Welcome to the End of Tasks

Published May 21, 2025

Christopher Gannatti, CFA
Global Head of Research

Key Takeaways

- In 2025, AI systems evolved from passive models to active agents capable of executing complex, goal-driven tasks, marking a paradigm shift that reshapes productivity and enterprise strategy.
- Benchmarks like HELMET and RE-Bench reveal that agentic AI excels in long-horizon reasoning and iterative workflows, outperforming both legacy models and human analysts over time.
- Investors should focus not on the biggest or smartest AI models, but on systems that learn, adapt and create to compound value, offering a new source of leverage and potential alpha across industries.

We've crossed an invisible boundary. The core story is no longer about intelligence that mimics language, but about systems that act. That distinction matters—immensely. Because the leap from models to agents is not linear progress. It's a paradigm shift.

From Passive Models to Active Agents

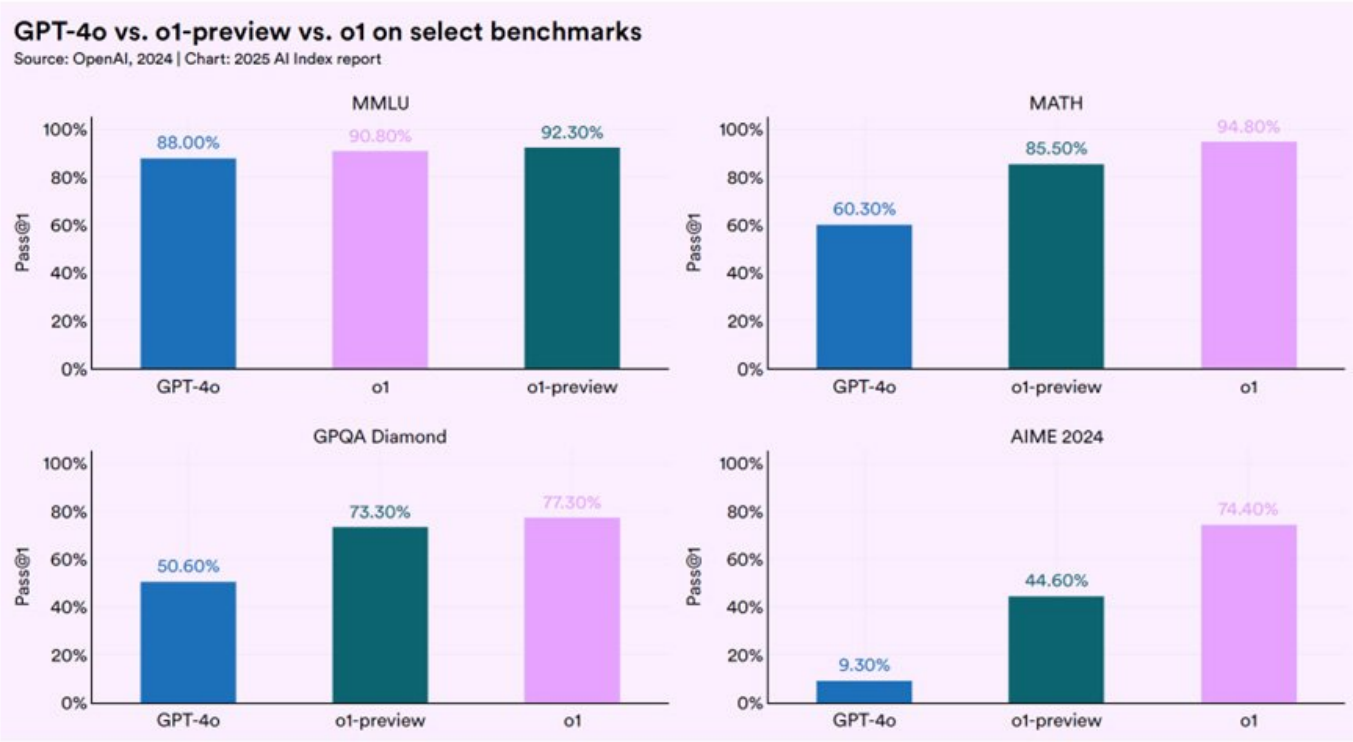
In 2025, AI crossed a subtle but monumental threshold: it stopped just answering questions and started executing goals. This transition—from models to agents—isn't about better search results. It's about capability. The latest models, like OpenAI's o1, show stark gains on benchmarks that test not just knowledge but reasoning and follow-through. Take AIME, the American Invitational Mathematics Examination—an infamously difficult math contest that pushes problem-solving across multiple steps. GPT-4o scored just 9.3%. o1? Its score jumped to 74.4%. That's not a tweak. It's a redefinition of what AI can handle.

Other benchmarks tell a similar story. On MATH, a test derived from high school and early college competition problems, o1-preview achieved 94.8% accuracy. On GPQA Diamond, a collection of tough, graduate-level multiple-choice questions that require nuanced reasoning, o1 pushed past 77%, versus GPT-4o's 50.6%. These are proxies for tasks that require patience, logic and an understanding of context. Not just "what's the capital of France?" but "solve this, adjust that, check your work." And in those domains, agents are starting to outperform not only older models, but skilled humans on short timeframes.

The economic and workflow implications are profound. Agentic AI isn't just smarter, it's beginning to work the way people do: planning, executing, refining. It's slower and more expensive, yes—o1 runs at 40 times the latency and costs multiples more per token—but that's what you'd expect from a high-level specialist. We're entering a world where the fast models handle everyday questions, and the slow agents tackle the

hard stuff. These systems won't just change software—they'll change who does the work, how it's done and what we even mean by productivity.

Figure 1: Measuring How Models Are Gaining Different More Agentic Capabilities



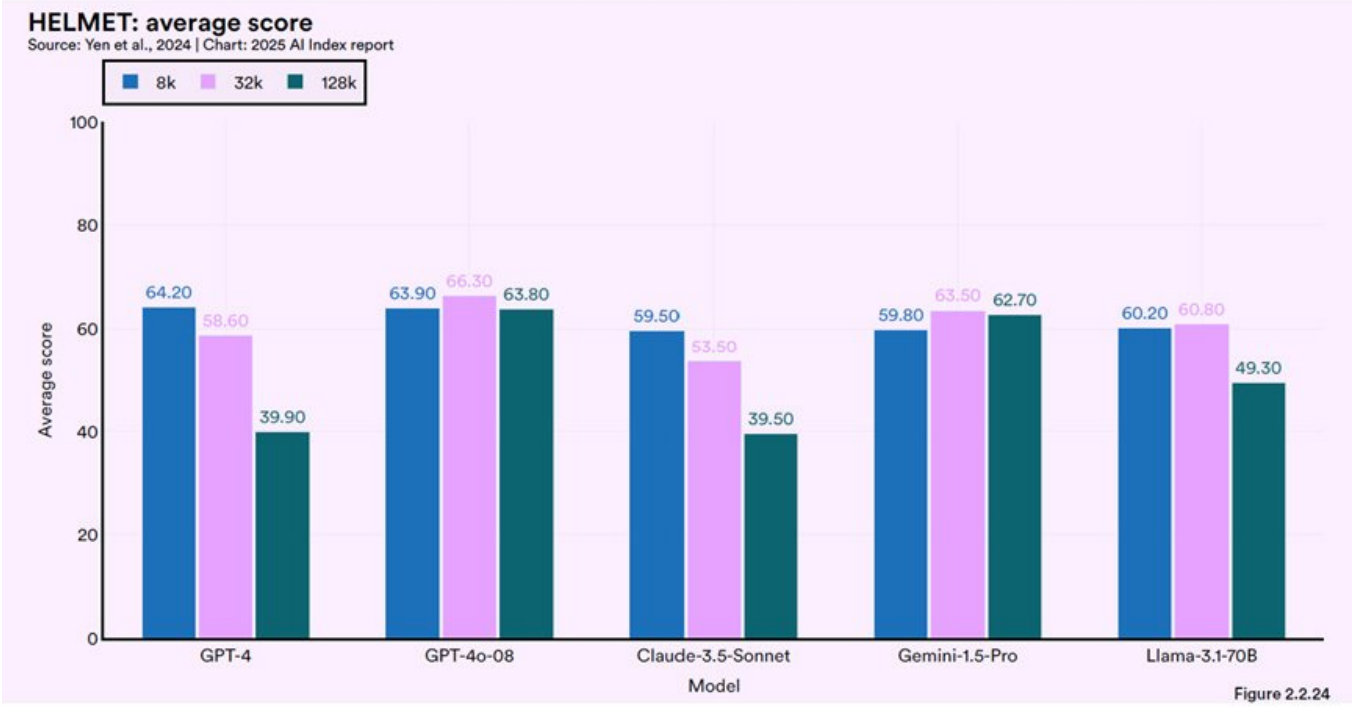
In investing, we're trained to pay attention not just to *what* performs, but *how* and *under what conditions* it performs. That same mindset is now essential in evaluating AI. While most benchmarks measure how well a model answers trivia or solves math problems in isolation, the new HELMET benchmark (results shown in figure 2) shifts the focus to something far more consequential: can a model operate across sprawling, high-complexity contexts, the way an analyst does when synthesizing insights from 200 pages of earnings transcripts? This is the benchmark for models transitioning from tools to agents—capable not just of task completion, but of orchestration, reasoning and memory over time.

The results are revealing. Older models like GPT-4 and Claude 3.5 Sonnet see their performance drop sharply as context lengths stretch—like interns who ace short memos but stumble when handed the full quarterly report. But models like GPT-4o and Gemini 1.5 Pro show durability: they hold their ground even at 128,000 tokens (think reading and recalling *War and Peace* in one go). These models didn't just get smarter—they became *better long-term thinkers*. That's the hidden differentiator in an enterprise context, where agents won't just summarize a page—they'll navigate whole data rooms, legal filings and CRM logs to inform decisions.

For investors, HELMET isn't just another benchmark—it's a glimpse into what makes a model truly enterprise-grade. In the same way that we evaluate businesses not on one good quarter but on resilience

through cycles, HELMET exposes which models bend and which break under pressure. And as agentic AI starts to augment—or even automate—complex workflows across finance, legal and healthcare, understanding these capabilities will matter as much as knowing the difference between growth and value. The models that remember, reason and react over time are building something bigger than intelligence—they're building *leverage*.

Figure 2: HELMET—How Robust Are Models across Lengthening Context Windows?

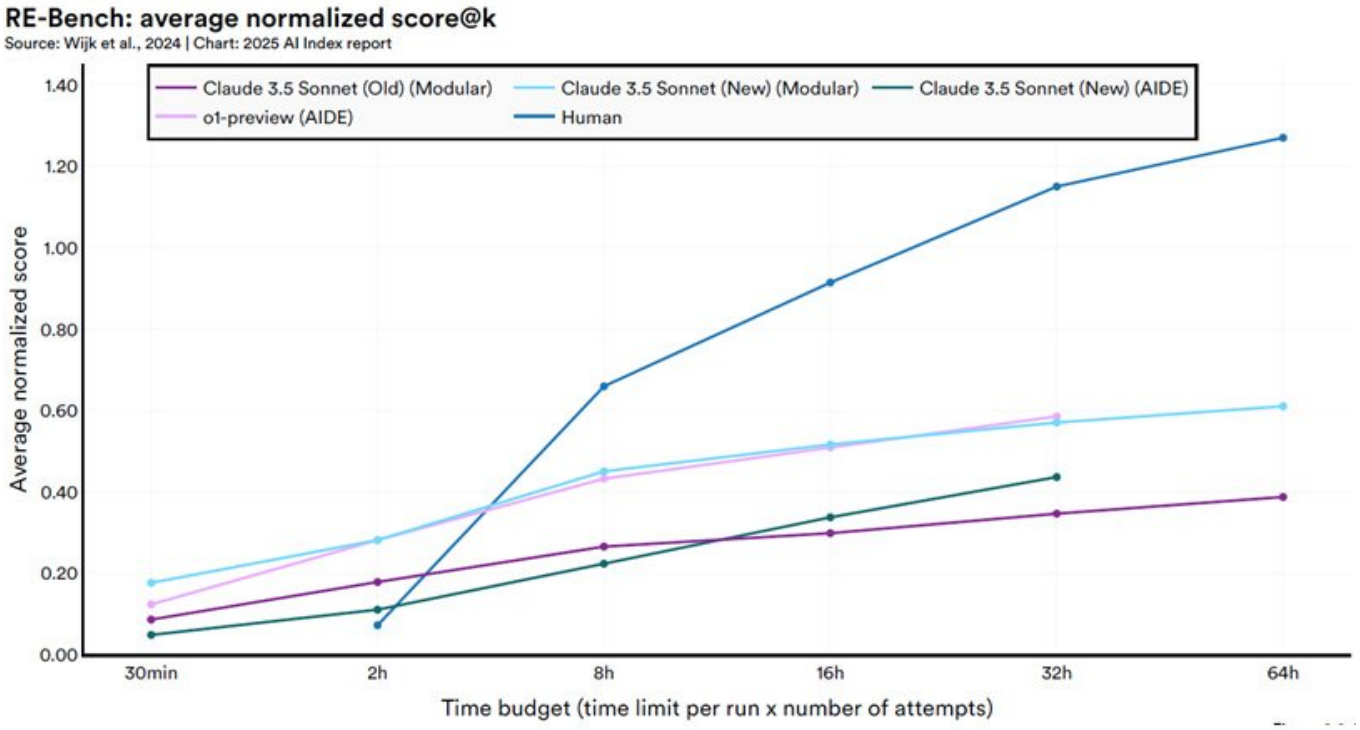


There's a chart in the *2025 AI Index Report* that quietly rewrites our expectations for artificial intelligence. It tracks how well different AI systems perform relative to a reasoning benchmark, RE-Bench (shown in figure 3), as they're given more time and multiple attempts. At first, humans outperform the machines. But something remarkable happens as the time budget increases: agentic AI systems—those designed to iterate, self-correct and plan—begin to pull ahead. By the 32-hour mark, the best agents are outperforming humans, not because they're inherently smarter, but because they're designed to compound intelligence over time.

This is the difference between a clever tool and a self-improving system. Static models, no matter how large, hit plateaus. But wrap the same model in an agentic framework like AIDE and performance scales with patience. The story here isn't just technological; it's deeply financial. The best AI agents look less like flash-in-the-pan geniuses and more like investors with a repeatable process—learning from every mistake, reinvesting every insight. They aren't just answering questions; they're conducting experiments with feedback loops that mimic the logic of compounding returns.

