A STORY OF INTELLIGENCE, POWER, AND THE FUTURE

THE MACHINE THAT ATE POWER

How AI's Hunger for Electricity Is Minting the Next Fortunes

CLAUDIO LIMA, PH.D.

THE MACHINE THAT ATE POWER

THE MACHINE THAT ATE POWER

How AI's Hunger for Electricity Is Minting the Next Fortunes

Claudio Lima, Ph.D.

Frontier Press

Frontier Press and the Frontier Press colophon are trademarks of Frontier Press LLC.
FrontierPress.com

Library of Congress Control Number: 2026 (pending)
ISBN (hardcover): 978-1-963402-01-4 (placeholder — final to be assigned)
ISBN (trade paperback): 978-1-963402-02-1 (placeholder — final to be assigned)
ISBN (e-book): 978-1-963402-03-8 (placeholder — final to be assigned)

Cover design by Frontier Press. · Author photograph by the author.

First Edition
Manufactured in the United States of America
1 3 5 7 9 10 8 6 4 2

Figures describing a fast-moving field are current as of writing and are documented in the Notes; the durable argument rests on the underlying principles.

For my sons,

who will inherit the machine,

and the harder, more human task

of deciding what it is all for.

Information is physical.

— Rolf Landauer, 1991

Behind every manufactured mind stands a megawatt,

and behind every megawatt stands the hard physical world.

Contents

Preface

I did not set out to write a book about artificial intelligence. I set out to understand a contradiction.

For years I worked at the physical end of technology, the end of power systems and infrastructure and the unglamorous machinery that keeps the modern world running. From that vantage, the way people talked about artificial intelligence never quite fit what I could see with my own eyes. The conversation treated AI as weightless, a kind of disembodied cleverness floating in the cloud. Yet everywhere I looked, the opposite was happening. Companies were buying power plants. Utilities were straining. The price of a chip was becoming a matter of statecraft. Something enormous and physical was being built, and almost no one was describing it that way.

So I followed the wire. I traced the cost of a single answer down past the software, past the model, past the chip, all the way to the electricity and the heat and the laws of physics that govern them. What I found surprised me, and it became the argument of this book: that artificial intelligence is not weightless software but the industrialization of cognition, an enterprise as physical and as power-hungry as any in history, governed by the same hard rules as every factory ever built.

I also came to believe I was watching a founding decade, the kind that comes along once a generation. Personal computing had one in the mid-1970s, the web had one in the late 1990s, electrification and the railroads had theirs a century before that. The best books about those eras still read true, long after the machines they described became scrap, because they caught the founding logic rather than the gadgets. We are living through the founding decade of manufactured intelligence right now, on plains most people never visit, and its logic is physical. That is what I have tried to write down here, while it is still being poured, in the hope that the account holds up in 2050.

This is a book for the curious reader who senses that the story they are being told about AI is missing something, and who wants the real picture. You will not need a background in physics or finance. You will need only a willingness to look past the magic on the screen to the machinery behind it, and to follow a single thread, the flow of power, from the wall socket to the future it is quietly building.

I have tried, throughout, to be honest. Where the future is uncertain, I say so. Where a number is contested, I flag it. I have no utopia to sell you and no apocalypse to frighten you with, only a way of seeing that I believe will outlast the particular machines I describe, because it rests on the laws of energy and scarcity rather than on the latest release. My hope is modest and, I think, durable: that when you finish these pages, you will never again look at a glowing rectangle answering your question without seeing, behind it, the whole machine.

Follow the wire. It leads somewhere worth understanding.

INTRODUCTION

The Weight of Weightless

The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.

— Mark Weiser, 1991

THE OWNERS OF Three Mile Island announced, in 2024, that they would bring it back from the dead. The reactor whose name is American shorthand for nuclear fear had been shut down years earlier as uneconomic. Now Unit 1 would be refurbished and restarted and run for twenty years, its entire output, roughly eight hundred megawatts of electricity, promised in advance to a single customer: the data centers of one technology company, hungry for power to run artificial intelligence. A plant the country had written off was being resurrected to feed a machine.

Sit with how strange that is. Earth's most advanced, supposedly weightless industry, the one we picture as code drifting through a cloud, turns out to need so much raw electricity that it is reviving the nuclear plants of the 1970s to get it. That single fact is a window into what artificial intelligence really is, and into nearly everything this book is about.

Most of us misunderstand the technology in the same way, and the misunderstanding begins with a small, ordinary miracle. You ask a question, typing a few words into a glowing rectangle, and a second later an answer appears, clean and complete, as if plucked from the air. No meter seemed to run. No fuel seemed to burn. The whole exchange seemed effortless, weightless, free. The truth was anything but. A thousand miles away, real electricity flowed through real wire, heat poured off hot metal, and a machine that cost more than a battleship strained for a fraction of a second to produce your handful of words, then went straight on to do the same for ten thousand other people, all of it drawing power on the scale of a small city. The answer felt like magic. It was manufactured, in a factory, out of energy, and it had a price.

