# AI Unveiled: Deep Research on the Most Important Discoveries and News in the World of AI from the Past 7 Days

## I. The Global Reasoning War: Architectural Divergence and the Agentic Pivot

The first week of December 2025 has unequivocally signaled the end of the "chatbot era" and the commencement of the "reasoning era." The discourse within the artificial intelligence community has shifted dramatically from metrics of conversational fluency to the rigorous quantification of agentic reliability, long-horizon task completion, and architectural efficiency. This transition is not merely semantic; it is structural. The leading research laboratories—OpenAI, Google DeepMind, Anthropic, and the disruptive newcomer DeepSeek AI—have simultaneously deployed systems that prioritize "thinking" over "predicting," fundamentally altering the technological landscape.

This period, defined by the "AI Unveiled" theme, has witnessed a bifurcation in development strategies. On one side, Western incumbents like OpenAI and Google are entrenching their dominance through vertical integration, embedding massive reasoning engines into proprietary ecosystems like Windows and Google Workspace. On the other side, the open-weight community, spearheaded by Chinese researchers, has introduced radical architectural innovations in sparse attention, challenging the assumption that frontier intelligence requires prohibitive computational resources. The result is a tripartite "Reasoning War" that has defined the news cycle from December 1 to December 7, 2025.

### 1.1 DeepSeek V3.2: The Efficiency Disruptor

Perhaps the most technically significant event of the week was the release of DeepSeek V3.2 and its specialized reasoning variant, V3.2-Speciale.[1] Emerging from the Chinese research lab DeepSeek AI, this release has reverberated through the industry not only for its performance capabilities—which benchmarks suggest rival or exceed those of Google's Gemini 3.0 Pro and OpenAI's GPT-5 series—but for its audacious architectural departure from standard Transformer designs.

#### The Architecture of DeepSeek Sparse Attention (DSA)

For years, the scaling laws of Large Language Models (LLMs) were constrained by the quadratic complexity of the attention mechanism. In a standard dense Transformer, every token in a sequence must calculate its relationship to every other token. As the context window length ($L$) increases, the computational cost grows by $O(L^2)$. This creates a

massive economic and latency barrier for tasks requiring long-context reasoning, such as analyzing entire codebases or processing legal repositories.

DeepSeek V3.2 introduces a novel mechanism termed **DeepSeek Sparse Attention (DSA)**, which effectively reduces this complexity to $O(Lk)$, where $k$ represents a small, dynamically selected subset of relevant tokens.[1] This is not a simple "sliding window" or fixed-pattern sparsity, which often degrades performance by missing long-range dependencies. Instead, DSA utilizes a sophisticated, content-aware routing system.

The mechanism operates through a two-stage process that fundamentally reimagines how neural networks "pay attention":

1. **The Lightning Indexer:** Rather than performing heavy matrix multiplications for all token pairs, the model employs a lightweight "lightning indexer." This component computes a preliminary index score ($I_{t,s}$) between a query token and preceding tokens. The indexer is computationally inexpensive, utilizing a reduced number of attention heads and operating in low-precision FP8 formats to maximize throughput.[3] It acts as a high-speed filter, identifying potentially relevant information without fully processing it.
2. **Fine-Grained Token Selection:** Based on the scores generated by the lightning indexer, the model executes a selection mechanism that retrieves only the key-value (KV) entries corresponding to the top-$k$ scores. The heavy "thinking"—the full attention computation—is then applied exclusively to this sparse, high-relevance subset.[3]
3. **Dynamic Training Protocol:** The efficacy of this system relies on a specialized training regimen. The model first undergoes a "Dense Warm-up" stage, where the lightning indexer is trained via KL-divergence loss to mimic the behavior of a standard dense attention mechanism. This ensures the router learns to identify "important" tokens accurately. This is followed by a "Sparse Training" stage, where the entire model optimizes for the sparse pattern, cementing the efficiency gains.[3]

The implications of DSA are profound. By decoupling sequence length from quadratic cost, DeepSeek has demonstrated that "frontier" intelligence does not necessarily require the massive data center footprint of a GPT-5 class model. This challenges the "compute moat" that Western hyperscalers have relied upon to maintain their competitive advantage.[1]

### V3.2-Speciale and "Thinking in Tool-Use"

Alongside the generalist V3.2, DeepSeek released **V3.2-Speciale**, a model fine-tuned exclusively for extreme reasoning tasks. This model has achieved gold-medal level performance in the 2025 International Mathematical Olympiad (IMO) and the International Olympiad in Informatics (IOI).[2]

The most critical innovation in the Speciale variant is the concept of **"Thinking in Tool-Use"**.[2] In previous generations of AI agents, the "Chain of Thought" (CoT)—the internal monologue the model uses to reason—was often broken or reset when the model had to pause to call an

external tool (like a calculator or Python interpreter). This resulted in a loss of context and reasoning depth.

