AI Neofeudalism and Intelligence Sovereignty: The Need for and Design of Decentralized Training, Computation, and Inference Infrastructure

Taewoo Park
Department of Physics, KAIST

This essay calls the concentration of models, compute, work memory, and decision rules in a small operating class "AI neofeudalism." The concentration unfolds as artificial intelligence becomes essential social infrastructure. The present AI industry is attempting to close the gap between enormous infrastructure investment and limited real demand through value based pricing and ecosystem dependence. Open weights do not provide an adequate exit either, because individuals lack the physical resources required to run them.

Within this structure, individuals use the value produced by AI while renting the means of production. As they accumulate their context and capabilities inside a particular system, they gradually lose the ability to leave. This essay compares that condition with the structure of company towns, where employment, housing, and consumption were governed together, and argues that modern dependence is completed not by physical coercion but by convenience, switching costs, and cognitive lock in.

As an alternative, this essay proposes decentralized training, computation, and inference infrastructure that joins heterogeneous devices owned by individuals and communities into shared means of production. Personal Resource Cells, replaceable job schedulers, Job Capsules that restrict purpose and authority, local data preservation, encrypted intermediate results, verifiable computation records, and contribution based settlement distribute the means of production, information, decision authority, and exit across different actors. Compute contribution is separated from citizenship in the network so that concentrated resources cannot become concentrated governance. DRIFT, the technical demonstration, connects the Apple MPS and NVIDIA CUDA backends in PyTorch into a single inference path. It is an initial implementation in which heterogeneous personal devices jointly process one AI request. Ultimately, decentralizing AI is not simply a question of publishing model files. It is a question of intelligence sovereignty, which returns to individuals and communities the ability to produce intelligence, verify it, and leave.

1 The Current Flow of AI Capital Cannot Sustain Itself Indefinitely

In 2024, Sequoia's David Cahn estimated that the gap between hyperscaler spending on AI infrastructure and the revenue actually earned by the AI ecosystem was already about $600 billion. Applying the same method to NVIDIA's 2026 data center revenue run rate of approximately $300.8 billion per year raises the threshold to about $1.2 trillion, while AI revenue attributable to actual final consumer spending is estimated at only $50 billion to $150 billion. Allianz Research calculated the divergence between AI capital expenditure (capex) and revenue growth at about 46%, already above the 32% recorded during the telecommunications overinvestment cycle of 2001.[1]

The problem is that the mechanism used to close this gap rests on a deeply circular loop. Rising valuations justify greater capital expenditure. That expenditure is then reinterpreted as evidence of explosive future demand, which supports valuations again. This contradictory cycle is driving the market. According to Moody's, hyperscalers have already signed about $662 billion in data center lease commitments. Since those leases have not yet commenced, they do not appear on balance sheets. These off balance sheet commitments equal about 113% of the same firms' combined adjusted debt.[2]

As a result, the free cash flow of the five largest technology companies turned negative for the first time in thirty five years. Capital expenditure intensity rose to 34% of revenue, more than twice the 15% peak of the 1990s internet boom. With NVIDIA even investing back into some of its own largest customers, reported revenue must be examined by distinguishing independent final demand from recycled capital. The same pressure helps explain why Samsung, SK Hynix, and Micron are discussing new fabrication capacity cautiously even though HBM prices have risen to nearly half the component cost of a Blackwell chip. AI capital expenditure forecasts have entered a stage in which the scale of inputs is no longer enough. Their sustainability must be demonstrated through demand and cash flow.[3]

By 2026, these conditions had become an official risk. The Bank of England warned in October 2025 that the earnings yield implied by the cyclically adjusted price earnings ratio (CAPE) was near a twenty five year low, while valuations were comparable to the peak of the dot com bubble. In June 2026, the Bank for International Settlements, often called the central bank of central banks, listed the sustainability of AI investment alongside persistent inflation, financial fragility, and fiscal deterioration as four major pressure points in its Annual Economic Report. It warned that disappointing returns could trigger an abrupt withdrawal of funding and turn the capital expenditure boom into a prolonged investment slump.[4] Demand side measurements are also negative. MIT Project NANDA reported that about 95% of the organizations it studied had obtained no measurable financial return from generative AI investment, and that only about 5% of custom enterprise AI tools had reached production.[5]

1.1 Companies Are Turning Retail Investors into a Channel for Late Stage Liquidity

As funding dries up, late stage capital and insiders look for an exit. The wave of enormous initial public offerings in 2026 is the result, and SpaceX is the textbook case. SpaceX went public at a valuation of about $1.77 trillion after reporting 2025 revenue of $18.674 billion, an operating loss of $2.589 billion, and a net loss of $4.937 billion. The offering itself was unusual. Instead of presenting a price range and letting demand determine the result, the company fixed a single "take it or leave it" price of $135, effectively bypassing price discovery.[6]

The retail allocation was more decisive. Retail investors ordinarily receive 5% to 10% of an initial public offering. SpaceX targeted as much as 30%, worth roughly $22.5 billion. Retail orders exceeded $100 billion, but the actual allocation remained at only about 20%. Most demand went unfilled, and many individuals requested hundreds or thousands of shares only to receive one or two.[6]

An initial public offering is a liquidity event for insiders and early investors. The expiration of lockups increases supply and transfers downside risk to retail investors who entered near the peak. This structure operates regardless of the intentions of any particular company. What is interesting is

why Musk expanded the retail allocation so unusually. Analysts argue that he sought the same effect seen at Tesla, where shareholders also become product customers. Individuals who own shares are more likely to subscribe to Starlink or use other products, which improves corporate cash flow. The structure combines the roles of investor and customer and financially reinforces commitment to the corporate ecosystem.

Individuals assume three roles at once: investor, consumer, and buffer that absorbs losses when valuations are corrected. When retail capital at this scale is tied to overvalued assets, losses during a correction suppress consumption through the wealth effect. That contraction returns as renewed pressure to revive corporate investment.

1.2 There Is No Individual Customer at the End of the Game of Chicken: Value Based Pricing and the Mirage of Open Weights

The industry's response to the pressure to revive investment is clear. It must end the game of chicken created by consumer facing loss making competition and generate revenue from the infrastructure cost base. This transition is not a forecast for 2030. It is already under way in 2026. OpenAI CFO Sarah Friar is pursuing it publicly. She proposes moving from cost plus pricing, where a margin is added to cost, toward value based pricing, where scarcity and customer value determine the price.

Her argument is that when compute is scarce, a provider can recover a larger share of the economic value the model creates for the customer. She further presented outcome based pricing, licensing, and intellectual property based contracts as explicit future models. Applied to drug discovery, this would allow the supplier to claim part of the economic value created by shortening the discovery period.[7] OpenAI is not alone in this transition.

