OpenAI Launches o3-pro at $80 Per Million Output Tokens on June 10—Beats Gemini 2.5 Pro and Claude 4 Opus on Math and Science Benchmarks

OpenAI just told developers that thinking costs money—$80 per million output tokens, to be exact. The company’s new o3-pro model doubles the price of its predecessor while simultaneously slashing costs on everything else by up to 80%.

The News: OpenAI Splits the Reasoning Market in Two

On June 10, 2025, OpenAI released o3-pro, its most capable reasoning model to date. The pricing tells the story: $20 per million input tokens and $80 per million output tokens—exactly double the base o3 model’s $10/$40 rates.

But here's what makes this launch strategically fascinating: OpenAI didn't just release a premium model. They released o4-mini at $1.10/$4.40 per million tokens on the same day, creating an 18× price spread between their cheapest and most expensive reasoning options.

The model is now available to ChatGPT Pro and Team users through the interface, and to all developers via the API. According to OpenAI's release notes, o3-pro is "designed to think longer and provide the most reliable responses," with responses typically taking "a few minutes" for challenging questions.

That latency figure matters. While most production AI applications optimize for sub-second responses, o3-pro explicitly trades speed for accuracy. OpenAI recommends it "for challenging questions where reliability matters more than speed, and waiting a few minutes is worth the tradeoff."

The Benchmark Numbers

OpenAI claims o3-pro outperforms Google's Gemini 2.5 Pro on AIME 2024 math benchmarks and Anthropic's Claude 4 Opus on GPQA Diamond, the PhD-level science evaluation. While the company hasn't disclosed specific scores for o3-pro, the base o3 model's numbers provide context for what "improvement" means here.

The base o3 already achieved 87.7% on GPQA Diamond and hit a 2727 Elo rating on Codeforces—compared to o1's 1891. On SWE-bench Verified, the standard coding benchmark, o3 scored 71.7% versus o1's 48.9%. These aren't incremental gains; they represent a generational leap in reasoning capability.

For o3-pro to measurably beat models that were already trailing base o3, we're looking at performance deltas at the extreme end of current AI capability.

Why This Matters: The Economics of Thinking

OpenAI just made an implicit argument: intelligence has a price, and that price should vary based on the difficulty of the task. This is a fundamental departure from the flat-rate API pricing that dominated the LLM market for years.

Consider what happens when you layer these three models into a production system:

- **o4-mini at $1.10/$4.40**: Classification, summarization, simple Q&A, high-volume tasks
- **o3 at $10/$40**: Standard reasoning, code generation, analysis tasks
- **o3-pro at $20/$80**: High-stakes decisions, complex scientific questions, problems where a wrong answer costs more than the compute

The spread between cheapest and most expensive is now 18× on input and roughly 18× on output. That's not a pricing quirk—it's a signal that OpenAI expects customers to route queries intelligently based on complexity.

Who Wins

Research organizations and enterprises with high-value decisions. If you're a pharmaceutical company using AI for drug interaction analysis, an $80/million token cost is trivial compared to the liability of a wrong answer. If you're a quant fund backtesting strategies, the accuracy premium pays for itself on the first trade.

Developers building tiered AI architectures. The simultaneous 80% price cut on existing models means the "standard" tier of AI applications just got dramatically cheaper. You can now afford to process 5× more volume at the same budget, reserving o3-pro for the queries that truly need it.

OpenAI's revenue mix. By creating a premium tier, OpenAI can capture more value from customers who previously maxed out on their highest-tier offering. Enterprise AI spending is increasingly bimodal—high-volume commodity tasks and low-volume critical decisions. This pricing structure maps perfectly to that reality.

Who Loses

Anthropic and Google in the enterprise reasoning market. If o3-pro genuinely beats Claude 4 Opus on GPQA Diamond and Gemini 2.5 Pro on AIME 2024, the "which model is best for hard problems" question just got a default answer. For CTOs evaluating vendors, benchmark supremacy matters—especially when the use case is mission-critical reasoning.

Startups building single-model applications. The cost complexity of routing between three reasoning tiers raises the bar for AI-native applications. You now need infrastructure to classify query difficulty, route appropriately, and handle the latency variance between sub-second o4-mini responses and multi-minute o3-pro

outputs.

The “AI is getting cheaper” narrative. Yes, OpenAI cut prices on existing models. But they also raised the ceiling on what “best” costs. The market is bifurcating into commodity AI (racing to zero) and premium AI (pricing on value).

Technical Depth: What’s Under the Hood

OpenAI hasn’t disclosed the architectural differences between o3 and o3-pro, but the observable behavior tells us something. The “designed to think longer” language suggests o3-pro allocates more compute per inference—likely through extended chain-of-thought reasoning, multiple internal verification passes, or both.

This matches the pattern we saw with o3-mini-high, which offered three “reasoning effort” settings. The high setting achieved 87.3% on AIME 2024 compared to lower scores at reduced compute levels. OpenAI’s o3 and o4-mini announcement from April 2025 established this compute-scaling approach as their core strategy for reasoning models.

