AI’s data scarcity problem can be solved through existing techniques for attribution-based control

OpenMined

Abstract

AI systems are running out of data, but not because the data doesn’t exist. On the contrary, we estimate that 6 orders of magnitude more data exists, but data owners are unable or unwilling to relinquish it for AI training. We diagnose the problem as one of use bundling. A data owner cannot choose which AI predictions they wish to support — and which ones to decline to support (e.g. violations of their privacy, copyright, legal constraints, intellectual property, etc.). Instead, data owners must decide whether to help every AI prediction an AI model might generate in the future — or none of them. The vast majority of data owners opt for the latter. However, if training data owners could enforce which AI predictions they wanted to support and which they declined to support (perhaps on an ongoing basis), we argue that most would participate in at least some predictions. That is to say, an AI user could call upon 6 orders of magnitude more data in the world to support their AI predictions, and AI’s present data scarcity problem would be averted. This position paper argues that existing techniques can offer precisely this capability: attribution-based control.

1 Introduction

The future of AI hinges on the acquisition of more high-quality data, as reinforced by scaling laws [Kaplan et al., 2020, Hernandez et al., 2022, Muennighoff et al., 2025]. While recent advances in large language models (LLMs) have been fueled by massive publicly available datasets, this approach is reaching its limits [Villalobos et al., 2024]. At the same time, frontier models like GPT-4 [OpenAI et al., 2024], LLaMA-3 [Grattafiori et al., 2024], and Qwen [Qwen et al., 2025] are using less than 0.03% of the data in existence because such non-public datasets remain largely inaccessible.

We argue that this is not a matter of data scarcity, but a design failure. Non-public data remains unused largely due to governance and attribution challenges. The dominant AI paradigm relies on two operations that undermine data governance: *copying* and *addition*. Training requires duplicating data into centralized datasets, stripping contributors of control—a problem known as the **copy problem** Trask et al. [2024]. Simultaneously, deep learning blends inputs via feature mixing and parameter updates, eliminating traceability—creating the **addition problem**. These mechanisms erase control and attribution, posing irreversible risks to data owners such as privacy, legal exposure, or loss of IP; ultimately excluding sensitive, high-value data from AI systems.

We claim that solving the data bottleneck requires rethinking the core assumptions of AI systems to enable attribution-based control (ABC)—the ability for data owners to retain fine-grained, enforceable control over how their data is used in both training and inference and the ability for users of AI to decide which upstream sources they wish to utilize. However, ABC is not achievable under either of the extreme ends of the current governance paradigms, private centralized models or their opposite, open-source weights; as both open/closed weight paradigms offer model holders unilateral control over model use.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

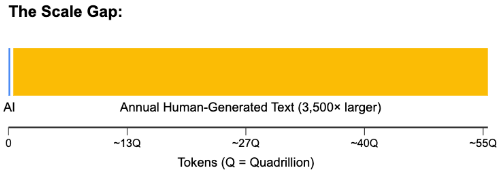

Figure 1: *The Scale Gap*: Annual human-generated text ($\approx 55Q$ tokens) to the largest dataset used for AI training ($\approx 20T$ tokens).

We survey a set of emerging technical mechanisms that pave the way toward ABC. Advances in differential privacy [Dwork, 2006], federated learning [McMahan et al., 2023], secure computation [Zhou et al., 2024], and retrieval-augmented architectures [Borgeaud et al., 2022, Izacard et al., 2022] show early promise in enabling data use without full disclosure. Moreover, architectures like PATE [Papernot et al., 2018] and RETRO [Borgeaud et al., 2022] suggest that it is possible to decouple parameter sharing from data sharing, enabling attribution to be preserved across systems.

Building on these ideas, we illustrate the path towards a networked approach for AI where models do not fully memorize data in centralized weights, but can access and query distributed, live, public and private data sources (and weights derived from them). This live approach allows data owners to opt in or out of specific predictions, enforces usage restrictions cryptographically, and enables dynamic governance across the AI supply chain. We argue that AI systems that implement attribution-based control can unlock the vast, underutilized data resources of the world; data that can power safer, more robust, and more equitable AI systems.

2 Barriers to further data access

The scale of non-accessible data Frontier models [Grattafiori et al., 2024, Qwen et al., 2025] have been trained on datasets of approximately 5–20 trillion tokens. Using RedPajama's token-to-byte estimate [Weber et al., 2024], Qwen2.5-7B's 18 trillion tokens require 90TB of storage [Qwen et al., 2025]. Yet this is minuscule compared to the volume of human-generated text. Cummins [2024] estimate that over 1850 trillion tokens are created daily via email and instant messages alone. In total, humans produce about 150 trillion tokens per day—over 55 quadrillion annually—roughly 3,500 times more than the largest training sets to date. This is still only a fraction of the global digital corpus, projected to reach 180 zettabytes by 2025 [Mider, 2024, Taylor, 2024]. By contrast, today's usable open datasets—Common Crawl (450TB, Wenzek et al. [2019]) and the Internet Archive's Wayback Machine[1] (100PB, [Kahle, 2024])—are orders of magnitude smaller. This raises a fundamental question: *why is so much valuable data left untapped by current AI systems?*

Risks & barriers that silo data There are significant barriers preventing data owners from contributing their data to AI systems, primarily due to the risks associated with granting access. Such risks include legal risks around privacy breaches, infringement of intellectual property rights, regulatory liability or downstream data misuse [Wang et al., 2024, Trask et al., 2024]. For example, medical institutions might wish to control who gets access to patient information and under what conditions, to adjust for privacy risk [Malin B, 2018]. Certain biological datasets carry inherent dual-use potential, requiring strict access controls to prevent both deliberate misappropriation and unintended harmful applications [Sandbrink, 2023]. Moreover, Longpre et al. [2024b] shows that access to creative works is in decline due to restrictive or unclear licenses, in an attempt to defend IP rights. This is a common pattern - Youssef et al. [2023] notes that many of the barriers to unlock medical data can be reduced down to two substantive concerns: maintaining attribution and control.

Today's AI systems fail to enable robust attribution [Huang et al., 2025]. In addition, the black-box usage of data at training and inference time prohibits data owners from observing, enforcing, and validating whether their data preferences (i.e., licenses, data policies) are appropriately applied [Katzy et al., 2024]. Currently, data owners must rely on and place their trust in AI operators to enforce blanket usage policies. They can only attempt to validate the efficacy of such enforcement through techniques like watermarking [Wei et al., 2024] or various black-box membership-inference attacks

[1] https://web.archive.org/

[Hu et al., 2022], which are often error-prone. In this current paradigm, data owners do not have enforceable, direct control over how an AI system uses their data (i.e. which prompts an AI model is allowed to respond to using information learned from their data) because once data owners grant access to a copy of their data to an AI operator, the AI operator then has full control over the use of that data in training or prediction. This presents two core problems for data owners:

- They cannot observe or approve of outputs involving their data.
- They cannot withdraw support for misaligned uses.

We argue that the lack of attribution-based control in both open-source and closed-source AI systems reinforces barriers for data owners, keeping much of their data inaccessible.

Lack of economic incentives In addition to preserving the barriers, the lack of attribution fails to motivate data owners to participate. In particular, as AI capabilities reach human-level performance on specific tasks [OpenAI et al., 2024], the fundamental premise of our existing intellectual property (IP) regime and its traditional profit mechanism are being disrupted. Creative workers have described this challenge as a *"double bind"*: the desire to embrace AI as a productivity amplifier, and the simultaneous fear of the same AI replacing them [Harvard Law Review, 2025, Tang et al., 2025]. In one such example, Hollywood writers went on strike, refusing to contribute their IP to AI systems out of fear of being replaced [The New York Times Editorial Board, 2023]. Publishers have also adapted a more defensive stance by specifying new restrictive rules to their content via robots.txt [Longpre et al., 2024b], and enforcing their rights through lawsuits [The New York Times Company, 2023]. These communities repeatedly surface a shared fear regarding their participation in AI systems: data owners cannot exercise attribution-based control.

Broken AI supply chain AI systems are becoming primary interfaces for information access [Zhang, 2024], distancing users from data owners [Arriagada and Ibáñez, 2020]. This disconnect breaks feedback and engagement loops which are crucial for improving content quality, and affects end-users who are exposed more to hallucinations or disinformation [Dziri et al., 2022, Rashkin et al., 2021, Vaccari and Chadwick, 2020]. Unlike journals or search engines that cite sources [Zuccon et al., 2023], AI systems often omit attribution. For example, a user asking about a medical condition might want to choose only peer-reviewed sources from scientific journals, excluding mentions on social media or public chat rooms [Gravel et al., 2023, Bhattacharyya et al., 2023]. Researchers whose work trains these models lose visibility and credit, undermining academic norms where citations signal value [Devriendt et al., 2021]. Without attribution, users doubt its authenticity, data owners lose incentives, and AI output declines—widening the gap between creators and their audiences.

3 The limitations of today's AI systems

In his work, Trask et al. [2024] explores the cause why such barriers persist for data owners and identifies the underlying problems that prohibit control and collaboration: information cannot be controlled once it is shared or bundled, introducing the *copy* problem. We investigate how this framework can be extended to today's AI systems and their lack of attribution and control.

3.1 The lack of attribution: *the addition problem*

The problem of attribution starts with data collection and processing, where source information is often not preserved [Longpre et al., 2023]. While better curation efforts can improve this premise, AI systems would still fail to maintain individual attribution. The reason emerges from deep learning's foundational premise: algorithms should learn everything from scratch through layers of (largely) unrestricted feature mixing on raw data [Goodfellow, 2017]. The root cause for the loss of traceability is addition within deep learning systems. For example, consider the source data to be 1 and 6, and observe the following sums:

$$1 + 6 = 7;$$

$$2 + 5 = 7$$

Addition obscures source identities by irreversibly entangling inputs. It is pervasive in training—used in feature merging, gradient aggregation, and weight updates—dispersing information across weights in complex patterns [Le et al., 2012]. Consequently, deep learning techniques erase attribution information during feature merging, gradient aggregation, and weight-updates. And attempting to track provenance across these operations becomes infeasible due to exponential growth in attribution paths, up to $\Omega(w * n)^t$ for w weights, n examples, and t steps, making brute-force tracing intractable.

Existing research efforts focus on overcoming this problem by tracing predictions back to training data through the use of influence function approximations [Koh and Liang, 2020, Schioppa et al., 2021], TracIn [Pruthi et al., 2020], or the removal of influential data points through machine unlearning [Nguyen et al., 2024]. However, they fundamentally struggle because they attempt to recover inforamtion which is already lost through addition. Consequentnly, influence functions and machine unlearning remain unsolved challenges in the literature [Nguyen et al., 2023, K and Søgaard, 2021].

3.2 The lack of control: *the copy problem*

When Party A gives information to Party B, Party A loses control over all future uses of that piece of information [Trask et al., 2024]. AI systems remain bound by this same fundamental limitation: to help train an AI model, data owners need to give a copy of their data to AI model operators, relinquishing control over how their data might be used or distributed after it is synthesized into the model. Similarly, when a model owner gives their model to someone else, they lose control over how that model might be used going forward.

Today, AI systems are distributed through two main paradigms: open-source and closed-source. The dichotomy between these two paradigms has become a proxy for broader concerns around privacy, disinformation, copyright, safety, bias, and alignment. However, we argue that while both prioritize specific needs of data owners, neither facilitates sufficient control or transparency for data owners:

Privacy and data rights: Privacy is one of the main barriers for data owners: closed-source proponents show that centralized control enables better privacy protection through careful data handling and access controls [Deng et al., 2024], while open-source proponents counter that transparency allows public verification of privacy measures [Hintersdorf et al., 2025, Wang et al., 2024]. However, as private data remains largely inaccessible, this fails to address the needs of data owners: they cannot control how their information is used and enforce their own boundaries between proper and improper use, as the control is delegated to external decision-makers or lost by design.

Copyright and intellectual property: Both closed and open-source models can uphold data owners' IP rights through licensing and usage restrictions. However, without any enforcing mechanism, this becomes a problem of trust and transparency: data owners need to trust such models follow compliance without a verification mechanism [Cen and Alur, 2024]. While open-source models distribute their models under highly permissive, unrestricted licenses that benefit the AI community, such licenses sometimes contradict the preferences of data owners, which are so vast and varied that a single license may not be able to fully express user desires [Longpre et al., 2024a]. These insufficiencies stem from delegating control or doing so without appropriate verifiable mechanisms.

Safety and misuse prevention: Closed-source proponents advocate that centralized oversight prevents harms [Deng et al., 2024], in contrast with open-source's claim that collective scrutiny better identifies risks [Hintersdorf et al., 2025]. In both cases, data owners cannot intervene if their data contributes to harmful outcomes to withdraw support: for closed-source, they trust AI operators for adjustments, while open-source AI's unrestricted distribution disables any post-hoc intervention.

Bias and representation: Closed-source teams employ careful curation to prevent bias [OpenAI et al., 2024]. Open-source teams suggest community oversight promotes fair representation and transparency [Eiras et al., 2024]. While complementary, neither open or closed source paradigms empower affected communities to control how their perspectives inform predictions. Furthermore, neither paradigm empowers AI end-users to decide on which sources they trust. Instead, both paradigms delegate control over bias and representation to whomever trains an AI model.

4 Enable attribution through concatenation and separability

Addition erases source attribution. We previously showed how this undermines traceability via feature mixing in deep learning. In this section, we present a simple alternative to addition - concatenation. Through various forms of concatenation, *undesired additions* can be replaced while preserving the dense information that is critical for AI systems to build its reasoning abilities.

4.1 Concatenation along natural boundaries

Concatenation, as opposed to addition, carries some properties that are particularly of interest to us: it preserves the source information while still allowing information to be combined later.

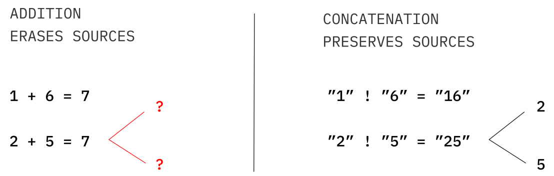

Figure 2: Illustrative example between addition and concatenation.

Notice how in Fig. 2, concatenation preserves source information, numbers ”1” and ”6”, but how their identity is destroyed when addition takes place. This erasure is the mechanism that removes attribution information within deep neural networks. However, we pose the following question: *Can concatenation reduce this problem in practice within deep learning systems?*

Deep learning’s central hypothesis would suggest we cannot. Addition is needed to densely mix features in order to learn the powerful correlations and representations that give deep learning its predictive capability. Presumably, deep learning maps multiple distinct-yet-related concepts into a shared feature (e.g., an image classifier combines features for detecting ears and fur to model specific animals) [Bau et al., 2017]. Yet, not all concepts require that representation power: certain concepts are densely intertwined, while others are near-orthogonal.

Certain information patterns (such as *grammar rules, logical operations, morphological patterns, edges in images*) appear across high percentages of data points. For these near-ubiquitous features, addition-based mixing is often suitable, and attribution is less critical since they act as shared computational tools (i.e. nobody owns English grammar or the principles of arithmetic — it is certainly in the commons). In contrast, sparse information—such as *the capital of France*, *chess rules*, or *the proprietary manufacturing process of a specific company*—tends to be source-specific and context-dependent. As noted by Chomsky [Chomsky, 1969], while common structures shape expression, the knowledge conveyed is often naturally partitioned by topic and source.

Assuming this is true, it means that a section of a neural network could be made sparse, namely the part which stores less common concepts (specific facts, domain expertise, semantic information, etc.), while a section needs to remain dense to store and synthesize concepts which are ubiquitous across a statistical distribution (logic, reasoning, grammar, etc.). This sparsity could drive attribution if addition is at least partially replaced with concatenation. The key question becomes: *how could one train a neural network which successfully partitions information into sparse and dense sections?*

4.2 Separable model architecture

To reduce undesired additions and replace them with concatenation, we need to achieve a successful partitioning of information into sparse and dense sections. To provide evidence that this is possible, we look into two important prior works: differential privacy and retrieval-based architectures.

Differential Privacy A technique inspired from the privacy space, differential privacy [Dwork, 2006], could provide a principled way to measure and control which features benefit from dense mixing versus sparse representation. Existing rich literature on differential privacy mechanisms, such as noise addition, gradient clipping or sensitivity bounds [Abadi et al., 2016], naturally filter out information unique to specific sources.

We suggest that the same mechanisms can be used to calibrate for the opposite effect that privacy erases: attribution. This is possible by leveraging its generic ability to systematically identify and filter out information that appears infrequently across many partitions of information which are being aggregated [Dwork, 2006]. We suggest that this ability could be applied to partitions of training data used in neural networks. A key architectural insight is that the model may learn correlations from information that appears consistently across sources and passes through privacy mechanisms, while source-specific knowledge is naturally isolated outside of the dense features.

Retrieval-based architectures Recent work, such as RETRO and ATLAS [Borgeaud et al., 2022, Izacard et al., 2022], provides initial evidence that knowledge naturally separates into general computational patterns: for instance, we can note how the Transformer reads its source-specific information from a database of vector embeddings that is hosted outside of the model's weights. In particular, these works show that memory can be outsourced and replaced by an external non-parametric knowledge source by employing a retrieval-augmented architecture without loss in performance.

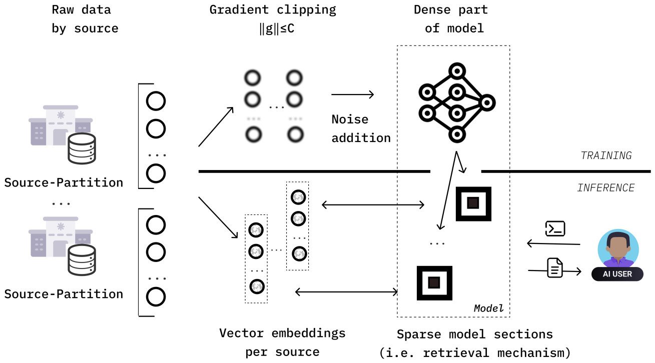

Figure 3: Example illustration of how we can separate the common and source-specific knowledge during training through differential privacy (upper part) and how this extends to the usage of the sparse, source-specific knowledge through retrieval mechanisms during inference (lower part).

These early pointers suggest that data owners might get a new lever of control through the tunable nature of the differential privacy mechanisms, as well as opt-in/opt-out ability enabled by the source-separation implemented by such model architectures (Fig. 3). However, we recognize that this framework builds on top of the following assumptions that will require empirical validation:

1. **General Separability Hypothesis:** Neural networks can separate common vs. source-specific information using privacy filters.
2. **Source-specific Separability Hypothesis**: Sparse model sections can be further partitioned by source, over which data providers can exercise their control.
3. **Source-specific Synthesizability Hypothesis:** Dense sections from different sources can be combined efficiently.

We consider these assumptions a requirement to technically enforce attribution and control. However, an immediate concern alongside these hypothesis might be *Could a separable model architecture achieve similar performance as traditional AI models?*.

4.3 Performance and attribution trade-offs

To understand the possible performance constraints of such a system, we investigate how this compares with existing model architectures deployed widely in the real-world and their existing support for attribution and control:

1. **Traditional deep learning systems** achieve state-of-the-art performance through unrestricted parameter sharing. On MNIST, non-private models reach 98.3% accuracy, while an identically parameterized model with differential privacy drops to 90% [Abadi et al., 2016]. However, this has another fundamental cost: attribution clarity is lost.
2. **Federated systems** maintain clear attribution boundaries by keeping data within its original institutional siloes. While this does not come with cost when data is IID, the ability to preserve attribution comes with a measurable cost for non-IID data: MNIST accuracy drops from 98.69% to 96.29% [Zhao et al., 2018] - an 2.4% degradation from simply partitioning non-IID data, training separate models, and averaging them. This surfaces a tension between attribution boundaries and the need to learn cross-source patterns.
3. **Pure memory-based approaches** like k-NN [Cunningham and Delany, 2021] provide perfect attribution by directly linking predictions to source examples. While these systems can achieve high accuracy on their training distribution through exact matching, they fundamentally cannot generalize beyond the patterns explicitly present in their memory banks despite achieving 97.2% accuracy [Grover and Toghi, 2019].

These examples suggest an unavoidable tradeoff between model performance, attribution, and generalization ability. If these tradeoffs are fundamental limits rather than engineering challenges, this would permanently constrain AI's access to siloed data. However, prior work [Borgeaud et al., 2022, Izacard et al., 2022] provides further evidence: RETRO (7.5B parameters) matches GPT3's (175B parameters) performance, while using 25x fewer parameters. This suggests that these properties can be achieved at the same: dense-level performance, dense-level generalization and memory-level attribution. Similar findings span other separable model architectures that employ various merging mechanisms, such as PATE, Federated RAG or Git Re-Basin.

PATE (Private Aggregation of Teacher Ensembles) [Papernot et al., 2018] achieves 98.5% accuracy on MNIST with differential privacy, close to the non-private 99.2%, and uniquely supports source attribution via individual teachers—offering stronger guarantees than RETRO or ATLAS. Federated RAG [Hou et al., 2025] also boosts both attribution and performance. Git Re-Basin [Ainsworth et al., 2023] further shows models trained independently on similar distributions can be merged without accuracy loss, challenging the need for joint training.

Together, these results point to a systematic advantage: architectures with explicit information paths can preserve attribution and enable transparency without sacrificing performance.

5 Enable control through structured transparency

The prior section illustrates how different architectures can preserve attribution while retaining performance. However, data owners still need to trust model operators that they will comply with their preferences (e.g., licenses). Even if a model preserves its attribution paths, it is still unilaterally controlled by whoever possesses a copy of the model. This creates a trust barrier that blocks the potential of such architectures.

We refer to the framework of structured transparency and a variety of recently proposed technologies from cryptography and distributed systems to address such problems:

1. **Input Privacy:** ***Preserve control and privacy during computation*** Federated learning [Gabrielli et al., 2023] addresses the copy problem by moving computation to the data, enabling models to train or query without centralizing information. Each data source could *run a web server they control*, deciding case-by-case whether to contribute. Complementary methods such as secure enclaves with attestation [Costan and Devadas, 2016], homomorphic encryption via key-value stores [Cheon et al., 2017], and secure multi-party computation (SMPC) [Xiong et al., 2021, Wagh et al., 2018] further protect data by ensuring it remains encrypted or partitioned throughout the computation process, preventing unwanted duplication or exposure.
2. **Output Privacy:** ***Avoid private data disclosure*** Whilst the above can ensure full privacy of the data during computation, the AI prediction may be vulnerable to reverse engineering attacks. To preserve output privacy, differential privacy is a great candidate for both training [Dwork, 2006], and retrieval [Koga et al., 2025].

3. **Input + Output Verification:** ***Verify authenticity & provenance*** If the above are properly implemented, an AI user is forced to rely upon vast collections of information they cannot see. To ensure information is real and verify its provenance, cryptographic techniques such as zero-knowledge proofs [Lavin et al., 2024] and attestation chains [Costan and Devadas, 2016] let data sources prove properties (e.g., peer review) without revealing content. Verified computations [Woodcock et al., 2024] attest that inputs stay unchanged, while public-key signatures let individuals validate claims (e.g., a journalist signing their article).
4. **Flow governance** The last challenge is ensuring that despite the large scale of the computation - between AI users and a large number of active sources - the right controls are distributed to the right parties. If such methods are successful, methods like SMPC can provide control which is both distributed and group-enforceable through techniques like additive secret sharing [Xiong et al., 2021].

Together, these five guarantees - input/output privacy, input/output verification and flow governance - lay the structured transparency foundation that can unlock ABC-enabled AI.

6 ABC-enabled AI: a new paradigm for distribution

To reduce the barriers that prohibit data owners to participate in AI and reinstate their incentives, we contend that it is necessary for AI models' to possess attribution-based control (ABC) (Fig. 4):

1. AI users control which data sources they rely upon for prediction
2. Data sources control which predictions they would like to support or decline

ABC implies that AI models's knowledge need to be partitioned by source, otherwise sources cannot independently exercise their control, neither during training or inference. We showed earlier that such architectures not only are possible, but there is evidence to show it can operate at scale. However, for such control to be possible, *we need to challenge the fundamental assumption that AI systems must exist as copyable files* - as existing paradigms (open vs closed) fail to enable the necessary control and transparency.

We propose the idea that, if ABC-enabled AI operated on a live, networked architecture, we can overcome this core trade-off between trust and control. Assuming each data source is queried or used directly from its original source, over a network, we prevent copying and can rely on structured transparency's guarantees to enable both visibility into how systems operate and precise control over how information is used. This resolves the false choice between open and closed source AI by creating a third option: **network-sourced AI systems** that are simultaneously transparent (through verification) and controlled (through cryptography), by implementing attribution-based control.

We believe this paradigm shift has far-reaching implications: if data owners can specify their enforceable preferences on an ongoing basis to control for privacy violations, copyright, legal constraints or IP, they could choose to participate in AI, unlocking six orders of magnitude more data.

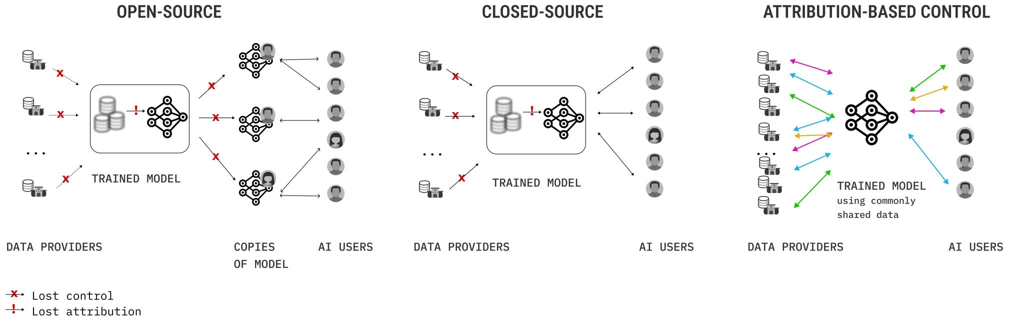

Figure 4: Visual comparison between open/closed-source AI systems and ABC-enabled AI.

7 Alternative views

The problem of accessing more data is pervasive across the industry. We recognize three alternative views that are most commonly encountered in the industry:

Centralized control. As discussed above, existing research efforts to address privacy and control are still concentrated on model architectures that are unilaterally controlled, with data owners' preferences being respected through AI owner benevolence. Without being enforceable, compliance is imperfect, with various precedents surfacing into the public space, including lawsuits [The New York Times Company, 2023] or outrageous findings [Wang et al., 2024]. We recognize that this approach has been successful in acquiring more data and extensive work has been deployed in creating effective guardrails and safety mechanisms for such models. However, these are still subject to privacy and security concerns [Wang et al., 2024] and while economic incentives are manually created for each data owner that contributes its data through direct licensing, we argue that this creates a barrier for the open advancement of AI and cannot be a scalable approach to capture the 6+ orders of magnitude of data that is currently siloed.

Use of exclusively public and synthetic data Existing efforts are being concentrated on generating synthetic data to overcome the scarcity of training data [Bauer et al., 2024], given to the privacy issues and expensive acquisitions. Additionally, there are complementary efforts to identify and correct licensing issues with public data, including lack of specified licenses or asymmetries between one's licenses and robots.txt [Longpre et al., 2024a], which are of great benefit to the entire AI community. However, this solution is temporary, as highlighted by Villalobos et al. [2024], because the existing LLM development trends will likely exhaust the public human data available in a short timeline. Prior work presents limitations of synthetic data, such as error amplification and diversity [Chen et al., 2024, Shumailov et al., 2024], but we find this line of work complementary with the efforts of unlocking data siloes, where synthetic data generation can be a viable solution in the technical framework of structured transparency.

Infrastructure-based data sharing Another emerging approach seeks to address access barriers by standardizing secure data-sharing infrastructure, often via federated learning frameworks [McMahan et al., 2023, Gabrielli et al., 2023] or data trusts. These systems aim to enable computation without raw data transfer by keeping data at rest and moving models to the data. This model can preserve privacy and offer a degree of control, particularly for institutional stakeholders. However, adoption remains limited due to high technical complexity, limited interoperability, and the need for strong governance frameworks to ensure compliance and fairness. Furthermore, these systems prioritize control during training, but ultimately delegate control of the aggregated model to a central party, falling down the path of unilateral control as data owners lose the ability to say which AI predictions they would like to support. Without attribution-based control, such systems still preserve the barriers for data owners, albeit in a slightly more distributed form.

8 Conclusion

At the heart of today's AI challenges—copyright, privacy, misattribution, disinfonrmation, and data scarcity—lies a deeper structural flaw: the absence of attribution-based control (ABC). Our current systems either centralize control behind corporate walls or abandon it entirely through open release. Both models fail to empower those who matter most: data owners and end users.

We argue for a new path. By replacing additive architectures with separable, attribution-preserving designs, and enforcing control through cryptographic guarantees (not trust) we unlock a new paradigm: AI as a live, distributed network. Not just a neural newtork, but a communication network — an interconnected network of neural networks. ABC transforms the acquisition of training data from a one-time extractive act into a sustainable, collaborative flow.

If adopted, we hold that this shift could unlock six orders of magnitude more data, enable safer and more aligned AI systems, and repair the broken supply chain between source and users. Attribution-based control is not just a technical fix—it may be the only sustainable path forward. As data grows more valuable and its misuse more consequential, enabling ABC becomes increasingly urgent. We believe the research community must begin treating attribution-based control as a first-class design constraint—not an afterthought. Building this foundation now is essential to ensuring the next generation of AI is collaborative, transparent, and aligned by design.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, October 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries, 2023. URL https://arxiv.org/abs/2209.04836.

Arturo Arriagada and Francisco Ibáñez. "you need at least one picture daily, if not, you're dead": Content creators and platform evolution in the social media ecology. *Social Media + Society*, 6: 205630512094462, 07 2020. doi: 10.1177/2056305120944624.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations, 2017. URL https://arxiv.org/abs/1704.05796.

André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. Comprehensive exploration of synthetic data generation: A survey, 2024. URL https://arxiv.org/abs/2401.02524.

M. Bhattacharyya, V. M. Miller, D. Bhattacharyya, and L. E. Miller. High rates of fabricated and inaccurate references in chatgpt-generated medical content. *Cureus*, 15(5):e39238, May 2023. doi: 10.7759/cureus.39238.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens, 2022. URL https://arxiv.org/abs/2112.04426.

Sarah H. Cen and Rohan Alur. From transparency to accountability and back: A discussion of access and evidence in ai auditing, 2024. URL https://arxiv.org/abs/2410.04772.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abdin. On the diversity of synthetic data and its impact on training large language models, 2024. URL https://arxiv.org/abs/2410.15226.

Jung Hee Cheon, Andrey Kim, Miran Kim, and Yongsoo Song. Homomorphic encryption for arithmetic of approximate numbers. In Tsuyoshi Takagi and Thomas Peyrin, editors, *Advances in Cryptology – ASIACRYPT 2017*, pages 409–437, Cham, 2017. Springer International Publishing. ISBN 978-3-319-70694-8.

Noam Chomsky. *Aspects of the Theory of Syntax*. The MIT Press, Cambridge, Massachusetts, March 1969. ISBN 9780262530071. Paperback edition.

Victor Costan and Srinivas Devadas. Intel SGX explained. Cryptology ePrint Archive, Paper 2016/086, 2016. URL https://eprint.iacr.org/2016/086.

Mark Cummins. How much LLM training data is there, in the limit? Educating Silicon (blog), May 2024. URL https://www.educatingsilicon.com/2024/05/09/how-much-llm-training-data-is-there-in-the-limit/.

Pádraig Cunningham and Sarah Jane Delany. k-nearest neighbour classifiers - a tutorial. *ACM Computing Surveys*, 54(6):1–25, July 2021. ISSN 1557-7341. doi: 10.1145/3459665. URL http://dx.doi.org/10.1145/3459665.

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. Masterkey: Automated jailbreaking of large language model chatbots. In *Proceedings 2024 Network and Distributed System Security Symposium*, NDSS 2024. Internet Society, 2024. doi: 10.14722/ndss.2024.24188. URL http://dx.doi.org/10.14722/ndss.2024.24188.

Thibaut Devriendt, Mahsa Shabani, and Pascal Borry. Data sharing in biomedical sciences: A systematic review of incentives. *Biopreservation and Biobanking*, 19(3):219–227, Jun 2021. doi: 10.1089/bio.2020.0037. Epub 2021 Feb 11.

Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-35908-1.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. On the origin of hallucinations in conversational models: Is it the datasets or the models? In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.387. URL https://aclanthology.org/2022.naacl-main.387/.

Francisco Eiras, Aleksandar Petrov, Bertie Vidgen, Christian Schroeder de Witt, Fabio Pizzati, Katherine Elkins, Supratik Mukhopadhyay, Adel Bibi, Botos Csaba, Fabro Steibel, Fazl Barez, Genevieve Smith, Gianluca Guadagni, Jon Chun, Jordi Cabot, Joseph Marvin Imperial, Juan A. Nolazco-Flores, Lori Landay, Matthew Jackson, Paul Röttger, Philip H. S. Torr, Trevor Darrell, Yong Suk Lee, and Jakob Foerster. Near to mid-term risks and opportunities of open-source generative ai, 2024. URL https://arxiv.org/abs/2404.17047.

Edoardo Gabrielli, Giovanni Pica, and Gabriele Tolomei. A survey on decentralized federated learning, 2023. URL https://arxiv.org/abs/2308.04604.

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks, 2017. URL https://arxiv.org/abs/1701.00160.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan,

Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The

llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jocelyn Gravel, Madeleine D'Amours-Gravel, and Esli Osmanlliu. Learning to fake it: Limited responses and fabricated references provided by chatgpt for medical questions. *Mayo Clinic Proceedings: Digital Health*, 1(3):226–234, 2023. ISSN 2949-7612. doi: https://doi.org/10.1016/j.mcpdig.2023.05.004. URL https://www.sciencedirect.com/science/article/pii/S2949761223000366.

Divas Grover and Behrad Toghi. Mnist dataset classification utilizing k-nn classifier with modified sliding-window metric, 2019. URL https://arxiv.org/abs/1809.06846.

Harvard Law Review. Artificial intelligence and the creative double bind. *Harvard Law Review*, 138(6):1585–1608, April 2025. URL https://harvardlawreview.org/print/vol-138/artificial-intelligence-and-the-creative-double-bind/. Developments in the Law—Artificial Intelligence, Chapter Two.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Ben Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. Scaling laws and interpretability of learning from repeated data, 2022. URL https://arxiv.org/abs/2205.10487.

Dominik Hintersdorf, Lukas Struppek, and Kristian Kersting. Balancing transparency and risk: An overview of the security and privacy risks of open-source machine learning models. In Bernhard Steffen, editor, *Bridging the Gap Between AI and Reality*, pages 269–283, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73741-1.

Charlie Hou, Mei-Yu Wang, Yige Zhu, Daniel Lazar, and Giulia Fanti. Private federated learning using preference-optimized synthetic data, 2025. URL https://arxiv.org/abs/2504.16438.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s), September 2022. ISSN 0360-0300. doi: 10.1145/3523273. URL https://doi.org/10.1145/3523273.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models, 2022. URL https://arxiv.org/abs/2208.03299.

Karthikeyan K and Anders Søgaard. Revisiting methods for finding influential examples, 2021. URL https://arxiv.org/abs/2111.04683.

Brewster Kahle. A Message from Internet Archive Founder, Brewster Kahle. https://archive.org/donate, 2024. Donation page outlining the Archive's mission, impact (99 PB+, 625 billion webpages, 38 million texts, 14 million audio recordings) and ways to support. Federal Tax ID: 94-3242767.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Jonathan Katzy, Răzvan-Mihai Popescu, Arie van Deursen, and Maliheh Izadi. An exploratory investigation into code license infringements in large language model training datasets, 2024. URL https://arxiv.org/abs/2403.15230.

Tatsuki Koga, Ruihan Wu, and Kamalika Chaudhuri. Privacy-preserving retrieval-augmented generation with differential privacy, 2025. URL https://arxiv.org/abs/2412.04697.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions, 2020. URL https://arxiv.org/abs/1703.04730.

Ryan Lavin, Xuekai Liu, Hardhik Mohanty, Logan Norman, Giovanni Zaarour, and Bhaskar Krishnamachari. A survey on the applications of zero-knowledge proofs, 2024. URL https://arxiv.org/abs/2408.00243.

Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean, and Andrew Y. Ng. Building high-level features using large scale unsupervised learning, 2012. URL https://arxiv.org/abs/1112.6209.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity, 2023. URL https://arxiv.org/abs/2305.13169.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi (Alexis) Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6(8):975–987, August 2024a. ISSN 2522-5839. doi: 10.1038/s42256-024-00878-8. URL https://doi.org/10.1038/s42256-024-00878-8.

Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, Kevin Klyman, Christopher Klamm, Hailey Schoelkopf, Nikhil Singh, Manuel Cherep, Ahmad Anis, An Dinh, Caroline Chitongo, Da Yin, Damien Sileo, Deividas Mataciunas, Diganta Misra, Emad Alghamdi, Enrico Shippole, Jianguo Zhang, Joanna Materzynska, Kun Qian, Kush Tiwary, Lester Miranda, Manan Dey, Minnie Liang, Mohammed Hamdy, Niklas Muennighoff, Seonghyeon Ye, Seungone Kim, Shrestha Mohanty, Vipul Gupta, Vivek Sharma, Vu Minh Chien, Xuhui Zhou, Yizhi Li, Caiming Xiong, Luis Villa, Stella Biderman, Hanlin Li, Daphne Ippolito, Sara Hooker, Jad Kabbara, and Sandy Pentland. Consent in crisis: The rapid decline of the ai data commons, 2024b. URL https://arxiv.org/abs/2407.14933.

Goodman K Malin B. Between access and privacy: Challenges in sharing health data, 2018.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2023. URL https://arxiv.org/abs/1602.05629.

Daniel Mider. Open source intelligence on the internet – categorisation and evaluation of search tools. *Internal Security Review*, 2024(Issue 31 (16)):383–412, 2024. ISSN 2080-1335. doi: 10.4467/20801335PBW.24.030.20807. URL https://ejournals.eu/en/journal/przeglad-bezpieczenstwa-wewnetrznego/article/open-source-intelligence-on-the-internet-categorisation-and-evaluation-of-search-tools.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. Olmoe: Open mixture-of-experts language models, 2025. URL https://arxiv.org/abs/2409.02060.

Elisa Nguyen, Minjoon Seo, and Seong Joon Oh. A bayesian approach to analysing training data attribution in deep learning, 2023. URL https://arxiv.org/abs/2305.19765.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning, 2024. URL https://arxiv.org/abs/2209.02299.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny

Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate, 2018. URL `https://arxiv.org/abs/1802.08908`.

Garima Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. Estimating training data influence by tracing gradient descent, 2020. URL `https://arxiv.org/abs/2002.08484`.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In Chengqing Zong, Fei Xia, Wenjie

Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.58. URL `https://aclanthology.org/2021.acl-long.58/`.

Jonas B. Sandbrink. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools, 2023. URL `https://arxiv.org/abs/2306.13952`.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. Scaling up influence functions, 2021. URL `https://arxiv.org/abs/2112.03052`.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, July 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07566-y. URL `https://doi.org/10.1038/s41586-024-07566-y`.

Yuying Tang, Haotian Li, Minghe Lan, Xiaojuan Ma, and Huamin Qu. Understanding screenwriters' practices, attitudes, and future expectations in human-ai co-creation, 2025. URL `https://arxiv.org/abs/2502.16153`.

P. Taylor. Amount of data created, consumed, and stored 2010–2023, with forecasts to 2028. `https://www.statista.com/statistics/871513/worldwide-data-created/`, 2024. Accessed: 2025-04-16.

The New York Times Company. Complaint, *N. Y. Times Co. v. Microsoft Corp.*, no. 23-cv-11195 (s.d.n.y. filed dec. 27, 2023). Case 1:23-cv-11195, Doc. 1, U.S. District Court for the Southern District of New York, December 2023. URL `https://hh-law.com/wp-content/uploads/2024/07/New-York-Times-complaint.pdf`. Court filing, 69 pp.

The New York Times Editorial Board. Hollywood's deal with screenwriters just rewrote the rules around A.I. *The New York Times*, September 2023. URL `https://www.nytimes.com/2023/09/29/opinion/wga-strike-deal-aijobs.html`. Opinion essay.

Andrew Trask, Emma Bluemke, Teddy Collins, Ben Garfinkel Eric Drexler, Claudia Ghezzou Cuervas-Mons, Iason Gabriel, Allan Dafoe, and William Isaac. Beyond privacy trade-offs with structured transparency, 2024. URL `https://arxiv.org/abs/2012.08347`.

Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6 (1):2056305120903408, 2020. doi: 10.1177/2056305120903408. URL `https://doi.org/10.1177/2056305120903408`.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data, 2024. URL `https://arxiv.org/abs/2211.04325`.

Sameer Wagh, Divya Gupta, and Nishanth Chandran. SecureNN: Efficient and private neural network training. Cryptology ePrint Archive, Paper 2018/442, 2018. URL `https://eprint.iacr.org/2018/442`.

Shang Wang, Tianqing Zhu, Bo Liu, Ming Ding, Xu Guo, Dayong Ye, Wanlei Zhou, and Philip S. Yu. Unique security and privacy threats of large language model: A comprehensive survey, 2024. URL `https://arxiv.org/abs/2406.07973`.

Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models, 2024. URL `https://arxiv.org/abs/2411.12372`.

Johnny Tian-Zheng Wei, Ryan Yixiang Wang, and Robin Jia. Proving membership in llm pretraining data via data watermarks, 2024. URL `https://arxiv.org/abs/2402.10892`.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data, 2019. URL https://arxiv.org/abs/1911.00359.

Jim Woodcock, Mikkel Schmidt Andersen, Diego F. Aranha, Stefan Hallerstede, Simon Thrane Hansen, Nikolaj Kuhne Jakobsen, Tomas Kulik, Peter Gorm Larsen, Hugo Daniel Macedo, Carlos Ignacio Isasa Martin, and Victor Alexander Mtsimbe Norrild. State of the art report: Verified computation, 2024. URL https://arxiv.org/abs/2308.15191.

Lizhi Xiong, Wenhao Zhou, Zhihua Xia, Qi Gu, and Jian Weng. Efficient privacy-preserving computation based on additive secret sharing, 2021. URL https://arxiv.org/abs/2009.05356.

Alaa Youssef, Madelena Y. Ng, Jin Long, Tina Hernandez-Boussard, Nigam Shah, Adam Miner, David Larson, and Curtis P. Langlotz. Organizational factors in clinical data sharing for artificial intelligence in health care. *JAMA Network Open*, 6(12):e2348422–e2348422, 12 2023. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2023.48422. URL https://doi.org/10.1001/jamanetworkopen.2023.48422.

Yukun Zhang. The influence of generative ai on content platforms: Supply, demand, and welfare impacts in two-sided markets, 2024. URL https://arxiv.org/abs/2410.13101.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. 2018. doi: 10.48550/ARXIV.1806.00582. URL https://arxiv.org/abs/1806.00582.

Ian Zhou, Farzad Tofigh, Massimo Piccardi, Mehran Abolhasan, Daniel Franklin, and Justin Lipman. Secure multi-party computation for machine learning: A survey. *IEEE Access*, 12:1–1, 01 2024. doi: 10.1109/ACCESS.2024.3388992.

Guido Zuccon, Bevan Koopman, and Razia Shaik. Chatgpt hallucinates when attributing answers. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '23, page 46–51, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400704086. doi: 10.1145/3624918.3625329. URL https://doi.org/10.1145/3624918.3625329.