In April 2026, Anthropic changed Claude Enterprise from a fixed price to dynamic usage based billing. Some analyses estimate that the change could raise costs for heavy users to two to three times their previous level. A base model plus metered billing based on the economic value of enterprise output is not a future business model but a present business model that frontier providers have already adopted.[8]

Unfortunately, this transition is effectively irreversible. One cause is the spread of agent workflows. Agent workflows have increased token use per session to ten to one hundred times the level of a simple conversational query. This breaks the simple assumption that adopting one company's model directly produces that company's revenue. Token prices continue to fall, but token consumption per session explodes even as prices fall, so unit economics naturally deteriorate under fixed pricing. From the provider's perspective, charging on the basis of output value can only become a compelling option.

Even apart from this underlying cause, when the valuations of frontier model companies are combined with large pools of retail capital through initial public offerings, pressure for real revenue becomes stronger. That pressure points toward high margin enterprise and value based pricing rather than a low cost consumer strategy.

Arguments of this kind usually end by presenting open weights as the only escape. Since open models such as DeepSeek, Qwen, Llama, and Mistral exist, the reasoning goes, individuals need not depend on centralized application programming interfaces. Yet this answer recedes as models

grow larger, because the hardware required to run even an open weights model is effectively beyond individual ownership.

DeepSeek is in fact distributed under the MIT License. The problem is the resources required to host it. The V4 family demands memory and accelerator resources beyond what a single consumer device can readily provide.[1] [9] The model is free. The hardware is not. This scale plainly lies beyond the range of personal computing resources. Cloud access cannot provide a final answer either.

Ollama Cloud, for example, has three plans: Free, Pro at $20 per month, and Max at $100 per month. Usage is measured not by token count but by graphics processor time, which varies with model size and request duration. Free is intended for light use, while Pro and Max support larger and more sustained workloads. The heaviest tier four model, deepseek-v4-pro, consumes more of the allowance over the same period. Pro and Max users can purchase a separate extra usage balance after reaching the default limit. Openness of the model file does not eliminate the marketization of computing costs.[10]

One break even analysis estimated that self hosting becomes preferable to the aggressively inexpensive DeepSeek-V4-Flash only after roughly 800 million tokens per month.[2] This greatly exceeds the usage of individuals and many small teams. In conclusion, under real constraints, the illusory label of open weights once again consigns individual customers to an equivalent payment system.[10]

2 Every Phenomenon Points toward the Strengthening of Dependence

The strange fact is that although all this information is public, we accept dependence. At the enterprise level, the condition is already real. In Zapier's 2026 enterprise survey, 81% of leaders were concerned about dependence on AI vendors, but only 6% said they could switch their primary provider without major disruption. Forty seven percent said that an outage at the primary vendor would stop core business functions. Deloitte's 2026 report stated that only one organization in five had mature governance for autonomous agents. Most organizations are accumulating dependence without a strategy for governing it.[11]

More important is the character of that dependence. It is not only functional but intrinsic. Once a team designs its workflow around a particular model's reasoning pattern, tone, and format, migration to another model requires the redesign of work procedures and evaluation criteria as well. The model itself becomes a commodity, while lock in compounds across the ecosystem around it, including partner networks and orchestration systems. When model, orchestration, data, and knowledge layers

[1] According to the official model cards, DeepSeek-V4-Flash activates 13B parameters per token out of 284B total parameters, while DeepSeek-V4-Pro activates 49B parameters per token out of 1.6T total parameters. The released weights use mixed precision, with FP4 for MoE experts and primarily FP8 for the remaining parameters. Actual accelerator memory requirements vary with the runtime, quantization, key value cache, context length, central processor offload, and tensor or pipeline parallel configuration.

[2] This estimate follows the comparison conditions published by CheckThat.ai in May 2026. Its table prices DeepSeek-V4-Flash output at $0.28 per million tokens. For self hosting, it assumes an RTX 5090 at 50% utilization running Llama 3.1 8B at 480 tokens per second with vLLM or 185 tokens per second with Ollama. A separate total cost of ownership example for a fifty person team includes thirty six month depreciation of a $7,000 workstation, electricity, cooling, storage, and engineering time. The break even point changes with the model, quantization, input and output ratio, electricity price, equipment cost, utilization, and operating labor.

become entangled at once, switching costs grow by multiplication, not addition. At the frontier, this is already being used as a highly systematic strategy.

Anthropic and Accenture, for example, announced a partnership to train as many as 30,000 Accenture professionals to use Claude. Separately, Anthropic made an initial investment of $100 million in the Claude Partner Network, which supports partner firms as they develop solutions based on Claude. When training and the partner ecosystem expand together, organizational procedures and external supply chains are reorganized around the same model. The cost of moving to another model rises with them.[12]

2.1 The Future the Individual Will Face Ten Years Later, in 2036

Corporate dependence culminates in a business problem, but individual dependence shakes the very capabilities of humanity. Dependence on AI reduces opportunities to exercise thought directly and weakens unaided performance. This is no longer only a metaphor. In an MIT Media Lab study, participants who wrote essays with a large language model showed lower brain connectivity than those who wrote without the tool. They also showed weaker recall and a weaker sense of ownership over their work. Lower engagement was also observed in a crossover session in which participants who had previously used a large language model later wrote without the tool.

In a survey by researchers at Carnegie Mellon University and Microsoft Research, workers reported investing less critical thought when they trusted AI more. A separate randomized learning experiment named the tendency to transfer cognitive burden to AI while reducing internal engagement and self regulation “metacognitive laziness.”[13]

Humanity has already witnessed classical precedents. Air France Flight 447 crashed in 2009, killing 228 people, after icing disengaged the autopilot and the pilots failed to understand the situation. The investigation found that long dependence on automation had allowed their manual flying skills to rust. Even an internal Air France report acknowledged weak fundamentals among some long haul pilots and a “general loss of common sense and general aviation knowledge.” As automation becomes more reliable, human operators have less room to contribute, and so when something goes wrong and automation hands control back, the operator is already out of the loop and cannot respond.

An old joke among pilots imagines a highly automated future cockpit occupied by a pilot and a dog. The dog bites the pilot whenever the pilot touches anything, and the pilot feeds the dog. The role left to the human at the end of dependence is already contained in that sentence. The same direction is visible in AI. Three months after adopting AI assistance, clinicians showed a 6% decline in their ability to detect tumors without it. These cases point to a clear result. As AI becomes more accurate, people have fewer opportunities to exercise and preserve their own capabilities and critical judgment.[14]

A junior developer can generate code from a prompt yet cannot retrace first principles to debug it when the model is wrong. A student can easily produce fluent output yet cannot regulate their own learning. They sit in the position of supervising a black box, but lack the internal capacity required for supervision. Taken together, these trends make a profoundly shocking conclusion imaginable.

The labor market then divides along a K shaped path. A small minority own, design, and direct AI. The majority are harnessed to AI and supervise its output. As value based pricing establishes a

structure in which created value is captured upward by infrastructure owners, the bargaining power of the latter group weakens structurally. This is the typical conclusion of an economy based on rent extraction. The results of production flow to the owners of the means of production, which here means compute and models, while those who labor on those means are fixed in the position of paying rent for them. To understand the future facing the laboring class after this division, a historical precedent must be revisited: the company town.[15]

2.2 The Company Town: We Are Voluntarily Subordinated

History repeatedly binds workers to those who own infrastructure. The name of that repetition is the company town. The crucial feature shared by historical company towns was that a company owned not only employment, but also housing, commercial space, the living environment, and the rules of the town. Overlapping ownership raised the cost of exit.

In the late nineteenth century, George Pullman built the planned town of Pullman, Illinois, beside his railroad car factory. The Pullman Company owned and rented the town's land and buildings, worker housing, and commercial space. It sought to set rents at a level that would recover 6% of its investment. For workers in company housing, rent was deducted from wages or paid through separate rent checks. When orders fell during the Panic of 1893, the company cut wages by an average of about 25% but did not reduce rents.

The employer and the landlord were the same actor, so workers could not readily escape the shock of falling wages through another housing choice. The conflict grew into the Pullman Strike of 1894 and a nationwide railroad boycott, which the federal government suppressed by deploying troops. The central fact of Pullman was that one company acted as employer, landlord, and de facto town administrator.[16]

More direct economic subjugation appeared clearly in isolated coal and logging company towns through scrip and company stores. Some companies issued workers scrip, a company currency usable only at the company store, either instead of cash or as an advance on the next wage payment. Where the company owned both housing and stores, rent, the prices of necessities, and store debt were tied to the same employment relationship. Practices differed by company. The Stearns Company in Kentucky, for example, paid workers cash every two weeks but provided wage advances in scrip before payday and deducted them from the next wage payment.

These precedents shared a result. Employers raised the cost of exit by controlling a worker's income, housing, consumption, and debt at once. That is why the song "Sixteen Tons," which describes the debt dependence of miners, survived as a cultural symbol. The Fair Labor Standards Act of 1938 introduced federal standards for the minimum wage and overtime pay, and required those wages to be paid in cash or a payment instrument redeemable at face value. Later federal wage regulations did not recognize scrip, tokens, or coupons as lawful wage payments. Scrip nevertheless persisted longer in some regions as wage advances, store credit, or an accounting record.[17]

This form recurs for a clear reason. It solves a problem in favor of the controller. It appears voluntary while making exit nearly impossible. Here lies the quieter trap of its modern form. Pullman workers could strike because they knew the company was simultaneously their employer, landlord, and town administrator. Coal town workers could see wages and advances, housing and consump-

Table 1. Structural comparison of company towns and the AI operating layer

Structure	Combined control	Cost of exit	Key distinction
Pullman	Employment, housing, town administration	Rent, rules of daily life, relocation	Cash wages
Coal and logging company towns	Wages, advances, housing, consumption, debt	Company stores, store credit, isolation	Use of scrip
AI operating layer	Compute, models, work memory, pricing	Workflows, personalization, nonportability	AI tokens are provider specific rights of use, not currency

tion return as debt to the company store. Today's dependent worker, by contrast, is well educated, well paid, and experiences dependence as convenience.

Cloud application programming interfaces and work platforms replace company housing and commercial space. Provider specific metered billing and closed rights of use replace scrip and company stores. Cognitive lock in replaces the town's rules of life. Operators can change the scope of provision and accessibility at any time. The cost of exit becomes unaffordable, yet the relationship is experienced as "voluntary." People tied to app stores, Amazon, and creator platforms already occupy this position. AI tokens are not currency and are not legally or economically identical to scrip. A narrower structural comparison is nevertheless possible: the issuer defines their permitted use and price, and they cannot be transferred between providers. Claude usage or credits cannot be moved to Codex. When a provider changes prices or rules of use, users must leave behind the workflows and memory accumulated in that ecosystem.

Terms such as "technofeudalism" and "rentier capitalism" describe only one face of the phenomenon. A more precise diagnosis is that the combined control of employment, housing, and town administration seen at Pullman, and the combined control of wages, advances, consumption, and debt seen in coal towns, now overlap in the form of cloud infrastructure and deeply rooted lock in. Modern AI dependence does not reproduce those earlier institutions exactly. It repeats a structure in which the same operator controls the means of production, work memory, prices, and exit.

The Fair Labor Standards Act of 1938 and subsequent wage regulations could require cash or payment instruments redeemable at face value, and could exclude scrip as a lawful wage payment, because wage deductions, debt to company stores, and housing control remained visible in contracts and ledgers.[17] Invisible dependence does not form the political constituency required to demand the same intervention.

2.3 The Substance of Neofeudalism

The first feature of neofeudalism is not that corporate owners rule like kings. It lies somewhere more ordinary. Workers rent the means of production required for survival, the memory of their work, their reputation, and their path of learning from a single operating layer. The corporation of 2036 does not merely employ people. It binds the order of information a worker sees, the permitted tools, the available models, the sentences sent to customers, and the judgments classified as risky into one

AI stack. Personal schedules, email, documents, code, meetings, and evaluations accumulate in the same context graph. That graph is a work assistant that knows the worker well. It is also a permission system that determines what the worker can do. The company owns that system, memory resides in the model, and the worker is competent only within that system.

This order needs no barbed wire. Memory and workflows that do not interoperate serve as its walls. The AI operating layer adds another layer to it. A data file may move, but the habits of the model that interpreted it, the authority structure among agents, the tacit exceptions accumulated in operation, and the reward system that imitates evaluator preferences do not move with it. The lock in of 2036 binds not only data, but relationships among data and organizational memory itself. Users choose a provider. The provider designs the range of choices available to the user.

The European Union Data Act reveals that this problem is already under way. It requires cloud providers to offer open interfaces, support for switching, and interoperability standards. Yet a legal right to move provides no meaningful exit without functional equivalence. A worker must be able to work as before after a complete migration. Only then is the exit real.[18]

If the value based pricing model envisioned by frontier companies is completed, workers will become tenants of the infrastructure that makes their labor possible, not merely sellers of their labor power. They will rent the means of production, then pay another share from the output made with those rented means. This is rent, the core of the feudal metaphor.

The worker of 2036 is not ordered to obey. Better choices are recommended at every moment. The system calculates which customer to address first, whose judgment to trust, when to rest, and which tone improves the prospect of promotion. People who follow its recommendations are predictable and productive. Those who depart from them become slow and risky exceptions. Loyalty emerges from this condition. Workers do not remain because they love the system. They remain because they cannot trust themselves outside it. We will take pride in doing the finest work by following a perfectly optimized workflow, but that is true only within the infrastructure of the company to which we belong. We are perpetuated “voluntarily.” This is the ultimate completion of neofeudalism.

3 Root Causes

3.1 Intelligence Is Used as a Public Good and Governed as Private Property

The first cause of this scenario lies in how intelligence is treated. Society is adopting AI as essential general purpose infrastructure, comparable to electricity and telecommunications, while leaving access, prices, memory, and rules to the private discretion of a few companies. Electricity does not form a user’s thoughts on the user’s behalf. AI is different. It mediates what is learned in school, what is approved in the workplace, and which evidence an individual reads. Greater necessity requires greater public control, but the current capital structure is designed so that suppliers’ capacity to collect rent grows as necessity grows. This is the first cause.

3.2 Openness Was Reduced to the Release of Files

The second cause is that the standard for openness has been too shallow. The ability to download weights was treated as decentralization. The Open Source Initiative’s Open Source AI Definition,

however, holds that freedom to use, study, modify, and share requires not only weights, but also training code and information about the data. The Initiative explicitly states that open weights alone cannot reproduce the training process or support deep modification. A further omission remains: the physical condition. A model that can legally be downloaded is not meaningfully open if it cannot run on personal equipment.

The code is open, but compute is closed. The weights are open, but electricity, memory, and networks lie outside the individual's reach. Forkability in the software era must be extended to runnability in the AI era.[19]

3.3 Companies Take the Gains and Individuals Carry the Risks

The third cause is the structure of distribution. Organizations own the time and labor costs saved through AI, while workers personally carry the risks of skill loss and transition. Model providers collect rent from usage and output value, while customers and workers bear the costs of service interruption, price changes, and mistaken judgments. Centralization gathered efficiency in one place and scattered risk downward. Under this structure, the more rationally individuals act, the deeper collective dependence becomes. Using the strongest model available is immediately beneficial to both individuals and companies. Yet when everyone chooses the same operating layer, alternative ecosystems wither and bargaining power shrinks at the next decision. Individual rationality destroys the common exit. This is a coordination failure.

3.4 The Value of Capability Is Not Measured Only by Output

The fourth cause is the definition of productivity. If the same report is produced in half the time, recorded efficiency rises. The lost opportunity to practice judgment, the time in which junior workers learn through failure, and the disappearance of tacit knowledge once distributed across an organization do not enter the ledger. Quarterly output has been placed before the reproduction of capability. The International AI Safety Report 2026 treats cognitive decline, automation bias, and the weakening of autonomy and agency as distinct risks associated with dependence on AI. This does not mean they occur equally for every user. Yet as long as the cost of maintaining human capability remains unmeasured, organizations will treat that cost as something to cut. Workers ultimately produce more results while owning less of the principle by which results are made.[20]

3.5 Exit Depends Too Heavily on Individual Purchasing Power

The final cause is the unit at which alternatives are imagined. The present choice is largely binary. A user rents the strongest centralized application programming interface, or buys expensive hardware to run a smaller model. The former deepens dependence. The latter accepts a gap from the frontier. The premise that one individual must compete with a data center is itself wrong. Personal compute has been combined before. BOINC has pooled idle computation from heterogeneous equipment owned by citizens to conduct research in astrophysics, biology, and climate science. The exit begins not with an enormous personal graphics processor, but with the recognition that coordination technology can turn the small resources held by individuals into shared means of production.[21]

4 Why Decentralized AI Is Necessary

The fundamental response to these problems begins with decentralization.

4.1 Using AI Is Not the Same as Being Able to Live Only upon AI

A good tool expands human capability. Calculators accelerate calculation. Search engines shorten the time required to reach information. AI assists writing, coding, analysis, and decision making. Using a tool is not itself dependence. The problem begins when the tool moves beyond assisting work and comes to own the memory and rules of work as well. A user's writing style, customer relationships, ways of avoiding failure, which colleague's judgment they trust, and which exceptions they have had approved accumulate inside models and agents. AI becomes more useful over time. At the same rate, the cost of leaving grows.

Neofeudalism therefore takes the form not of a concentration camp but of an office where everything works well. The worker is employed, performs well, and receives better recommendations. Yet the real assets that compose this productivity do not belong to the worker. Accumulated context, personalized information, rights of access to internal data, and detailed capabilities maintained by the model are bound to the operating layer. In this order, obedience is not made by commands. It is the very fact that, the moment one leaves the system, one becomes slower, loses memory, and must rebuild one's reputation. The walls are erected not by contracts but by switching costs.

4.2 Convenience Is Purchased with Rent, and Even the State Cannot Resolve the Problem

A centralized system appears highly rational. Large data centers deploy equipment efficiently, optimize networks, and manage models and tools together. Individuals and small organizations can use powerful models immediately without building complex infrastructure themselves. Yet the relationship changes when efficiency becomes necessity. An irreplaceable supplier can demand a share not merely of token cost but of the value created by AI output. If a model wins a contract, the supplier can price a share of the contract value. If it identifies a drug candidate, the supplier can take a share of the discovery value. If it raises a worker's output, the supplier can claim part of the increase. Rent expands beyond a monthly subscription into a share of production.

At the same time, the cost of failure flows downward. Customers and individuals bear service interruptions, price changes, model hallucinations, leaks of personal data, and worker deskilling. The operating layer takes the gains and leaves the risks with workers. The convenience created by centralization becomes the reason it cannot be left. That inability to leave strengthens pricing power in return. In such cases, a general solution can be incorporation into public infrastructure operated by the government. Yet government is itself an enormous organization with simultaneous access to law, administration, telecommunications networks, and systems of identity.

Political change, emergency measures, claims of national security, borders, and sanctions can become authority to block models and data. When one state controls compute, models, and citizen identity together, it can terminate access and identify participants across a broader domain than a company can.

Publicness is therefore not the same as state ownership. Public compute here means compute that no actor can close alone, whose participants can amend its rules together, and which can fork into another operating arrangement. Ideally, individuals, user cooperatives, independent foundations, research communities, local nonprofit operators, and common protocol funds would own different layers of resources, while none would hold the power switch for the whole network.

4.3 Decentralized AI Is Necessary

If AI were only a writing tool, changing providers would be enough. When AI mediates education and employment, medical information and administration, organizational memory and judgment, the operating policy of the supplier becomes a rule of life. A structure in which one company determines the access, price, memory, and recommendations of essential infrastructure cannot be repaired by the virtue of that company. The arrangement of authority itself must change. The first condition is that no actor can alter access to the whole system or its rules alone. Authority to select models, assign work, verify results, and finalize payment must be separated. Companies, governments, and large resource providers must not hold several of these powers in the same decision.

Even when weights can be downloaded, a model is not open to an individual if it requires hundreds of graphics processors, electricity, and network capacity. To turn openness into a substantive right, the ability to execute must be held in common alongside code and weights. The second condition is therefore to combine dispersed physical resources into shared means of production. Home graphics processors, laptop neural processing units, servers owned by small organizations, equipment jointly purchased by cooperatives, long term storage held by independent foundations, and spare personal bandwidth and electricity can operate under the same protocol. Heterogeneous resources owned by many people can then be assembled when needed.

BOINC has already shown an early form of this principle by pooling idle computation from heterogeneous equipment owned by citizens for research in astrophysics, biology, and climate science. The next step can move beyond a structure in which one research institution consumes donated computation. Participants can solve their own problems and jointly own both the models and the right to operate them.[21]

Pooling personal devices alone does not change ownership. If a central platform assigns work, sets prices, and pays device owners only a minimum fee, it is merely a distributed data center and a new chain of subcontracting. The third condition is therefore to verify who created which value and carried which risk, then settle the same event exactly once. The contributions of those who completed computation, communities that supplied data, resources held ready for failure, people who verified results, and maintainers of protocols and public interest models are recorded as distinct items.

Failures, insurance, and refunds are tied to the same economic event so that duplicate claims cannot occur. Compute contribution must not become governing authority. A person with more equipment may receive more payment for computation, but cannot buy more votes. Economic reward and political rights must remain separate so that a distributed market does not harden into another domain of capital.

Another crucial domain is autonomy and the system that makes autonomy possible. Exit is currently too expensive for the individual. Leaving the strongest central application programming inter-

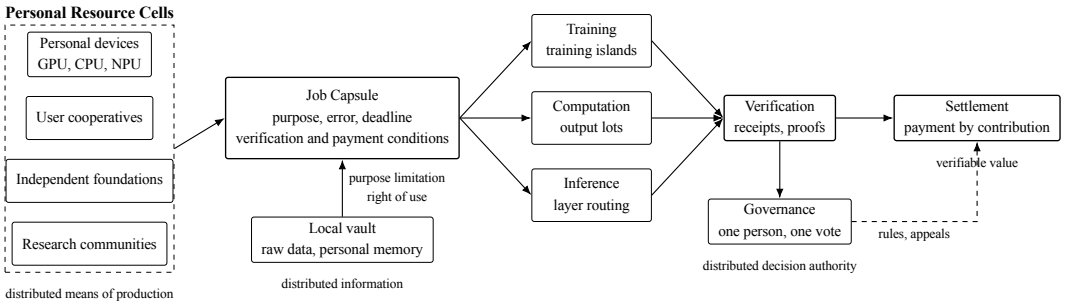

Figure 1. Decentralized AI infrastructure from Personal Resource Cells through training, computation, inference, verification, and settlement

face means accepting a smaller model and slower hardware. After requiring one person to compete with a data center, responsibility for being unable to leave is assigned to that person's choice. The final condition is therefore a collective exit in which production can continue when many people leave together. Participants must be able to carry not only models and code, but also execution resources, evaluation rules, settlement records, common funds, and the state of personalized workflows. This is the final key to leaving feudalism.

Decentralized AI can now be defined. Decentralization is not the placement of computers in many regions. It is not the issuance of a cryptocurrency, nor the construction of a vast public data center by the state. It is a structure that distributes the means of production, personal information, decision authority, and exit across different actors, so that no one can close the system alone and anyone can continue production through another combination. Decentralization of production combines resources owned by many people. Decentralization of information leaves raw data and personal memory with their owners. Decentralization of decisions separates work assignment, verification, payment, and rule making.

Decentralization of exit makes it possible to fork with the model and its operating assets in hand. These four axes must operate together. Distributing computation alone creates a cheap subcontracting network. Keeping data local alone allows the central scheduler to become a new gatekeeper. Only together do these axes form a system of decentralized AI.

5 Designing Decentralized AI Infrastructure

5.1 Step 1: Combining Personal Resources into a Shared Compute Resource

Resources owned by individuals differ widely. They include graphics processors, central processors, neural processing units, memory and storage, bandwidth, spare electricity and thermal capacity, sensors, local data, and even the time required to judge a particular language or domain. The first task is to register them as Personal Resource Cells. A Personal Resource Cell does not transfer ownership of a device to the network. It grants only limited authority over a defined scope of use.

A resource can be divided into precise units: "12 GB of graphics memory available from 1 a.m. to 5 a.m.," "neural processing unit available while the battery remains above 70%," or "only this statistical operation is permitted on these data." Providing a graphics processor does not open the file

system. Providing storage does not grant authority to read its contents. Resources are shared. The provider's life is not.

A fixed wallet address combined with a device model, region, and connection time can reveal what one person does and when that person is at home. The commons therefore does not request detailed device identity. It requests only proof that the resource satisfies the memory, numerical precision, availability period, and failure conditions required by the present task.

If every model must be trained only on personal devices, resources disconnect easily and long term storage is difficult. If only large devices are admitted, control returns to a small number of operators. The possible layers of cells therefore begin with personal laptops, desktop computers, mobile devices, and home servers. Above them could stand graphics processors and storage jointly purchased by user cooperatives in a future where decentralized AI has taken root. Equipment entrusted to independent foundations, research communities, and nonprofit consortia could support long term preservation and large scale training.

Under this structure, these institutions are not subordinate agencies of government ministries. Citizens and users of the protocol constitute their boards and budgets in a distributed manner. Large equipment supplies performance. Small equipment broadens ownership. No layer can train, verify, and settle a model alone. A large cell responsible for long term storage has no authority to assign work, and a personal device need not contribute additional resources to obtain citizenship.

The conditions satisfied by each resource are crucial when dispersed resources are assembled. Every task is therefore defined as a Job Capsule. A Capsule contains the model and code to be used, the formats of inputs and outputs, the purpose of data use, permissible error, privacy limits, the method of verification, the deadline, and failure recovery conditions. The network then reserves all required resources together. A task does not begin when compute nodes have been secured but no verifier is available, or when data are ready but no payment has been placed in escrow. Inputs, computation, networking, recovery resources, storage, verification, and payment must all be ready before execution.

Work assignment does not place fixed identities on an open auction board. Offers made within the same period are collected into a large set, and price, performance, regional distribution, and conflicts of interest are evaluated while the offers remain encrypted. A participating resource learns only its own role. The identities of its competitors and the relationships that caused an exclusion are not exposed. The scheduler therefore does not become another lord. It remains a replaceable tool that computes a combination of resources satisfying a public contract. Several schedulers can interpret the same Capsule, and another implementation can continue the task if one scheduler stops.

5.2 Step 2: Different Forms of Distribution for Training, Computation, and Inference

Training comes first. The greatest bottleneck in large scale training will be the communication required for distant devices to synchronize the same state at every step. Personal equipment differs in speed and availability and often leaves the network. If synchronous data center training is copied directly, the slowest device determines the speed of the whole system. The solution is to form small training islands among nearby devices. Each island performs several local steps, then shares only a summarized update.

DiLoCo reported performance comparable to synchronous training across eight loosely connected workers while reducing communication by about 500 times. OpenDiLoCo extended the method across two continents and three countries and recorded compute utilization of 90% to 95% in practice. Photon reported experiments that reduced communication by 64 to 512 times in federated pretraining of models up to 7B parameters.[22]

This system extends the same principle to ownership. Raw data remain in repositories owned by each individual and community. Only encrypted updates leave an island. Each update is bound to the model version from which it began, the data policy it followed, and the interval of training it completed. Every valid update received before the deadline is combined under a rule fixed in advance. Operators cannot inspect the results and then select only favored regions or groups. Device departure is not treated as an exception.

Go With The Flow reported reductions in training time of as much as 45% by rerouting streams of microbatches in a heterogeneous environment spanning ten regions, where resources repeatedly joined and left.[23] Departure is therefore treated as an ordinary condition rather than an error. When one island disappears, another continues the work. If a training round fails to attract enough participants, it terminates instead of relaxing its privacy threshold. This is one general form of the process.

Many other computations stand between training and inference. Data cleaning, embedding, search index construction, evaluation, safety testing, model compression, and adapter merging continue throughout the system. If these tasks are fixed to function calls for one graphics processor and one framework, devices become tied to the same provider again. That is not complete decentralization. FusionLLM represents a model as a computation directed acyclic graph and has tested fine grained work placement and adaptive compression across links ranging from 8 Mbps to 10 Gbps in an environment of 48 graphics processors.[24]

This system generalizes that distributed form by dividing work into small output lots with defined input and output hashes, computational meaning, and permissible error. Heterogeneous cells such as an Apple neural processing unit and a CUDA graphics processor need not execute the same code. They need only satisfy the same contract for results. Resource providers cannot charge an unlimited amount of graphics processor time. The requester places the price of the result in escrow in advance. When a verified output lot is completed, the price is divided among the participants in the actual path.

Using a slower and less efficient device to produce the same result therefore does not raise the total price. The unit sold is not resource consumption but a verified result. The contract formed at the preceding stage makes this possible.

Inference is the most important stage. When a user submits a question, fast response and continuity of conversational state matter. If no single device can hold the entire model, different devices must divide its layers. One device computes an intermediate representation, or hidden state, and sends it to the next. The final device produces probabilities for the next token. The technical demonstration described below, DRIFT, or Decentralized Routed Inference For Tokens, implements and publishes this inference path.[25]

5.3 Step 3: Managing the Security Risks of Distribution

Devices owned by strangers operate one model together. Sending the original text and personal context without protection would recover the means of production at the price of privacy. The answer is not to copy and collect data, but to limit the purpose and scope of computation permitted to access those data. Personal documents and conversations, schedules, medical records, work memory, and personalization adapters remain in an encrypted local vault. Instead of the original data, the network receives only an authorization defining the permitted purpose, operation, duration, and output.

For example, a user could permit only gradient computation and aggregate statistics for disease classification instead of uploading an entire medical record to a training server. The model computes where the data reside. The user's manner of speech, relationships, and work memory remain personal assets that can move to any model. Even when raw data do not move, individual gradients and updates can reveal information. The groups to be aggregated, the minimum number of participants required before results are disclosed, and the amount of noise to be added must be fixed before the task begins. A privacy budget must have a ceiling that no participant can exceed. If people in need of money can sell more of their privacy, the information market becomes another class market.[26]

Privacy does not mean hiding only numbers inside ciphertext. Work assignments, payments, disputes, and participation in a fork can also identify a person. In this system, the recipient, timing, form of ciphertext, and privacy cost are recorded as one disclosure object. Normal payments and requests to recover from censorship are sent through shielded inboxes of the same size. Yet the particular technical mechanism is not the central proposition. From the generation of data, through computation, to deletion, the system must remain within the purpose limitation established at the beginning.

5.4 Step 4: Contributions of Computation and Settlement

A network of personal resources cannot persist on goodwill alone. Electricity, equipment depreciation, storage, and verification all cost money. The system must prove who did what and return the resulting value. As in blockchain systems, participants who do not know one another need methods to prove performance and settle many small transactions together. Prior research exists at this point. Proof-of-Learning proposed a structure that verifies actual training through intermediate states and stochastic trajectories in stochastic gradient descent (SGD), and zkPoT set out an approach for proving that a committed model was correctly trained on a committed data set without revealing additional information about either the model or the data set.[27]

Every claim that changes model promotion or payment must therefore be reproducible or decidable by a verification group selected in advance. Evidence of contribution cannot merely be stored. It must be possible to reopen and inspect it.

Computation, data, standby resources, and public goods create value in different ways. Combining them into a single reputation score would allow those with the most equipment to control standards of verification and governance as well. Payment must therefore remain separate. Execution receives the price of verified results. Standby nodes receive compensation for availability and actual recovery. Code maintenance and safety evaluation draw from a public goods budget. One completed task must also generate payment only once. When a task begins, its economic obligation and escrow are created first. The original execution, retries, failover, and insurance claims are then all bound under that

obligation and escrow. Filecoin's proofs of storage, Akash's bidding and escrow, and Ethereum's optimistic rollups and state channels provide precedents for verification, exchange, disputes, and small value settlement. This essay combines these components into a settlement structure that hides who sent a particular prompt while making it publicly verifiable that no new money was created and the same labor was not paid twice.[28]

5.5 Step 5: Separating Compute Volume from Citizenship

A greater contribution must never increase the size of a participant's vote. If the same group controls work assignment, model evaluation, payment, and changes to rules, thousands of nodes will once again labor for a single fiefdom. Economic rewards created by compute volume must be separated from governing authority that belongs to persons.

Citizenship does not come from the number of graphics processors or a token balance. A user who contributes no device still has one person, one vote. Agenda setting, evidence, approval, execution, and appeal belong to different roles. The same controlling group cannot hold two roles in one decision. A large graphics processor cooperative cannot evaluate the model it trained. A foundation entrusted with long term storage cannot terminate payment or citizenship. Performance creates prices. Citizenship remains with persons. If people who contribute no device receive a basic allocation of access as part of basic intelligence, that access is not a reward for contributing resources.