The Benchmark Context

Let’s put these numbers in perspective:

GPQA Diamond is designed to stump experts. Questions require PhD-level domain knowledge in physics, chemistry, and biology. The base o3’s 87.7% score was already remarkable—human experts achieve roughly 65-70% without access to references. For o3-pro to “beat Claude 4 Opus” on this benchmark, we’re talking about performance in the 88-91% range, which represents near-saturation of the test.

AIME 2024 (American Invitational Mathematics Examination) tests competition-level mathematical reasoning. These aren’t textbook problems; they require creative problem-solving and multi-step proofs. The o3-mini-high model achieved 87.3% on AIME 2024. For o3-pro to beat Gemini 2.5 Pro “on AIME 2024,” the score must exceed whatever Google’s model achieves—likely putting o3-pro in the 90%+ range.

Codeforces Elo of 2727 for base o3 places it at the Grandmaster level—top 0.5% of competitive programmers. This isn’t “writes working code”; it’s “solves

algorithmic puzzles that most senior engineers can't."

SWE-bench Verified at 71.7% for base o3 means the model successfully resolves 71.7% of real-world GitHub issues from major open-source projects. The jump from o1's 48.9% to o3's 71.7% represents a 47% relative improvement. That's the difference between a model that's "helpful for coding" and one that can meaningfully contribute to engineering workflows.

Tool Integration Capabilities

o3-pro launches with access to ChatGPT's core tool suite: web search, Python code analysis, and vision reasoning. Notably absent are image generation and Canvas support—suggesting OpenAI is positioning this as a pure reasoning model, not a multimodal creative tool.

The Python integration matters for technical users. o3-pro can write and execute code as part of its reasoning process, enabling iterative problem-solving where it validates its own solutions programmatically. Combined with multi-minute reasoning times, this suggests workflows where the model genuinely "works on" hard problems rather than producing instant outputs.

The Contrarian Take: What the Coverage Gets Wrong

Overhyped: The Benchmark Wars

Every o3-pro headline focuses on beating Gemini and Claude on specific benchmarks. Here's what that coverage misses: benchmarks measure capability at the frontier, but production value depends on reliability at scale.

A model that scores 90% on GPQA Diamond but produces inconsistent results across multiple runs is less valuable than one scoring 85% with high reproducibility. OpenAI's emphasis on "reliability" in their positioning suggests they understand this—but we don't yet have independent data on o3-pro's variance characteristics.

The benchmark comparisons also obscure use-case specificity. Claude 4 Opus might trail on GPQA Diamond but outperform on tasks requiring nuanced communication or ethical reasoning. Gemini 2.5 Pro might lose on AIME 2024 but offer superior

multimodal integration for computer vision applications. The “best model” framing is misleading because best-for-what matters more than best-on-average.

Underhyped: The Routing Problem

The real story isn’t o3-pro’s capabilities—it’s the infrastructure challenge of deciding when to use it.

With an 18× cost spread between o4-mini and o3-pro, query routing becomes a first-class engineering problem. Get it wrong in one direction, and you’re burning money sending simple questions to the expensive model. Get it wrong in the other direction, and you’re delivering garbage answers by sending complex queries to the cheap model.

This creates demand for a new category of tooling: meta-models that classify query complexity and route appropriately. Some teams will build this internally; others will wait for third-party solutions. Either way, the “one model for everything” approach is dead.

The Latency Trade-off Most Teams Aren’t Ready For

“A few minutes” for challenging questions sounds acceptable in a press release. In production, it creates architectural headaches that most coverage ignores.

User-facing applications can’t block for minutes. You need async processing, progress indicators, webhook callbacks, or message queues. Your error handling must account for timeouts that are 100× longer than typical API calls. Your rate limiting and concurrent request strategies need complete revision.

For internal tools—research assistants, analysis pipelines, code review systems—the latency is manageable. For customer-facing products, it requires fundamental UX rethinking. The applications that thrive with o3-pro will look different from current AI interfaces.

Practical Implications: What You Should Actually Do

Immediate Actions for Engineering Leaders

1. Audit your current OpenAI spend by query type.

Before chasing o3-pro, understand what you're currently paying for. The 80% price cuts on existing models mean your baseline costs are about to drop. Segment your API calls by complexity: what percentage are simple classification tasks that could run on o4-mini? What percentage are complex reasoning queries where o3-pro's accuracy premium would justify the cost?

Most organizations will find 70-80% of their queries belong in the cheapest tier. The remaining 20-30% split between o3 and o3-pro based on stakes, not just difficulty.

2. Build or buy a query router.

Here's a simple heuristic to start:

- Queries under 100 tokens with clear intent → o4-mini
- Queries requiring multi-step reasoning, code generation, or analysis → o3
- Queries involving scientific accuracy, legal/medical domains, or where error cost exceeds $100 → o3-pro

This is crude. You'll refine it based on your domain. The point is to start routing now, because flat-rate "send everything to one model" approaches leave money on the table.

