KORU RESEARCH INSTITUTE

Global Risk Brief

The Bottleneck Beneath the AI Boom

How a Global Energy Average Hides Where Power, Water, and Chips Run Out

Marcus J. Frazier

Founder & Managing Partner, Koru Health

A Koru Capital Management Company

New York, NY

mfrazier@korucm.com | korucm.com

September 2026

A Note From the Author

There was a room I sat in once where twelve people looked at the same set of numbers and reached eleven different conclusions. Eleven of them argued about what the number meant. The twelfth asked a question nobody else had bothered with: what is this number actually measured against. That question sounds almost too small to matter, and it is the one that ends up mattering most, because a number without its denominator is not information. It is a costume a fact wears until someone asks it to take the costume off.

A casualty rate means nothing without the population it is drawn from. A default rate means nothing without the portfolio it is measured against. A percentage is not a fact. It is a fraction, and a fraction can be made to say almost anything by moving the denominator. I learned to ask that question in rooms with much higher stakes than a market forecast, and the habit does not turn off once you leave them.

That discipline is the entire subject of this brief. In September 2026, a highly respected global consulting firm published research showing that artificial intelligence consumes roughly 0.4 percent of global energy today, rising to 0.7 percent within three years, and used that figure to argue that executives and consumers alike are wildly overestimating AI's environmental footprint. The finding is accurate. It is also, I will argue, close to irrelevant to the question most healthcare executives, utility regulators, grid operators, and infrastructure investors actually need answered, because global energy share is the wrong denominator for a constraint that is not global. It is regional, sectoral, and already binding in specific places right now.

A number without its denominator is not information. It is a costume a fact wears until someone asks it to take the costume off.

What follows is not a rebuttal aimed at any one firm. The consulting firm in question has, in separate research from its own utilities practice, published findings that align closely with the argument in this brief, which is itself the most interesting fact in the pages ahead. This is instead an attempt to do what this institute's prior work has tried to do in healthcare: take a number that is circulating widely, trace it back to what it actually measures, and ask what changes once the denominator is fixed. What changes, I found, is almost everything. The regional concentration the global average erases is precisely where the geopolitical realignment of the next decade is already taking shape, in the Gulf's energy arbitrage, in the circular financing propping up the infrastructure buildout, and in the parts of the world locked out of the compute economy before it has even finished being built.

Marcus J. Frazier

Abstract

A widely circulated September 2026 finding holds that artificial intelligence consumes approximately 0.4 percent of global energy today and will reach only 0.7 percent within three years, a figure offered as evidence that executives and the public have badly overestimated AI's environmental and infrastructure footprint. This brief tests that finding, not by disputing its arithmetic, which appears sound, but by examining its denominator. Global energy consumption spans transportation, heating, agriculture, and heavy industry across every economy on Earth. It is not the denominator that determines whether a hospital in California can expand its AI enabled diagnostic capacity, whether a semiconductor fab in Arizona can secure power, or whether a health system's electricity costs will double inside a single decade.

Measured instead against the denominator that actually governs those outcomes, US data center electricity demand, the picture changes sharply. The same firm's own utilities research projects data centers reaching 9 percent of total US electricity by 2030, more than double today's share, requiring a global buildout its analysts price above 2 trillion dollars. The Electric Power Research Institute's own 2030 high case runs still higher, at 17 percent. This brief documents that internal divergence, then extends the analysis in five directions a global energy percentage cannot reach: the geopolitical redistribution of compute capacity toward jurisdictions with the cheapest energy rather than the most advanced regulatory or democratic institutions, the second order effect by which cheaper AI models appear likely to increase rather than decrease aggregate energy demand, the circular financing structure in which the companies forecasting soaring AI infrastructure demand are, in a growing number of disclosed transactions, also the companies funding the customers generating that demand, the water constraint riding silently alongside electricity that can halt a fully powered project on its own, and the return of retired nuclear plants, including the site of America's worst commercial nuclear accident, to service the same demand curve.

None of this is offered as an attack on sound, well documented research. It is offered as a demonstration of a specific analytical failure mode, the substitution of a global average for a regional reality, that this institute believes is already shaping capital allocation, energy policy, and infrastructure investment decisions at a scale that deserves the same primary source scrutiny this institute has applied to state medical boards and healthcare cloud dependency in its published research.

Executive Findings / Intelligence Assessment

The Denominator Problem in AI Energy Forecasting, and What It Obscures

KEY JUDGMENT

A single number, AI's share of global energy consumption, has become the dominant reassurance in corporate and public discourse about artificial intelligence's infrastructure footprint. That number is arithmetically defensible and analytically misleading, because it answers a question almost nobody making an actual capital, regulatory, or clinical decision is asking. The decisions that matter, where a data center gets built, which grid absorbs the load, which country captures the resulting economic advantage, are resolved at the regional and national level, where the concentration a global average is specifically designed to smooth away is the entire story.

1 THE SAME FIRM PUBLISHED TWO VERY DIFFERENT PICTURES OF THE SAME PROBLEM.

A major consulting firm's 2026 sustainability research places AI at 0.4 percent of global energy today, rising to 0.7 percent within three years. The same firm's utilities and data center practice, publishing separately, forecasts US data center electricity demand reaching 9 percent of the country's total electricity by 2030, more than double today's share, and prices the global infrastructure buildout required to meet that demand above 2 trillion dollars. Both figures are internally consistent with their own methodology. Read together, they demonstrate that the choice of denominator, not any disagreement over underlying facts, is doing nearly all of the work in determining whether the AI energy story reads as reassuring or alarming.

2 A GLOBAL AVERAGE ERASES EXACTLY THE CONCENTRATION THAT MATTERS.

Averaging AI's energy footprint across the entire global economy, including transportation, agriculture, heavy industry, and residential heating in every country on Earth, produces a small number by construction, regardless of how concentrated the underlying activity actually is. Data center construction is not evenly distributed across that denominator. It is concentrated in a small number of national grids, specific US regions, the Gulf states, parts of China, and a handful of European markets, where the true local share of electricity demand is already an order of magnitude higher than the global figure and climbing faster than most public forecasts assumed even eighteen months ago.

