OpenAI Deep Research Launches February 2 with 26.6% on Humanity’s Last Exam—O3-Powered Agent Autonomously Researches for 5-30 Minutes, Beats DeepSeek R1 by 177%

While the AI industry spent January in collective panic over DeepSeek’s efficiency claims, OpenAI quietly shipped something more significant: an agent that thinks for thirty minutes before answering and outperforms the competition by 177% on expert-level questions.

The News: OpenAI’s First Autonomous Research

Agent Goes Live

OpenAI launched Deep Research on February 2, 2025, marking the company's first production deployment of an autonomous AI agent designed for multi-step internet research. The system runs on a fine-tuned version of O3, OpenAI's latest reasoning model, optimized specifically for web browsing, data analysis, and processing text, images, and PDFs.

The core mechanics: Deep Research operates autonomously for 5-30 minutes per query, browsing hundreds of sources across the open web before synthesizing findings into fully cited reports. This isn't retrieval-augmented generation with a few documents. It's an agent that constructs and executes a research plan, pivots when it hits dead ends, and produces analyst-grade output.

Initial availability was limited to ChatGPT Pro subscribers at $200/month with 100 queries per month. By February 25, 2025, OpenAI expanded access to Plus users, and by April 24, 2025, the company restructured limits entirely: Pro tier jumped to 250 queries/month, Plus/Team/Enterprise got 25/month, and Free users received 5/month.

The geographic rollout followed quickly. Three days after the US launch, Deep Research expanded to the UK, Switzerland, and the European Economic Area. A July 17, 2025 update added visual browser capabilities, integrating the feature into ChatGPT's broader agent mode.

The Benchmark That Actually Matters

Deep Research scored 26.6% on Humanity's Last Exam, a benchmark designed to be unsolvable by current AI systems. For context: GPT-4o scored 3.3%. DeepSeek R1, the model that dominated headlines for its cost efficiency, managed 9.4%.

That 177% improvement over DeepSeek R1 deserves scrutiny. Humanity's Last Exam isn't your typical multiple-choice benchmark padded with questions that leak into training data. It's curated specifically to test expert-level reasoning across domains where existing AI systems consistently fail—the kind of questions that require synthesizing information across multiple sources, understanding implicit context, and reasoning through novel problem formulations.

A 26.6% score still means the system fails nearly three-quarters of the time on the hardest questions humans can devise. But the gap between 3.3% and 26.6% isn't incremental improvement. It's a categorical shift in capability.

> The difference between 3% and 27% on impossible questions isn't about being slightly better—it's about crossing the threshold from "occasionally lucky" to "frequently useful."

The improvement comes not from a smarter base model alone, but from the agentic architecture. Deep Research doesn't just retrieve and summarize. It executes multi-step research workflows: formulating search strategies, evaluating source credibility, cross-referencing claims, identifying gaps, and iterating until it has enough information to construct a coherent answer.

Technical Architecture: What Makes Deep Research Different

The O3 Foundation

Deep Research runs on a fine-tuned variant of O3, OpenAI's reasoning model that succeeded O1. The fine-tuning targets three specific capabilities: web navigation, multi-document synthesis, and extended reasoning chains that can span the 5-30 minute execution window.

The multi-modal processing is significant. The system handles text, images, and PDFs natively, meaning it can parse academic papers, analyze charts in earnings reports, and process technical documentation without requiring users to pre-extract content. This matters for real research workflows, where critical information often lives in formats that pure text models struggle to access.

Agentic Execution vs. Single-Shot Inference

Traditional LLM queries follow a straightforward pattern: user submits prompt, model generates response, interaction complete. Deep Research inverts this. The user submits a research question, and the system enters an autonomous execution loop that can run for up to thirty minutes.

During this window, Deep Research:

- Formulates an initial research plan based on the query
- Executes web searches and navigates to relevant sources
- Evaluates content relevance and credibility
- Identifies information gaps and formulates follow-up searches
- Synthesizes findings across multiple sources
- Generates citations and constructs the final report

This isn’t a hardcoded workflow. The O3 model makes decisions at each step about what to do next, when to pivot, and when to conclude that it has sufficient information. The “5-30 minutes” range reflects this variability—simple queries resolve faster, complex ones consume the full window.

The Compute Economics

The pricing tells a story. At $200/month for 100 queries (now 250), each Deep Research query costs roughly $0.80-$2.00 in user-facing pricing. Given that each query involves 5-30 minutes of GPU-intensive O3 inference plus real-time web browsing, OpenAI is likely operating at thin margins or subsidizing adoption.

Compare this to standard ChatGPT queries, where the same $200 buys effectively unlimited usage. Deep Research represents a fundamentally different cost structure—one where extended reasoning and real-time internet access make per-query economics unavoidable.

This explains the tiered rollout strategy. Pro users at $200/month are effectively paying for early access and higher limits. The April 2025 expansion to lower tiers with 5-25 queries/month suggests OpenAI found an acceptable cost floor for broader adoption while reserving the heavy-usage corridor for premium subscribers.