DeepSeek V3.2-Speciale maintains its reasoning state *through* the tool call. It can formulate a hypothesis, write code to test it, execute the code via an API, analyze the returned error or result, and refine its hypothesis—all without breaking the continuity of its internal reasoning chain. This capability, honed on a synthesis pipeline of over 85,000 complex instructions [2], allows for robust self-correction in software engineering and mathematical proof generation, areas where "one-shot" models frequently fail.

## 1.2 OpenAI GPT-5.1-Codex-Max: The Agentic Workhorse

While DeepSeek focused on architectural efficiency, OpenAI responded to the competitive threat by doubling down on reliability and integration with the release of **GPT-5.1-Codex-Max** on November 19, 2025, with wider availability rolling out through early December.[5] This release was framed internally as a "Code Red" response to the rapid advancements of rivals like Google and Anthropic.[6]

### The Compaction Mechanism and Long-Horizon Persistence

GPT-5.1-Codex-Max represents a departure from general-purpose chat models. It is explicitly architected for "long-horizon" agentic tasks—complex workflows that may span hours of computation and millions of tokens of context.[5] The defining technical feature enabling this is **"Compaction"**.[5]

In standard LLM interactions, context is ephemeral; as the conversation exceeds the window limit, earlier information is truncated. "Compaction" is a native capability where the model actively synthesizes its own history. Instead of simply dropping old tokens, Codex-Max recursively summarizes and compresses previous reasoning steps, decisions, and architectural states into a dense latent representation. This allows the model to work coherently across multiple context windows, effectively simulating an infinite memory for the duration of a specific project.[5]

This mechanism is particularly vital for enterprise software development. A developer can task Codex-Max with a "project-scale refactor"—a task that involves understanding dependencies across hundreds of files, planning an architectural migration, and executing changes incrementally. The compaction ensures that the model does not "forget" the initial architectural constraints or the user's specific requirements as it moves from file to file.[5]

### Operationalizing the Agent: Windows Integration and the CLI

OpenAI has positioned Codex-Max not merely as a text generator, but as an active participant in the development environment. It is the first model in the GPT family natively trained to operate within **Windows environments**.[5] This training data includes specific tasks related to

the Windows terminal, file system manipulation, and PowerShell command execution, making it a "better collaborator" in the Codex CLI (Command Line Interface).[5]

By optimizing the model for the specific nuances of the world's most dominant enterprise operating system, OpenAI is creating a lock-in effect. Codex-Max is designed to live inside the developer's workflow, acting as a headless autonomous agent that can navigate the directory structure, run tests, and debug errors without human intervention.[9] The shift is clear: the product is no longer "ChatGPT" (the chat interface); the product is the **agent** that lives in the terminal.

## 1.3 The "Big Three" Comparative Analysis

The nearly simultaneous updates from DeepSeek, OpenAI, and Google (with Gemini 3.0 Pro) have created a distinct hierarchy of capability for December 2025. While marketing narratives often obscure technical realities, a forensic analysis of the available benchmarks and technical reports reveals clear differentiation strategies.

