AI Crash Memo Stress Test

Idea In Short

A macro research shop wrote a fictional memo from 2028, imagining an AI-driven crash that wipes 38% off the S&P 500 and pushes unemployment past 10%. It went viral, investor Michael Burry amplified it to roughly 16 million views and it briefly erased billions in market value before drawing a public response from the White House. The forecast itself does not hold up: Goldman Sachs projects unemployment drifting to about 4.5% by the end of 2026, not a collapse. But the memo's underlying mechanism is real. AI agents are collapsing the switching costs that many software and services businesses depend on for pricing power. This piece treats the scenario as what it actually is, a stress test for identifying which business models an AI agent can route around and which ones it has to run through

-

A Scenario Mistaken For a Forecast

In February 2026, a macro research shop called Citrini Research published a piece titled "The 2028 Global Intelligence Crisis." The authors wrote it as financial fiction, a memo dated June 2028 looking back at a hypothetical AI-driven economic collapse. It described a stock market down 38%, unemployment above 10% and a self-reinforcing loop where AI displaces workers, spending falls and companies cut deeper to protect margin.

Nothing about the piece claimed to be a forecast. It was a thought experiment built to explore one version of how agentic AI could reshape the economy, not a prediction anyone should trade against. The internet treated it as a weather report anyway.

What Actually Happened

The scenario moved markets regardless of its intent. After investor Michael Burry shared it, the memo drew roughly 16 million views, a public response from the White House and a real selloff across software, payments and delivery stocks. IBM fell about 13% in a single trading

session, its steepest one-day drop in more than two decades.1

Citadel Securities, the market maker founded by Ken Griffin, published a formal rebuttal within days, calling the scenario a misread of macro fundamentals. The reaction said less about AI economics than about how fast a well-written scenario moves money when few people check the label first.

The Ghost GDP Thesis

Citrini Research was founded in 2023 by James van Geelen and this memo was co-authored with Alap Shah. Its central idea, dubbed "Ghost GDP", describes an economy where corporate output keeps rising while the income that used to circulate through payrolls disappears. Consumer spending drives roughly 70% of US gross domestic product [GDP], so a world where AI raises profit without raising paychecks produces a deflationary drag no interest-rate cut can fix, according to the memo's own logic.

The idea is clean, internally consistent and built to spread. That is also its weakness: a single-variable story travels faster than a messy, accurate one and this one was engineered to travel.

The Mechanism Worth Taking Seriously

Strip away the macro narrative and one observation survives scrutiny: AI agents are collapsing switching costs. When software can compare vendors, complete a task or replicate a workflow on its own, the friction that used to lock a customer into a subscription starts to disappear. That friction was never incidental. For a large share of software and services businesses, it was the entire moat.

Once switching costs fall, whoever sat in the middle of a transaction because they were embedded, not because they were the best option, loses pricing power. That repricing has nothing to do with whether unemployment hits 10% by 2028. It is happening to business models right now, one renewal cycle at a time.

Sorting Business Models Into Two Piles

Read as a business diagnostic instead of a doomsday clock, the memo sorts companies into

two groups. The exposed group includes software-as-a-service [SaaS] products with thin differentiation, whose renewals depended on switching being annoying rather than the product being irreplaceable. The same logic hits workflow-embedded intermediaries whose entire pitch was convenience.

The resilient group includes businesses an agent has to work through rather than around: systems of record, deeply embedded workflows and products built on proprietary data a competitor cannot cheaply copy. Owning the billing relationship is not enough anymore; owning a position an agent cannot substitute is what matters. That shift is already visible in the numbers. AI became the single most-cited reason employers gave for job cuts in March 2026, according to the Challenger, Gray & Christmas layoff report, a sign the repricing is further along than the memo's fictional 2028 timeline suggests.2

The Skeptics' Case

The case against taking the memo literally is strong. Citadel Securities argues its core error is conflating recursive technology with recursive adoption: a model can improve every quarter, but organizations do not restructure every quarter. Org inertia, retraining lag, regulation and ordinary human friction all sit between a new capability and an actual layoff.

Frank Flight, the strategist behind Citadel's response, frames technology adoption as a curve rather than a straight line, with integration costs and diminishing returns eventually flattening it out.3 Displacing white-collar work at the scale the memo describes would also require far more computing capacity than current AI systems use today, which sets a natural ceiling on how fast the loop can spin.

What the Data Actually Shows

The real 2026 numbers back the skeptics far more than the memo. Goldman Sachs economist Pierfrancesco Mei expects US unemployment to drift from 4.3% to about 4.5% by the end of the year, with faster AI adoption adding at most another 0.3 percentage points.4

Job losses in AI-affected industries have been visible but moderate

The Budget Lab at Yale found no macro-level sign of mass displacement either. The share of workers in high, medium and low AI-exposed occupations has stayed remarkably steady since generative AI tools became mainstream.5 Even Citrini hedges its own scenario with a line worth keeping: repricing is not the same as collapse.

The Real Takeaway for Operators

Klarna offers the clearest preview of how this plays out inside a real company. The buy now, pay later firm replaced hundreds of customer service roles with an AI assistant, then found the AI could handle volume but not complexity and started rehiring humans for cases that needed judgment.6

> We focused too much on efficiency and cost. The result was lower quality and that's not sustainable

The lesson generalizes past customer support. Net revenue retention, not headline unemployment, is the number that will show switching-cost erosion before it hits a stock price, because expansion and churn move long before headlines do. Executives who track that signal will see repricing coming; the ones staring only at annual recurring revenue [ARR] will get surprised by it.

Key Partners

CoreWeave's business depends on a small number of deep partnerships rather than a broad supplier base. Nvidia sits at the center, supplying the graphics processing units [GPUs] that define CoreWeave's entire value proposition and shaping its roadmap around each new chip generation. Microsoft, OpenAI, Meta and Anthropic function as both anchor customers and demand partners whose multi-year commitments justify new data center buildouts. Debt and equity investors, including lenders who accept GPU fleets as collateral, complete the partner base by funding the capital spending the model requires.

Key Activities

CoreWeave's core activity is building and operating GPU-dense data centers fast enough to keep pace with AI demand. That means procuring Nvidia hardware at scale, securing power capacity years in advance and standing up new facilities on a continuous basis. The company also develops MLOps software that lets customers orchestrate large training and inference workloads across its infrastructure. Raising the equity and debt needed to fund tens of billions in annual capital spending is itself a constant, active workstream.

Key Resources

The single biggest resource is the GPU fleet itself, more than 250,000 Nvidia chips spread across roughly 33 data centers. Long-term power purchase agreements covering 4.6 gigawatts of capacity are just as critical, since power availability, not chip supply, increasingly limits how fast new capacity comes online. Proprietary orchestration and MLOps software separates the raw hardware from a commodity rental. Engineering talent capable of running GPU clusters at extreme density rounds out the resource base.

Value Propositions

CoreWeave sells access to scarce, high-end GPUs faster than customers could typically get from a general-purpose hyperscaler. Its infrastructure is purpose-built for AI training and inference rather than adapted from general cloud computing, which shows up in utilization and performance. Customers get flexible capacity at a scale few providers can match, backed by integrated MLOps tooling for managing large model workloads. For AI labs racing to train frontier models, that combination of speed, scale and specialization is the entire pitch.

Customer Relationships

CoreWeave's largest relationships run on multi-year contracts rather than pay-as-you-go usage, with Microsoft alone accounting for roughly two-thirds of fiscal 2025 revenue. Anchor customers such as OpenAI, Meta and Anthropic receive dedicated technical support and, in some cases, co-investment style arrangements tied to specific capacity buildouts. Smaller developers and startups access capacity through more self-serve arrangements. The relationship model overall favors depth with a handful of accounts over breadth across many.

Channels

Direct enterprise sales handle the largest contracts, since deals of this size involve custom capacity planning and negotiated terms rather than a self-serve checkout. Nvidia's own partner ecosystem generates referrals, given how closely CoreWeave's growth is tied to Nvidia's hardware roadmap. A developer platform with application programming interface [API] access serves smaller customers who want GPU capacity without a dedicated sales process.

Customer Segments

Frontier AI model developers such as OpenAI and Anthropic represent the highest-value segment, since their training runs consume capacity at a scale few other customers match. Hyperscalers like Microsoft use CoreWeave as overflow capacity when their own data centers cannot keep up with demand. Enterprises building AI products internally, Meta among them, form a third segment, alongside quantitative finance firms and AI startups that need burst capacity without owning infrastructure.

Cost Structure

Capital spending dominates CoreWeave's cost base, with fiscal 2026 guidance of $31 billion to $35 billion for GPUs and data center buildouts. Servicing roughly $21 billion in long-term debt, much of it secured against GPU collateral, adds a second major cost layer, alongside power costs that scale with every new facility. Rapid depreciation on GPUs, which lose value as newer chip generations arrive, compounds the pressure on margins.

Revenue Streams

Revenue comes primarily from usage-based and contracted GPU cloud computing fees charged to AI labs, hyperscalers and enterprises. Long-term reserved capacity agreements make up the bulk of the business, reflected in a contracted backlog exceeding $66 billion as of mid-2026. Individual multi-year deals, including a $4 billion commitment from OpenAI and a $21 billion expansion with Meta, anchor future revenue well beyond the current fiscal year.

- 1The 2028 Global Intelligence Crisis
- 2AI Blamed Heavily For March Layoffs, Report Says

- 3Citadel Demolishes Viral Doomsday AI Essay
- 4How Will AI Affect The US Labor Market
- 5New Data Show No AI Jobs Apocalypse For Now
- 6The Great AI Layoff Is Turning Into The Great AI Rehire

Summary

The 2028 crash memo fails as a prediction. The 2026 labor data looks nothing like a spiral and Citrini never claimed otherwise. As a business-model diagnostic, though, it earns its virality. It names the single most overrated moat of the software era, customer inertia. Any company whose retention depends on switching being annoying, rather than the product being irreplaceable, is carrying revenue on borrowed time as agentic AI shrinks that friction. The fix is not to brace for a specific year. It is to ask, today, what customers would still pay for once an AI agent can compare, switch and negotiate on their behalf. Businesses built on real differentiation, proprietary data or workflows an agent must pass through rather than around will absorb the shock. Businesses built on convenience and lock-in are the ones that should be worried