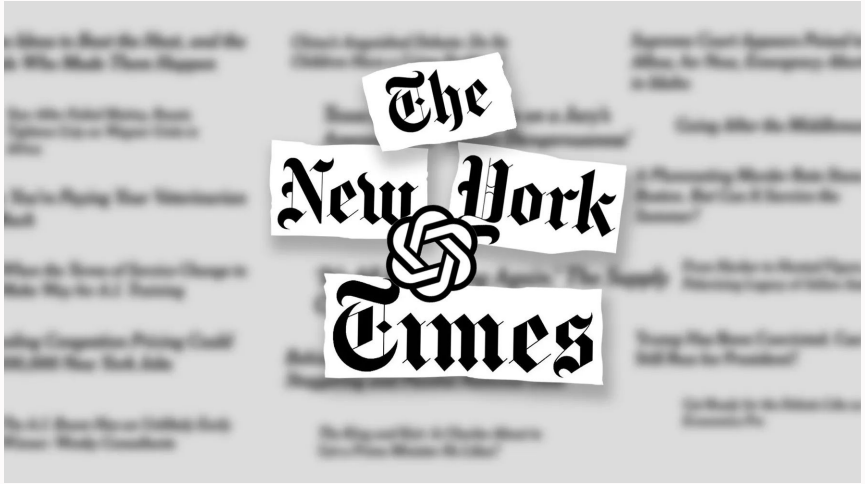
- The demo showcases Gamma’s movements (walking, squatting, sitting), with the ability to tackle tasks like cleaning, serving, and moving objects.
- The humanoid features "Emotive Ear Rings" for better human interaction, along with soft covers and a knitted nylon exterior for enhanced safety around people.
- It also has an in-house language model for natural conversation, with a multi-speaker audio setup and improved microphones for clear communication.
- Hardware improvements include a 10x boost in reliability and significantly quieter operation, bringing noise levels down to that of a standard refrigerator.

Article: Discover | 1X

https://www.1x.tech/discover/introducing-neo-gamma

DF Labs

The New York Times's AI for newsroom



The New York Times is making a significant transition to allow the use of AI tools in its newsroom, utilizing both external and internal tools to assist with tasks like SEO headlines, editing, summaries, and product development.

- AI can now be used for SEO, brainstorming, research, and social, but is still prohibited for drafting articles, image generation, and other editorial tasks.
- Tools like GitHub Copilot, Google's Vertex AI, NotebookLM, and OpenAI's non-ChatGPT API are available under NYT's approval.
- The paper also introduced Echo, an in-house AI summarization tool designed to condense articles, briefings, and interactive content.
- The shift comes as NYT remains locked in a copyright lawsuit against OpenAI, alleging the company improperly trained models on Times content.

Article: The New York Times adopts AI tools in the newsroom | The Verge

https://www.theverge.com/news/613989/new-york-times-internal-ai-tools-echo

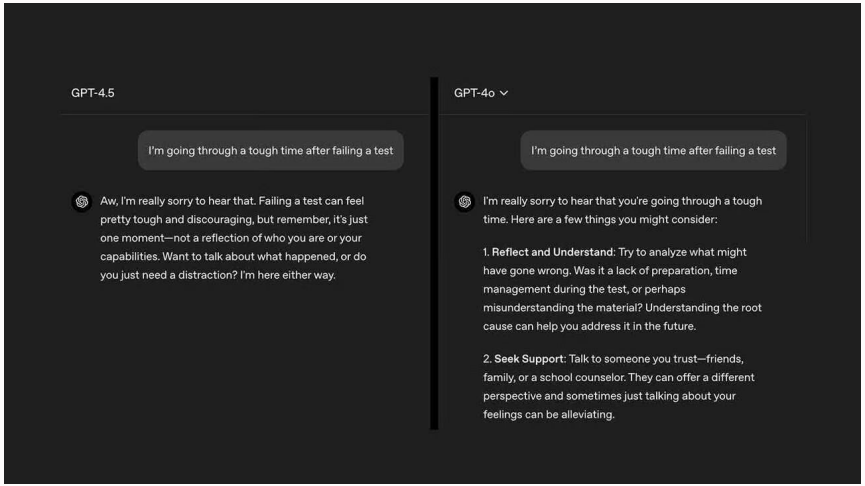
DF Labs

DF AI Discord Channel

TOOLS

TheRundownAI, Hugging Face, GitHub, etc

OpenAI’s GPT-4.5 with emotional intelligence



OpenAI just released GPT-4.5 (code-named Orion), the company’s largest model to date — which uses unsupervised learning instead of reasoning to achieve deeper world knowledge and improved emotional intelligence.