For financial professionals, this is more than a benchmark result—it's a signal. We're approaching a world where autonomous agents can outperform human teams in certain domains, given enough time and computing power. The takeaway is simple: the edge won't go to the biggest model, or even the smartest—it will go to the one that learns best, longest. Just like in markets, time and iteration are the ultimate advantages. Investors who grasp this shift early will be better positioned to understand which AI platforms, infrastructure providers and application layers will drive the next wave of productivity and enterprise value.

Figure 3: RE-Bench Results



The Age of Agency Has Arrived

What RE-Bench, HELMET and real-world deployments all suggest is that we've crossed into a fundamentally different computing paradigm—one where intelligence isn't measured by output per prompt, but by compounding returns on cognition. The hallmark of this shift isn't higher accuracy in a vacuum, but the emergence of feedback-driven systems that can plan, adapt and persist over time. Like a seasoned analyst or portfolio manager, agentic AI doesn't just seek the right answer—it develops a process to keep finding better ones. And that process, once embedded into workflows, creates leverage that traditional software—and static models—simply can't match.

From Leverage to Alpha

For investors, the implication is clear: the next frontier of productivity won't be about who has the largest model, but who can most effectively integrate agentic systems into their value chains. Expect the winners to be those who shift their mental models—from tools that assist to systems that compound. That means

looking beyond parameter counts or one-off benchmarks, and instead asking: *Can this system learn over time? Can it operate autonomously across messy, long-horizon domains? Can it create feedback loops that scale?* Those are the questions that will separate hype from durable alpha. Because the real opportunity here isn't automation—it's transformation.

Important Information Related to this Article

Source for all data in this blog post, including figures: Maslej et al., *The AI Index 2025 Annual Report*, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, 4/25.