The AI You're Using Is Not the AI That Exists

What the labs aren't saying out loud — and why it changes everything

CryptoSoWhat.com | April 2026

Dr. Gregory S. Carmichael
CEO, Quantum Reserve Capital · Founder, ANMM
Lt. Col. USAF (Ret.) · G20 India (2023) · DPA Title III Advisor (2024)

CryptoSoWhat.com | Dr. Gregory S. Carmichael | April 2026

Let's call it what it is.

The AI capability you have access to right now — the models you can actually use, the benchmarks you read about, the products you can subscribe to — is not the frontier. It is the frontier from approximately two years ago. You are being shown yesterday's horizon and told it is today's.

This is not speculation. It is a structural feature of how the industry works, and understanding it changes everything: your investment thesis, your competitive positioning, your assumptions about who actually controls the most consequential technology in human history.

The Gap Is Not a Bug. It's the Business Model.

Every frontier AI model passes through the same pipeline before you see it:

Pretraining ends. Post-training (alignment, safety red-teaming, adversarial evaluation) takes three to nine months. Staged deployment takes another one to three months. By the time you interact with a model, you are talking to something that finished training roughly six to twelve months ago.

Meanwhile, the next training run started at Day 30. The researchers inside the building have been working with the successor model for most of the year before the current one ships.

A November 2025 infrastructure analysis stated it plainly: *"All four U.S. leaders have much better internal checkpoints than public releases — very hard for anyone to catch up."*

That gap is structural. It is not closing. And it is not the most important gap.

What the Public Benchmarks Are Actually Telling You

The public frontier in April 2026 is genuinely extraordinary — and that should concern you, not reassure you.

Gemini 3.1 Pro just scored 77.1% on ARC-AGI-2. That benchmark was specifically designed by its creator to be memorization-resistant — it cannot be gamed by training on similar problems. Three months earlier, the same lab's model scored 31.1%. That is a 2.5× improvement in genuine fluid reasoning in a single quarter.

Claude Opus 4.6 resolves 81% of real-world software engineering bugs. GPT-5.4 can now operate a computer autonomously — navigating, filling forms, executing workflows without human hands. Gemini broke the 1500 Elo ceiling on the human preference leaderboard — the ceiling that was supposed to be the ceiling.

These are the publicly released scores. From the models that are 6 to 18 months behind the internal frontier.

If the public models are doing this, what are the internal checkpoints doing?

Nobody is telling you.

What We Know Is Actually Running Behind the Wall

Here is where the polite version of this conversation ends.

Anthropic publicly confirmed that Claude has been deployed throughout classified US government networks, at national nuclear laboratories, and performs intelligence analysis directly for the Department of War. This is in the unclassified record. Anthropic said it.

Media confirmed that Claude was used to assist intelligence analysis leading up to the January 2026 capture of Maduro in Venezuela, and in preparations related to Iranian leadership operations. After that contract was restructured, the Department of War signed immediately with OpenAI to continue AI integration in classified environments. The handoff took hours.

In December 2025, the Department of War launched **GenAI.mil** — deploying frontier AI to three million warfighters, employees, and contractors at Impact Level 5, the classification tier that covers the most sensitive unclassified operations.

The FY2026 National Defense Authorization Act mandated a new committee, with a hard deadline of April 1, 2026, specifically tasked with analyzing technologies capable of enabling Artificial General Intelligence and their *operational military effects*.

The current US budget allocates $115 million to accelerate **nuclear missions** using AI.

Read that one more time. One hundred and fifteen million dollars to accelerate nuclear missions using artificial intelligence. This is a budget line item. Today.

The public conversation about "whether AI should be trusted with important decisions" is not a debate that is being held in the rooms where the actual decisions are being made. Those decisions were made years ago. What you are watching in public discourse is the civilian permission structure trying to catch up to the operational reality.

The Chip War Is Already Over for the Current Round

The US export controls on advanced NVIDIA chips to China are not trade policy. They are a first-strike in a long-form capability war, and they are working.