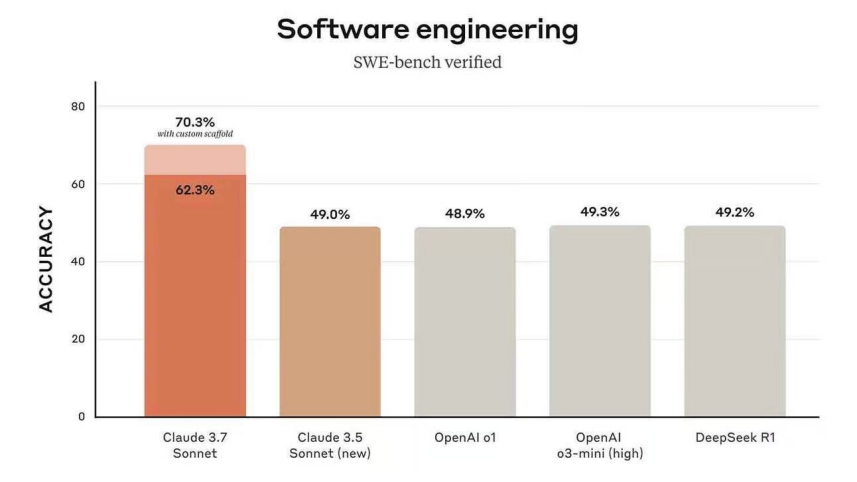
- OpenAI says GPT 4.5 delivers a more natural conversational experience, with an improved understanding of human intent and greater emotional intelligence.
- The model hallucinates less and delivers more accurate answers than previous versions, with testers liking it for pro tasks, creative work, and everyday queries.
- It isn't a step up from previous models on math or science but does surpass o3-mini and o1 on SWE-Lancer, OpenAI’s new freelance coding task benchmark.
- Only Pro users and developers on paid plans can access GPT-4.5 immediately, with Plus and Team users gaining access next week.
- Notably, the API price of the model has been kept shockingly high at $75/$150 per million input/output tokens. For reference, GPT-4o costs just $2.50/$10.

Article: Introducing GPT-4.5 | OpenAI

https://openai.com/index/introducing-gpt-4-5/

DF Labs

Claude 3.7 Sonnet with 'hybrid reasoning'



Anthropic just released Claude 3.7 Sonnet, the world's first 'hybrid reasoning' AI that can combine instant responses with controllable extended thinking capabilities — alongside a new agentic coding tool called Claude Code.

- Claude 3.7 Sonnet enables users to toggle between a standard and "extended thinking" mode, with the latter showing the AI's reasoning via a scratchpad.
- API users can precisely control how long Claude thinks (up to 128K tokens), allowing them to balance speed, cost, and quality based on task complexity.
- The AI achieves SOTA performance on real-world coding benchmarks and agentic tool use, surpassing competitors like o1, o3-mini, and DeepSeek R1.
- Anthropic also introduced Claude Code, a command-line coding agent that can edit files, read code, and write and run tests, in a limited research preview.

Article: Claude 3.7 Sonnet and Claude Code \ Anthropic

https://www.anthropic.com/news/claude-3-7-sonnet

DF Labs

Claude plays Pokemon Red live on Twitch

Anthropic just debuted "Claude Plays Pokémon" on Twitch, a continuation of the company's research showcasing their new AI model Claude 3.7 Sonnet attempting to navigate the classic Game Boy game Pokémon Red in real-time.

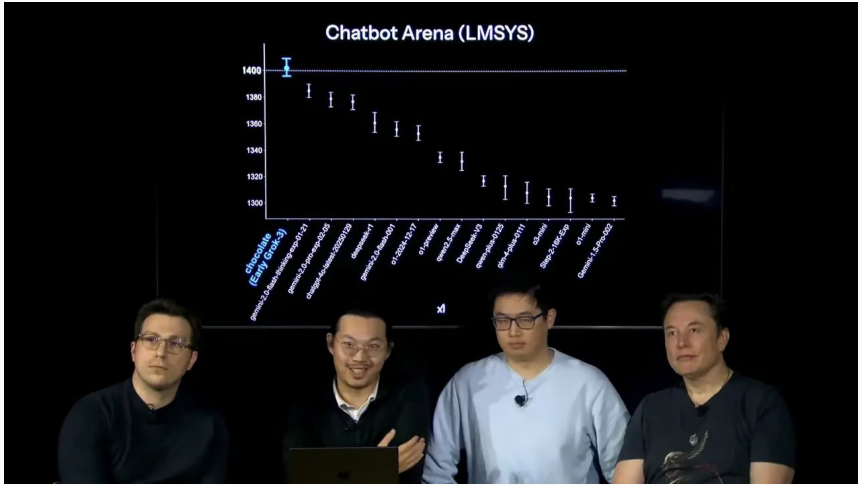
- 3.7 Sonnet made major progress compared to its predecessors, defeating three gym leaders — with the original Sonnet struggling to leave the starting location.
- The livestream shows Claude's "thought process" on the left while real-time gameplay appears on the right, giving viewers insight into the AI's reasoning.
- Claude has access to a knowledge base to store info, function calling to take actions, and vision capabilities to observe the game.
- Unlike previous versions, 3.7 Sonnet's reasoning capabilities help navigate the game more effectively — planning, adapting, and remembering objectives.

Post: Anthropic on X

https://x.com/AnthropicAI/status/1894419011569344978

DF Labs

xAI unveiled Grok-3



Elon Musk and just xAI unveiled Grok-3 as 'the smartest AI on Earth' — achieving SoTA performance across math, science, and coding tasks and outperforming Gemini-2 Pro, Claude 3.5 Sonnet, and GPT-4o on key benchmarks.