Why This Matters: The Shift from Chat to Autonomous Work

The End of the “Assistant” Paradigm

Every major AI product since ChatGPT’s launch has operated on the assistant model: user asks, AI responds, user follows up, AI responds again. The user drives

the interaction. The AI reacts.

Deep Research breaks this pattern. The user provides a research objective, then steps away. The AI drives the interaction. The user receives a deliverable.

This isn’t a subtle distinction. It’s the difference between a tool and an employee. Tools require constant operation. Employees require clear objectives and then execute independently.

> The moment you can describe what you need and walk away for thirty minutes, AI stops being a tool and starts being labor.

Competitive Implications

Deep Research positions OpenAI directly against Perplexity, which built its business on AI-powered search with citations. But Perplexity’s approach is fundamentally synchronous—query, retrieve, respond in seconds. Deep Research trades latency for depth, spending minutes to produce what Perplexity produces in moments.

These aren’t competing for the same use case. Perplexity wins when you need a quick answer with sources. Deep Research wins when you need a junior analyst’s day of work compressed into half an hour.

Google faces a different threat. Deep Research represents the first credible alternative to the research workflow that’s driven Google search traffic for decades: the iterative loop of search, scan results, click through, read, refine query, repeat. Deep Research collapses that loop into a single interaction.

The Pro Tier Validation

When OpenAI launched the $200/month Pro tier in late 2024, skeptics questioned whether any consumer would pay that premium. Deep Research provides the answer: exclusive access to capabilities that don’t exist at any price point elsewhere.

The strategy mirrors enterprise software pricing. You’re not paying for “more ChatGPT.” You’re paying for a different category of capability that happens to live inside the same interface. The 250 queries/month limit at Pro tier suggests OpenAI

expects heavy users to run 8-10 deep research sessions daily—plausible for knowledge workers whose job is research-intensive.

The Contrarian Take: What Everyone Gets Wrong

The DeepSeek Comparison Is Misleading

Headlines emphasizing the “177% improvement over DeepSeek R1” miss the fundamental asymmetry. DeepSeek R1 is an open-weights model designed for efficiency and local deployment. Deep Research is a cloud-hosted, agentic system with real-time internet access running on OpenAI’s most expensive inference infrastructure.

Comparing their benchmark scores is like comparing a calculator’s math accuracy to a financial analyst’s forecasting ability. They’re optimizing for different objectives. DeepSeek’s achievement was cost efficiency—doing more with less compute. Deep Research’s achievement is capability expansion—doing things that weren’t possible regardless of cost.

The 177% number is real, but it tells us more about what agentic architectures enable than about OpenAI’s model superiority over DeepSeek’s.

The Benchmark Overhype

Humanity’s Last Exam was designed to create a benchmark that current AI couldn’t game. A 26.6% score represents genuine progress. But the benchmark’s design—expert-level questions across specialized domains—may not correlate with the tasks most users actually need.

A system that scores 26.6% on questions designed to stump PhDs might still struggle with messy real-world research tasks: synthesizing conflicting sources, identifying when information is outdated, recognizing domain-specific credibility signals, and knowing when to admit uncertainty.

Early user reports suggest Deep Research excels at tasks with clear answers discoverable across multiple sources, but struggles when research requires subjective judgment or domain expertise that isn’t well-represented on the public web.

The Underappreciated Innovation: Time-Based Billing

The real precedent Deep Research sets isn’t benchmarks or agentic architecture. It’s the introduction of compute-time as a user-facing concept.

When you submit a Deep Research query, you’re told it will take 5-30 minutes. This is unprecedented for consumer AI products. Users have been trained to expect instant responses. Deep Research resets expectations, establishing that some AI work requires time—and by extension, that time has cost.

This opens the door for future pricing models based on compute duration rather than flat subscriptions or per-token fees. Imagine an AI product priced per minute of agent execution time. Deep Research is training users to accept that model.

Practical Implications: What to Actually Do

For Engineering Leaders

If your team spends significant time on research tasks—competitive analysis, technical documentation, market sizing, regulatory research—Deep Research deserves evaluation. The 25 queries/month at Plus tier ($20/month) provides enough runway for meaningful testing.

Start with tasks where you know what “good” looks like. Run queries on topics where your team has already done manual research. Compare the output quality, citation accuracy, and time savings against the manual baseline. This will tell you whether Deep Research meets your quality bar before you integrate it into workflows.

The architecture also hints at where autonomous agents are heading. If you’re building systems that will eventually incorporate agentic capabilities, Deep Research offers a production reference implementation to study. The 5-30 minute execution window, the query-based pricing, the trade-off between latency and depth—these are design patterns you’ll encounter when building your own autonomous systems.