**Table 1: Comparative Analysis of Leading Reasoning Models (December 2025)**

| **Feature / Metric** | **DeepSeek V3.2 / Speciale** | **GPT-5.1-Codex-Max** | **Gemini 3.0 Pro** | **Claude Opus 4.5** |
|---|---|---|---|---|
| **Primary Architecture** | Mixture-of-Experts (MoE) with **Sparse Attention (DSA)** | Transformer with **Native Compaction** | Multimodal Native (Transformer variant) | Dense Transformer (High Reasoning focus) |
| **Core Philosophy** | **Efficiency & Access:** Breaking the cost curve via architectural sparsity. | **Persistence & Reliability:** Long-horizon tasks via memory compaction. | **Integration & Scale:** Massive context and ecosystem dominance. | **Safety & Precision:** Maximizing "correctness" in code generation. |
| **Benchmark: SWE-bench Verified** | Competitive (Frontier Class) | **77.9%** (Highest Reported) | 76.2% | 77.2% |

| | | | | |
|---|---|---|---|---|
| **Benchmark: Terminal-Bench 2.0** | N/A | 58.1% | 54.2% | **61.3%** |
| **Benchmark: LiveCodeBench (Elo)** | High (Gold Medal Math Performance) | **2,243** | ~1,487 | 1,418 |
| **Output Speed** | ~50-80 tokens/sec (Cost-optimized) | ~150+ tokens/sec (Performance-optimized) | ~100+ tokens/sec | ~63 tokens/sec |
| **Context Handling** | Sparse Routing ($O(Lk)$) | Multi-Window Compaction | Massive Dense Window (1M+) | High-Accuracy Dense Window |
| **Access Model** | **Open Weights** (MIT License) | **Closed API** / Enterprise Only | **Closed Ecosystem** | **Closed API** |

**Market Insight:** The data indicates a convergence in top-line performance metrics (SWE-bench scores are within a 1.7% margin). However, the *utility* profiles are diverging. DeepSeek wins on raw price-to-performance, threatening to commoditize the reasoning layer for developers who can host their own models. OpenAI wins on *project persistence*, making it the superior choice for enterprise workflows that require maintaining state over days. Anthropic's Claude Opus 4.5 retains a slight edge in "safe" and verified terminal usage, appealing to security-conscious sectors.[10]

### 1.4 The "Code Red" and Strategic Delays

The intensity of this competition has forced strategic recalibrations. Reports confirm that OpenAI CEO Sam Altman issued an internal "Code Red," resulting in the delay of consumer-facing products—such as the "Pulse" personalized update feature and new advertising models—to focus all resources on regaining the reasoning crown from Google and DeepSeek.[6] This prioritization underscores that for the AI giants, the "reasoning capability" is the existential metric; all other product features are secondary to maintaining the smartest model on the market.

## II. The Quantum Leap: Google's "Willow" and the Threshold of Verification

While the generative AI market fights over percentage points on coding benchmarks, Google has claimed a milestone that may eventually render classical silicon bottlenecks obsolete. In early December 2025, Google unveiled the **"Willow" quantum chip**, a breakthrough that marks the transition of quantum computing from theoretical physics to engineering reality.[13]

## 2.1 Crossing the Error Correction Threshold

For three decades, the fundamental obstacle to useful quantum computing has been "noise." Qubits are notoriously fragile; interaction with the environment causes them to lose their quantum state (decoherence), introducing errors into calculations. Historically, adding more physical qubits to a system increased the total noise, making the computer *less* reliable, not more.

The "Willow" chip is historic because it is the first system to demonstrate **exponential error reduction below threshold**.[13] This means that as Google scaled up the number of physical qubits used to create a single "logical" (error-corrected) qubit, the error rate *decreased* exponentially ($10^{-2} \rightarrow 10^{-6}$).[14]

This is the "Wright Brothers moment" for quantum computing. It proves that the "surface code"—the algorithm used to correct errors by spreading information across many physical qubits—actually works at scale. It validates the roadmap toward building a fault-tolerant quantum computer capable of running indefinitely without crashing due to noise.[13]

## 2.2 The 10 Septillion Year Benchmark

To quantify this leap, Google executed a standard random circuit sampling benchmark on Willow. The chip completed the computation in **under five minutes**. Google's researchers estimate that executing the same calculation on one of the world's fastest classical supercomputers (such as Frontier or Aurora) would take **10 septillion ($10^{25}$) years**—a duration vastly exceeding the age of the universe.[13]

While critics often argue that random circuit sampling is a synthetic benchmark with limited real-world utility, its purpose is to demonstrate "quantum supremacy"—the point where a quantum machine performs a task that is physically impossible for a classical machine. Willow has moved the goalposts from "difficult for a supercomputer" (the Sycamore chip's claim in 2019) to "physically impossible for any classical computer that could ever exist."