3. Redesign latency-sensitive workflows.

If you have production systems that could benefit from o3-pro's accuracy but can't tolerate minute-long latencies, architect around it:

- Batch complex queries during off-peak hours
- Pre-compute answers for predictable high-stakes queries
- Implement "draft/refine" patterns where o4-mini provides instant responses while o3-pro generates higher-quality versions asynchronously

Code Patterns Worth Testing

Try this architecture for coding assistance workflows:

> User query → o4-mini (instant response, catches 70% of requests) → If confidence < threshold, escalate to o3 → If o3 flags high complexity, queue for o3-pro review

For research and analysis pipelines:

> Initial analysis request → o3 (standard reasoning, 1-2 minute response) → o3-pro validation pass (3-5 minutes) → Human review of discrepancies

The key insight: o3-pro doesn't have to answer every question. It can serve as a verification layer for o3's outputs on critical queries.

Vendors and Tools to Watch

LLM routing and orchestration platforms. Martian, Not Diamond, and similar startups offer automated model selection. With the cost spread now 18×, their value proposition sharpened considerably. Expect rapid feature development targeting the o3/o3-pro decision boundary.

Observability tools with cost attribution. Helicone, LangSmith, and Weights & Biases now need granular cost tracking per model tier. The days of "average cost per query" metrics are over; you need distribution views showing which queries hit which tier.

Async API frameworks. FastAPI's background tasks, Celery, and similar tools become essential for o3-pro integration. If your stack is synchronous request-response only, you'll need to add async capabilities.

Forward Look: Where This Leads in 6-12 Months

The Tiered Intelligence Market Crystallizes

OpenAI just established the template. Expect Anthropic and Google to follow with explicit model tiers at distinct price points. The ambiguity of "which model should I use" gives way to clear market segmentation: commodity reasoning, standard reasoning, premium reasoning.

Within 12 months, most enterprise AI deployments will run across multiple tiers from multiple providers. Vendor lock-in decreases as interoperability increases, but operational complexity rises. The “AI platform” market shifts from providing models to providing orchestration.

The Routing Layer Becomes Table Stakes

Query classification and model routing will transition from competitive advantage to basic infrastructure. Just as CDNs became ubiquitous for web applications, intelligent routing will become ubiquitous for AI applications.

This creates acquisition targets. The standalone routing companies either get acquired by cloud providers (AWS, Azure, GCP) or by the model providers themselves. Google already has routing infrastructure through Vertex AI; OpenAI could build or buy similar capabilities.

The “Accuracy Premium” Expands

If o3-pro at $80/million tokens finds market acceptance, expect $200-500/million token offerings within 18 months. The principle that “more thinking costs more money” scales upward. For applications where wrong answers cost millions—clinical diagnostics, legal discovery, financial modeling—$500/million tokens is still cheap compared to the alternative.

This creates an interesting dynamic: the ceiling on AI pricing rises while the floor approaches zero. The market stratifies into high-volume commodity use cases (price sensitivity dominates) and high-stakes specialized use cases (accuracy sensitivity dominates).

Multi-Minute Latencies Become Standard for Premium Models

o3-pro’s “few minutes” response time normalizes the expectation that premium AI takes time. Future models will likely push this further—10-minute, 30-minute, or even hour-long “deep reasoning” sessions for sufficiently complex queries.

This shifts the AI UX paradigm. Instead of instant answers, users will “submit” hard problems and receive notifications when results are ready. Email-like interfaces for AI interaction emerge alongside chat-like interfaces. The metaphor shifts from “assistant” to “analyst.”

OpenAI’s Competitive Moat Becomes Pricing Structure

By controlling three tiers at three price points, OpenAI captures value across the entire demand curve. Competitors offering single models at single prices look unsophisticated by comparison. The ability to route customers to the right tier automatically becomes a moat—customers don’t need to choose between providers for different use cases if one provider covers the entire spectrum.

Industry analysts have noted that this pricing strategy resembles cloud computing’s evolution from flat-rate compute to sophisticated tiered offerings. OpenAI is following the AWS playbook: own the full stack, price on value, make switching costly through convenience.

The Real Question Nobody’s Asking

The coverage of o3-pro focuses on benchmarks and prices. Here’s the question that matters more: what happens to human expertise in domains where AI achieves superhuman performance?

GPQA Diamond tests PhD-level science knowledge. If o3-pro scores 90%+ while human experts achieve 65-70%, what’s the role of the human? In mathematics, coding, and scientific reasoning—the domains where o3-pro excels—AI capability now exceeds average expert capability for narrow question-answering tasks.

This isn’t about job replacement. It’s about cognitive augmentation becoming mandatory for professional competence. The scientist who doesn’t use o3-pro operates at a disadvantage to one who does. The programmer who doesn’t route hard algorithmic questions to AI is less productive than one who does.

The $80/million token price isn’t just a cost to manage—it’s an investment in capability that humans can’t match independently. Organizations that treat it as a budget line item will lose to those that treat it as a competitive weapon.

o3-pro’s real significance isn’t that AI got more expensive—it’s that human-only reasoning just became a liability.