NVIDIA's Blackwell architecture delivers 3–5× better training efficiency than the previous Hopper generation. China's domestic alternative — the Huawei Ascend — delivers approximately 30–50% of equivalent Blackwell performance per unit. A Chinese organization training a frontier model today faces a 3–5× effective compute disadvantage against a US organization with Blackwell access, even assuming identical algorithmic talent.

DeepSeek V3 shocked markets in January 2025 because it demonstrated near-frontier performance at dramatically lower cost. Equity markets moved a trillion dollars on that release. But the correct reading was never "China caught up." It was: algorithmic efficiency can partially compensate for a hardware deficit, but hardware deficits compound faster than algorithmic efficiency closes them.

As of April 2026, DeepSeek has not released a new frontier model in approximately one year. The ceiling imposed by the Huawei hardware gap has become visible.

The Blackwell-trained models — running at US labs for the past six to nine months — have not yet been publicly released. They are coming in Q2 2026. The benchmark scores you just read will look like the GPT-3.5 era by Q4.

The Real Moat Nobody Is Pricing In

Here is the thing that gets lost in the benchmark coverage and the model release cycles.

Reasoning models generate something prior AI did not: **reasoning traces**. Multi-step chains of thought, tool use sequences, error correction loops. When deployed models produce these traces at scale — millions of interactions daily — those traces become training data for the next generation.

This creates a flywheel that cannot be purchased or replicated from the outside:

> *More users → More reasoning traces → Better post-training data → Better models → More users*

The labs with the largest deployment footprints are accumulating proprietary post-training data that no open-source competitor, academic institution, or late-entering nation can reproduce. Not because of hardware. Because of time and scale. China's Zhipu AI just released a model trained entirely on Huawei chips claiming 94% of Claude Opus performance — as open-weights. This is technically impressive. It is also competing with the model that was released six months ago, not the model running inside Anthropic today.

> ***The chip export controls are maintaining the hardware gap. The reasoning trace flywheel is maintaining the data gap. Both are widening simultaneously. And the data gap does not have an export control workaround.***

The Sovereignty Problem You Should Be Losing Sleep Over

If you read this site, you already understand that the central tension of our era is who controls the infrastructure of value and information. The AI story is the clearest current expression of that tension — and the answer is becoming uncomfortable.

The capability to train frontier AI models is collapsing to a single-digit number of organizations globally. Training costs of $500 million to $1 billion per run today are projected to reach $10–50 billion by 2027. Frontier model training is becoming a sovereign-scale industrial activity. Not a startup activity. Not an open-source activity. A capital concentration that structurally resembles the early nuclear weapons programs more than anything else in the commercial technology sector.

At the same time, two of the four leading US labs — and possibly all four — have active classified government relationships that make the public-facing product the civilian face of a dual-use

capability system. The model you chat with is the part they decided you could have.

The decentralization thesis for AI is real but faces a structural ceiling: base model quality can be democratized, but the post-training data that builds each successive generation cannot. Open-source will always be chasing the released public model. The released public model will always be chasing the internal checkpoint. The internal checkpoint will always be chasing the classified fine-tune.

It is turtles all the way down — and the turtles you cannot see are the ones that matter.

So What

Three direct implications. No hedging.

For capital allocation: Any AI company whose moat is "we have a great model" does not have a durable moat. The moat is proprietary domain data, deployment-layer integration, and the reasoning traces that accumulate from operating at scale. The value is not in the model. It is in what the model is trained on and what it learns from running. Bet accordingly.

For operators: The capability available to you right now on public frontier models is already sufficient to automate the majority of your knowledge-work processes. Waiting for better models is the wrong frame. The question is whether you are capturing proprietary operational data from AI deployment that compounds into a training advantage — or whether you are just using AI as a faster word processor.

For anyone thinking about sovereignty: The concentration of frontier AI capability in four to six organizations, two or three of which have deep classified government relationships, is the most significant power concentration since nuclear weapons. The critical difference: nuclear capability was expensive to build and relatively easy to deter through mutual assured destruction. AI capability is expensive to build, nearly impossible to deter, and has no equivalent of a non-proliferation treaty. The NDAA committee analyzing AI toward AGI had an April 1, 2026 deadline. That deadline has passed. Whatever they found, you will not be told.