## 2.3 Implications for the Future of AI

The convergence of "Willow-class" quantum processors and Artificial Intelligence is the next major frontier. The "Quantum Echoes" algorithm demonstrated on Willow [15] hints at a future where quantum computers serve as specialized accelerators for specific AI workloads.

1. **Chemical Simulation:** The immediate application is simulating molecular interactions for drug discovery and materials science—a task that is essentially a quantum mechanical

problem. AI models like AlphaFold have approximated this, but a quantum computer acts as a native simulator of nature.[13]

2. **Training Optimization:** Long-term, quantum linear algebra could revolutionize the training of neural networks. The optimization landscapes of massive models are complex high-dimensional spaces; quantum algorithms (like Quantum Grover Search or HHL) theoretically offer quadratic or exponential speedups in navigating these spaces.

While Willow will not be powering ChatGPT next year, it signals that the hardware ceiling for computation is about to be shattered. As classical scaling laws (Moore's Law) slow down, quantum scaling is just beginning its exponential ascent.

# III. The Industrialization of Agency: Amazon's Strategic Pivot

While Google pushed the boundaries of physics and DeepSeek redefined attention mechanisms, Amazon Web Services (AWS) utilized the first week of December to fundamentally restructure its AI offering. Moving away from the perception of being "behind" in the model wars, Amazon launched the **Nova** family, a suite of models and services designed to industrialize the deployment of AI agents in the enterprise.[16]

## 3.1 The Nova Model Family: Specialized Utility

Amazon's strategy eschews the "one model to rule them all" approach in favor of a diversified portfolio of specialized models, each optimized for a specific point on the price/performance curve [17]:

- **Nova Pro:** The "frontier intelligence" model, designed for complex reasoning, multi-step agentic tasks, and coding. It is the direct competitor to GPT-5.1 and Claude Opus.
- **Nova Lite:** A multimodal workhorse optimized for speed and cost. It is designed to process massive streams of documents, images, and videos in real-time, making it ideal for "backend" processing where latency is critical.
- **Nova Micro:** A text-only model focused on extreme low-latency responses, suitable for simple classification or routing tasks where cost is the primary constraint.

## 3.2 Nova Act: Operationalizing "Computer Use"

The most disruptive operational release from AWS is **Nova Act**.[16] This service represents the commoditization of "Computer Use"—the ability of an AI agent to interface with software through the Graphical User Interface (GUI) rather than an API.

Unlike fragile screen-scraping bots of the past, Nova Act utilizes a custom-trained Nova 2 Lite model that "sees" the screen. It separates perception (visualizing the DOM and UI elements) from execution (clicking, typing, scrolling).

- **The Web Gym:** To achieve reliability, Amazon trained these agents in synthetic environments called "web gyms"—simulated internets where the agents practiced booking flights, filling forms, and navigating complex enterprise dashboards millions of times.[16]
- **Reliability at Scale:** AWS claims a task reliability rate of over 90%, a significant improvement over the 60-70% rates often seen in open-source agent frameworks.[16]

**Economic Implication:** Nova Act effectively turns every legacy enterprise application into an API. Companies running ancient ERP systems or mainframe front-ends no longer need to embark on costly multi-year refactoring projects to modernize their stack. They can simply deploy Nova Act agents to "operate" the legacy software, creating an AI automation layer on top of existing technical debt.

### 3.3 Nova Forge: The Rise of "Novellas"

**Nova Forge** addresses the desire for enterprises to own their intelligence. It is a service that allows organizations to fine-tune and customize Nova models, creating bespoke versions referred to as **"Novellas"**.[18]

- **Checkpoint Access:** Uniquely, AWS is providing access to pre-training, mid-training, and post-training checkpoints. This allows companies to inject their data at the foundational level of the model, rather than just "polishing" the surface with standard fine-tuning.[19]
- **Data Blending:** To prevent "catastrophic forgetting"—where a model trained on medical data forgets how to speak English—Forge allows for "data blending." Enterprises can mix their proprietary data with Amazon's curated "replay buffers" of general knowledge, ensuring the model retains its general reasoning capabilities while becoming a domain expert.[18]
- **Reinforcement Fine-Tuning (RFT):** Forge democratizes Reinforcement Learning from Human Feedback (RLHF). Companies can define their own reward functions—for example, a chemistry lab could reward a model for suggesting molecules that are synthesizeable—and train the model to maximize that specific metric.[20]

This shifts the market dynamic from "renting intelligence" (API calls) to "manufacturing intelligence." Corporations are now building their own intellectual property in the form of model weights, moving from prompt engineering to model engineering.

## IV. Generative Physics and Spatial Intelligence

The innovation in text reasoning was matched by significant leaps in generative media. The first week of December saw the release of models that do not just "hallucinate" pixels, but demonstrate a coherent understanding of physics, time, and 3D space.

### 4.1 Google Veo 2: Physics-Aware Video Generation

Google released **Veo 2**, a state-of-the-art video generation model available via Vertex AI.[21] While previous video models were plagued by "morphing" artifacts (where objects would inexplicably change shape or vanish), Veo 2 is distinguished by its **natural motion generation** and physical consistency.[22]

- **Technical Specs:** The model generates 720p video at 24 frames per second (FPS) for durations of 5 to 8 seconds.[23]
- **Causal Reasoning:** Veo 2 demonstrates an understanding of cause and effect. In generated videos, a character does not just fall; they fall *because* they slipped on a wet surface. This "physics-awareness" suggests that the model is learning an internal world model, not just pixel statistics.[24]
- **Cinematic Control:** The model supports advanced camera controls, allowing users to specify lens types, pans, and zooms. This positions Veo 2 not just as a consumer toy, but as a professional tool for storyboarding and pre-visualization in the film industry.[22]

### 4.2 Meta SAM 3: Segmenting the Physical World

Meta advanced the field of computer vision with the release of the **Segment Anything Model 3 (SAM 3)** on November 19 (impacting the Dec 1-7 news cycle).[25] SAM 3 represents the bridge between Large Language Models and the physical world.

- **Temporal Segmentation:** Unlike SAM 2, which focused on images, SAM 3 excels at video. It can track objects across time, maintaining a persistent identity for a "car" or "person" even as they move behind obstacles (occlusion) or change orientation.[24]
- **Text-to-Segmentation:** SAM 3 introduces an "Open Vocabulary" interface. Users can prompt the model with text (e.g., "highlight all the red hats") rather than manual clicks. This allows LLMs to "see" and "query" the visual world, a critical capability for robotics.[26]
- **3D Reconstruction:** The model supports inputs for 3D reconstruction pipelines, enabling the creation of "digital twins" of physical spaces.[27]

### 4.3 Amazon Nova Reel and Canvas

Amazon joined the generative media market with **Nova Reel** (video) and **Nova Canvas** (image).[28]

- **Nova Reel:** Generates 6-second videos at 720p/24fps. It is explicitly designed for e-commerce, with features to turn a static product image into a rotating video showcase.[29]
- **Nova Canvas:** An image generation model with professional editing capabilities, such as "outpainting" (expanding an image) and "inpainting" (changing specific objects) via text prompts.[30]

## V. Scientific AI: The GenCast Weather Model

While commercial AI focused on media and code, Google DeepMind demonstrated the power of AI in the natural sciences with **GenCast**.[31]

## 5.1 Diffusion for Meteorology

GenCast is a machine learning model that predicts weather globally for up to 15 days. Unlike traditional Numerical Weather Prediction (NWP) models, which solve complex fluid dynamics equations on supercomputers, GenCast is trained on decades of historical weather data (ERA5) and uses a **diffusion process** to generate forecasts.[32]

## 5.2 Speed and Ensemble Forecasting

The primary advantage of GenCast is efficiency. It can generate a 15-day forecast in **one minute** on a single Cloud TPU v4. A traditional physics-based model would require hundreds of CPU nodes and significantly more time.[32]

This speed allows for massive **ensemble forecasting**. Instead of running one forecast, meteorologists can run thousands of slightly different variations to determine the *probability* of extreme events. GenCast has proven superior to the gold-standard ECMWF (European Centre for Medium-Range Weather Forecasts) in predicting the tracks of cyclones and the risks of extreme heatwaves.[31] This capability is critical for renewable energy grids, which need accurate probabilistic forecasts of wind and solar output to balance the electrical load.[31]

# VI. Theoretical Frontiers: NeurIPS 2025 Highlights

The Conference on Neural Information Processing Systems (NeurIPS), held in early December 2025, provided the theoretical counterpart to the week's product releases. Two papers, in particular, highlighted the systemic risks and architectural opportunities of the current AI wave.