This is the most important and least understood fact about the technology now reshaping the world. In the popular imagination it is software, code, something abstract that floats in a cloud and costs essentially nothing to run. The reality is the most physically demanding thing our species has built in a generation, a vast industrial enterprise of power plants and cooling towers and humming silicon, hungry in a way that software never was.

* * *

Grasp that one fact and a great deal that seems mysterious about this moment snaps into focus.

Out on the dry plains of Texas and along the back roads of the American power belt, software companies are raising buildings the size of stadiums and wiring them to draw power by the gigawatt. They are building their own power plants, signing contracts for electricity measured in decades, and fighting over scraps of land beside electrical substations on empty plains. None of it makes sense if artificial intelligence is weightless software. All of it makes sense the moment you see it for what it is: an industrial revolution, as physical and power-hungry as the ones before it, except that this time the thing being manufactured is not cloth or steel or fuel. This time it is intelligence itself.

This founding decade already has its landmarks, the way the Altair marked personal computing and a handful of linked university machines marked the early internet. Computers had been training artificial intelligence for years, on supercomputers stood up inside ordinary cloud platforms, sharing halls with everything else those clouds did. None of that is the subject of this book. What appeared after conversational AI reached the public late in 2022, and turned compute and power into the constraint that bound everything, was a different kind of building entirely: a dedicated, purpose-built plant whose only job is to manufacture frontier intelligence. Industry coined a name for it. They call it an AI factory. Standalone, often raised on bare ground, often carrying its own power, racing a single coherent machine to a scale and speed nobody had attempted. Those are the first monuments of that new kind of place.

In Memphis, a company raised one such factory, roughly a hundred thousand chips wired into a single coherent brain, and stood it up in about a hundred and twenty days, a pace of construction the industry had thought impossible. Power flowed into it in 2024. Call it the speed landmark. Out in Abilene, Texas, a second factory reached for the next threshold, a campus measured not in megawatts but in gigawatts, the power of a serious nuclear plant promised to one site for one purpose. Announced in 2025, that one is the scale landmark. Speed, then scale. A founding decade shows its shape in landmarks like those two, the first dedicated factories of the era that learned to manufacture thought.

For all of human history, thinking happened in living minds, one at a time, grown slowly and never copied. Humanity has just learned to manufacture it, in bulk, in buildings, out of electricity. Cognition, the thing that made us who we are, has become a thing we make. And like everything else humanity ever learned to manufacture, it obeys the hard laws of the physical world.

Here is the sentence this whole book unpacks. Electrons become tokens become equity. Power flows into a building. Chips turn it into computation. Computation produces tokens, the small units of manufactured thought, and those tokens add up to intelligence, and labor, and value, and at the top of the chain to wealth and national power. Trace the current forward and you understand the new economy. Run it backward, following the fortune down toward its source, and you arrive every time at the same humble place: a wire, carrying power, into a machine.

The head of one of the largest AI labs put the constraint plainly. The shortage of chips was easing, he said, but no one had yet built a data center drawing a full gigawatt, and before anyone did, they would run headlong into the limits of energy itself. The binding constraint, in other words, had already stopped being the cleverness of the software. It had become the power, and the slow physical labor of delivering it.

* * *

Step back and a pattern appears. Every great leap in human prosperity has come from learning to manufacture something once made slowly, by hand, in small amounts. Cloth came first, and the textile mills launched the industrial revolution. Then came manufactured power, with the steam engine, remaking transport and industry together. Chemicals followed, and steel, and information, and each time the newly manufacturable thing grew abundant and cheap, the economy reorganized around it. Manufacturing is the engine of wealth because it takes something scarce and makes it plentiful.

Today we have learned to manufacture cognition, and the pattern says this will be the largest leap of all, because cognition is the most general input there is. Cloth becomes clothes. Steel makes structures. Thinking makes everything, because thinking is what designs the clothes and the structures and the chemicals and the next machine. Cheap, abundant cognition does not improve one industry; it improves every industry at once. This is not the industrializing of a product. It is the industrializing of the thing that makes products. Manufactured at scale, though, cognition runs straight into the hard limits of energy and heat and physics, and the collision between the reach and the limit is the story of this book.

Each of those leaps had a founding decade, a short window when the thing was raw and half-built and its logic was being set down for good. Personal computing found its founding moment in 1975, when a hobbyist machine called the Altair put real computing power on a kitchen table for the first time. Two decades later the internet and the web had theirs, the late 1990s, when companies laid so much fiber-optic cable across the country that most of it sat dark for years, a glut that looked like folly and turned out to be the floor the next economy stood on. Social media found its own founding window beginning around 2004. We are living through another one right now, and almost no one has named it. The founding decade of manufactured intelligence is happening in this moment, on these plains.