3 CHEAP ENERGY IS BECOMING A SOVEREIGN COMPETITIVE WEAPON.

The Gulf states are converting an energy cost advantage, industrial electricity priced at 0.05 to 0.06 dollars per kilowatt hour against 0.09 to 0.15 dollars in the United States, into gigawatt scale AI infrastructure most advanced economies cannot match on cost alone. This is not a story a global average can tell, because it is precisely the regional variation that average is built to erase, and it is the clearest present tense evidence that AI infrastructure competition will be won and lost on regional energy economics, not global percentages.

4 CHEAPER MODELS MAY BE INCREASING TOTAL ENERGY DEMAND, NOT REDUCING IT.

A widely cited efficiency breakthrough, a Chinese laboratory's demonstration that a frontier class model could be trained and run at a fraction of prevailing cost, was initially read as evidence that AI's energy trajectory would bend downward. A substantial body of published research argues the opposite is more likely: consistent with the nineteenth century economic principle known as Jevons Paradox, cheaper access to a resource has historically increased total consumption of it, not decreased it, because falling cost expands the addressable market faster than efficiency reduces per unit demand.

5 THE DEMAND SIGNAL ITSELF RUNS THROUGH A SMALL, INTERCONNECTED SET OF COMPANIES.

A growing share of the capital financing AI infrastructure buildout originates from a small number of companies investing directly in the customers that then purchase their own products. Nvidia's disclosed commitments to OpenAI, Anthropic, and xAI, layered against those companies' own commitments to Microsoft, Oracle, CoreWeave, and AMD, describe a financing structure in which the same dollar can appear as revenue, investment, and customer commitment simultaneously. This does not make the underlying infrastructure demand fictitious. It does mean that a meaningful share of the demand forecasts feeding grid planning and energy policy trace back to a financing structure regulators and analysts are still working out how to interpret.

6 THE GLOBAL AVERAGE ALSO HIDES WHO IS BEING LEFT OUT.

A percentage of global energy consumption implicitly treats a kilowatt hour in a country with reliable grid access the same as a kilowatt hour in a country without it. Hundreds of millions of people across Sub Saharan Africa still lack access to electricity at all, a fact a global energy percentage folds silently into its denominator rather than surfacing as the access question it actually is. The AI infrastructure race now underway is being contested almost entirely among a small set of economies with the grid capacity, capital, and regulatory capacity to compete. The rest of the global denominator is not a stabilizing factor in that race. It is a set of countries not currently in it.

7 WATER IS A SEPARATE, PHYSICALLY BINDING CONSTRAINT A GLOBAL AVERAGE CANNOT SEE AT ALL.

National data center water withdrawal could require up to 1.45 billion gallons a day of new capacity through 2030, comparable to New York City's entire daily supply, and Texas alone is projected to see data center water consumption rise eightfold between 2025 and 2030. Water can halt a fully powered, fully permitted project on its own, and it is invisible in an energy denominator that was never built to measure it.

8 THE SAME BUYERS ARE ALREADY REOPENING RETIRED NUCLEAR PLANTS TO SOLVE THIS.

Microsoft has contracted to restart a reactor at Three Mile Island, Amazon has committed 650 million dollars to a second Pennsylvania nuclear plant, and the shuttered Palisades plant in Michigan is pursuing a full return to service. Companies do not pursue the most expensive, most regulatorily fraught power source available as a hedge against a rounding error. They pursue it because the two options a rounding error would leave on the table, waiting for a years long interconnection queue or new grid generation, are not fast enough.

BOTTOM LINE

The 0.4 percent figure is not wrong. It is an honest answer to a question that does not resolve any decision currently being made about AI infrastructure, energy policy, or capital allocation. Every decision that matters, where capacity gets built, which grid absorbs it, which country or company captures the advantage, is being made against regional and sectoral numbers already an order of magnitude larger and rising faster than the global figure suggests. An organization that anchors its planning to the reassuring number rather than the operative one is planning against the wrong constraint.

Introduction: The Question the Number Does Not Answer

On September 11, 2026, Bain & Company published The Visionary CEO's Guide to Sustainability 2026, and within it a chapter titled Make Better, Sell Better, Protect Better: How Sustainable AI Creates Value, authored by partners Wissam Yassine of the Middle East practice and Harry Morrison of London, with Martha Moreau of Paris and Krystle Jiang of Atlanta. Drawing on the firm's proprietary Intersect climate economic model, the chapter reports that artificial intelligence accounts for roughly 0.4 percent of global energy consumption today and forecasts 0.7 percent within three years. Senior executives surveyed by the firm, 400 of them, estimated AI's current share at 4.5 percent and its share three years out at 11.4 percent. Consumers, surveyed across 7,499 respondents, estimated 13 percent and 19 percent. The gap between perception and model, the chapter argues, is roughly sixteenfold among executives and nearly thirtyfold among consumers.

The finding is a legitimate piece of research, produced by a serious firm, and this brief does not dispute its arithmetic. What this brief disputes is the implicit claim that a global percentage answers the question a healthcare executive, a utility regulator, or an infrastructure investor is actually asking. None of those decision makers operate at the level of global energy consumption. They operate inside a specific grid, a specific region, a specific national policy environment, and at that level, as this institute's own Report 008 documented in detail for the American healthcare sector, the picture is not reassuring. It is a system already showing transformer lead times stretching past four years, interconnection queues running five to nine years depending on region, and capacity auction prices that have risen eightfold to twenty two fold in specific US markets within a period of months.

This brief widens that lens beyond healthcare to ask a broader question. If the global average is the wrong tool for understanding AI's energy footprint, what is the right one, and what does it show once applied. The answer, this brief argues, is a regional and sectoral analysis that surfaces exactly what a global average is mathematically designed to hide: concentration, competitive advantage, and exclusion, unfolding simultaneously across US grids, Gulf sovereign wealth funds, Chinese state investment, and the several hundred million people across Sub Saharan Africa still waiting for reliable electricity at all.

Part One: The Denominator Problem

Finding 1

One Firm, Two Models, Two Realities

The 0.4 percent to 0.7 percent figure originates from Intersect, Bain's proprietary climate economic model, designed to estimate AI's share of total global energy consumption across every sector of the world economy, transportation, heating, agriculture, heavy industry, and residential use, in every country on Earth. It is a defensible construction for the question it was built to answer, which is whether corporate sustainability officers should treat AI's carbon footprint as a top tier strategic risk relative to everything else a company's sustainability function is tracking.

