WORKING PAPER
Submitted to Knight Georgetown Institute Digital Competition Conference
January 2026

Layered Data, Layered Power:
Rethinking Digital Competition in Generative AI Markets

Dr. Courtney C. Radsch
Director, Center for Journalism
Open Markets Institute

Abstract

Generative AI is reshaping digital competition by elevating data access, control, and reuse to a central source of market power. Dominant technology firms increasingly derive competitive advantage not from any single product or dataset, but from layered configurations of data that span multiple markets, infrastructures, and stages of the AI lifecycle. Yet prevailing competition and platform governance frameworks remain oriented toward discrete product markets and static assessments of dominance, limiting their ability to detect or address data-rooted forms of entrenchment. This paper develops a conceptual framework of layered data power to explain how different categories of data – first-party, proprietary, licensed, extracted, and synthetic – operate together as reinforcing inputs across generative AI ecosystems. It distinguishes between static data and data possession, governance over data flows, and dynamic data feedback loops, showing how market power increasingly lies in structuring future data generation and access rather than merely accumulating datasets. Drawing on analysis of real-world industry dynamics and using the case of Google's illegal Search monopoly as a case study, the paper illustrates how layered data power raises barriers to entry, enables exclusionary integration, and obscures competitive harm when analysis is confined to single markets or lines of business. The paper concludes that existing competition approaches are limited in their applicability to data-driven AI ecosystems and the prevention or dismantling of data monopolies.

1. Introduction

Generative artificial intelligence is reorganizing the competitive dynamics of digital markets and economic and information ecosystems. As AI systems are embedded across search, advertising, social platforms, productivity tools, and cloud services, they increasingly integrated into our economic and national security strategies, meaning data has become both a critical input and a strategic lever through which dominant tech corporations consolidate and extend their power. These shifts, however, are not fully captured by prevailing approaches to digital competition, which continue to assess dominance primarily within discrete products, services and markets rather than across integrated data ecosystems. As a result, they do not adequately address the economics of generative AI and typical remedies to market power are in inadequate, inaccessible, or unimaginable.

Competition analysis has long recognized that digital markets exhibit distinctive characteristics including multi-sided interactions, non-price competition, and rapid innovation cycles that are shaped by extreme returns to scale, network effects, and the pivotal role of data. Nearly a decade ago the OECD underscored how "competition in the digital economy is also increasingly a competition between ecosystems," the interconnected networks of services and devices in which dominant platforms leverage their control over data to create lock-in effects, strengthen their market position, and expand into adjacent markets, reinforcing their dominance through data-driven feedback loops and economies of scope (Crémer et al., 2019, p. 33). Yet courts, competition authorities, and regulators continue to struggle with how to address the role of data in securing and perpetuating dominance throughout digital ecosystems amid the transition to a generative AI-centric economy, as underscored by the recent antitrust case against Google Search in the United States.

The rise of generative AI intensifies a different challenge: competitive advantage increasingly flows from the ability to combine, govern, and continuously generate and regenerate data across markets, infrastructures, and ecosystems. Dominant firms such as Google, OpenAI, Microsoft, Meta, and xAI do not compete in generative AI in a single market. Instead, they draw on layered access to user data, proprietary operational data, licensed third-party datasets, large-scale extraction of web-based and rights-restricted content, and increasingly, synthetic data generated by AI models.

Existing competition and platform governance frameworks struggle to address this configuration of power. Market definition remains anchored in present-day products, while remedies often target individual services or forms of conduct. Even when structural remedies are proposed, they face opposition and reticence due to what some perceive as the technical difficulty or even impossibilities of implementing structural remedies and the degradation that could be caused to other parts of the digital ecosystem (Hovenkamp, 2023). Even newer gatekeeper regimes like the UK's Strategic Market Status designation tend to enumerate services rather than interrogate how data moves across them. And without complementary transparency legislation to ensure access to the information needed to assess such movement, the information asymmetries between digital platforms and everyone else continues to impeded meaningful reform. As a result, competitive harm rooted in cross-market data reuse, feedback loops, copyright infringement, and future data generation is difficult to observe, hard to prove, and harder still to remedy.

This paper argues that these limitations stem in part from how data is conceptualized in competition analysis. Rather than treating data as an asset or a generalized scale advantage, I advance a multidimensional framework of layered data power to capture how different data types and governance mechanisms interact across markets and over time. Drawing on research in the field of artificial intelligence and documented practices across major AI platforms, the paper proceeds as follows. Section 1 situates the analysis within existing debates on digital competition, platform power, and data governance, with particular attention to the limits of market-by-market analysis in technology ecosystems. Section 2 identifies four primary categories of data inputs relevant to generative AI systems that seek to provide a comprehensive framework for all types of data in the ecosystem. These four categories are (1) first-party proprietary data, including user-generated and operational data; (2) third-party proprietary data obtained through licensing and/or partnerships; (3) external public data typically obtained through APIs or web scraping, which may be public or rights-restricted content; and (4) synthetic data generated by AI system outputs. I show how these data types function not in isolation, but as layered and mutually reinforcing inputs that amplify incumbents' advantages across training, deployment, and continuous model improvement. Section 3 develops the layered data power framework and clarifies its analytical distinctions, emphasizes how market power increasingly lies in governance, in other words, by structuring the conditions under which data is produced, accessed, and reused throughout the AI ecosystem. The conclusion suggests diagnostic questions that regulators should consider when seeking to address data-rooted dominance in AI.

2. Digital Competition, Platform Power, and Data Governance

2.1 Market definition and ecosystem dynamics

Competition policy has traditionally relied on market definition to structure analysis of dominance and harm. The proliferation of technology platforms that operate multiple services and products, intermediating interactions between users, advertisers, developers, and content providers in multiple markets has highlighted the limitations of this approach and underscored the need for interventions that do not rely on illegal harms to have *already* been committed.

Although some analysts and regulators have increasingly emphasized ecosystem dynamics (Jacobides, 2021; Jenny, 2021; Kira et al., 2021), antitrust law in the United States remains backwards looking and dominated by debates over market definition that often fail to address the reality of contemporary corporate power. Antitrust has also become politicized and turned into a political weapon as independent regulatory and law enforcement agencies have been gutted and reconstituted under the authority of the executive branch, though this is not a dynamic covered in this paper. An ecosystemic perspective encompasses the idea that firms operate as integrated systems of products, services, and infrastructures whose competitive significance cannot be assessed in isolation. This is particularly important to accurately understand power and dominance in the field of artificial intelligence. Within this framework, dominance may arise not from market power within a single service, but from the ability to leverage assets such as user relationships, technical infrastructure, and/or data, across markets.

Generative AI systems are not standalone products, but rather infrastructures and capabilities embedded across services, trained and refined through data collected and generated in diverse contexts, and dependent on cloud infrastructure, chips and talent that are in limited supply

(Lynn et al., 2023; C. C. Radsch et al., 2025). Evaluating competition within any one service or market risks missing how advantage accumulates at the ecosystem level, particularly with respect to data.

2.2 Data as a competitive asset and its limits

Data has increasingly been recognized as a competitive asset that contributes to market power in digital markets, and as an essential differentiating factor in the GAI boom. Access to large volumes of user and operational data can improve product quality, personalization, and learning, potentially raising barriers to entry for competitors while securing the incumbent's ability to attract and retain users, often on "free" services (Iansiti, 2021; Just, 2018; Khan, 2017; OECD, 2024b). At the same time, debates persist over whether data advantages are durable or self-correcting, particularly where users multi-home or data exhibits diminishing returns (Hagiu & Wright, 2025)

Furthermore, some scholars argue that data feedback loops are unlikely to translate into AI market dominance since not all proprietary data is unique or irreplaceable, and many data sources can be substituted, which may limit the extent of market concentration (Abbott & Marar, 2025). But when the nature of learning affects the feedback loops (e.g. recommendations) and there are feedback loops that combine both across-user and within-user learning, dominance is more likely.

Much of this discussion, however, treats data as a relatively undifferentiated input. Analyses often focus on how much data firms possess or whether rivals can access similar datasets, without fully examining how data is governed, combined, or redeployed across services. This perspective is especially limited in the context of generative AI, where different data types serve distinct functions at different stages of development, deployment, and improvement and are used across product lines in ways that strain traditional market conceptions.

2.3 Infrastructure, platforms, and AI deployment

Recent work on AI governance has highlighted the role of infrastructure, such as cloud services, in shaping competition and innovation (Lynn et al., 2023; Narayan, 2022; van der Vlist et al., 2024). Control over compute resources, development tools, and deployment environments can create dependencies that reinforce platform power. Yet infrastructure alone does not explain how competitive advantage persists once AI systems are deployed and integrated into user-facing services.

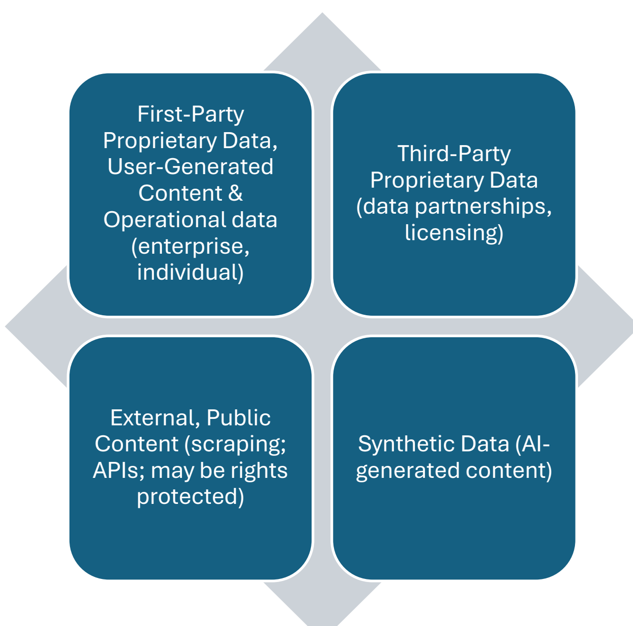
Data governance bridges this gap. The ability to collect data through deployed systems, govern access to that data, and feed it back into model improvement links infrastructure, platforms, and markets into a single competitive system. Understanding this interaction is essential for assessing competition in generative AI ecosystems and underscores many of the limitations of current antitrust enforcement.

3. Conceptual Framework: Layered Data Power

Competitive advantage in generative AI does not derive from any single category of data in isolation but rather emerges from the ability of dominant firms to create, observe, combine, govern, and recycle multiple layers of data in ways that shape both present performance and future data generation. In this section I outline four categories of data based on control and ownership of the data: (1) first-party proprietary data, including user-generated and operational data; (2) third-party proprietary data obtained through licensing and/or partnerships; (3) external public data typically obtained through APIs or web scraping, which may be public or rights-restricted content; and (4) synthetic data generated by AI system outputs. Furthermore, market power in generative AI increasingly arises from the interaction between these categories of data across various dimensions, rather than from any one alone. The framework distinguishes three analytically separable dimensions, which are relevant to all three categories of data:

1. Static data holdings: These include archives, historical user data, proprietary operational datasets, licensed third-party data, and curated corpora of real-world or synthetic data used for training or grounding AI models.
2. Governance power over data flows: This refers to the ability to determine who can access data, under what conditions, and for what purposes, through terms of service, licensing agreements, exclusivity arrangements, APIs, and technical architectures.
3. Dynamic data feedback loops: Deployed AI systems generate new data through user interactions, reinforcement learning, and retrieval-augmented generation. Firms that control deployment environments can capture this data and use it to continuously improve systems, reinforcing their market position.

1 Four Categories of Data Inputs to AI Systems



Layered data power emphasizes the interaction between data types and governance mechanisms across the AI lifecycle, rather than any single factor in isolation. Data inputs are continuously generated, refined, and redeployed across products and services, often under conditions governed by dominant firms themselves. Control over these processes enables incumbents to translate existing advantages into durable and expanding forms of market power, even in the absence of exclusive ownership or access over all relevant data sources, and to potentially derive even greater power from each category of data than rivals even if they have assets in one category. The concept of layered data power captures how multiple forms of data operate together as reinforcing sources of competitive advantage across markets and over time and is distinct from:

- Traditional data advantages, which focus on the size or uniqueness of datasets within a given market, tend to discuss data advantages in terms of volume, velocity, variety, and veracity within specific product or service markets and are asset-based.
- Economies of scale or scope, which explain cost efficiencies typically with respect to production and the near-zero additional costs involved in providing digital products/services but do not address control over data generation or access, and are focused on the supply side
- Network effects, which describe user-driven feedback and network size but not data governance choices, for example about data access, portability, or interoperability and are focused on the demand side.

Competitive advantage in generative thus AI arises from the ability of firms to combine and govern multiple categories of data across three analytically distinct dimensions. Static data holdings provide the material basis for model development; governance power determines access, exclusion, and reuse; and dynamic feedback loops generated through deployment enable continuous data capture and system improvement. Market power increasingly emerges from interactions across these layers rather than from any single data category in isolation.

3.1 Category one: first-party proprietary, user-generated, and operational data

The first category consists of first-party proprietary data generated by the firm (FPP data); data generated by users through their interaction with the firm's services and user generated content (UGC); and internal operational data used to optimize products and systems. This includes, for example, search queries and social engagement metrics, advertising performance data, usage logs, archives, and internal evaluation and ranking data. The largest tech firms at the forefront of AI development have billions of users, significant market share, and a range of products and services that generate training data and fine-tuning that can be used both in AI development and feedback into finetuning the service for users, or potentially to generate even better data.

Dominant firms deploy generative AI across multiple products, enabling data generated in one context to improve performance in others for relatively low marginal cost and thus attract more users. Search queries inform language models; usage of productivity tools refines AI assistants; social and advertising data shape content moderation and targeting systems. These crossproduct flows blur market boundaries and allow firms to leverage established positions into emerging AI-enabled markets. Google, for example, has more than 500 million users on each of 15 distinct platforms and can leverage data generated on one to develop AI that can be applied across its product ecosystem including to the 2 billion search users in 200 countries (Melton, 2025; Times & News, 2025).

3.1.2 Data Generation and Accumulation as Governance Power

The growing recognition of the value of proprietary data is evidenced by Big Tech companies quietly granting themselves permission to scrape the content and mine the data created by users of their products, conditioning usage, especially in the free tiers of service, on their right to train AI systems. Most of these platform companies updated their terms of service without explicit notification to users and although some give users the right to opt-out of AI training, the default is typically opt-in (Tan, 2024). Some, like Meta, which owns Facebook, Instagram and WhatsApp, did not even allow for that option until a backlash from users and lawsuits claiming the failure to obtain consent violated the EU's personal data protection law gave Europeans an opt-out option (Lazzaro, 2024; Wrona, 2024).

Meta used public Instagram and Facebook text and video posts to train its models, and has used a database of billions of Facebook posts and images in more than 100 languages (Wrona, 2024). Microsoft likewise used content from its subsidiaries like LinkedIn, Bing and control over GitHub, one of the world's largest repositories of code, to train its AI systems while GitHub Copilot simultaneously reinforces Microsoft's cloud and developer ecosystem. Although framed as internal integration, the arrangement functions analogously to a proprietary data license with foreclosure effects for rival model developers; while this may or may not lead to market concentration in one market, it is only one of several factors that tend to favor ecosystem dominance (Hagiu & Wright, 2025). X, formerly Twitter, shared user data with Elon Musk's xAI company to train its AI model, Grok (*Elon Musk's X Will Use Public Data to Train AI Models*, 2023; Pandey, 2023). Google is able to leverage the data generated through its unrivaled dominance in search as well as the more than 500 hours of video uploaded to YouTube every minute, often with accompanying descriptions and transcription, to train it AI models which in turn are used to enhance its Search experiences (Wiggers, 2022). The richness of multimedia content and accompanying textual descriptions make it a valuable source of training data for large language and multimodal AI models, but the company has told other companies that such use is a violation of YouTube's terms of service (TOS) (Fielding, 2024; Patel, 2023).

It appears to be a trend toward platforms prohibiting third-party scraping and attempting to secure their exclusive use of this proprietary data and/or their ability to monetize access. X, which was acquired by Elon Musk's xAI startup, put in place restrictions on use for AI training and fine-tuning and limited access to its API in order to retain exclusive access to a valuable feed of human-generated data or offer it for sale, like Reddit, another social network, did. Reddit, for example, made $60 million deals with both OpenAI and Google to allow them access to a regular supply of real-time, fresh data created by Reddit's 50 million users while locking in Reddit's use of Google's VertexAI cloud and OpenAI as an "advertising partner" and demanding that Microsoft pay as well (Heath, 2024; Maiberg ·, 2024).

This category has been the primary focus of competition analysis to date. Authorities and scholars have examined how control over large user bases enables firms to accumulate high-volume, high-velocity data that can improve service quality and reinforce user lock-in. In generative AI contexts, such data plays a critical role in model fine-tuning, evaluation, personalization, and deployment optimization.

Within the layered framework, however, the significance of this data extends beyond scale. First-party and operational data form the *baseline layer* upon which other data advantages are built. They also serve as the primary source of dynamic feedback, allowing firms to continuously refine models based on real-world use. Importantly, these data are often inaccessible to rivals because of both formal exclusivity as well as because they are generated within closed ecosystems controlled by dominant platforms.

3.2. Category two: Third-party Proprietary Data: Data "Partnerships" & Licensing

The second category consists of third-party proprietary data obtained through licensing arrangements, partnerships, or commercial agreements. This includes datasets owned by publishers, content platforms, data brokers, and other firms, as well as enterprise data contributed through cloud or productivity services. Unlike first-party data (Category 1) or publicly accessible data (Category 3), third-party proprietary data allows dominant firms to convert external informational resources into governed inputs through contract rather than ownership or acquisition, thereby extending control across the AI value chain without formal vertical integration. As access to high-quality public web data becomes increasingly contested legally, technically, and politically (C. C. Radsch, 2024), leading AI developers and digital platforms have expanded their access to third-party proprietary data obtained through licensing agreements and strategic partnerships. These have emerged as a central mechanism through which incumbent firms reinforce advantages in model quality and downstream market power while deterring litigation and reshaping bargaining dynamics across media, cultural, and information industries. These arrangements thus function as a form of private governance, defining permissible uses, technical conditions, and access rights over data in ways that structure competition and innovation. By converting contested data uses into contractually authorized private arrangements, licensing bypasses courts and regulators and deters questions of legality and fairness through confidential bilateral agreements. As a result, data licensing operates not only as an AI input procurement strategy, but as a governance mechanism that reallocates control over information flows while remaining largely invisible to conventional merger or market-concentration analysis.

While often presented as arms-length market transactions and evidence that the market is working in the absence of regulation and legal clarity, these arrangements are shaped by significant asymmetries in information and bargaining power. Corporations with the deepest pockets and biggest user base are best positioned to coerce and cajole publishers, record labels, social media companies, and content creators to strike licensing deals, or what the industry calls partnerships. Hyperscalers and AI incumbents consistently appear on the demand side, while content producers, who are often fragmented and do not have access to information about the use and value of their data in AI systems, occupy the supply side, reinforcing structural bargaining imbalances (C. Radsch, 2024a). Confidentiality and individualized commercial negotiations prevent content producers from benchmarking terms or meaningfully assessing the value extracted from their data, further exacerbating asymmetries.

Within the layered data power framework, generative AI systems can be understood as operating across several analytically distinct layers, including access to data, model training and fine-tuning, deployment infrastructure such as cloud and compute, and downstream distribution through search, assistants, or productivity tools. In principle, competition could occur independently at each layer. In practice, however, licensing deals frequently involve broad, multi-purpose rights that include model training, fine-tuning, retrieval-augmented generation (RAG), and real-time product integration, often with cloud exclusivity, API credits, and technical assistance (Thomas & Kretschmer, 2025). Microsoft's "partnership" with OpenAI, furthermore, made it the company's exclusive cloud provider, meaning that OpenAI deals that included publisher integration and API credits may have also tied those publishers to Microsoft's Azure cloud, reinforcing Microsoft's dominance in AI (Lynn et al., 2023; C. C. Radsch et al., 2025). As a result, the data access agreement often functions as a form of vertical tying or bundling that allows AI firms with existing dominance in one layer, particularly cloud or platform distribution, to extend that dominance into other layers of the AI tech stack, reinforcing market power without formal vertical integration or explicit exclusivity (C. C. Radsch et al., 2025).

These commercial deals thus have the potential to reinforce monopolistic dynamics in other parts of the AI tech stack, specifically with respect to cloud, contributing to a mutually reinforcing monopoly broth. As data improves model performance, demand for affiliated cloud resources increases, further enhancing bargaining leverage to secure additional data access creating a feedback loop that entrenches market power across layers. Furthermore, by selectively internalizing external data sources through contracts, dominant firms extend control over information flows while avoiding the scrutiny traditionally associated with mergers or acquisitions.

3.2.1 Data exclusivity and uniqueness

Dominant AI firms are able to secure access to high-quality or real-time data, sometimes on semi-exclusive terms, while rivals lack comparable resources or leverage. In the wake of the Reddit-Google deal, only Google's search engine had access to Reddit, which blocked other crawlers from indexing its site, giving Google exclusive access its robust community of human commentators while giving Reddit what appears to be preferential access in search results (Goodwin, 2024; Lebow, 2024). Because Reddit's continuously updated, real-time discussions are especially valuable for model alignment and evaluation, exclusive or preferential access to such data confers advantages that cannot be readily replicated through static or scraped datasets. Empirical research indicates that uniqueness is a factor that contributes to non-incremental and potentially significant improvements, yet thus far my review of the information publicly available about the deals show that exclusivity between the content/data provider and the AI developers is extraordinarily rare. Furthermore, the emerging AI licensing market includes a variety of licensing intermediaries that facilitate the transaction (in various ways), from dominant technology platforms to startups with a range of business models and value (C.

Radsch, 2024b). Moreover, there is a lot of information in the world that but has not yet been datified, not least of all publisher and cultural archives and low-resourced digital languages. This would indicate that access to third party data is unlikely to be a decisive factor in anticompetitive technological development if it remains non-exclusive.

However, even though few agreements include exclusivity, multiple forms of functional exclusivity are at play which mirror historical tying strategies in search, mobile operating systems, and ad tech (*Comment to the French Competition Authority on Competition and Generative AI*, 2024). These include cloud-linked licensing, where access to proprietary datasets is conditioned on hosting, compute credits, or preferred deployment on a dominant cloud platform; product ecosystem bundling, in which content/data licensors receive AI credits, tools, analytics, or distribution advantages that raise switching costs and encourage lock-in; and contractual non-disclosure, limiting transparency, accentuating information asymmetries, and preventing collective bargaining by data suppliers. Platforms that are vertically integrated within the AI ecosystem can also generate the money needed to pursue data acquisition, including for low resources digital languages, and favorable policy.

Within the layered framework, licensed data functions as an *amplifier*. It enhances the performance and legitimacy of models trained on first-party data while reinforcing advantages across deployment, distribution, and governance. Licensing thus operates not as a substitute for extractive data practices but may instead contribute to selectively stabilizing and legitimizing them, thus consolidating advantage at the top of the market.

3.2.2 Licensing as leverage rather than market exchange

Licensing arrangements for AI training and grounding data are often framed as voluntary, bilateral market transactions. In practice, however, they frequently reflect asymmetric power dynamics produced by pre-existing platform dominance. Firms such as Google, Microsoft, and OpenAI are able to offer selective and discretionary licensing deals, most often with large publishers, data brokers, or content platforms (particularly those most likely to sue), while smaller firms and new entrants lack both the capital and the strategic position to secure comparable access (Duffy, 2025; Guaglione, 2026).

These licensing arrangements could function as exclusionary tools in at least three ways. First, they could lock up high-quality or real-time data sources through exclusivity or preferential terms, raising rivals' costs and limiting the availability of legally usable data for competing AI developers. Second, these arrangements could normalize a two-tiered data ecosystem in which dominant incumbents enjoy privileged access while others remain dependent on legally uncertain, lower quality, or other suboptimal data sources. Third, they could allow incumbents to present themselves as compliant or "responsible" data users while having already amassed vast quantities of unlicensed data during earlier periods of regulatory ambiguity, enabling them to argue against formal regulation while securing their advantage with respect to less capitalized firms or startups.

This dynamic reinforces what this paper describes as the ability of dominant and gatekeeper firms to combine historical data accumulation with present-day governance over future data access. Licensing thus does not replace extraction, it sits alongside it, selectively legitimizing access for dominant firms while leaving structural asymmetries intact.

3.3 Category three: external public and rights-restricted data accessed through scraping or APIs

The third category encompasses external data that originates outside an AI developer (and its affiliates) and is drawn from broadly accessible sources on the open web or through publicly available systems, including government repositories, open access academic archives, and some news media sites, social platforms, and user-generated content portals. This category straddles an important distinction between genuinely public domain data such as datasets published by governmental agencies or openly licenses scientific corpora, and rights-restricted content that, while publicly accessible, remains subject to intellectual property rights, contractual terms of service, and privacy interests. Examples of the latter include news articles, creative works, and user-generated content hosted on third-party sites. The salience of this category of data arises from the gap between public accessibility and legal or contractual permission, where developers operate under conditions of uncertainty and legal ambiguity. These conditions favor well-resourced firms able to withstand potential litigation over smaller competitors or startups that require legal clarity to avoid potentially litigation and/or attract investment.

3.3.1 Data Intermediation as Governance Power:

Thus far, the biggest technological advances in frontier AI have been drive by the accumulation of data, which was possible because of several reasons. First, the legal ambiguity over how and whether to apply copyright. Second, the lag time between the development of large language and frontier models and their deployment online and through the platforms we use every day, with the time required for legal, technical and political processes, particularly democratic ones. Third, the relative impunity to regulation and liability big tech corporations enjoy globally coupled with market caps outsizing many countries. The companies headquartered in Silicon Valley and their Chinese rivals have also offered inadvertent protection to startups that also want to develop AI using the biggest datasets possible, and are offering up proof of concept with every new AI integration. This gives them ecosystemic influence so that even if they do not dominate or even succeed in a specific product market, they are still setting

Access to this data is governed less by ownership than by intermediation. This external data is typically obtained through web scraping via crawlers or APIs and firms that control gateways to the web such as content delivery networks (CDNs), search engines, browsers, app stores, or hosting infrastructure can shape the terms under which content is discovered, indexed, and reused. Cloudflare, for example, routes about 20 percent of global web traffic and recently implemented default blocking of AI crawlers to give its users greater control over how their data is used (C. Radsch, 2025).

In some cases, online content providers face coercive choices: participation in AI training or retrieval may be effectively required to maintain visibility or traffic, as in the case of Google crawlers for search and AI. Site owners worry that blocking Google AI crawlers could negatively impact where they appear in search results since there are no rules against such retaliation.

Within the layered framework, this category highlights the role of governance power over data flows. