FAST 5

Generative AI Headlines This Week

March 22, 2025

1

Nvidia’s AI & Robotics Advances at GTC

Brief:

Nvidia CEO **Jensen Huang opened GTC 2025 with a two-hour keynote, calling it “AI’s Super Bowl” and unveiling major updates on chips, robotics, and autonomous vehicles**.

Breakdown:

- Nvidia’s upcoming GPUs include Blackwell Ultra (late 2025), Vera Rubin (2026), and Feynman (2028), each promising significant performance gains.
- Huang emphasized that scaling is not slowing down and computational demand for AI is “100x more than we expected a year ago.”
- He introduced Isaac GR00T N1, the first open humanoid robot foundation model, alongside a comprehensive physical AI training dataset.
- The new DGX Spark and DGX Station bring data center-grade AI computing to personal workstations, with Huang calling it “the computer for the AI age.”
- A new robotics physics engine, Newton, developed with Google DeepMind and Disney was showcased through ‘Blue,’ a Star Wars-style robot.
- Nvidia also announced a partnership with General Motors to build the company’s first fleet of self-driving cars.

Why It’s Important:

Huang called AI’s progress an ‘inflection point.’ Nvidia’s advancements reinforce its dominance in AI infrastructure across industries. If its roadmap holds, AI acceleration isn’t slowing down anytime soon.

Claude Gets Real-Time Web Search

Brief:

Anthropic added web search capabilities to Claude, **enabling real-time information access and closing a significant feature gap with competitors like ChatGPT and Gemini**.

Breakdown:

- Web search is integrated directly into Claude 3.7 Sonnet, automatically triggering when more current or accurate information is needed.
- Claude provides direct citations for web-sourced information, making it easy for users to verify sources and fact-check responses.
- The feature is available to all paid Claude users in the U.S., with plans for international and free-tier access soon.
- Users can also enable the feature by toggling the ‘Web Search’ tool in their profile settings.

Why It’s Important:

It took Claude longer than its rivals to get web access, but Anthropic’s models remain among the best, and real-time web access could take them to the next level in an increasingly competitive market.

OpenAI Enhances Voice AI with Custom Personality

Brief:

OpenAI launched its next-gen **API-based audio models for text-to-speech and speech-to-text, allowing developers to customize AI speaking styles and providing improved speech recognition** across multiple languages.

Breakdown:

- The new gpt-4o-mini-tts model adjusts its speaking style based on simple text prompts, such as "speak like a pirate" or "use a bedtime story voice."
- The GPT-4o-transcribe speech-to-text models achieve state-of-the-art performance in accuracy and reliability, surpassing existing Whisper models.
- OpenAI also introduced openai.fm, a public demo platform where users can test different voice styles and experience the new models.
- These models are available via OpenAI's API, with integration support through the Agents SDK for developers building voice-enabled AI assistants.

Why It's Important:

Customizing voice outputs means more dynamic and natural AI interactions and a wider range of applications. However, OpenAI's demos appear to lag behind competitors like ElevenLabs in terms of human-like voice quality.

NVIDIA Launches Open Reasoning Models

Brief:

Nvidia released its Llama Nemotron family of open-source reasoning models, d**esigned to accelerate enterprise adoption of agentic AI capable of complex problem-solving and decision-making**.

Breakdown:

- The new model family includes Nano (8B), Super (49B), and Ultra (249B), each optimized for different deployment scenarios.
- Early benchmarks show the Super version outperforms Llama 3.3 and DeepSeek V1 across STEM and tool testing.
- The models feature a hybrid toggle, allowing them to switch between intensive reasoning and direct responses based on the task.
- Post-training improvements include 20% better accuracy than base Llama models and 5x faster speed than rival open reasoners.
- Nvidia is releasing an "AI-Q Blueprint" framework in April to help businesses connect AI agents with their systems and data sources.

Why It’s Important:

Nvidia's reasoning models, along with many other announcements from its GTC conference, positioned the company with all the necessary components to compete across the AI stack from advanced hardware to high-performance reasoning models.

5

Baidu Launches Ultra-Low-Cost AI Models

Brief:

Baidu unveiled two **ultra-low-cost AI models: ERNIE 4.5, a major upgrade to its foundational model, and ERNIE X1, a deep-think-capable model**.

Breakdown:

- ERNIE 4.5 features enhanced EQ, language skills, hallucination prevention, logical reasoning, and coding capabilities.
- Baidu claims ERNIE 4.5 outperforms GPT-4o on multiple benchmarks while costing just 1% of its price, around $0.55 and $2.20 per million input and output tokens.
- ERNIE X1, Baidu's first reasoning-focused model, rivals China's DeepSeek R1 at half the cost.
- Like DeepSeek R1, ERNIE X1 uses a step-by-step "thinking" approach, excelling in complex calculations and tasks like document understanding.

Why It's Important:

China continues to drive AI costs towards zero, with ERNIE 4.5 at 1% of GPT-4o's price and ERNIE X1 undercutting DeepSeek R1. This AI price war could push Western firms to slash rates, making advanced AI more accessible worldwide.

WANT TO GO **BEYOND THE HEADLINES**

VIEW MY NEWSLETTER

Generative AI Enterprise

WELCOME, EXECUTIVES

In today's edition:

- Will Agentic AI eliminate or enhanc
- Sequoia on adopting the stochastic
- OpenAI's web agent brings enhanc
- Galileo shares AI agent evaluation
- Transformation and technology in t
- Career opportunities & events.

Read time: 4 minutes.

PLATFORM STRATEGY

Will Agentic AI eliminate o

ENTERPRISE TRANSFORMATION

Accenture envisions the future with gen AI agents

BEST PRACTICE INSIGHT

Gartner's 10 best practices for scaling GenAI

ACCELERATORS & CASE STUDIES

BCG AI playbooks for senior executives

Image source: Boston Consulting Group

Brief: Boston Consulting Group (BCG) released two new playbooks, totaling nine, to guide CEOs and senior executives (CFOs, COOs etc.) on AI transformation, including GenAI. Each ~25 slide playbook draws from 1,000+ AI programs, offering strategies, roadmaps, and case studies to drive enterprise value.

Breakdown:

- For Finance Leaders, the 23 slide Finance playbook explores AI opportunities, evolution of processes, operating models, and more.
- For Technology Leaders, the 24 slide Data & Digital Platforms playbook covers maximizing data value, evolving tech stacks for AI, and more.
- For Operations Leaders, the 20 slide Supply Chain playbook highlights GenAI applications and starting points, while the 24 slide Customer Service playbook explores economics, teams, and more.
- For Risk Leaders, the 32 slide Risk & Compliance playbook details AI driven risk management, capability evolution, and more.
- For Sales Leaders, the 23 slide Customer Engagement playbook covers ideation, personalization, and communication, while the 23 slide B2B Sales playbook focuses on team evolution, strategies, and more.
- For People Leaders, the 22 slide HR playbook outlines future structures, tools, skills, and GenAI performance gains.

Why it's important: These playbooks for leaders provide actionable strategies and real world learnings to increase the likelihood of success in realizing tangible value from AI investments.

BEST PRACTICE INSIGHT

McKinsey's AI transformation blueprint for banking

Follow Lewis Walker for more!