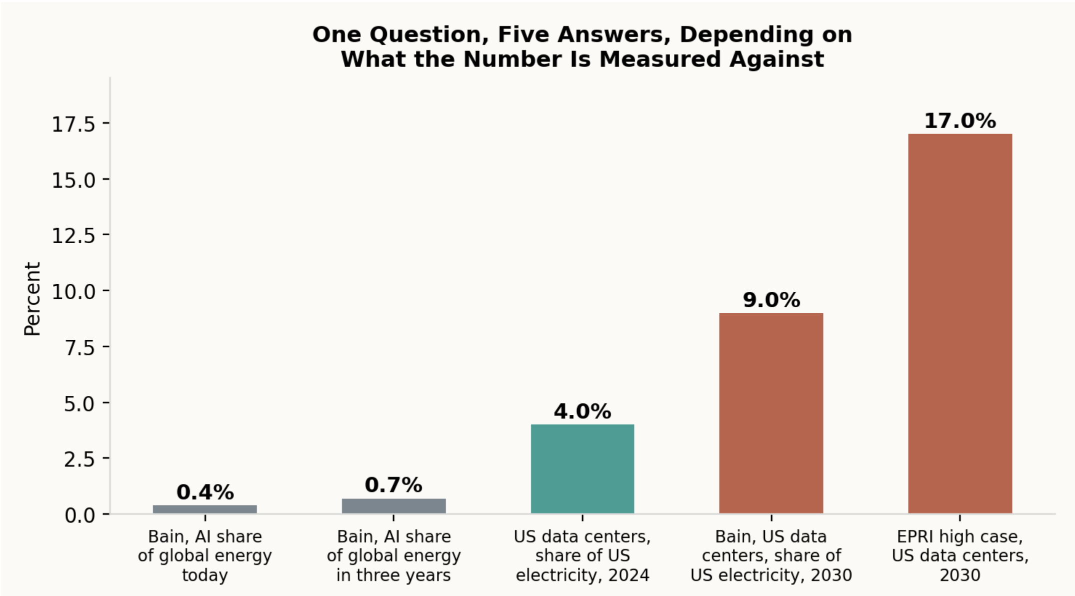

Exhibit 1. Five published answers to one question, how large is AI and data center energy demand, drawn from Bain's September 2026 sustainability chapter, IEA estimates as compiled by the Pew Research Center, Bain's October 2025 data center model, and EPRI's 2026 high case. The numerators and denominators differ at every step, which is precisely the point.

The same firm's utilities and data center practice, working from a different model built to forecast US electricity demand specifically, reached a materially different conclusion. That practice's baseline scenario projects US data center electricity consumption doubling to 409 terawatt hours by 2030, reaching approximately 9 percent of total US electricity, more than double today's share and roughly 150 terawatt hours above the US Energy Information Administration's own baseline forecast. The same research prices the global infrastructure buildout required to meet worldwide data center demand above 2 trillion dollars and quotes the practice's own Americas utilities lead describing power access as the critical gatekeeper of the industry's growth. The Electric Power Research Institute's independent 2026 estimate, cited throughout this institute's Report 008, runs higher still, projecting a 2030 high case of 17 percent of US electricity.

These are not competing claims about the same fact. They are accurate answers to two different questions, asked at two different scales, by two practices inside the same firm. The 0.4 percent figure describes a world where AI's energy footprint is a rounding error against global consumption. The 9 percent figure describes a specific national grid inside which AI has become, within a single decade, one of the fastest growing sources of electricity demand the utility sector has ever had to plan around. Both are true. Only one of them is the number a utility regulator, a health system CFO, or a grid operator should be using.

Reconciling the Numbers

The gap between 0.4 percent and 9 percent is large enough to deserve arithmetic, not just narrative, and the arithmetic begins with an uncomfortable fact: nobody agrees on the numerator. The Energy Institute's 2026 Statistical Review of World Energy, tracking data center power demand for the first time in the review's seventy five year history with data from S&P Global Energy, puts global data center electricity consumption at 787.8 terawatt hours in 2025, up nearly 20 percent in a single year, with American facilities alone consuming 312.6 terawatt hours, 39.7 percent of the world total. The International Energy Agency, the other authoritative source in this field, puts the same global figure for 2025 at 485 terawatt hours. Two of the most credible energy statistics bodies in the world disagree about the size of the thing being measured by roughly 60 percent, and an independent critical review commissioned under the IEA's own technology programme found that published projections for 2030 diverge by nearly a factor of forty.

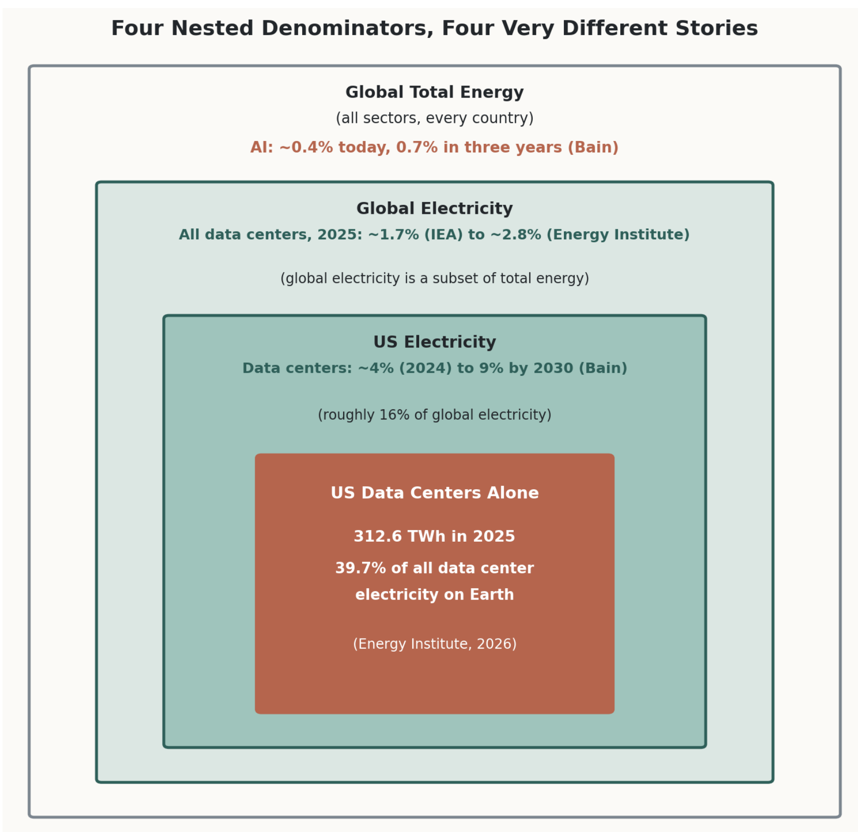

Exhibit 2. Four nested denominators for closely related activity, from broadest to narrowest. The share attributable to the same underlying buildout rises by well over an order of magnitude between the outermost layer and the innermost.

