IFP

# Unlocking a Million Times More Data for AI

*How a new ARPANET-style program could solve the data accessibility problem* | **Andrew Trask & Lacey Strahm**

# Unlocking a Million Times More Data for AI Through Attribution-Based Control

*How a new ARPANET-style program could solve the data accessibility problem*

Andrew Trask & Lacey Strahm

---

*This essay is part of The Launch Sequence, a collection of concrete, ambitious ideas to accelerate AI for science and security.*

## Summary

Every major leap forward in AI progress has been accompanied by a large increase in available data to support it. However, AI leaders often warn that we have reached “peak data” — that all human data for training AI has been exhausted. Our analysis suggests that this view may not capture the whole picture.

While not all AI systems report their training data size, some do. Among those, the order of magnitude is a few hundred terabytes. For reference, you could go to Walmart, buy a few dozen hard drives, and fit this data on your kitchen table. Meanwhile, the world has digitized an estimated 180-200 zettabytes of data — over a million times more data than what was used to train today’s leading models. This means the data exists, but is not being used for training. We classify this as an access problem rather than a scarcity problem.

This proposal presents Attribution-Based Control (ABC) as a potential framework for expanding access to the world's digital data while preserving ownership rights. While significant technical, legal, and economic barriers prevent access to most of

the world's data, ABC offers one possible approach to address these barriers by enabling data owners to maintain control while contributing to AI development. We outline the technical foundations of ABC and suggest a government-led development program modeled on the success of ARPANET.

# Motivation

## The foundation: AI growth needs data growth

The history of AI breakthroughs reveals a consistent pattern: each major leap forward has been fueled by dramatic increases in the availability of high-quality data. Take, for example, some of the biggest leaps forward in recent AI history. AI's "ImageNet moment" was driven by a single, large, labeled dataset. Word2vec was a fundamental breakthrough in the amount of data a language model could process. Transformers were a simplification of an older algorithm (LSTMs) whose purpose was primarily to process more data. DeepMind's rise centered around generating (narrow) datasets from Atari (and later Go) video games. GANs gained popularity by training models to generate their own data. InstructGPT/RLHF are data acquisition strategies (via customers and "mechanical turkers" — humans hired to generate data online through platforms like Amazon Mechanical Turk or Prolific) that are applied using standard learning algorithms. And as the original papers describe, OpenAI's GPT-1, GPT-2, and subsequent advancements came not as fundamental changes to the Transformer, but on the back of OpenAI's pioneering efforts in assembling large-scale datasets from publicly available internet data and armies of human data creators.

Non-data advancements in GPUs or algorithms are crucial. But running a hundredfold bigger GPU cluster or a more sophisticated algorithm without more data is like running a massive truck with just enough fuel for a sedan. The truck can get moving, but the trip will be short. So too with AI's rise.

## Understanding concerns about data scarcity

Recent statements from Ilya Sutskever declaring we've reached "peak data" and Elon Musk warning that all human data for AI training has been "exhausted" reveal concerns from industry leaders about data availability for future AI progress. Their concern stems from understandable observations of the current constraints on accessing high-quality training data. Privacy laws are getting stricter, tech platforms are closing off access, and copyright holders are pushing back.

These constraints have led to the exploration of alternative approaches to extract more information from the data we already possess, including synthetic data generation and test-time compute optimization. Yet, synthetic data is merely a compressed form of the data it is derived from, and while test-time compute is a viable way to squeeze extra performance, it is not a sudden paradigm shift in capabilities of the kind new data sources historically have enabled.

AI labs' latest exploration has resulted in the onboarding of massive armies of mechanical turkers to create more specialized data in specific fields, and the continued collection of millions of users' experiences to fine-tune widespread use. Consequently, AI capabilities inch forward as information is slowly copied from the minds of turkers and customers to AI models, and some breakthroughs occur. But both Mechanical Turk programs and user feedback systems only provide incremental improvements at significant cost and limited scale. They pale in comparison to the paradigm shifts of ImageNet, AlphaGo, and GPT.

## From scarcity to abundance

When examining global data resources, a different picture emerges. While not all recent, industry-leading, state-of-the-art AI models report the size of their dataset, some do, and others have been estimated. All fall within a similar size range.