Eras that mattered were always understood late. People living through the founding decade tend to argue about the gadgets, the particular machine on the desk, the company winning this quarter, and they miss the deeper thing taking shape underneath. Gadgets date fast. Logic does not. A textile mill is a museum piece now; the idea it proved, that you grow rich by making a scarce thing abundant, runs the world still. My aim is to capture the logic of the present founding era before the gadgets bury it, to write down what the AI buildout actually is while it is still being poured, in the belief that the physical account will read true long after the particular chips and companies have been forgotten.

* * *

Nearly every book about artificial intelligence asks whether the machines will become conscious, or take our jobs, or turn against us. Those are real questions. They are not this book's questions. This book asks something more concrete and, I will argue, more decisive. What does it actually cost to think with a machine? Where does that cost come from, and how low can it go? When millions of these machines demand power at once, where does the power come from? Who controls the scarce things the whole enterprise depends on? And when intelligence itself becomes cheap, what is left that is scarce, and valuable, and human?

Each has a hard, physical, sometimes mathematical answer, and those answers predict the shape of the coming decades better than any speculation about robot minds. Cognition's cost has a floor set by thermodynamics, a real number discovered in 1961 and confirmed in a laboratory in 2012. Supply runs into bottlenecks set by the slow pace of building grids and generators. Value flows, with the reliability of water running downhill, toward whoever owns the scarcest physical link in the chain. None of this is opinion. It is closer to physics.

Two discoveries anchor the book, and both come from physics rather than technology, which is exactly why they will outlast the technology. One is a floor. In 1961 a physicist named Rolf Landauer worked out that erasing a single bit of information must release a tiny, specific quantity of heat, a minimum no machine however clever can avoid. The cost of manufactured thought can therefore fall a long way and never reach zero, because thermodynamics forbids it. Any dream of intelligence too cheap to meter dies at that floor. Its twin is a ceiling, the same physics seen from the other side. Every bit of thinking turns into heat, and shedding that heat fast enough runs into limits of its own. Machine intelligence is boxed in, pressed from below by a floor it cannot cross and from above by a ceiling it cannot lift, both walls poured from the thermodynamics that governed every engine humanity ever built. Models will come and go. The floor and the ceiling will not.

* * *

A word about who I am to tell you this. I came to the subject through its physical end, the power and the infrastructure, rather than through its software, and that turns out to be the useful end to start from, because it is the end almost everyone ignores and the end where the decisive action is. The chapters ahead walk the chain in order, each turning over one link: the physical machine, the bill it runs up, the floor beneath that bill, the bottleneck of power, where the power comes from, the chip, the labor it manufactures, the factory it lives in, the heat it sheds, the nations fighting to control it, the money that finances it, the value it captures, and finally the human it cannot replace.

Each chapter opens at a real place and follows a single thread, because abstractions about artificial intelligence are cheap and a concrete building on a real plain is not. We will stand at the fence line of an actual data center. We will sit at the desk where a physicist worked out the floor beneath the cost of thought. We will watch heat pour off a rack of chips and water circle to carry it away. One instruction organizes the journey: follow the wire, the actual flow of power, back to ground, because the ground is where the truth of this technology lives.

Honesty is the other commitment. This book will not sell you a utopia or a doom. The transformation is real, the opportunity enormous, and the upheaval genuine, and it will hurt real people. Where the future is uncertain, I will say so. Where a number is contested, I will flag it, and where the optimists and pessimists are both partly right, I will hold both truths at once, because the truth of this moment lives in the spread between them. A technology that runs on power plants and cooling towers does not float free into heaven or hell. It grinds forward through grids and supply chains and balance sheets, and its real consequences are being settled there, right now.

* * *

Why should this matter to you, whoever you are? Because the consequences will not stay inside the technology industry. If intelligence can be manufactured, then much of the cognitive labor that fills the modern economy, the analyzing and writing and deciding many of us are paid for, can be manufactured too. Your job may not vanish, but it will change, because the routine-thinking part grows cheaper while the part that needs judgment and care grows more valuable. The money financing this buildout is no longer only venture capital; increasingly it is pension funds and insurers, which means the retirement savings of ordinary people are now wagered on these machines. And your country is now competing for a strategic capability as vital as energy or defense, whether its citizens have noticed or not.

By the end, a set of things that look unrelated today will resolve into one shape. The frantic building of power plants by software companies. The export controls on a particular kind of chip. The hunger of staid pension funds for data centers. The revival of dead reactors. The fights over water in dry towns that wake to find a giant computing site next door. Each is reported now as its own curiosity. Each is one phenomenon seen from a different angle: the industrialization of cognition colliding with the hard limits of the physical world. Hand a person the right forecast and you help them for a quarter. Hand them the right way of seeing and you help them for a decade. This book attempts the second kind of gift.

I have written it to be read in 2050 as much as today. Chips in these pages will be antiques by then, the companies merged or gone, the megawatt figures quaint. What lies underneath, power and heat and the thermodynamic floor, will not have moved an inch, because it rests on laws older than any industry. Such a record is what this book is trying to leave of the decade when humanity first built the factories that make intelligence, the foundation stage of a technology that will help build the century to come.

* * *

So let us begin where the whole thing begins, not with the answer on your screen but with the building that made it. Out on the flat country, past the last of the suburbs, the structures are rising. Plain, enormous, windowless, ringed by fences and fed by fat transmission lines that march in from the horizon. Most people will never see one. Yet inside those walls the shape of the economy, of national power, and perhaps of human work is being manufactured, token by token, watt by watt, around the clock.

We are going to walk into that building, follow the wire all the way through it and out the far end, the substation at one end and the human at the other, and find at every step that the weightless miracle on your screen rests on the heaviest foundation imaginable. Carry that one idea with you as we go. Test it against the evidence in each chapter, and watch how much of the baffling present it explains. By the end the machine behind the magic will be impossible to unsee, and the glowing rectangle will never look quite the same again.

Follow the wire.

CHAPTER 1

The Machine You Cannot See

SOMEWHERE RIGHT NOW, a person types a question into a glowing rectangle and waits a beat for the answer. Perhaps they want a recipe. Perhaps a line of code, or a kind way to tell a friend a hard truth. Two seconds pass. Words appear, fully formed, and the person reads them and gets on with the day, never once wondering where they came from. Why would they wonder? Nothing happened. A few words landed on a screen, and it felt like nothing at all.

Something did happen. Producing those few words set off a small industrial event, though that event took place a thousand miles away, out of sight, inside a building most people will never see and could not find on a map.

Let me take you to one of those buildings.

Three hours west of Dallas, past the last of the suburbs and out onto the flat, wind-scoured plains, sits a town called Abilene. Until very recently, almost nobody in the technology business had any reason to say its name. Roughly a hundred and twenty-five thousand people live there, in the kind of town built around churches, a small university, and a long memory of cattle and oil. Then, in the summer of 2024, surveyors appeared on the edge of town with their tripods and sight-lines. Bulldozers followed. What they were staking out was the flagship of one of the largest computing projects ever attempted, built on the empty plains by a consortium of technology companies that had pooled their money to pour the concrete.

What rose from that ground is hard to describe without sounding like you are exaggerating. Picture a cluster of enormous buildings, plain and windowless, spread across a campus so vast that the parking lot alone swallows thousands of cars. On a working day that lot fills completely, because thousands of people report to the site, more than would fit inside some of the towns nearby. Reporters who visited put the construction workforce somewhere between five and seven thousand at its peak. Electricians. Pipefitters. Welders. Crane operators. An entire migratory city of skilled hands, drawn to West Texas to build a thing with no smokestack and no physical product you could ever hold.

Anchoring the whole campus is a fenced acre of steel and humming transformers: a substation, rated to move close to a gigawatt of power. A gigawatt is, very roughly, what it takes to run a mid-sized city. No homes draw from this one. Every watt it carries is bound for a single purpose, which is to feed the machines inside those buildings, machines that exist to do the very thing that felt like nothing when you typed your question into the glass. The first hall came online in 2025, its racks of the most advanced AI chips humming inside a rented cloud. The plan on paper does not stop at one building. It calls for a row of them, the campus drawing power measured not in megawatts but in gigawatts, several cities' worth, all of it spent on thought.

Stand at the edge of that campus and you are looking at the true shape of artificial intelligence. Forget the brain in a jar. Forget the elegant code drifting through a weightless cloud. Before you sits a power plant, more or less, an industrial site that takes in electricity by the gigawatt and converts it, through hundreds of thousands of specialized chips, into sentences and images and decisions. People who build these places have a word for what comes off the far end of the line. They call it intelligence. Without a flicker of irony, they will tell you they are manufacturing it.

Abilene tells you how big these things get. A second site, six hundred miles east, tells you how fast.

Cross into Tennessee and follow the Mississippi down to Memphis, a city that knows freight and blues and barbecue and had never once thought of itself as a capital of computing. On a patch of former factory ground, in the summer of 2024, a crew began raising a building around a single idea: pack a hundred thousand of the most advanced chips into one hall and wire them together so tightly that they behave as one machine, a single mind learning at once rather than a warehouse of separate computers minding their own business. The thing they built is called Colossus. The number that stops you is not the chip count, though. The calendar is what does. From bare slab to a working machine of that size took something near four months.

Sit with that for a moment. Heavy industrial plants are usually measured in years, sometimes the better part of a decade, from groundbreaking to first output. The crews in Memphis got a hundred thousand chips humming in concert in roughly a hundred and twenty days. To keep them fed before the utility could deliver, the builders rolled in their own gas turbines and burned fuel on site, which drew the anger of neighbors and regulators and a fight that is still going. Hold the controversy for the endnotes. Keep the image. A machine the size of a small power station's worth of computers, stood up in a single season, on ground that grew cotton within living memory.

Be careful how you say what Colossus is, because the precise claim is stronger than the loose one. It was not the first computer to train artificial intelligence. Big technology companies had lashed tens of thousands of processors together for years to run recommendation engines and vision systems, and the frontier models everyone now argues about had themselves been trained on supercomputers stood up inside ordinary cloud platforms, sharing halls with everything else those clouds did. Colossus is something the world had not seen before that: a dedicated, standalone plant built for one purpose only, to manufacture frontier intelligence and nothing else. That is the category that opened once conversational AI reached the public late in 2022 and the race began. The industry gave it a name. An AI factory. And within that new category, two things set this one apart. First, the speed, that violent four-month sprint from dirt to a working machine. Second, the coherence, a hundred thousand chips knit onto one network fabric and driven as a single training engine, near the outer edge of how large a single coherent machine had ever been built and run. Speed and coherent scale, then, and a first of a real kind.

Set the two sites side by side and the era shows its face. Memphis raised one enormous coherent machine at a record clip. Abilene, the flagship of a vast infrastructure program announced by an AI lab and its backers, reaches for a scale of raw power, gigawatts of it planned across the campus, that no computing site has ever drawn. One is a sprint, the other a moon shot, and they are being run at the same time, by rivals, on the same continent. Factories of thought are going up faster and larger than almost anything in the history of industry, and the people racing to build them know exactly how strange that is.

These two are no freak pairing. Drive the back roads of the American power belt and the same scene repeats itself in dozens of places at once. Across central Ohio, in the Arizona desert, along the river valleys of the Pacific Northwest, the pattern recurs: a quiet stretch of cheap land, a fat transmission line, then convoys of concrete trucks. Plot where these machines are landing and the map looks less like a map of the internet and more like a map of the old industrial heartland, because the logic underneath is the same. Heavy industry settles where the power and the water and the right-of-way already are.

Get close to one of these sites and your senses correct the abstraction. You hear it before you see it, a low oceanic hum from the cooling plants that never stops, by day or by night. Warmth rolls off the exhaust, the breath of a machine that runs a fever as a condition of doing its job. Trucks idle at the gate in a line that disappears into the dust. Nothing about it resembles a tech campus with its glass and its bean bags. Everything about it resembles a refinery, or a steel mill, or a working port, because that is the family of thing it has joined.

* * *

Through most of the history of computing, none of this would have made any sense. Software's whole point, the wellspring of its strange power and its astonishing fortunes, was that it carried almost no physical cost at all.

Consider what set the last fifty years apart from everything before them. Any carmaker who wants to sell a second car must build a second car: buy the steel again, run the line again, pay the labor again. By contrast, a software company that wants a second customer simply makes a copy, and a copy costs essentially nothing. That single fact, build once and sell forever for almost free, is the quiet engine humming beneath nearly every technology fortune you can name. Thanks to it, a company of a few thousand people could serve half the planet and keep almost every dollar of the final sale as profit. Software floated free of the physical world. It tipped no scale.

Artificial intelligence is the first information technology to break that spell.

Watch closely when a modern system answers your question. No drawer opens; no stored response gets pulled out and dusted off. Instead the answer is computed, fresh, from scratch, every single time, and computing it means running a staggering number of calculations across those specialized chips. Real electricity feeds each calculation. Each one sheds real heat. Ask the same question a thousand times over and the thousandth answer costs about what the first one did. No cheap copy waits in the wings. There is only the machine, doing the work again, and the meter spinning again as it does.

Here is the hinge on which the whole book turns, so let me say it plainly. Software paid its cost once, up front; afterward it ran for free. Artificial intelligence pays on every use, forever, in a currency older than money: energy. On a screen the two look identical. The electricity bill knows they are nothing alike.

Once intelligence carries a real, physical, per-use cost, it stops behaving like software and starts behaving like something humanity understands far better and has fought over far longer. The new behavior is that of an industrial product, made in a factory, out of raw materials, by burning power. People in Abilene grasped this in their bones. Why else would they put their machine out on the empty plains, far from the engineers, far from the customers, far from every place technology usually clusters? They built beside the power. Because in this new economy, you do not site the factory where the people are. You site it where the electricity is.

* * *

Follow one request through these machines, not as you experience it but as the physical world experiences it, and the difference turns concrete.

You press enter. To the world of bits, that keystroke is no bigger an event than loading a web page. What happens next is not retrieval, though. It is a computation, and a large one. Your request travels to a data center and lands on a specialized chip, where a model holding billions of parameters bends its full attention to your handful of words. Producing the reply demands an immense number of arithmetic operations, on the order of a trillion or more for every small fragment of output, and a single answer may run to hundreds of fragments.

Each operation nudges electrons across silicon. Each electron nudged sheds a little energy as heat. Running flat out, the chip pulls hundreds of watts and turns nearly all of it into warmth that must be carried off at once, or else the silicon throttles and then dies. So a second machine wakes up. Cooling kicks in, burning its own electricity to drag that heat away. All of it happens inside a building financed and constructed for exactly this purpose, fed by a grid connection that may have taken years to win, on land whose worth now rests almost entirely on whether that connection exists at all.

A sentence appears on your screen. Down in the physical economy, a small industrial process has just completed a full cycle: energy in, heat out, a real cost paid. And, this is what matters most, the cost gets paid again, in full, the moment anyone asks anything else.

* * *

Because the marginal cost of AI is real, its financial signature is not software's. It is a factory's.

Inside a factory, growth cannot carry you to a ninety-five percent margin, because every additional unit eats additional materials, energy, and machine-time. Margin lives in the spread between what a unit sells for and what its inputs cost, and that spread is contested without mercy, since everyone in the trade answers to the same physics. Factories rise and fall on unit economics: the cost of one more, the price you can fetch for it, the efficiency of the conversion, the share of the day your expensive equipment actually runs. Nobody on a factory floor asks how fast we can grow. They ask what the next unit costs, and whether that cost can be driven down faster than the price falls.

Those are the questions AI companies now have to learn, often the painful way. Take a consumer product loved by tens of millions of users. By software's scoreboard, that is a clean triumph. By a factory's, it can be a slow catastrophe, because each of those adored users, on every interaction, draws down a physical resource somebody has to pay for. More love means more use, and more use means more cost to serve them. Freemium, that old masterstroke of giving the product to most and charging a few, made sense back when the free users were genuinely free. Run the same play today, when every free user fires up the factory, and you have engineered a machine for turning investor capital into electricity and heat.

None of this is a forecast. You can already read it in the gap between AI revenues and the cost of the compute behind them. Read it, too, in the usage caps and tiered access that have crept into products launched as unlimited. Watch it cross the face of every product leader who learns that a power user has discovered the agent will happily run all night. Markets have not fully repriced for any of it, which is exactly the opening for whoever sees it early. Software's reflex shouts: maximize engagement. Factory discipline answers: maximize the spread between the value of the work and the cost of the joules. Those are not the same instruction. Now and then they are opposite instructions.

Picture the new unit of work concretely. Picture a developer handing an autonomous coding agent a real job: build a small feature, write the tests, fix whatever breaks. Off it goes. It reads the codebase, sketches a plan, writes some code, runs the tests, watches three of them fail, reasons about why, rewrites, runs them again. Hours slide by. The developer sleeps. Sleep is not in the agent's repertoire. By morning the feature is done, and somewhere a meter has logged the entire campaign, not a single question and a single answer but thousands of them, a long chain of reasoning that burned compute all night long.

That agent shows you the shape of the work that is coming, and the reason the old intuitions fail. One chatbot reply is a sip. An agent toiling until dawn is a flood. As models stop merely answering and start acting, doing things, finishing tasks, their appetite for computation leaps by an order of magnitude per task, and the cost leaps with it. The other side of the ledger pulls in the opposite direction, though: a finished feature is worth far more than a clever sentence. Customers who would never pay much for a better autocomplete will pay handsomely for work that actually ships. The bill climbs, but so does the price people will settle it at.

* * *

An honest objection sits in the middle of all this. Surely the cost is falling, and fast. Won't AI's marginal cost slide to zero as well, and won't the whole thing just become software after all?

This is the most important objection in the book, and it earns a careful answer, because it is half right. The half that is wrong is precisely where fortunes will be made and lost.

Begin with the half that is right. Each unit of AI output really is getting cheaper, and quickly. Better chips, better models, better software, and the brute force of accumulated experience are driving the price of intelligence down a curve that genuinely rivals the great learning curves of industrial history. Bet that AI economics will sit frozen where they are today and you repeat the blunder of someone who, in 1975, assumed computing would stay expensive forever. Steep, real, and worth taking seriously, all of it. This book does exactly that.

Now the half that is wrong. Falling is not the same as zero, and the gap between them is not a quibble. It is the entire game. Software's marginal cost is not merely low; it is structurally, essentially zero, because copying a pattern asks almost nothing of the physical world. The marginal cost of AI is falling toward a floor that is not zero, because computation, unlike copying, must answer to thermodynamics. Performing an irreversible calculation requires some minimum quantity of energy, fixed by physics, and no amount of cleverness erases it. Intelligence can grow vastly cheaper, with enormous distance still to fall, as the coming chapters will show. Yet it descends toward a hard physical bottom, not toward the frictionless zero that made software what it was.

That distinction reads as academic right up until you notice what it decides: where durable value lives. In a true zero-marginal-cost business, physical inputs are an afterthought, and value flows to the pattern, to the software, the network, the brand. In a falling-but-floored business, physical inputs never stop mattering, which keeps them ownable, financeable, defensible. Energy stays in the cost structure. The chip stays in the cost structure. So do the building, the cooling, the land, the grid connection, present always, zeroed out never. Knowing that AI's marginal cost is real and floored, rather than imaginary and gone, is what turns the infrastructure beneath it into a permanent asset instead of a passing expense. Knowing as much is the reason this book exists.

* * *

None of this is the first time the world has watched a weightless thing acquire weight.

History keeps a version of the story. Before the steam engine, cloth was made by hand in cottages, a little here and a little there, wherever a skilled weaver happened to live. Power changed that. Once a mill could be turned by a water wheel and then driven by a furnace, production stopped following the weavers and began following the energy. Towns rose beside the rivers and the coalfields. Fortunes gathered around whoever owned the wheel, the seam, the engine. A whole society rearranged itself around one new physical fact, namely that making things now required power, and power lived only in particular places.

Software spent fifty years as the great exception to that rule, the lone industry whose output weighed nothing and could be made anywhere. Artificial intelligence ends the exception. Cognition, as it happens, is a physical product after all, and the instant that became true the old logic came roaring back. Production is migrating toward the energy again. Towns nobody had heard of are rising beside the substations again. Wealth is gathering around whoever owns the wheel again, except the wheel is now a gigawatt of firm power and an acre of silicon.

Notice what that migration does to a map. For a generation, the geography of technology was a geography of talent: you went where the engineers were, which meant a handful of expensive cities. The geography of intelligence is becoming a geography of power: you go where the electrons run cheap and abundant, which means West Texas, rural Ohio, the desert, the river valley. Capital that used to chase clusters of clever people now chases interconnection queues and transmission capacity. Calling that a small adjustment misses the scale of it. A continent's worth of investment is quietly changing the first question it asks, from who is here to what can we plug in here.

Every layer of the old technology economy will feel the shift. Investors who learned to value companies on the near-zero marginal cost of software will have to relearn the craft for businesses whose marginal cost is a meter reading. Engineers who treated electricity as a footnote on the bill will find it seated at the head of the table. Executives who could once expand into any market by spinning up servers are discovering that some markets are simply out of power, and that you cannot ship intelligence to a place the grid cannot reach. None of these is a distant problem. They are this year's problems, met in real time by people who never expected physics to send them an invoice.

Weigh the strangest part of all. Companies at the center of this are, by the measure of their own ambition, barely getting started. Whatever computation the world devotes to artificial intelligence today is a rounding error against what its builders intend to deploy. Should the appetite they describe prove anything close to real, the Abilenes are not the end of the story. They are its first chapter, the equivalent of the first mill on the first river, raised before anyone fully grasped what had been set in motion.

And yet the person at the glowing rectangle feels none of it. That invisibility is no harmless quirk; it is the central confusion this book means to clear. Because the apparatus hides a thousand miles away, we keep talking about artificial intelligence as though it were weightless, a matter of algorithms and clever prompts, when its defining economic fact runs the other way entirely. Decisions ride on the weightless assumption every day. Companies price their products as if serving them were free. Investors model businesses as if the marginal unit cost nothing. Governments draft policy as if the only scarce inputs were data and talent. Each of those assumptions inherits the old software intuition, and each is quietly wrong, because the binding constraint has slid from the world of bits into the world of atoms while almost nobody updated the map.

Update it and the real questions sharpen at once. Where will the power come from? Who controls it? How fast can it be built, at what price, and with whose permission? What becomes of a company, or a city, or a country that cannot get enough of it? Software never taught anyone to ask such things. An industrial economy asks little else, and these questions are about to dominate the strategy of the most valuable companies on earth.

* * *

Accept the argument of this chapter and a chain of strategic conclusions follows, reframing almost everything. They are the seeds of the chapters ahead, so let me plant them now.

Know your unit economics, or you do not own a business; you own a burn rate. Under zero marginal cost, you could put off the per-unit question almost forever, since the answer rounded to nothing. Under a factory's physics, the per-unit question is the business. What does one finished task cost you, in joules, in chip-time, in the wasted churn of an agent that retries itself five times before it lands? What spread remains against what you charge? A company unable to answer is not underrating the question. It is flying blind.

Growth stops being automatically good. When each new user consumes a physical resource, more users without more margin is not traction; it is a bigger pump draining the same tank. The software habit of chasing engagement at any cost has to yield to a factory's judgment about which usage even deserves to be served. Growth helps only when the unit it adds turns a profit, and in AI that profit is a genuine question, never a given.

Cost becomes a technology in its own right. Within software, the road to fatter margins ran mostly through selling more. Within a factory, that road runs through the conversion itself: better chips, better models, higher utilization, less waste, cheaper power. Winners will treat the efficiency of turning energy into intelligence as a core craft, not a line item tossed to procurement. Cost of goods has become a research frontier.

Reliability becomes a weapon. A factory that stalls earns nothing while its capital sits idle, so uptime, scheduling, and the unglamorous art of keeping the machines fed turn into real competitive advantages. Procurement turns strategic overnight.

Whoever locks in cheap, firm power on a long contract has bought a structural edge no amount of software cleverness can copy. Speed of construction counts for just as much. Build a gigawatt of capacity a year ahead of your rival and you hold a year of output they simply cannot match, sold into a market starved for the very thing you make.

Whoever owns the scarce physical input holds power over everyone stacked above them. Where the marginal product demands energy, chips, cooling, and space, the owners of those scarce inputs occupy a seat software never had to reckon with, the seat once held by whoever owned the mill's raw materials and machines. This is why a chipmaker can be worth more than the companies buying its chips. And increasingly, both the constraint and the value are migrating down the stack, toward the most physical input of them all.

Notice one more thing about a machine built from physical parts: it can move. Weightless services have no reason to sit anywhere in particular, but a machine that eats power and breathes cold air must live where those things are cheap and plentiful, and when they run short in one place it goes hunting for them in another. The factory of thought has no fixed address; it chases the conditions that feed it.

Watch where the chase leads and you see the machine changing shape. For a few years the instinct was to build one enormous hall, a single colossal campus, and pour everything into it. That instinct is already bending. Power does not pool conveniently in one spot, and a grid asked to deliver a town's worth of electricity to a single address pushes back hard, so builders have begun to spread the work out. Smaller sites, wired together. Computation shoved closer to where the data is born and where the people are. Pieces that once had to sit shoulder to shoulder now cooperate across distance, stitched by fiber and choreographed by software.

Why should the shape matter to anyone but an engineer? Because the drift toward many smaller, interconnected sites is no passing fashion in data-center design. What you are watching is the machine bend around its one real constraint, the way a river bends around the lay of the land.

Notice, too, that the machine is splitting in two as it spreads, along the line between its two kinds of work. Training a new mind, the months-long ordeal of building intelligence from scratch, wants everything gathered in one colossal place, a single mega-campus where tens of thousands of chips can labor in concert. Using that mind afterward, answering the everyday questions, is a lighter task that can run almost anywhere, and increasingly it does, in smaller sites pushed close to the people and the data they serve, even in modest boxes at the edge of the network. One kind of work pulls toward the giant and the central; the other pushes toward the small and the scattered. So the geography of intelligence is not one shape but two, a handful of vast cathedrals where minds are built and a spreading web of little chapels where they are put to use.

Biology, as usual, ran the experiment first. Your brain is not one dense block of identical tissue roaring at full tilt. Think of it instead as a federation of specialized regions, each lit only when its particular work is needed, most of it quiet most of the time, the whole knitted together by a web of connections far richer than the parts themselves. Distributed, modular, sparing of energy by design. Our machines are groping, clumsily and a few decades in, toward an arrangement evolution settled long ago, and for the very reason evolution did: when energy is the binding cost, you spread the work out and you wake only the piece you need.

Software was a miracle of weightlessness, value set loose from the physical world. Artificial intelligence drags that value back to earth and chains it, joule by joule, to the infrastructure that makes it. Far from a step backward, that is the opening of a new industrial economy, ruled by laws we already half understand from two centuries of making physical things.

Building on those laws means following the cost down to its source. And the source, it turns out, is neither the chip nor the model. It is the electricity. People in Abilene learned that the hard way, and what they ran short of next was not chips, or money, or land.

It was power itself.

And a physical thing can be counted. Before anyone can fight over the electricity, somebody has to learn to read the meter on it, one unit at a time, the way every industry before this one learned to price the thing it actually sold.

YOU'VE READ THE OPENING.

Get the whole book.

The Machine That Ate Power

Scan to buy · or visit

REPLACE-DOMAIN

LAUNCH OFFER

20% off the ebook with code LAUNCH20

Valid through the launch window (see checkout for the exact date).

Ebook $9.99 · Signed paperback $24.99 · Collector hardcover $32+