## 6.1 The "Artificial Hivemind" Effect

Winning a Best Paper Award, *"Artificial Hivemind: The Open-Ended Homogeneity of Language Models"* presented a sobering analysis of AI diversity.[33]

- **The Study:** The authors introduced "Infinity-Chat," a dataset of 26,000 diverse, open-ended user queries. They tested over 70 state-of-the-art models from different providers (OpenAI, Anthropic, Meta, etc.).[34]
- **The Finding:** The study revealed a pronounced "Artificial Hivemind" effect. Not only do individual models exhibit "intra-model repetition" (saying the same thing repeatedly), but there is massive **"inter-model homogeneity."** Despite different architectures and training data, the models largely produce the same generic, "safe" responses.[34]
- **The Cause:** The paper identifies Reinforcement Learning from Human Feedback (RLHF) as the culprit. Current reward models are calibrated to a narrow "average" of human

preference, penalizing quirkiness or diverse viewpoints.
- **The Implication:** If the world relies on these models for ideation and information, we risk a "civilizational mode collapse," where human creativity is flattened by a feedback loop of homogenized machine output.[33]

### 6.2 Solving the "Attention Sink"

Another Best Paper, *"Gated Attention for Large Language Models,"* addressed a mechanical inefficiency in Transformers.[33]

- **The Phenomenon:** In standard attention, the Softmax function forces the model to assign attention *somewhere*, even if the current context is irrelevant. Models often dump this "waste" attention on the first token of the sequence (the "attention sink").
- **The Solution:** The researchers proposed **Gated Attention**, which adds a sigmoid gate to the attention heads. This allows the model to output a "zero" attention score, effectively ignoring irrelevant information completely.[33]
- **The Result:** This simple change improves training stability and allows models to scale to deeper layers (more "thinking" capacity) without the noise of forced attention. It is a key enabler for the next generation of massive reasoning models.[33]

## VII. Geopolitics, Safety, and Regulation

The technological advancements of the week have unfolded against a backdrop of intensifying geopolitical tension, particularly focused on the rise of DeepSeek.

### 7.1 The NIST Report and Security Flaws

In early December 2025, the U.S. National Institute of Standards and Technology (NIST) released a critical report on DeepSeek's models.[36] The report highlighted "security shortcomings," including unprotected databases that could grant administrative control to attackers.[37] More broadly, it warned that the widespread adoption of Chinese-origin models in Western codebases creates a supply chain risk, potentially introducing subtle vulnerabilities or "backdoors" into critical infrastructure software.[36]

### 7.2 The "Safe Chips Act" and Export Controls

Simultaneously, U.S. lawmakers introduced the **"Safe Chips Act"** and updated export control rules.[38]

- **Tier III Designation:** The new rules place China, Russia, and North Korea in "Tier III," creating a presumption of denial for licenses to export not just chips, but **advanced closed-weight dual-use AI model weights**.[38]
- **The Paradox:** The success of DeepSeek V3.2 challenges the efficacy of these controls. By innovating on software architecture (Sparse Attention), DeepSeek has managed to train frontier-class models presumably using restricted or older hardware. This suggests

that the "chip war" is insufficient to stop AI proliferation if algorithmic efficiency continues to improve at its current pace.

# Conclusion

The week of December 1–7, 2025, represents a pivotal moment in the history of artificial intelligence. The industry has moved decisively beyond the "wow factor" of generative text and into the hard engineering of **agentic reliability**, **physical grounding**, and **architectural efficiency**.

The "Global Reasoning War" has produced three distinct paradigms:

1. **DeepSeek's Efficient Intelligence:** Proving that algorithmic innovation (Sparse Attention) can challenge the brute-force compute dominance of the West.
2. **OpenAI's Persistent Agents:** Focusing on long-horizon reliability and OS-level integration (Codex-Max) to capture the enterprise workflow.
3. **Amazon's Industrial Infrastructure:** Commoditizing the "factory floor" of AI with specialized models (Nova) and agentic tools (Nova Act) that modernize legacy systems.

Simultaneously, **Google's Willow** chip offers a glimpse of a post-silicon future, while **Veo 2** and **SAM 3** are giving AI eyes and a visual cortex. However, the warnings from **NeurIPS** regarding the "Artificial Hivemind" remind us that as these systems become more powerful, they also risk becoming dangerously homogenous.