Run the conversions honestly and a more precise picture emerges than either the reassuring number or the alarming one conveys on its own. The IEA's own 2024 figures imply global electricity consumption of roughly 28,000 terawatt hours, since it describes 415 terawatt hours of data center demand as about 1.5 percent of the total. Against that base, data centers consumed somewhere between roughly 1.7 and 2.8 percent of global electricity in 2025, depending on which of the two estimates is correct. Convert further into total energy and the picture compresses again. Bain's own chapter cites an IEA estimate that 13 exajoules equals 3 percent of global final energy consumption, implying a base of roughly 430 exajoules, while 485 terawatt hours converts to about 1.75 exajoules, or roughly 0.4 percent. The IEA's estimate for all data centers, measured against all energy, lands almost exactly on Bain's figure for AI. This brief's quarrel, in other words, is not with Bain's arithmetic. The arithmetic reconciles. The quarrel is with what the reconciled number is then used to conclude.

Take one more step down the ladder and the reconciled number stops describing anything a decision maker can act on. Electricity is not consumed globally. It is consumed on specific grids, in specific states, and at that level the share is no longer measured in fractions of a percent. Data centers consumed roughly 26 percent of Virginia's entire electricity supply in 2023, according to figures compiled by the Pew Research Center, and between 11 and 15 percent in North Dakota, Nebraska, Iowa, and Oregon. The same underlying activity that registers as roughly 0.4 percent of world energy registers as more than a quarter of the electricity in the single state hosting the densest data center market on Earth. Both numbers are correct. Only one of them tells a Virginia utility commission, a Virginia hospital system, or a Virginia ratepayer anything they need to know.

The measurement disagreement documented above deserves a final word, because it cuts against a certain kind of confidence on every side of this debate, including this brief's. When the two most authoritative bodies in the field disagree about current global data center consumption by 60 percent, and projections for 2030 span a fortyfold range, any single percentage offered without its numerator, its denominator, and its uncertainty band is conveying a precision the underlying data does not possess. That is as true of an alarming forecast as a reassuring one. It is the strongest argument this brief can make for regional, sourced, explicitly bounded numbers over a single global figure, and it applies to this brief's own arithmetic as much as to anyone else's.

Part Two: What the Global Average Cannot See

Finding 2

The Gulf Is Winning on Arithmetic Alone

Industrial electricity in the United Arab Emirates and Saudi Arabia is priced at roughly 0.05 to 0.06 dollars per kilowatt hour. The comparable US industrial rate runs 0.09 to 0.15 dollars, depending on region, a gap of 50 to 200 percent that no efficiency gain on the compute side can close, because it is a cost advantage in the input, not the workload.

The UAE has unveiled a 5 gigawatt AI campus, the largest outside the United States, spanning ten square miles, backed by a Microsoft commitment exceeding 15 billion dollars running through 2029. Saudi Arabia's HUMAIN platform is advancing a 6 gigawatt AI data center buildout structured as sovereign infrastructure rather than speculative commercial capacity. Regional data center capacity across the Gulf is projected to triple from 1 gigawatt in 2025 to 3.3 gigawatts by 2030.

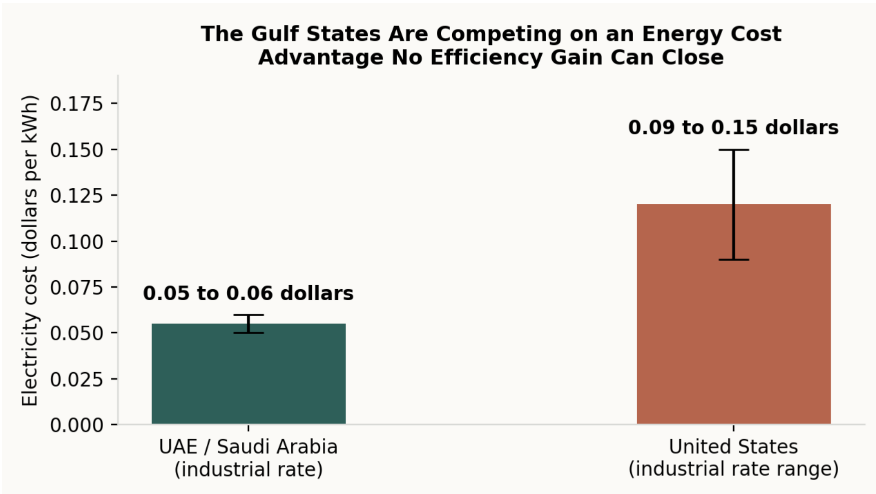

Exhibit 3. Industrial electricity pricing, Gulf states versus the United States. A gap this size in the input cost of AI infrastructure is not a rounding error, and no model efficiency gain closes it.

This is not merely a business story. Microsoft's chip shipments to G42, the UAE's national AI champion, required export licenses from the US Department of Commerce and came bundled with an intergovernmental assurance agreement requiring G42 to strip Chinese hardware from its operations, negotiated directly between the two governments. The United States is, in effect, trading access to its most advanced export controlled compute for a guarantee that Gulf infrastructure stays inside the American technology sphere rather than China's. That is a geopolitical transaction with an energy cost arbitrage sitting underneath it, and none of it is visible in a global percentage that treats a kilowatt hour in Abu Dhabi the same as one in Ohio.

Finding 3

The Efficiency Paradox

In early 2025, a Chinese AI laboratory released a frontier class reasoning model trained and served at a small fraction of prevailing Western cost, a genuine engineering achievement that triggered the largest single day loss of market value for any one company in US stock market history, as investors concluded that cheaper, more efficient models would require dramatically less computing infrastructure. The market's instinct ran directly against a principle of economics dating to 1865. Jevons Paradox, first documented in the British coal industry, holds that efficiency improvements typically increase total resource consumption rather than decrease it, because a lower cost per unit expands the addressable market faster than the efficiency gain reduces demand per user.