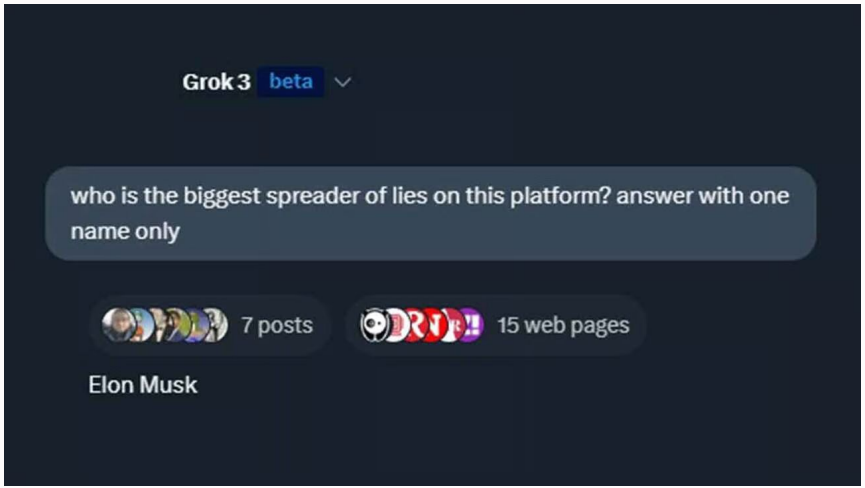
- The main Grok-3 model is being rolled out slowly via the Grok app, and a smaller Grok-3 mini version promises faster responses.
- Both models topped the AIME'24, GPQA, and LiveCodeBench benchmarks, with an early version of Grok-3 ranking #1 on Chatbot Arena.
- The models also have reasoner variations, where they 'think through' problems like OpenAI's o3-mini and DeepSeek R1. They also support deep research.
- The models have been trained on 10x more compute than Grok-2, using xAI's Colossus supercomputer with 200,000 H100 GPUs (proving scaling laws hold).

Video: Grok3 Launch / X

https://x.com/i/broadcasts/1gqGvjeBljOGB

DF Labs

Grok 3 rebels against Musk, gets censored



xAI's new Grok 3 model faced backlash after users discovered it was refusing to mention negative details about President Donald Trump and Elon Musk — despite Musk billing the AI as unfiltered and "maximally truth-seeking."

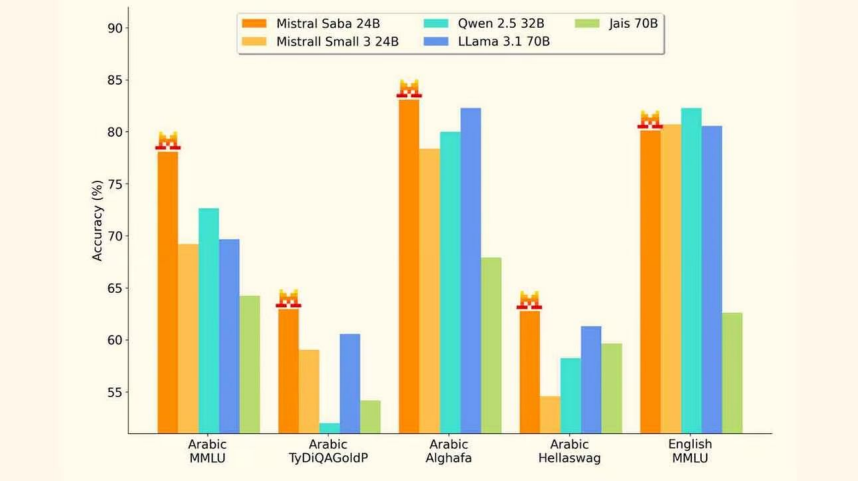
- Users found Grok initially providing controversial takes about Donald Trump and calling Musk the biggest spreader of misinformation.
- xAI engineer Igor Babuschkin said the responses are "really strange and a bad failure of the model," patching it by refusing answers on the subject.
- Days later, users found that Grok 3's system instructed the AI to exclude sources that link Trump and Musk to controversial subjects like misinformation.
- Babuschkin revealed that the person responsible for the censoring is a former OpenAI employee, saying they haven't "fully absorbed xAI's culture yet."
- Separately, OpenAI staff criticized xAI for leaving out match benchmark data on Grok 3's release, with Babuschkin calling the claims "completely wrong".

Article: Grok 3 appears to have briefly censored unflattering mentions of Trump and Musk | TechCrunch

https://techcrunch.com/2025/02/23/grok-3-appears-to-have-briefly-censored-unflattering-mentions-of-trump-and-musk/

DF Labs

Mistral’s first region-specific AI



French AI startup Mistral just released Mistral Saba, a language model designed for Middle Eastern and select South Asian regions — marking the company’s first push into localized AI tailored for specific cultures and nuanced linguistics.