As we move toward 2026, the question is no longer "What can AI say?" but "What can AI do, how much does it cost, and whose values does it represent?" The events of this week have set the stage for answering those questions.

### Works cited


1. DeepSeek 3.2 Unleashes a New Era of Efficient and Open AI, Challenging Industry Giants, accessed December 7, 2025, https://markets.financialcontent.com/wral/article/tokenring-2025-12-6-deepseek-32-unleashes-a-new-era-of-efficient-and-open-ai-challenging-industry-giants
2. DeepSeek-V3.2 Release | DeepSeek API Docs, accessed December 7, 2025, https://api-docs.deepseek.com/news/news251201
3. DeepSeek-V3.2: Pushing the Frontier of Open Large ... - arXiv, accessed December 7, 2025, https://arxiv.org/abs/2512.02556
4. DeepSeek unveils new AI models rivalling GPT-5 and Gemini 3 Pro, accessed December 7, 2025, https://indianexpress.com/article/technology/artificial-intelligence/deepseek-unveils-new-ai-models-rivalling-gpt-5-and-gemini-3-pro-10398473/
5. Building more with GPT-5.1-Codex-Max | OpenAI, accessed December 7, 2025, https://openai.com/index/gpt-5-1-codex-max/
6. OpenAI to reportedly launch GPT-5.2 on December 9 - Times Of AI, accessed

December 7, 2025, https://www.timesofai.com/news/openai-could-launch-gpt-5-2-on-december-9/
7. OpenAI Delays Some Products Amid Competition From Google and Anthropic, accessed December 7, 2025, https://www.pymnts.com/news/artificial-intelligence/2025/openai-delays-some-products-amid-competition-from-google-anthropic
8. Building more with GPT-5.1-Codex-Max - Simon Willison's Weblog, accessed December 7, 2025, https://simonwillison.net/2025/Nov/19/gpt-51-codex-max/
9. Codex changelog - OpenAI for developers, accessed December 7, 2025, https://developers.openai.com/codex/changelog/
10. GPT 5.1 vs Claude 4.5 vs Gemini 3: 2025 AI Comparison - Passionfruit SEO, accessed December 7, 2025, https://www.getpassionfruit.com/blog/gpt-5-1-vs-claude-4-5-sonnet-vs-gemini-3-pro-vs-deepseek-v3-2-the-definitive-2025-ai-model-comparison
11. AI News & Trends December 2025: Complete Monthly Digest - Humai.blog, accessed December 7, 2025, https://www.humai.blog/ai-news-december-2025-monthly-digest/
12. Claude 4.5 Opus vs. Gemini 3 Pro vs. GPT-5-codex-max: The SOTA coding model, accessed December 7, 2025, https://composio.dev/blog/claude-4-5-opus-vs-gemini-3-pro-vs-gpt-5-codex-max-the-sota-coding-model
13. Meet Willow, our state-of-the-art quantum chip - Google Blog, accessed December 7, 2025, https://blog.google/technology/research/google-willow-quantum-chip/
14. Google Quantum AI, accessed December 7, 2025, https://quantumai.google/
15. Google claims it has made a 'major breakthrough' in quantum computing with an algorithm ... - PC Gamer, accessed December 7, 2025, https://www.pcgamer.com/hardware/google-claims-it-has-made-a-major-breakthrough-in-quantum-computing-with-an-algorithm-13-000x-faster-than-a-traditional-equivalent-although-not-everyone-is-convinced/
16. Amazon Nova - Generative Foundation Model - AWS, accessed December 7, 2025, https://aws.amazon.com/nova/
17. The Amazon Nova family of models: Technical report and model card - Amazon Science, accessed December 7, 2025, https://www.amazon.science/publications/the-amazon-nova-family-of-models-technical-report-and-model-card
18. AWS introduces Nova Forge for training bespoke 'Novella' frontier models, accessed December 7, 2025, https://siliconangle.com/2025/12/02/aws-introduces-nova-forge-training-bespoke-novella-frontier-models/
19. Introducing Amazon Nova Forge: Build your own frontier models using Nova - AWS, accessed December 7, 2025, https://aws.amazon.com/blogs/aws/introducing-amazon-nova-forge-build-your-own-frontier-models-using-nova/
20. Amazon Nova Forge: Custom Foundation Models Are No Longer Just for Tech