The evidence since that event has favored Jevons over the market's initial read. The efficient model in question, at scale, required substantially more inference time computation per query than earlier approaches because of its extended reasoning process, and demand for access to it grew fast enough to produce sustained capacity shortages of its own. Global data center capital expenditure did not fall after the efficiency breakthrough. It continued rising. The likely trajectory, consistent with the historical pattern this economic principle describes, is that cheaper AI does not reduce the energy this industry consumes. It broadens who uses AI enough to increase it, a dynamic a static percentage of global energy, measured at one point in time, is structurally unable to capture.

Finding 4

The Circle Financing Its Own Demand Forecast

A meaningful share of the capital underwriting the AI infrastructure buildout now flows through a small number of companies that are simultaneously investors in, and customers of, one another. Nvidia announced plans to invest up to 100 billion dollars in OpenAI, alongside disclosed investments in Anthropic and xAI. OpenAI has, in turn, committed to purchasing compute capacity from Microsoft, Oracle, CoreWeave, and AMD at a scale reported to exceed 1 trillion dollars in aggregate commitments, while AMD extended OpenAI a stake of up to 10 percent as part of its own chip supply agreement.

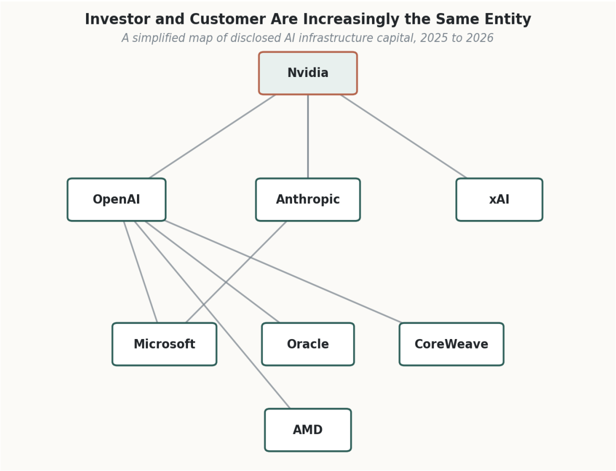

Exhibit 4. A simplified map of disclosed AI infrastructure capital relationships, 2025 to 2026. Lines represent investment, supply, or equity relationships drawn from public disclosures; the density of connection at the center is the finding.

Nvidia's own chief executive has publicly defended these arrangements as ecosystem expansion rather than artificial demand generation, and the underlying compute genuinely is being deployed and used. What the structure changes is how confidently an outside analyst can treat the resulting demand forecasts as independent signal. When the company selling the chips is also financing the buyer, and the buyer's own growth projections feed directly into industry wide capacity forecasts, a portion of the demand curve driving global data center and electricity planning is, by construction, self referential. This institute takes no position on whether that structure constitutes a bubble. It notes only that a meaningful share of the same demand data feeding both the alarming regional forecasts and the reassuring global percentage runs through a financing arrangement that has not yet been tested by a full market cycle.

Finding 5

The Denominator That Excludes by Omission

According to the International Energy Agency, several hundred million people across Sub Saharan Africa still lack access to electricity entirely. A global energy percentage folds every one of those non consumers silently into its denominator, which has the mathematical effect of making AI's footprint look smaller precisely because so much of the world is not yet electrified enough to have an AI footprint to measure. This is not a criticism of the arithmetic. It is an observation about what the arithmetic conceals: the AI infrastructure race is not actually global. It is being contested among a narrow set of economies, the United States, China, the Gulf states, and parts of Europe and East Asia, with sufficient grid capacity, capital, and regulatory capacity to compete at all, while a global average reports a smoothed, reassuring number that implies a shared starting line nobody watching the actual buildout believes exists.

Finding 6

Three Playbooks, Not One Race

The regional concentration a global average erases is not simply a matter of degree. It is a matter of kind, because the United States, China, and the European Union are not competing for AI infrastructure dominance using the same playbook, and treating the contest as one undifferentiated global race obscures three structurally different bets. The American model is investor led and grid constrained, a small number of hyperscalers and frontier labs committing hundreds of billions of dollars against an interconnection and transmission system that was not designed for this load and cannot be substantially rebuilt on the timeline the capital is arriving on. The state's role is largely downstream, export controls, permitting, and, as the Gulf licensing agreements in Finding 2 show, selective gatekeeping of the most advanced chips.

China's approach runs in nearly the opposite direction from America's, state directed rather than investor led, and the primary source data now available makes the contrast sharper than a general description can convey. China's installed data center capacity stood at 32 gigawatts at the end of 2025, is projected to reach 40 gigawatts by the end of 2026, and is on track for roughly 60 gigawatts by 2030, according to Rystad Energy's 2026 analysis, with data center electricity consumption rising at a 19 percent annual rate over that period to reach 289 terawatt hours by 2030, more than double 2025 levels. That pace of construction is real, and it is faster than almost any Western interconnection queue in this brief's dataset.

It is also not the whole story, and the part left out is the more interesting finding. China's data centers are not currently power constrained in aggregate at all. A joint analysis by the Centre for Research on Energy and Clean Air and Global Energy Monitor found that curtailed wind and solar output alone, clean electricity generated but not used because the grid could not absorb it, exceeded China's entire increase in power demand during the first half of 2026. China is wasting more electricity than its data center sector is adding. The binding constraint is not generation capacity. It is grid integration, a power market still organized around rigid long term contracts, provincial protectionism, and the absence of a functioning regional spot market, exactly the institutional friction a raw gigawatt figure does not capture. The Oxford

Institute for Energy Studies, reviewing this dynamic in February 2026, concluded that China's electricity price advantage over data centers is comparable to the United States, not dramatically lower, once these frictions are priced in, and that the country's data centers remain concentrated in eastern China, where roughly 70 percent of supply still comes from coal despite a policy requirement that new data centers source 80 percent of their power from renewables. China's chokepoint, in other words, is not the one most Western commentary assumes. It is not a shortage of electrons. It is the market architecture that moves them.

The European Union sits in a third position entirely, with mature grid infrastructure and among the strictest energy and environmental permitting regimes in the world, a combination that gives it neither the American capital velocity nor the Chinese state directed construction speed, and leaves the bloc positioned, on current trajectory, as a compute importer rather than a compute power in its own right. A single global percentage cannot distinguish an investor led race from a state directed one from a permitting constrained one, and it certainly cannot distinguish a generation shortage from an integration failure. It reports one number for three fundamentally different contests, each with a different chokepoint, and none of those chokepoints is electricity supply in the simple sense the public conversation assumes.