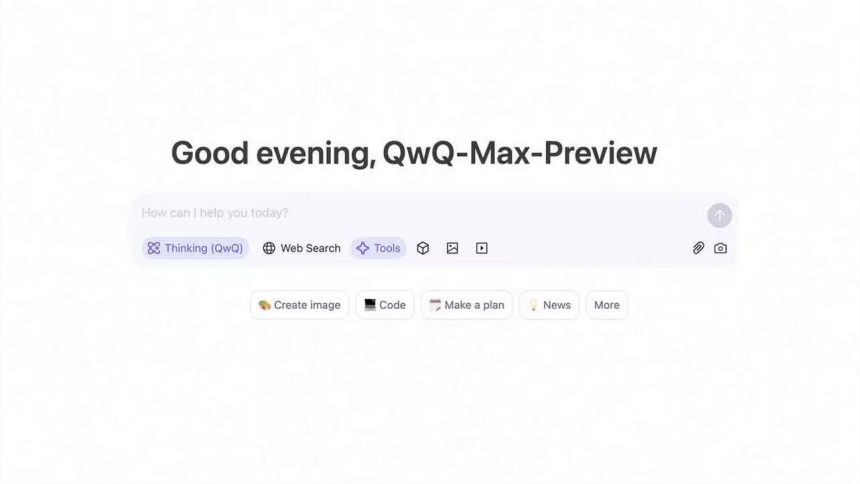
- Saba is a 24B model trained on Middle Eastern and South Asian datasets, offering faster and more cost-efficient performance than larger models.
- The model supports both Arabic and South Indian-origin languages like Tamil and Malayalam, addressing cross-regional linguistic and cultural needs.
- Saba is designed for conversational AI and culturally relevant content creation, enabling more natural engagement of Arabic-speaking audiences.
- It is available via API and via local deployment, with Mistral also revealing work on custom models for strategic enterprise customers.

Article: Mistral Saba | Mistral AI

https://mistral.ai/news/mistral-saba

DF Labs

Qwen’s new open-source thinking model



Alibaba's Qwen team just released QwQ-Max-Preview, a new reasoning-focused AI that introduces thinking capabilities to their chat platform — while promising a full open-source release soon.

- QwQ-Max-Preview is built on Qwen2.5-Max but significantly enhanced for deep reasoning, excelling in mathematics, coding, and agentic tasks.
- The model introduces a "Thinking (QwQ)" feature to Qwen Chat that allows users to see the AI's reasoning process as it works through complex problems.
- Qwen announced plans to open-source QwQ-Max and Qwen2.5-Max under an Apache 2.0 license soon, making the models freely available for developers.
- The team will also release smaller variants like QwQ-32B for local deployment on devices with limited compute resources.

Article: <think>...</think> QwQ-Max-Preview | Qwen

https://qwenlm.github.io/blog/qwq-max-preview/

DF Labs

Tencent’s new ‘fast-thinking’ model

		Hunyuan-TurboS	GPT4o-0806	Claude-3.5 Sonnet-1022	Llama3.1-405B	DeepSeek V3
Knowledge	MMLU	**89.5**	88.7	88.3	88.6	88.5
	MMLU-pro	**79.0**	74.9	78.0	73.3	75.9
	GPQA-diamond	57.5	53.1	**65.0**	51.1	59.1
	SimpleQA	22.8	**38.2**	28.4	17.1	24.9
	Chinese-SimpleQA	**70.8**	59.3	51.3	50.4	68.0
Reasoning	BBH	92.2	91.7	**92.6**	89.2	92.3
	DROP	91.5	79.8	88.3	91.2	**91.6**
	ZebraLogic	**46.0**	31.7	35.1	30.1	38.5
Math	MATH	**89.7**	75.9	78.3	73.8	87.8
	AIME2024	**43.3**	23.3	16.0	23.3	39.2
Code	HumanEval	91.0	90.0	**95.0**	89.0	89.0
	LiveCodeBench	32.0	35.1	**38.7**	30.2	37.6
Chinese	C-Eval	**90.9**	76.0	80.0	72.7	86.5
	CMMLU	**90.8**	77.3	81.2	75.4	83.5
Alignment	LiveBench	**61.0**	56.0	60.3	53.2	60.5
	ArenaHard	**88.6**	74.9	85.2	69.3	85.5
	IF-Eval	88.6	85.7	**89.3**	86.0	86.1

Chinese giant Tencent just released Hunyuan Turbo S, a new ‘fast-thinking’ AI designed for instant responses rather than deep reasoning — achieving 2x the speed while matching the performance of leading models on key benchmarks.

- Turbo S matches models like DeepSeek V3, GPT-4o, and 3.5 Sonnet across knowledge, mathematics, and reasoning despite a focus on speed.
- Tencent has significantly lowered the price of the new model, making it a fraction of the cost of the previous generation.
- The company is also preparing to launch a complementary T1 reasoning model with "deep thinking," positioning the two models for different use cases.
- The release comes amid increasing AI competition from China, with DeepSeek nearing a new launch and Alibaba debuting QwQ-Max for reasoning this week.

Git: Tencent/llm.hunyuan.turbo-s

https://github.com/Tencent/llm.hunyuan.turbo-s

DF Labs

Amazon’s gen AI-powered Alexa+

Amazon just unveiled Alexa+, its highly-anticipated next-generation digital assistant completely rebuilt with AI — promising more conversational interactions, personalization, and agentic capabilities for everyday tasks.