Giants, accessed December 7, 2025, https://workos.com/blog/amazon-nova-forge-custom-foundation-models-are-no-longer-just-for-tech-giants
21. Bring your ideas to life: Veo 2 video generation available for developers, accessed December 7, 2025, https://developers.googleblog.com/en/veo-2-video-generation-now-generally-available/
22. State-of-the-art video and image generation with Veo 2 and Imagen 3 - Google Blog, accessed December 7, 2025, https://blog.google/technology/google-labs/video-image-generation-update-december-2024/
23. Veo 2 | Generative AI on Vertex AI - Google Cloud Documentation, accessed December 7, 2025, https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/veo/2-0-generate
24. 10 Key AI Research Breakthroughs from 2025 So Far - Fueler, accessed December 7, 2025, https://fueler.io/blog/key-ai-research-breakthroughs-from-so-far
25. SAM 3 Introduces a More Capable Segmentation Architecture for Modern Vision Workflows, accessed December 7, 2025, https://www.infoq.com/news/2025/11/meta-sam3/
26. SAM3: A New Era for Open-Vocabulary Segmentation and Edge AI, accessed December 7, 2025, https://www.edge-ai-vision.com/2025/11/sam3-a-new-era-for-open%E2%80%91vocabulary-segmentation-and-edge-ai/
27. New Segment Anything Models Make it Easier to Detect Objects and Create 3D Reconstructions - About Meta, accessed December 7, 2025, https://about.fb.com/news/2025/11/new-sam-models-detect-objects-create-3d-reconstructions/
28. What is Amazon Nova? - Amazon Nova - AWS Documentation, accessed December 7, 2025, https://docs.aws.amazon.com/nova/latest/userguide/what-is-nova.html
29. The Amazon Nova Family of Models: Technical Report and Model Card - arXiv, accessed December 7, 2025, https://arxiv.org/html/2506.12103v1
30. The Amazon Nova Family of Models: Technical Report and Model Card, accessed December 7, 2025, https://assets.amazon.science/96/7d/0d3e59514abf8fdcfafcdc574300/nova-tech-report-20250317-0810.pdf
31. GenCast predicts weather and the risks of extreme conditions with state-of-the-art accuracy, accessed December 7, 2025, https://deepmind.google/blog/gencast-predicts-weather-and-the-risks-of-extreme-conditions-with-sota-accuracy/
32. GenCast: Diffusion-based ensemble forecasting for medium-range weather - arXiv, accessed December 7, 2025, https://arxiv.org/html/2312.15796v1
33. Announcing the NeurIPS 2025 Best Paper Awards, accessed December 7, 2025,

https://blog.neurips.cc/2025/11/26/announcing-the-neurips-2025-best-paper-awards/
34. Artificial Hivemind: The Open-Ended Homogeneity of Language ..., accessed December 7, 2025, https://openreview.net/forum?id=saDOrrnNTz&referrer=%5Bthe%20profile%20of%20Liwei%20Jiang%5D(%2Fprofile%3Fid%3D~Liwei_Jiang2)
35. Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free - OpenReview, accessed December 7, 2025, https://openreview.net/pdf?id=1b7whO4SfY
36. CAISI Evaluation of DeepSeek AI Models Finds Shortcomings and Risks | NIST, accessed December 7, 2025, https://www.nist.gov/news-events/news/2025/09/caisi-evaluation-deepseek-ai-models-finds-shortcomings-and-risks
37. Security Concerns due to China's DeepSeek Artificial Intelligence Model - United Service Institution of India, accessed December 7, 2025, https://usiofindia.org/pdf/SecurityConcernsduetoChinasDeepSeekArtificialIntelligenceModel.pdf
38. U.S. Export Controls and China: Advanced Semiconductors | Congress.gov, accessed December 7, 2025, https://www.congress.gov/crs-product/R48642
39. US senators unveil bill to keep Trump from allowing AI chip sales to China - Al Jazeera, accessed December 7, 2025, https://www.aljazeera.com/economy/2025/12/4/us-senators-unveil-bill-to-keep-trump-from-allowing-ai-chip-sales-to-china