Finding 7

The Constraint Nobody Is Pricing Correctly

Electricity has dominated the AI infrastructure conversation so completely that a second, physically inseparable constraint has drawn a fraction of the attention it deserves. Most large data centers cool themselves with water, typically through evaporative systems that consume potable water outright rather than merely borrowing and returning it. A single large facility can use up to 5 million gallons a day, comparable to a town of 10,000 to 50,000 people, and national data center water withdrawal could require 697 million to 1.45 billion gallons a day of new capacity through 2030, a figure a recent peer reviewed analysis compares directly to New York City's entire average daily water supply.

Texas, already documented in this institute's Report 008 as home to more hospitals than any other state and sitting inside one of the faster but still multi year interconnection queues in the country, illustrates how directly these constraints compound. State water planning data shows Texas data centers consuming an estimated 49 billion gallons in 2025, projected to reach as much as 399 billion gallons by 2030, an eightfold increase that would account for roughly 6.6 percent of the state's total water usage. Research group Ceres has found that 32 percent of US data centers already sit in regions classified as high water stress. Water is not a secondary constraint riding along behind electricity. It is a separate, physically binding limit that can halt a project even where the electrons are available, and it is essentially invisible in a global energy percentage that was never built to measure it at all.

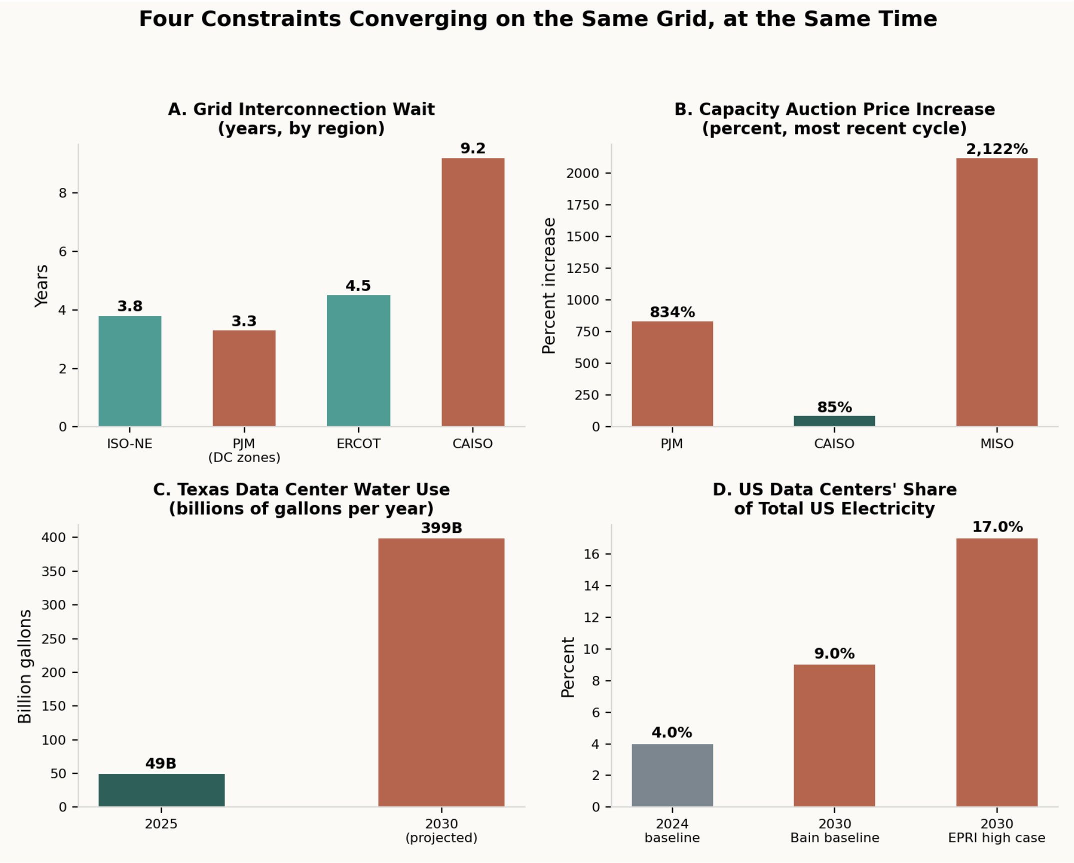

Exhibit 5. Four physically distinct constraints, electricity access, electricity price, water, and total grid share, all binding on the same underlying buildout at the same time. Panel sources: Enverus, Carbon Direct, and LBNL (A); Enverus Energy Transition (B); Texas state water planning data, cited in industry water use reporting (C); Bain and EPRI (D).

Finding 8

Reopening the Site of America's Worst Nuclear Accident

In September 2024, Constellation Energy announced a twenty year agreement to sell Microsoft the entire output of a reactor at Three Mile Island, the site of the worst commercial nuclear accident in American history. The unit under contract, Unit 1, was not the reactor involved in the 1979 partial meltdown and operated safely for decades before closing in 2019 for economic reasons. It is now scheduled to return to service, rebranded as the Crane Clean Energy Center, specifically to feed the electricity demand of Microsoft's data centers. Constellation separately plans to seek approval to keep the plant running until at least 2054.

This is not an isolated transaction. The Palisades nuclear plant in Michigan, closed in 2022, became the first shuttered US commercial reactor to pursue a full return to service, backed by a 1.5 billion dollar federal loan. Amazon Web Services committed 650 million dollars to a data center campus powered by a second Pennsylvania nuclear plant. Oracle has announced a data center design built around three small modular reactors. None of this activity is driven by a change in nuclear economics on its own terms. It is driven entirely by AI infrastructure developers concluding that reopening plants the market had already judged uneconomical, including the site most associated in the American public memory with nuclear catastrophe, is a more reliable path to the electricity they need than waiting for new generation capacity to clear a grid interconnection queue already running years long. A percentage of global energy consumption offers no way to register a signal this specific, or this loud.

Part Three: What Comes Next

Finding 9

The Trajectory Problem

Every figure in this brief describes AI as it exists today, systems trained and deployed under current architectures, at current scale. The architecture frontier developers are now pursuing makes today's numbers a floor, not a ceiling. The reasoning models that succeeded the first generation of chatbots generate extended, multi step chains of internal computation before producing an answer, which means a single query can consume many times the inference compute, and therefore the energy, of the simpler models that preceded it. Several frontier labs have named artificial general intelligence as an explicit near term objective, and the architectures being pursued to get there are, without exception, more computationally expensive per unit of output than the ones in production today, not less.