For Technical Founders

Deep Research’s launch clarifies competitive dynamics. If your product competes

with quick-answer AI search (Perplexity, Google AI Overviews), you’re not directly threatened. If your product serves use cases that benefit from extended research and synthesis, you’re now competing with a $200/month incumbent.

The opportunity lies in the gaps. Deep Research is a horizontal tool—it handles any research task without domain specialization. Vertical applications that combine Deep Research-style agentic research with domain-specific knowledge, workflows, and integrations remain viable.

Consider: Deep Research can produce a market analysis report, but it can’t update your CRM with the findings, schedule follow-up meetings based on insights, or integrate with your existing competitive intelligence workflow. The agentic research capability becomes a component, not a complete solution.

For Practitioners

If you have access, focus your Deep Research queries on tasks that meet three criteria: 1) the information exists on the public web, 2) synthesizing multiple sources adds value, and 3) speed matters less than depth.

Bad fit: “What’s the current stock price of Apple?” (single source, real-time data, speed matters)

Good fit: “What are the main technical approaches to battery thermal management in electric vehicles, and how do the major manufacturers differ in their implementations?” (multiple sources, synthesis required, depth matters)

Export and save the outputs. Deep Research generates reports that may be useful later but exist only in your chat history by default. Treat these as documents worth preserving, not disposable chat messages.

The Six-Month Outlook

Capability Expansion

The July 2025 visual browser update indicates OpenAI’s roadmap: Deep Research will progressively gain capabilities that expand what “research” means. Expect integration with code execution for data analysis, ability to interact with authenticated web services (your Salesforce, your analytics dashboards), and longer

execution windows for more complex tasks.

The 30-minute ceiling is a product constraint, not a technical one. O3 can reason longer. OpenAI is calibrating user expectations and cost structures before extending runtime. By early 2026, expect “Deep Research Pro” or similar tiers offering multi-hour research sessions at premium pricing.

Competitive Response

Google will ship something comparable. They have the models, the web access, and the distribution. The question is whether they can overcome institutional inertia that prefers protecting search revenue over cannibalizing it with AI research agents.

Anthropic’s Claude already has strong reasoning capabilities and recently gained web access. An agentic research mode is technically feasible and strategically logical. If Anthropic doesn’t ship their version within six months, it’s a strategic choice, not a capability gap.

Perplexity will likely respond by extending their own session capabilities—longer reasoning, more sources, depth options that compete with Deep Research at lower latency. The “instant answer vs. deep research” dichotomy may blur as both approaches gain features from the other.

Market Implications

The success of Deep Research’s tiered pricing will determine how the industry prices agentic products. If users accept query-based limits and premium pricing for autonomous capabilities, expect every major AI product to introduce similar structures. If adoption stalls at the Pro tier, the industry will need different models.

Enterprise adoption will tell the real story. 25 queries/month per seat is insufficient for teams with serious research needs. Expect OpenAI to launch enterprise-specific Deep Research tiers with higher limits, SSO integration, and data governance features by Q4 2025.

The broader implication: we’re entering an era where AI product differentiation comes not from model capability alone, but from agentic architecture—how the system uses its capabilities autonomously over time. Fine-tuning models is becoming table stakes. Building effective agents is the new competitive frontier.

The Labor Question

Deep Research’s outputs compare favorably with junior research analyst work. Not senior analyst judgment—the system still lacks the domain expertise and contextual understanding that experienced researchers bring. But the mechanical work of finding sources, extracting key information, and synthesizing findings into readable reports is now automatable.

This has implications for hiring, team structure, and professional development in research-intensive organizations. The question isn’t whether AI replaces research analysts—it doesn’t, fully. The question is what ratio of senior to junior analysts you need when juniors have a tool that 10x their research throughput.

Organizations that figure out how to effectively pair Deep Research with human judgment will outcompete those that either ignore the tool or try to replace human researchers entirely. The winning formula is likely augmentation: human experts defining research questions and evaluating outputs, AI systems doing the mechanical work in between.

What This Tells Us About OpenAI’s Strategy

Deep Research launched three days after DeepSeek’s R1 announcement dominated AI headlines. The timing wasn’t coincidental. While the industry debated whether cheaper models were “enough,” OpenAI demonstrated what expensive models enable that cheaper ones can’t.

The message: there’s a frontier beyond efficiency. Cheaper inference is valuable. But autonomous agents that work for thirty minutes and produce deliverables—that’s a capability class that scales with compute, not against it.

OpenAI is betting that the market will stratify. Cost-conscious applications will use efficient models. High-value applications will pay premium for agents that do work, not just answer questions. Deep Research is OpenAI’s proof point that the premium tier has defensible value.

The bet seems reasonable. In a world where GPT-4 quality is commoditizing, OpenAI’s moat was shrinking. In a world where agentic capabilities require massive inference compute, proprietary infrastructure, and real-time web access, OpenAI’s lead extends again.

Deep Research doesn’t just change what ChatGPT can do—it redefines what AI products are: from tools that respond to queries to agents that complete work, from conversation partners to autonomous labor.