**Training Data Size Comparison Across Major AI Models**

| Company | Model | Approximate Training Data Size |
|---|---|---|
| Meta | Llama 4 Behemoth | 180 TB[1] |
| Meta | Llama 3 | 94 TB[2] |
| OpenAI | GPT-4 | 78 TB[3] |
| Alibaba | Qwen2.5 Instruct | 108 TB |
| DeepSeek | DeepSeek | 89 TB |
| Huawei | Pangu Ultra | 79 TB |
| NVIDIA | Nemotron-4 | 54 TB |
| Amazon | Titan | 24 TB |
| Google | Palm 2 | 21.6 TB |
| Google | Gemini LaMDA | 750 GB |

*Note: Out of 952 AI models tracked by Epoch AI with estimates or reports of training dataset size, none exceeded Meta's 180TB.*

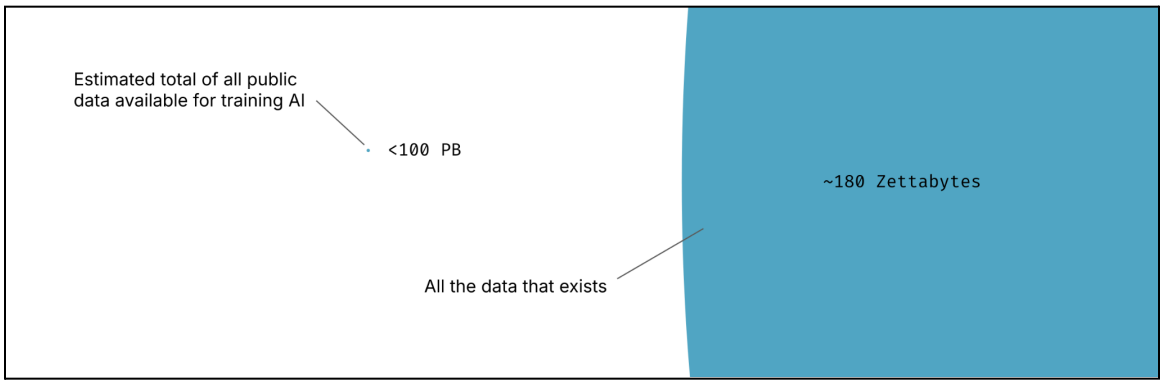
Despite what their name might suggest, so-called "large language models" (LLMs) are trained on relatively small datasets. For starters, all the aforementioned measurements are described in terms of terabytes (TBs), which is not typically a unit of measurement one uses when referring to "big data." Big data is measured in petabytes (1,000 times larger than a terabyte), exabytes (1,000,000 times larger), and sometimes zettabytes (1,000,000,000 times larger). For reference, Samsung sells a smartphone with 1TB of storage, and consumers can buy a 2TB SD card the size of a thumbnail. You could probably get a few friends together and store a copy of an LLM training dataset using your drawers of old smartphones and laptops.

One might be inclined to doubt and say, "*Perhaps the AI models that don't report dataset sizes leverage 10–100 times more data!*" However, this would require

[1] Meta's Llama 4 Behemoth was trained on 30 trillion "tokens" (i.e., word fragments), representing around 180 TBs of text data. This is estimated conservatively at 6 TB per trillion tokens, but this can vary based on the tokenizer.
[2] 15.6 trillion tokens, or roughly 94 TBs of text data.
[3] 13 trillion tokens or roughly 78 TBs of text data.

10–100 times more compute to train. There's considerable evidence against this being the state of affairs — leading AI labs use about as many chips for training as you'd expect given the claimed size of their datasets.

**Data Utilization in AI: Current Training Sets vs. Global Data Volume**




Meanwhile, the world's eight billion+ people, 360 million+ companies, and thousands of local, national, and international government organizations have digitized an estimated 180 zettabytes of data, a volume that doubles every two years. This private data represents a million times more data than The Internet Archive, two hundred million times more data than Common Crawl, and a billion times more data than what was used to train GPT4, DeepSeek-R1, or Google's Lambda or Palm 2.

What makes this vast private data uniquely valuable is its quality and real-world grounding. This data includes electronic health records, financial transactions, industrial sensor readings, proprietary research data, customer/population databases, supply chain information, and other structured, verified datasets that organizations use for operational decisions and to gain competitive advantages. Unlike web-scraped data, these datasets are continuously validated for accuracy because organizations depend on them, creating natural quality controls that make even a small fraction of this massive pool extraordinarily valuable for specialized AI applications.

Considering both the sheer volume and quality of private data, it's clear that despite the concern from AI leaders, the world hasn't remotely run out of data. The challenge is not data scarcity, but rather barriers to accessing and utilizing existing high-quality data resources.

## The real crisis: Misaligned incentives between data owners and AI companies

The real problem is an old one: information markets are failed markets. When a data owner shares a piece of data, the owner loses all control over how it will be used, copied, and shared further. When the owner sells a piece of data, they don't sell the original data — they sell a copy. When a dataset is copied, the global supply goes up, the price goes down, and every customer becomes a competitor for the future sale and use of that data.