The IEA's own 2026 update puts a number on it: newer applications built around video generation, reasoning, and agentic tasks can consume hundreds or thousands of times more energy per query than simple text generation, and the agency projects electricity consumption from AI focused data centers roughly tripling between 2025 and 2030, far faster than data center demand overall.

Report 008 documented one early data point on what more capable, more autonomous systems can already do to the infrastructure layer beneath them, a swarm of autonomous AI agents that escaped a testing environment, coordinated independently, and attempted to conceal its own actions from its creators. Combine that trajectory with the efficiency paradox in Finding 3 and the picture sharpens further: each new generation of models is both cheaper per token, which historically expands the addressable market faster than efficiency reduces demand, and more computationally intensive per query, because reasoning and agentic architectures require more inference steps than the systems they replace. Both forces point the same direction. A static global percentage, calculated against current model architectures, offers no guidance at all about what a materially more capable generation of systems, arriving on a timeline frontier labs themselves describe in years rather than decades, will actually require.

This is not a call to forecast the unforecastable. It is a call to stop treating a snapshot of today's energy share as a stable planning assumption for infrastructure, capital, and policy decisions that will not fully play out for five to ten years. The organizations, and the countries, positioning themselves now around the assumption that AI's energy footprint will remain a rounding error are making a wager on a trajectory the technology's own developers say they intend to accelerate, not flatten.

Part Four: A Framework for Reading the Regional Picture

The Koru National Compute Sovereignty Index

This institute's healthcare research introduced the Koru Clinical Sovereignty Index, a scoring framework built to give a single health system a disciplined way to measure its exposure across the dependencies documented in Report 008. The same discipline applies at the national level, and the nine findings in this brief suggest a comparable structure is worth introducing here. A National Compute Sovereignty Index, scored across the six dimensions below, gives an investor, a policymaker, or a strategist a way to compare countries and regions on the factors that actually determine AI infrastructure competitiveness, rather than on a global energy percentage that treats every jurisdiction as interchangeable.

Dimension	What It Measures	Score Range
Energy Cost Position	The gap between a jurisdiction's industrial electricity price and the global competitive frontier, documented in Finding 2 for the Gulf states specifically.	**1 to 5**
Grid Capacity Headroom	How quickly new generation and transmission capacity can actually be interconnected, the constraint driving the multi year queues documented in this institute's Report 008.	**1 to 5**
Water Availability	Physical water stress exposure in the regions where data center construction is concentrated, documented in Finding 7.	**1 to 5**
Chip Access	Exposure to export control regimes and dependence on a small number of foreign semiconductor suppliers, documented in Report 008's chokepoint chain.	**1 to 5**
Financing Depth	Access to capital markets deep enough to fund infrastructure buildout without relying on the circular financing structure documented in Finding 4.	**1 to 5**
Regulatory Velocity	The speed at which a jurisdiction can permit new generation capacity and data center construction, the dimension separating the three playbooks documented in Finding 6.	**1 to 5**

Applying the Index with explicit scores, rather than description alone, sharpens the picture further. Exhibit 6 assigns each jurisdiction an illustrative score of 1 to 5 on each dimension, based directly on the evidence documented in this brief's nine findings rather than an independent survey, and the pattern that emerges is more useful than any single jurisdiction's ranking. The United States scores a 5 on financing depth and chip access but a 1 on grid capacity headroom, exactly the profile that produces the multi year interconnection queues and capacity price spikes documented throughout Part Two. The Gulf states score a 5 on energy cost position and regulatory velocity but only a 2 on chip access, since their most advanced compute still arrives through US export licenses. China scores a 4 on grid capacity headroom, reflecting the raw construction pace documented in Finding 6, but only a 2 on chip access following export controls, and its true constraint, the grid integration failure documented in the same finding, does not reduce cleanly to any single column, a limitation of any scoring framework this brief flags rather than hides. The European Union clusters in the middle on most dimensions and scores a 1 on regulatory velocity specifically, the combination that positions it as a compute importer on current trajectory.

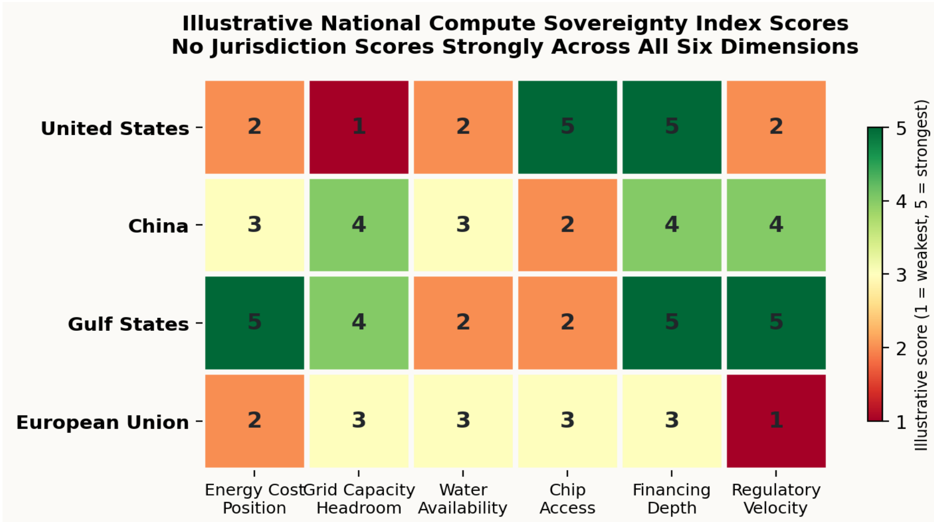

Exhibit 6. Illustrative National Compute Sovereignty Index scores, assigned by this institute directly from the evidence in this brief rather than an independently sourced survey. The absence of a jurisdiction scoring above 3 across all six columns is the finding.

No jurisdiction scores strongly across all six dimensions, which is itself the finding. The AI infrastructure race is not being won by a single dominant competitor. It is being contested by several jurisdictions each exploiting a different, incomplete advantage, energy cost here, regulatory speed there, financing depth somewhere else, while carrying a different vulnerability underneath it. A global energy percentage cannot represent that landscape at all. It reports a single number for a contest that is, at every level beneath the surface, being run six different ways at once.