People who lack equipment and electricity, or decline to participate because their data are sensitive, must still have access to basic AI guaranteed by the common protocol fund and must hold equal governing authority. Only then can an individual refuse to provide data and devices. Users must be able to disable recommendations, choose manual workflows, and export personal memory, agent graphs, evaluation rules, and adapters in a machine readable form. The paradox in which deeper personalization makes exit more difficult is resolved by removing switching costs. True sovereignty is secured through the right to leave.

6 DRIFT: Decentralized Routed Inference For Tokens

DRIFT, or Decentralized Routed Inference For Tokens, is a technical demonstration that turns the question "Can heterogeneous personal devices jointly produce the inference of one model?" into an executable inference path. It divides the decoder layers of one language model between a Mac using Apple MPS and a personal computer using NVIDIA CUDA. Each socket worker materializes the weights of its assigned decoder layers and the session specific key value cache for those layers. In thin head mode, the first and final nodes additionally hold the embedding and the normalization and language model head, respectively. In `--chain` mode, the hidden state passes directly from one node to the next for each token. In the default star mode, every hop passes through the head. Both modes connect the MPS and CUDA backends in PyTorch through a wire contract composed of a four byte length prefix, a msgpack dictionary, and a tensor byte payload.

In an in process split measurement using Qwen2.5-1.5B-Instruct, Apple MPS, and fp16, all 411 tokens generated from six prompts matched the full model baseline on the same Mac. In a cross device measurement, Mac MPS held the first fourteen layers and a Colab NVIDIA T4 CUDA instance held the remaining fourteen. All 130 tokens generated from three prompts matched the baseline execution.

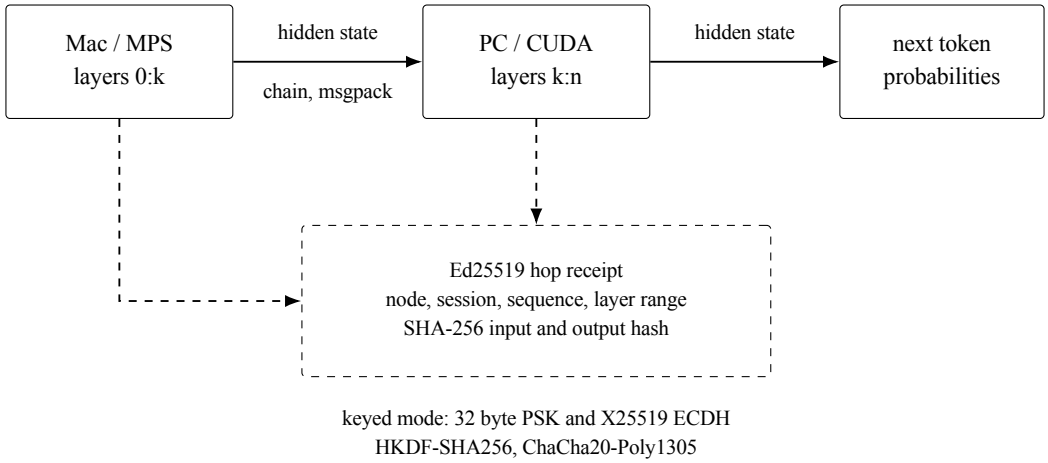

Figure 2. Chain mode, keyed wire, and hop receipts implemented in DRIFT

In a failover integration test generating 48 tokens, one worker was terminated during generation. The model was repartitioned across the surviving worker and a spare, and prefill was repeated over the prompt and all previously generated tokens. The final 48 tokens matched an uninterrupted baseline execution. The present demonstration remains far too slow to rival frontier model services. Yet by joining different kinds of personal devices to process one request in practice, it marks a first step from which the possibility of decentralized inference can be seen.

Each worker signs its node public key, session, sequence, mode, assigned layer range, and SHA-256 hashes of inputs and outputs with an Ed25519 identity key. For every token, the head verifies signatures, adjacency of hashes between hops, the anchors for the first input and final output, and continuity of layer ranges. When a 32 byte network preshared key is configured, each connection derives directional keys with HKDF-SHA256 by mixing ephemeral X25519 elliptic curve Diffie Hellman with the preshared key. It then encrypts and authenticates each msgpack frame with ChaCha20-Poly1305. Connections without a network key use plaintext. The receipt verifies the execution path and continuity of signatures. It does not prove the correctness of the computational result itself. The receipt journal aggregates tokens and layer tokens for each node. It does not perform payment or settlement.[25]

DRIFT is an early prototype that implements the inference data plane of decentralized AI. The present code uses a central orchestrator to initiate work and assign layers, self asserted node membership, replay based failover, and a contribution ledger. Wire encryption protects data in transit, but participating workers process decrypted input identifiers and hidden states. The next design layer consists of a replaceable Job Capsule scheduler, admission that resists Sybil attacks, redundant execution and audits through recomputation, a payment and settlement protocol, and a data privacy mechanism that separates prompts and intermediate representations from participating workers. DRIFT demonstrates that heterogeneous personal devices can divide the inference of one model in practice and record that path through keyed wire transport and signed receipts. Its demonstrated scope extends no further.

7 Decentralization Does Not Reject Trust. It Decomposes the Power to Block

AI neofeudalism does not follow from the emotional claim that one company is evil. It follows from a structure in which society uses intelligence as essential infrastructure while binding the means of production, personal memory, evaluation, and exit inside one operating system. Every repetition of the most convenient choice dries out alternative ecosystems. Rational individual choices erase the common exit. Serious discussion of decentralization is therefore necessary now, in 2026, while speculation still burns. The unilateral power to block must be decomposed wherever essential infrastructure is privately governed.

The present trajectory is moving toward granting transnational corporations and governments comprehensive control over, and the power to block, the infrastructure of intellectual activity. These developments must be understood on the same plane as our awareness of sovereignty. In the system proposed here, the state is not the final operator. Neither a company, a large cooperative, nor an independent foundation owns the whole. Small resources owned by individuals and communities overlap across different layers, while a public protocol forms the contracts between them. No one can switch off the whole system alone. Anyone can continue through another combination. To own intelligence is not to possess a model file.

It also means holding the resources required to execute together, defining the boundary of personal information, verifying results, receiving a share of created value, participating in rules with an equal vote, and ultimately leaving for another operating system. In a new era, these are also the substance of sovereignty.

The end of a borrowed future is not a cheaper subscription. It is the ability to produce one's own share without borrowing it.

References

[1] David Cahn, “AI’s $600B Question,” Sequoia Capital (2024), [link] ; NVIDIA, “NVIDIA Announces Financial Results for First Quarter Fiscal 2027” (2026.5.20), [link] ; Allianz Research, *AI Capex Cycle* (2026.3), [link] ; Anomaly Investments, “This Obviously Is an AI Bubble. The Math Says So” (2026.6.7), [link]

[2] Moody’s Ratings, “US Hyperscalers: Off-Balance-Sheet Future Leases Mask Significant Debt-Like Obligations” (2026.2.23), [link]

[3] Man Group, “The AI Bubble: Hidden Risks and Opportunities” (2026), [link] ; Allianz Research, *AI Capex Cycle* (2026.3), [link] ; Anomaly Investments, “This Obviously Is an AI Bubble. The Math Says So” (2026.6.7), [link]

[4] Bank of England, “Financial Policy Committee Record — October 2025,” [link] ; Bank for International Settlements, “I. Progress and Peril,” *Annual Economic Report 2026* (2026.6.28), [link]

[5] Aditya Challapally, Chris Pease, Ramesh Raskar, and Pradyumna Chari, *The GenAI Divide: State of AI in Business 2025*, MIT Project NANDA (2025.7), [link]

[6] Space Exploration Technologies Corp., Form S-1/A, U.S. Securities and Exchange Commission (2026), [link] ; SpaceX, “Announces Pricing of Initial Public Offering” (2026), [link] ; CNBC, “SpaceX IPO Explained” (2026.6.9), [link] ; Associated Press, “SpaceX’s IPO Gives Retail Investors an Unusually Large Allocation” (2026), [link] ; Bloomberg, “SpaceX IPO Draws Over $100B in Retail Orders” (2026.6.11), [link] ; CNBC, “SpaceX IPO Leaves Retail Investors with Too Few Shares” (2026.6.15), [link]

[7] Sarah Friar, “A Business That Scales with the Value of Intelligence,” OpenAI (2026), [link]

[8] Anthropic, “How Am I Billed for My Enterprise Plan?” Claude Help Center, [link] ; PYMNTS, “Anthropic Switches to Usage-Based Billing for Enterprise Customers” (2026), [link]

[9] DeepSeek-AI, “DeepSeek-V4-Flash” and “DeepSeek-V4-Pro,” official model cards, Hugging Face (2026), [Flash] and [Pro]

[10] Ollama, “Pricing,” [link] ; KK Kavishka Karunanayake, “Ollama Pricing 2026: Total Cost & Competitors Compared,” CheckThat.ai (updated 2026.5.29), [link]

[11] Zapier, “AI Vendor Lock-In Survey” (2026), [link] ; Deloitte, *The State of AI in the Enterprise* (2026), [link]

[12] Anthropic, “Anthropic and Accenture Announce Landmark Partnership” (2025), [link] ; Anthropic, “The Claude Partner Network” (2026), [link]

[13] Nataliya Kosmyna et al., “Your Brain on ChatGPT,” MIT Media Lab (2025), [link] ; Hao-Ping Lee et al., “The Impact of Generative AI on Critical Thinking,” Microsoft Research / CHI 2025, [link] ; Yizhou Fan et al., “Beware of Metacognitive Laziness,” *British Journal of Educational Technology* / arXiv:2412.09315 (2024), [link]

[14] Risk Engineering, “Air France Flight 447: Confusion on the Flight Deck,” [link] ; *International AI Safety Report 2026*, pp. 89–90, [link]

[15] Jathan Sadowski, “The Internet of Landlords: Digital Platforms and New Mechanisms of Rentier Capitalism,” *Antipode* 52 (2020), 562–580, [link] ; Virginia Commonwealth University Libraries, “Company Towns: 1880s to 1935,” [link]

[16] U.S. National Park Service, “The Town of Pullman,” [link] ; U.S. National Park Service, “Fact or Fiction: Did Pullman Use Scrip?”, [link] ; U.S. National Park Service, “Park Brochure — Pullman National Historical Park,” [link] ; U.S. National Park Service, “The Strike of 1894,” [link] ; United States Strike Commission, *Report on the Chicago Strike of June–July, 1894*, S. Ex. Doc. 53-7 (1894), [link]

[17] U.S. National Park Service, “Scrip—A Coal Miner’s Credit Card,” [link] ; Yale Energy History, “Coal Mining and Labor Conflict,” [link] ; Shaun Richman, “Company Towns Are Still with Us,” *The American Prospect* (2018), [link] ; U.S. Department of Labor, “Fair Labor Standards Act of 1938: Maximum Struggle for a Minimum Wage,” [link] ; Electronic Code of Federal Regulations, 29 CFR 531.27, 531.34, [link] and [link]

[18] European Union, Regulation (EU) 2023/2854, *Data Act*, [link]

[19] Open Source Initiative, “Open Source AI Definition 1.0,” [link] ; Open Source Initiative, “Open Weights: Not Quite What You’ve Been Told,” [link]

[20] *International AI Safety Report 2026*, [link]

[21] David P. Anderson, “BOINC: A Platform for Volunteer Computing,” *Journal of Grid Computing* 18 (2020), 99–122, [link]

[22] Arthur Douillard et al., “DiLoCo: Distributed Low-Communication Training of Language Models,” arXiv:2311.08105 (2023), [link] ; Sami Jaghouar et al., “OpenDiLoCo,” arXiv:2407.07852

(2024), [link] ; Lorenzo Sani et al., “Photon: Federated LLM Pre-Training,” arXiv:2411.02908 (2024; revised 2026), [link]

[23] Nikolay Blagoev et al., “Go With The Flow: Churn-Tolerant Decentralized Training of Large Language Models,” arXiv:2509.21221 (2025), [link]

[24] Zhenheng Tang et al., “FusionLLM: A Decentralized LLM Training System on Geo-Distributed GPUs with Adaptive Compression,” arXiv:2410.12707 (2024), [link]

[25] TaewoooPark, “DRIFT: Decentralized Routed Inference For Tokens,” GitHub repository, source code and README, commit `d363287`, [link]

[26] National Institute of Standards and Technology, “Privacy Attacks in Federated Learning” (2024), [link] ; Keith Bonawitz et al., “Practical Secure Aggregation for Privacy-Preserving Machine Learning,” CCS 2017, [link]

[27] Hengrui Jia et al., “Proof-of-Learning: Definitions and Practice,” arXiv:2103.05633 (2021), [link] ; Kasra Abbaszadeh, Christodoulos Pappas, Jonathan Katz, and Dimitrios Papadopoulos, “Zero-Knowledge Proofs of Training for Deep Neural Networks,” IACR ePrint 2024/162 (2024), [link]

[28] Filecoin Docs, “Storage Proving,” [link] ; Akash Network Docs, “Deployments,” [link] ; Ethereum.org, “Optimistic Rollups” and “State Channels,” [link] and [link]