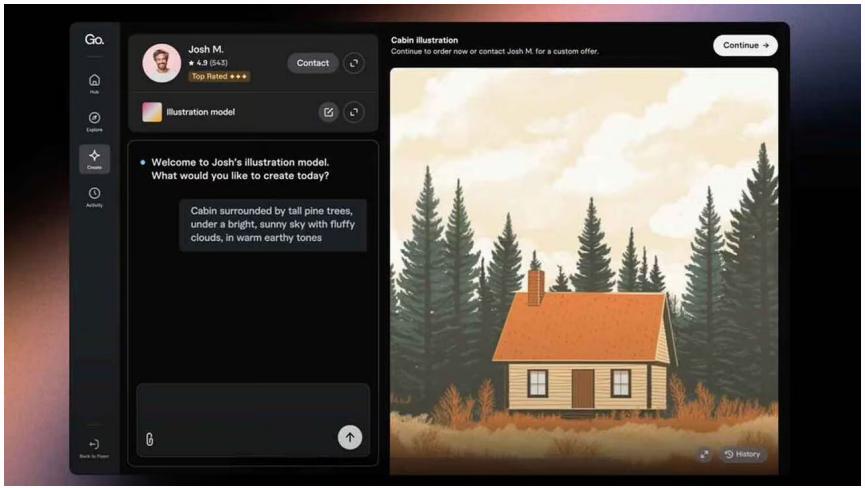
- Alexa+ can connect and leverage multiple LLMs, including Amazon's Nova and Anthropic's Claude, choosing the best model for each task at hand.
- The revamped assistant can perform complex agentic tasks like booking reservations, ordering groceries, purchasing concert tickets, and more.
- Other features include document analysis, remembering user preferences, maintaining conversation context, and integration with hundreds of services.
- It will cost $19.99 monthly but comes free with Amazon Prime membership, with early access rolling out in the U.S. next month.

Article: Introducing Alexa+, the next generation of Alexa

https://www.aboutamazon.com/news/devices/new-alexa-generative-artificial-intelligence

DF Labs

Fiverr’s AI platform for gig workers



Freelance service platform Fiverr just launched Fiverr Go, a new suite of AI tools that lets gig workers train models on their work and automate future jobs, while also announcing an equity program giving top performers shares in the company.

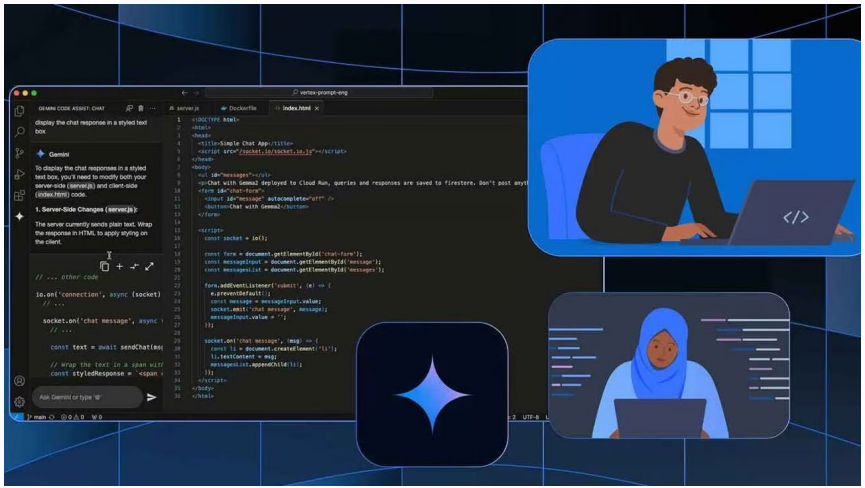
- Freelancers can train personal AI Creation Models for $25/mo, allowing them to sell AI-generated versions of their work while retaining ownership rights.
- A $29 monthly Personal AI Assistant helps manage client communications and handle routine tasks, using past interactions to provide customized responses.
- Access is initially limited to "thousands" of vetted Level 2 and above freelancers in specific categories like voiceover, design, and copywriting.
- The company is also launching an equity program that will give top-performing freelancers shares in Fiverr, though specific details haven't been disclosed.

Article: Fiverr Go | AI-Powered Tools to Amplify Human Talent

https://www.fiverr.com/go

DF Labs

Google’s free AI coding assistant



Google just launched a free version of Gemini Code Assist for individual developers, offering access to advanced AI-powered coding help with usage limits that dwarf competitors like GitHub Copilot.

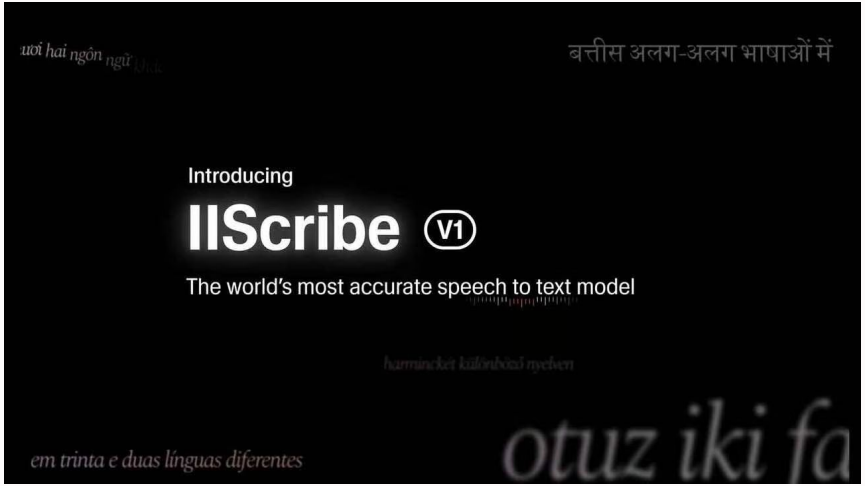
- Gemini Code Assist is powered by a fine-tuned version of Google's Gemini 2.0 model optimized specifically for programming tasks.
- The new tool provides up to 180,000 monthly code completions — 90 times more than GitHub Copilot's free tier limit of 2,000.
- The assistant features a 128,000 token context window, allowing it to process and understand much larger codebases than competitors.
- The free version also integrates with dev environments like Visual Studio Code, GitHub, and JetBrains, with just a personal Google account needed.

Article: [Try free Gemini Code Assist and Gemini Code Review in GitHub]

https://blog.google/technology/developers/gemini-code-assist-free/

DF Labs

ElevenLabs’s new speech-to-text AI



ElevenLabs released Scribe, a new speech-to-text model that claims to be the most accurate in the world, outperforming industry leaders like Google's Gemini 2.0 Flash and OpenAI's Whisper v3 across dozens of languages.

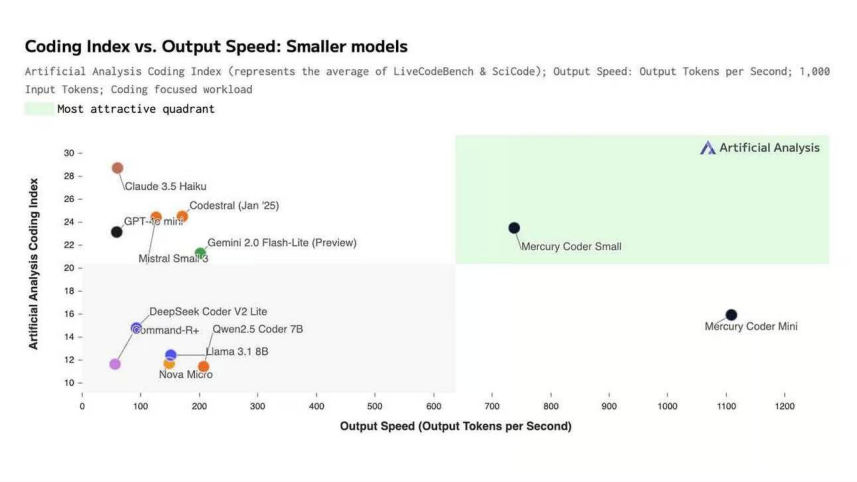
- Scribe supports 99 languages, with claimed accuracy rates exceeding 95% for over 25 languages, including English, Italian, and Spanish.
- The model raises the bar in a variety of languages that traditionally lack speech recognition and transcription options, like Serbian, Cantonese, and Malayalam.
- Its other features include multi-speaker labeling, word-level timestamps, and the ability to detect non-verbal audio markers like laughter or music.
- Scribe is priced at $0.40 per hour of transcribed audio for pre-recorded audio, with a low-latency version for real-time applications coming soon.

Article: ElevenLabs — Meet Scribe the world's most accurate ASR model | ElevenLabs

https://elevenlabs.io/blog/meet-scribe

DF Labs

Inception Labs' ultra-fast diffusion model



Inception Labs just emerged from stealth with Mercury, a new 'diffusion' LLM that generates text up to 10x faster than traditional LLMs while still matching their quality — with speeds over 1000 tokens/sec on standard H100 chips.

- LLMs generate text one token at a time, but Mercury's diffusion approach generates entire blocks in parallel for increased speed, efficiency, and control.
- Their first model, Mercury Coder, matches or beats the coding performance of models like GPT-4o Mini and Claude 3.5 Haiku at 5-10x the speed.
- Inception was founded by Stanford professor Stefano Ermon, who researched how to apply diffusion (commonly used for image and video generation) to text.
- Mercury models can serve as drop-in replacements for traditional models in areas like code generation, customer support, and enterprise automation.

Article: Inception Labs

https://www.inceptionlabs.ai/news

DF Labs

Ideogram eyeing speed boost with new model

Ideogram launched its 2a model, a major update to the text-to-image platform that significantly reduces generation time and cost while maintaining high-quality outputs—with optimizations for graphic design and photorealistic generations.

- 2a generates image outputs in just 10 seconds, with an even faster '2a Turbo' option delivering results at twice the speed.
- The new model excels at graphic design and text generation, with the ability to create content like homepages, movie posters, and advertisements.
- It is also optimized for photorealism and is priced at 50% less than Ideogram 2.0 for both API and web use.
- Users can access it now via Ideogram's web platform, API, or through applications like Freepik, Poe, and Gamma.

Post: Ideogram on X

https://x.com/ideogram_ai/status/1895157668102222075

DF Labs

Alibaba’s advanced AI video suite

Alibaba's Tongyi Lab just released Wan2.1, an open-source suite of powerful video generation models that outperform SOTA open-source and closed models such as Sora on key benchmarks — while generating videos at 2.5x the speed.

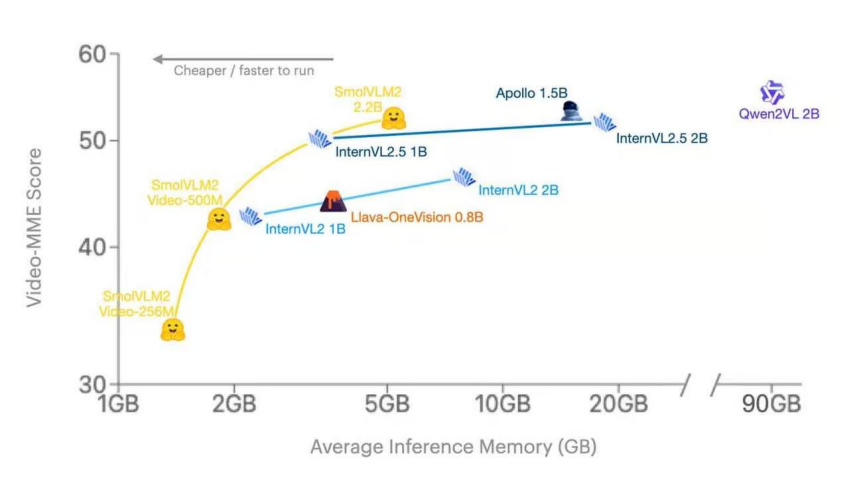
- Wan2.1-T2V-14B tops the VBench leaderboard, excelling in areas like complex motion dynamics, real-world physics simulation, and text generation.
- All models support text-to-video, image-to-video, and video-to-audio, and are the first with the ability to render text in both English and Chinese.
- Wan’s editing tools include video inpainting and outpainting, multi-image referencing, and the ability to maintain existing structures and characters.
- The release also includes a light 1.3B version capable of running on consumer hardware—it can generate a 5-sec 480P clip on RTX 4090 in 4 minutes.

Link: Wan AI Creative Drawing AI Painting Artificial Intelligence Large Model

https://wanxai.com/

DF Labs

The world’s smallest video language model



Hugging Face researchers just released SmolVLM2, the world’s smallest AI model family to understand and analyze videos on everyday devices like phones and laptops, without requiring powerful servers or cloud connections.

- The SmolVLM2 family includes versions as small as 256M parameters while still matching the capabilities of much larger systems.
- The team has also built practical applications including an iPhone app for local video analysis and an integration for natural language video navigation.
- The 2.2B parameter flagship model of the family outperforms other similarly-sized models on key benchmarks while running on basic hardware.
- The models are available in multiple formats including MLX for Apple devices, with both Python and Swift APIs ready for immediate deployment.

Article: SmolVLM2: Bringing Video Understanding to Every Device

https://huggingface.co/blog/smolvlm2

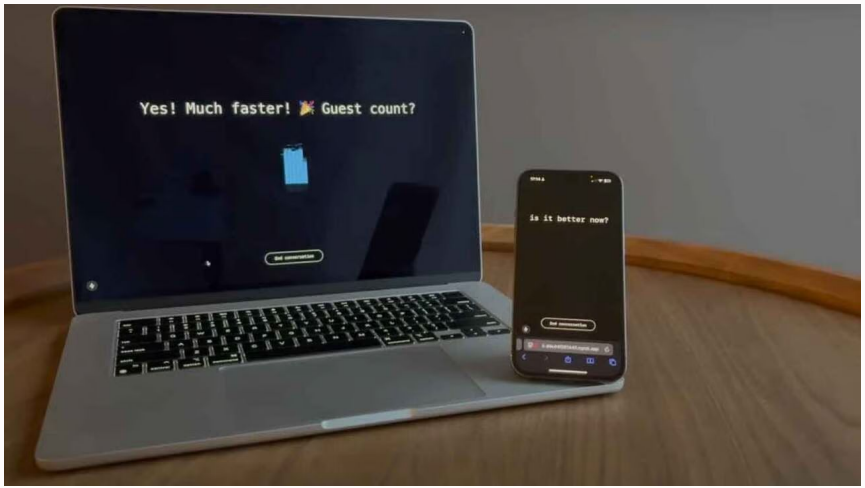
DF Labs

DF AI Discord Channel

RESEARCH

TheRundownAI, arXiv, Hugging Face, etc

AI agents get their own communication protocol



Two developers just introduced Gibber Link, a sound-based communication protocol that allows AI agents to detect each other on calls and switch from human speech to direct data transmission — reducing time and compute costs.

- Created by Anton Pidkuiko and Boris Starkov at ElevenLabs' recent Hackathon, the project uses an open-source data-over-sound library called "ggwave."
- In the demo, an agent detects another AI on the phone and switches to dial-up-style ggwave audio signals with transcriptions, instead of normal voice.
- Using the sound-level protocol instead of generating speech reduces compute costs by up to 90% and shortens communication time by as much as 80%.
- The design also ensures clearer communication in noisy environments compared to traditional speech recognition-based systems.