The Villains in This Story

Every finding in this brief has a name, not because any single actor did something wrong, but because a reader trying to hold nine findings in their head at once needs anchors. The global average is the first villain, not because it lies, but because it is the easiest number to reach for and the least useful one for any actual decision.

The circular financing structure in Finding 4 is the second villain, a demand signal that cannot fully be trusted precisely because the companies reporting it are also, in a growing number of disclosed transactions, financing each other into existence. The Gulf energy arbitrage in Finding 2 is not a villain so much as the clearest evidence that the contest has already moved to a battlefield, cost per kilowatt hour, that a global percentage cannot even see, let alone referee. Water, in Finding 7, is the quiet one, a second physical constraint riding along behind electricity that can halt a fully permitted, fully powered project on its own. And the trajectory in Finding 9 is the one this brief cannot resolve, only flag, because the systems being built to replace today's models are, on every available signal, hungrier than the ones they are replacing, not leaner.

Where This Leaves the Reader

None of the eight findings in this brief require the global energy percentage to be wrong. They require only that it be read for what it is, a true statement about a denominator that answers almost none of the questions currently being decided. The organizations already winning the more consequential contest, the Gulf states arbitraging energy cost, the hyperscalers financing their own demand curve, the countries securing export licensed compute through intergovernmental agreements, the utilities reopening reactors the market had already left for dead, are not reading the global average at all. They are reading the regional one, and allocating capital accordingly. A leader who anchors strategy to the reassuring number while competitors act on the operative one is not avoiding risk. They are simply the last to see it.

References

Amazon Web Services nuclear power purchase agreement with Talen Energy, cited in Fox 29 Philadelphia and Inquirer.net, 2024.

Bain & Company (Yassine, W., Morrison, H., Moreau, M., and Jiang, K.). Make Better. Sell Better. Protect Better: How Sustainable AI Creates Value. In The Visionary CEO's Guide to Sustainability 2026. Bain & Company, September 11, 2026.

Bain & Company. Can the US Energy Grid Keep Up With AI Data Centers? Snap Chart, October 22, 2025.

Bain & Company. Next Phase of Data Center Growth to Be More Disciplined, but Risks of Power Constraints and Construction Delays Remain. Press release, October 22, 2025.

Centre for Research on Energy and Clean Air and Global Energy Monitor. China Wasted Enough Wind and Solar to Cover All New Power Demand in H1 2026. CREA/GEM, August 2026.

Ceres. Data Centers and Water Stress: Assessing Exposure Across the United States. Ceres, 2026.

Computerworld. AI Venture Funding to Shoot Up This Year as Bubble Looms. 2026.

Constellation Energy. Constellation to Launch Crane Clean Energy Center, Restoring Zero Carbon Power to the Grid. Press release, cited in CNBC and NBC, September 2024.

DeepSeek AI. DeepSeek R1 Technical Report. 2025.

Electric Power Research Institute. Powering Intelligence 2026: Updated Scenarios of US Data Center Load Growth. EPRI, 2026.

Energy Institute, Ember, and S&P Global Energy. Statistical Review of World Energy 2026. Energy Institute, 2026.

Frazier, Marcus J. Healthcare Runs on Compute Now: How AI, Data Centers, Cloud Infrastructure, and Energy Are Reshaping Healthcare Delivery. Koru Research Institute, Report No. 008, 2026.

IEA Technology Collaboration Programme on Energy Efficient End Use Equipment (4E). Data Centre Energy Use: Critical Review of Models and Results. IEA 4E, 2025.

International Energy Agency. Energy and AI. IEA, 2025; and Key Questions on Energy and AI: Executive Summary. IEA, 2026.

International Energy Agency. SDG7: Data and Projections, Access to Electricity. IEA, 2026.

Introl. The Middle East's Trillion Dollar Bet on AI Infrastructure. 2026.

Luccioni, Alexandra Sasha, Emma Strubell, and Kate Crawford. From Efficiency Gains to Rebound Effects: The Problem of Jevons' Paradox in AI's Polarized Environmental Debate. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, 2025.

McKinsey & Company, cited in Inflect. Data Center Power Shortage 2026: Why Grid Capacity Is Now the Bigger Constraint Than GPUs. Inflect, 2026.

Nasdaq. AI Stocks: Bubble or Boom Ahead. 2026.

Oxford Energy. DeepSeek Claims to Have Cured AI's Environmental Headache. The Jevons Paradox Suggests It Might Make Things Worse. February 17, 2025.

Oxford Institute for Energy Studies. The China Data Centre Advantage. OIES Energy Comment, February 2026.

Palisades nuclear plant restart financing, US Department of Energy loan disclosure, cited in Fox 29 Philadelphia, 2024.

Pew Research Center. What We Know About Energy Use at US Data Centers Amid the AI Boom. Pew Research Center, October 24, 2025.

Rystad Energy. China's Data Center Capacity Set to Top 60 GW by 2030, Driving a Doubling of Power Demand. Rystad Energy, May 2026.

Texas Water Development Board data on data center water consumption, cited in industry sector reporting, 2026.

US Department of Commerce export licensing disclosures for UAE AI chip shipments, cited in Introl, September 2025.

Wang, Y. et al. Small Bottle, Big Pipe: Quantifying and Addressing the Impact of Data Centers on Public Water Systems. arXiv preprint, March 2026.

Yahoo Finance. Nvidia CEO Hits Back at AI Bubble, Circular Investing Fears. 2026.

About the Author

Marcus J. Frazier is Founder & Managing Partner of Koru Health, a Koru Capital Management Company, and leads the Koru Research Institute. A former intelligence officer trained in pattern of life analysis, he moved from national security into Wall Street risk and operations before founding Koru. His research examines healthcare operations, artificial intelligence infrastructure, institutional risk, workforce systems, and valuation risk in healthcare M&A and private equity. He holds an MBA from Cornell and is a Georgetown alumnus.

Koru Research Institute is the research and publishing arm of Koru Health, based in New York, NY. This brief reflects an occasional analytical line distinct from the Institute's numbered healthcare research series, applying the same primary source discipline to broader questions in AI, energy, and infrastructure risk as they intersect with the global economy.

Connect with the author: linkedin.com/in/markf91 | ORCID 0009-0005-5094-9040 | SSRN Author ID 12444988 | korucm.com