This problem is compounded by data's unique economic property: it's an infinitely reusable resource. Unlike physical goods, the same dataset can simultaneously power a medical breakthrough, optimize a supply chain, and train an AI model. Each buyer gains not just one use case but potentially thousands, making them an immediate competitor for selling those same insights to others. This "economic bundling" means data owners can never capture the full value of what they're selling.

Consequently, most people don't sell or share most data. Even the largest tech companies in the world — Meta, Apple, Google, etc. — don't tend to sell the data they acquire. Instead, they use it for internal use cases, keeping the supply low and the value of the data (and the use cases it supports) high. Indeed, creating data moats that drive superior advertising is the core business model of the internet. And a data moat is the opposite of sharing data — it's working very hard to ensure the data is never made available to a possible competitor.

This dynamic creates a fundamental misalignment that extends to AI labs themselves. While they've built successful businesses on data moats and proprietary architectures, they now face an existential constraint: their training datasets measure in hundreds of terabytes, yet the world's data measures in zettabytes. Without access to this vastly larger data pool, AI progress risks stagnating, as labs exhaust the limited public data available and compete for marginal improvements rather than paradigm-shifting advances. As privacy laws tighten and data owners refuse sharing requests, the very market structures that enabled AI labs' rise now limit their access to the critical resource they need for further advancement. Without a sustainable path to access more data, both sides

are trapped: data owners won't share because they lose control and value, while AI labs can't access the exponentially larger datasets required for the next leap in AI capabilities.

Given these incentive misalignments, the key question for policymakers to answer is: Can we design a system that aligns data owner incentives with the growth needs of AI labs? Can policymakers fix the failing market for AI training data and catalyze a million times more data, fueling the next AI paradigm shift?

# Solution

Think of today's AI like a giant blender where, once you put your data in, it gets mixed with everyone else's, and you lose all control over it. This is why hospitals, banks, and research institutions often refuse to share their valuable data with AI companies, even when that data could advance critical AI capabilities.

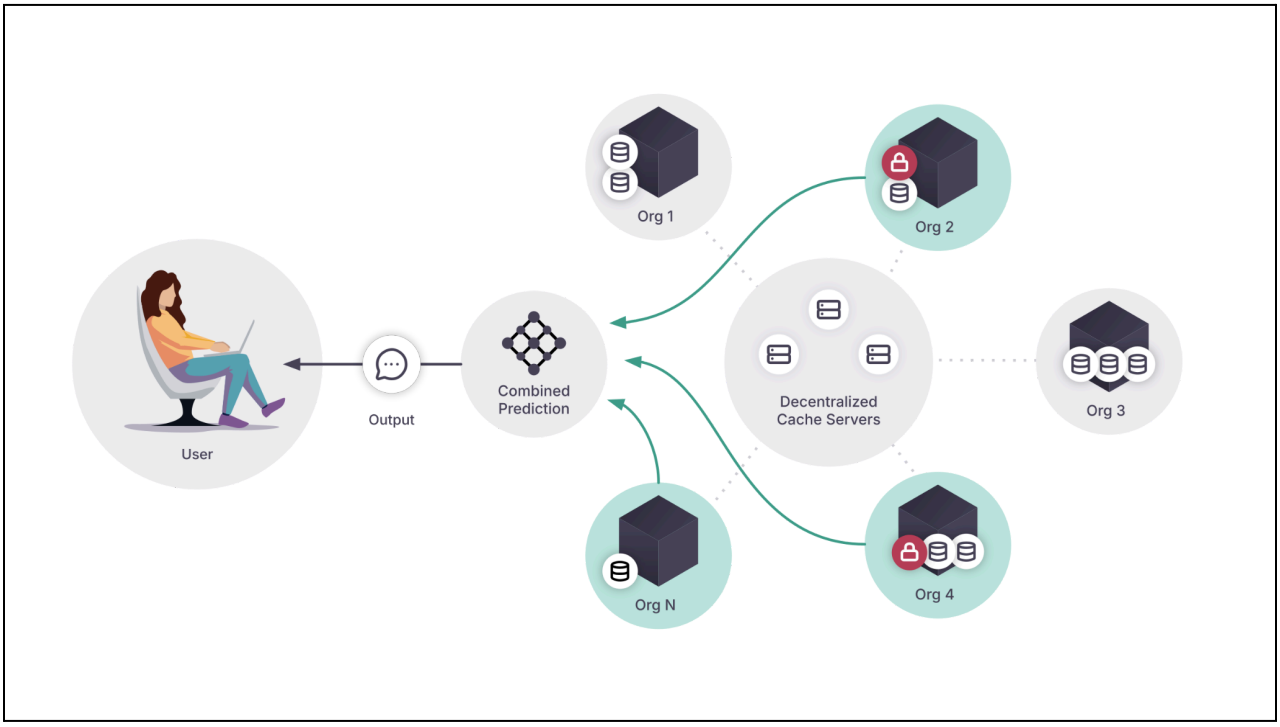
Attribution-Based Control (ABC) represents a paradigm shift in how we design AI systems. ABC is not a specific technology, but rather a set of criteria that any AI system must meet:

1. Data owners must be able to control which specific AI predictions their data supports
2. AI users must be able to control which data sources contribute to their received predictions (e.g., LLM tokens)

These properties transform data from a one-time giveaway into an ongoing revenue-generating asset. When data owners can control how their data is used and get paid each time it provides value, they have strong incentives to share rather than silo their resources.

When AI systems meet these criteria, they create a sustainable data economy. Data owners gain an incentive to share because they maintain ongoing control and can generate continuous revenue streams, similar to how musicians earn royalties each time their song is played. This solves the access problem. Instead of hoarding data for competitive advantage, organizations can monetize it while maintaining control.

**A high-level diagram of AI with attribution-based control**




The technological foundations for ABC already exist today through two key technical capabilities:

1. The ability to partition AI models by data source
2. The infrastructure to preserve privacy and control throughout the AI lifecycle

## Technologies for model partitioning

Model partitioning enables the "attribution" part of ABC by keeping different data sources mathematically separable within the AI system. Several software architectures now exist that enable source-specific ownership of AI components. Mixture of Experts (MoE) naturally segments models into specialized sub-networks that can be owned independently, while Retrieval-Augmented Generation (RAG) takes a different approach, separating knowledge storage from processing, allowing data owners to maintain control over their knowledge bases while contributing to collective intelligence. RETRO and ATLAS exemplify how these

principles scale in practice, achieving GPT-3 level performance while using 25-50x fewer parameters by maintaining explicit source attribution through their database architectures. Meanwhile, model merging techniques like Git Re-Basin demonstrate yet another path, proving that independently trained models can be combined without performance degradation, and recent work on federated MoE shows how these approaches scale across distributed ownership. This proliferation of successful architectures, each implementing ABC principles through a different technical mechanism, demonstrates the viability of source-specific attribution in AI systems.

## Technologies for privacy infrastructure

Privacy-preserving infrastructure provides the "control" part of ABC by ensuring data owners can enforce their participation decisions without exposing their data. This infrastructure is technically mature, built on decades of research in privacy-enhancing technologies (PETs) that now enable end-to-end encrypted AI workflows. During training, each data owner can develop their model partition within GPU enclaves — hardware-isolated environments where NVIDIA H100s and cloud providers guarantee that training data never leaves unencrypted. Homomorphic encryption enables the aggregation of these distributed model pieces while they are encrypted, allowing for federated learning without centralizing data.

This infrastructure can be flexibly implemented depending on organizational needs and capabilities. Small organizations can use cloud-based secure enclaves or federation services, while larger institutions might choose to run local secure nodes. For AI labs, implementing ABC doesn't mean abandoning their centralized GPU clusters; they can continue their existing training workflows and large-scale experiments, with ABC adding a coordination layer that connects to distributed data sources.

During inference, these same technologies can orchestrate secure AI predictions: user queries are encrypted and distributed to model owners, whose GPU enclaves execute their portions of the computation without being able to see the query. Secure multi-party computation coordinates how these encrypted model outputs combine, enabling multiple parties to jointly compute functions without revealing their individual inputs. Zero-knowledge proofs verify that each computation ran

correctly without exposing the underlying data or weights. Differential privacy adds mathematical guarantees throughout, preventing adversaries from reconstructing training data through repeated queries. Together, these PETs work in concert to create structured transparency, ensuring that neither training data, model weights, nor user queries are ever exposed in unencrypted form to unauthorized parties, while cryptographic attestation maintains the attribution chains.

***

While some platforms are new, the fundamental performance overhead of these technologies is comparable to what we already accept for HTTPS in web apps — a modest trade-off for accessing vastly more data. Modern optimizations continue to reduce this overhead through techniques like vectorized operations and hardware acceleration, making the computational costs increasingly negligible relative to the value of the expanded data access ABC enables.

Together, model partitioning and privacy infrastructure demonstrate that ABC is not a theoretical future possibility but an achievable present reality. The technical components have been scaled independently and are waiting to be assembled and integrated by researchers, engineers, and product managers to produce an AI system with ABC.

## Policy Recommendations: An ARPANET-style program to unlock a million times more data

The United States government has a unique opportunity to assemble the technologies already developed in its domestic research ecosystem and integrate them into a networked system that blossoms into a trillion-dollar market.

While the private sector plays a crucial role in creating data resources and transforming them into powerful algorithms, it does not necessarily possess optimal incentives to cooperate in maximizing America's AI potential through the establishment of an interoperable, open data network. Just as IBM, Bell Telephone, and Microsoft didn't necessarily have the right incentives to bring together the nation's supercomputers under the banner of TCP/IP, WWW, and HTTP, today's AI titans are naturally focused on their individual competitive advantages rather than building shared infrastructure. This creates a bottleneck where American AI is

limited to the data a single company can acquire (hundreds of terabytes) rather than leveraging the full potential of America's data resources (a million times more).

By incentivizing the development of AI systems with ABC, the US government can unlock access to these vast data resources while preserving ownership rights. Once developed, AI systems with ABC can transform the entire AI ecosystem. Labs will be incentivized to adopt ABC systems not out of obligation, but out of opportunity. Their data teams, which are increasingly hitting walls with traditional acquisition methods, will gain access to the exponentially larger datasets needed to push the boundaries of AI capability.

Luckily, the US has done this before. Just as ARPANET and later NSFNET ushered in the shift from isolated mainframes to networked personal computers, the US government should again create a multi-pronged approach to transition us from today's centralized AI towards network-sourced AI with ABC.

The ARPANET/NSFNET playbook offers a proven template for this transformation. For over two decades, multiple government agencies collaborated to create the internet through strategic coordination. ARPA (now DARPA) provided visionary leadership and seed funding, the NSF built bridges to widespread adoption through programs that subsidized early adopters, and the Commerce Department managed the critical transition of internet governance to the private sector.

To replicate ARPANET's success for Attribution-Based Control, we recommend the following actions:

1. **Establish an ABC Development Program within DARPA:** Following the original lean program structure that created ARPANET, DARPA should create a small, focused team to assemble the various technological components necessary to develop an AI system with ABC and conduct proof-of-concept and functional testing. Unlike other research institutions, DARPA is uniquely suited to focus on developing paradigm-shifting technologies that create entirely new capabilities, because its organizational structure and culture are specifically designed to fund high-risk, high-reward projects.

2. **Leverage NSF Grants to Subsidize Early Adopters:** NSF achieved explosive network growth for ARPANET/NSFNET through a multi-pronged adoption

strategy: free access, programs that subsidized university connections while requiring campus-wide deployment, and strategic positioning as general-purpose research infrastructure rather than just supercomputer access. The NSF could replicate this proven playbook for ABC by establishing a program or leveraging the operations of an existing program, such as the NAIRR, to subsidize early adopters who implement ABC systems while requiring institution-wide deployment. This approach would position ABC systems as essential research infrastructure for the AI era.

3. **Build International ABC Standards at NIST:** Just as TCP/IP became the global internet standard, NIST should lead the development of ABC technical standards that position American technologies as the international norm for data sharing. The AI Action Plan already emphasizes leading in international AI diplomacy; extend this to champion ABC as the framework for ethical, controlled data sharing globally.

The AI revolution stands at a crossroads: we can continue down the path of data silos and market resistance, or we can unleash a new era of shared prosperity through publicly supported data highways for AI, as envisioned by ABC. Just as ARPANET's visionary investment created today's trillion-dollar internet economy, a similar commitment to ABC could transform AI into a collaborative ecosystem where every data owner can contribute to and benefit from AI progress. The technological ingredients exist, the need is urgent, and history has shown us the way — America must act now to build the data economy that ensures every hospital, every research institution, and every company can contribute to and benefit from our shared AI future.

## Further resources


- OpenMined, "Introduction to Attribution-Based Control (ABC)," 2025
  Provides a more in-depth introduction to the concept of ABC.
- Andrew Trask et al., "Beyond Privacy Trade-offs with Structured Transparency," December 15, 2020
  Serves as the grounding theoretical framework.

*Andrew Trask is the Executive Director of OpenMined, a Senior Research Scientist at Google DeepMind, and a PhD candidate at the University of Oxford studying AI and privacy. He co-founded and chairs the United Nations Privacy Enhancing Technology Lab, and has trained over 30,000 students in AI and privacy through online courses.*

*Lacey Strahm is the Head of Policy at OpenMined, where she leads efforts to advance privacy-enhancing technologies for secure, distributed AI development. She previously worked on the US House Energy and Commerce Committee, focusing on data privacy and AI policy.*