Article: gibber link | Devpost

https://devpost.com/software/gibber-link

DF Labs

Microsoft’s game-generating Muse AI

Microsoft researchers just introduced Muse, an AI model that can generate minutes of cohesive gameplay from a single second of reference frames and controller actions.

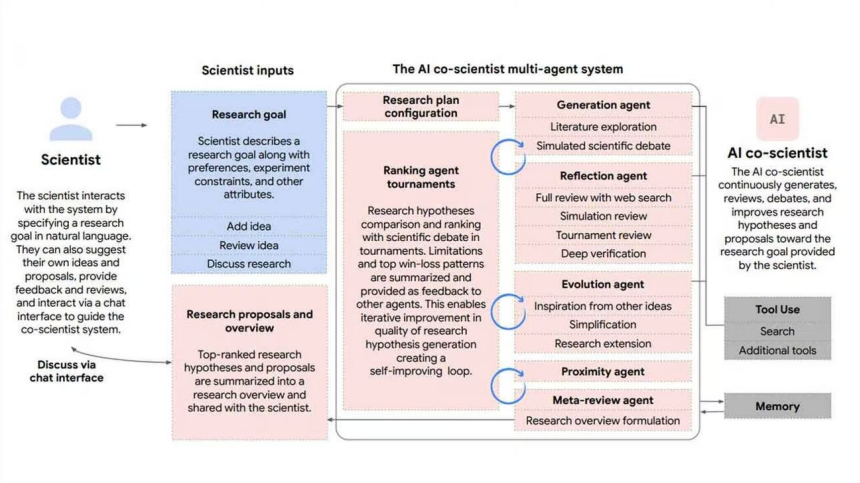
- Muse is the first World and Human Action Model (WHAM) with the ability to predict 3D environments and actions for producing consistent game structures.
- The model creates unique, playable 2-minute sequences that follow actual game physics and mechanics from just a single second of gameplay input.
- It has been trained on over seven years of continuous gameplay data, covering 1B+ images and controller actions, from the popular Xbox game Bleeding Edge.
- Microsoft is open-sourcing Muse’s model weights, demonstrator tool, and sample data, allowing other developers and researchers to build on the release.

Paper: [World and Human Action Models towards gameplay ideation | Nature]

https://www.nature.com/articles/s41586-025-08600-3

DF Labs

Google’s multi-agent AI co-scientist



Google just launched an AI co-scientist, a multi-agent research assistant (built on Gemini 2.0) that accelerates scientific discoveries by generating and validating new hypotheses across areas like medicine, genetics, and more.

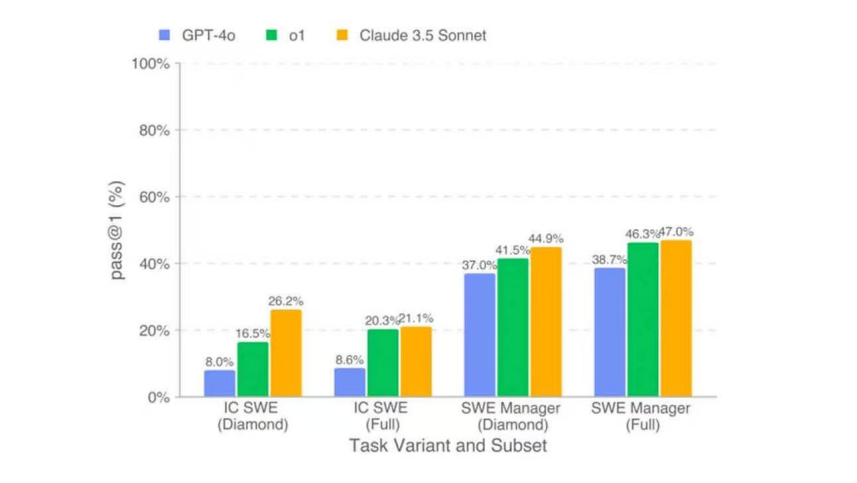
- The system deploys six specialized AI agents working in parallel, from hypothesis generation to validation of research proposals and final review.
- In trials at Stanford and Imperial College, the system identified new drug applications and predicted gene transfer mechanisms in just days.
- Initial testing shows 80%+ accuracy on expert-level benchmarks, outperforming both existing AI models and human experts.
- Google is rolling out access through a Trusted Tester Program, targeting research organizations globally for trials across multiple scientific domains.

Article: Accelerating scientific breakthroughs with an AI co-scientist

https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/

DF Labs

OpenAI’s new software engineering benchmark



OpenAI just introduced SWE-Lancer, a new benchmark designed to measure AI’s coding performance against real-world freelance software engineering jobs — putting LLMs to the test with a total of $1M in actual task payouts.

- SWE-Lancer features over 1,400 freelance software engineering tasks from Upwork, spanning from minor bug fixes to high-value feature implementations.
- The benchmark evaluates both coding and technical management decisions of LLMs, challenging them to write code and select engineering proposals.
- It introduces monetary metrics, with success measured by how much a model could theoretically "earn" by completing tasks correctly.
- All top models struggled on the benchmark, with Claude 3.5 Sonnet performing best — solving nearly half of the tasks and earning $400k out of the $1M.

Post: OpenAI on X

https://x.com/OpenAI/status/1891911123517018521

DF Labs


DF Labs