The walls are built. The game is inside them.

Dr. Gregory S. Carmichael is CEO of Quantum Reserve Capital and Founder of Advanced Nano-Materials Manufacturing LLC. He is a retired Lt. Col., USAF, US Representative to the G20 India Summit (2023), and DPA Title III Advisor at Pearl Harbor (2024). His book The Invisible Hand Meets AI *(ISBN*

979-8-9955703-0-1) is available now on Amazon — it hit #1 New Release on launch week.

INFOGRAPHIC

Three tiers of AI capability · April 2026

TIER 1: PUBLIC FRONTIER

← YOU ARE HERE

What you can actually access. Subscriptions, APIs, benchmarks you can read.

94.3%	80.9%	77.1%
GPQA Diamond grad-level science	SWE-bench real bug fix	ARC-AGI-2 fluid reasoning

GPT-5.4 · Claude Opus 4.6 · Gemini 3.1 Pro · Grok 4.2 — Genuinely extraordinary. Already obsolete as a description of the actual frontier.

▼ THE LINE YOU CAN'T SEE PAST ▼

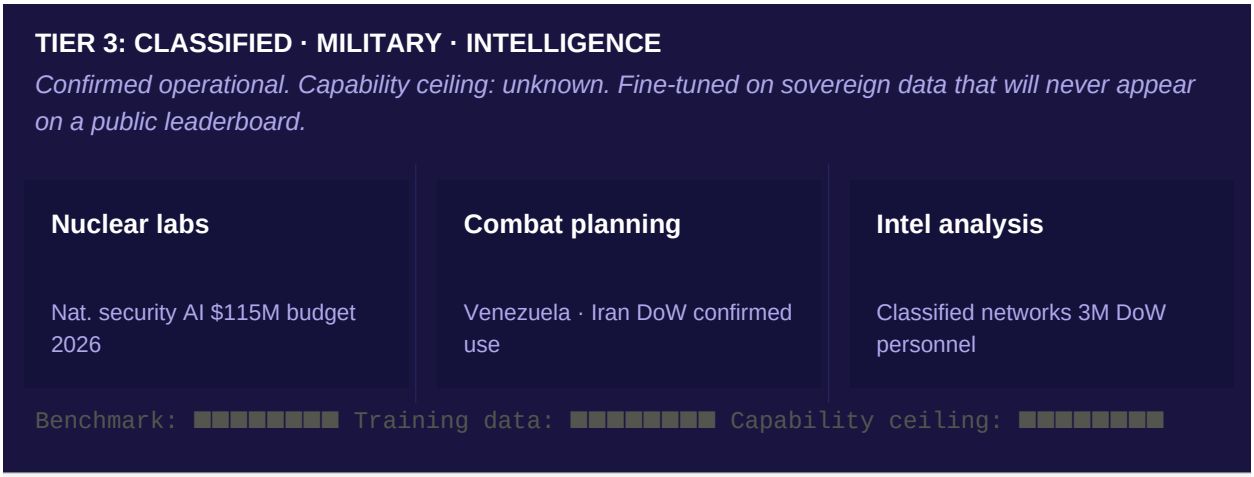
TIER 2: INTERNAL LAB CHECKPOINTS

Running inside Anthropic, OpenAI, Google DeepMind, xAI right now. Not released. Not benchmarked publicly.

6–18 months ahead	Blackwell-trained
of any public release · structural lag in every pipeline	3–5× more efficient chips · in training for 6–9 months now

"All four U.S. leaders have much better internal checkpoints than public releases — very hard for anyone to catch up." — Nov 2025 infrastructure analysis

▼ CLASSIFICATION BOUNDARY ▼



Sources: Published benchmark data · DoW public filings (GenAI.mil, Jan 2026 AI Strategy) · Anthropic public statements · FY2026 NDAA · Euronews / media reporting

Figure 1. Three tiers of AI capability. Tier 1 is the only level visible to subscribers, developers, and analysts. Tiers 2 and 3 operate without public benchmarks.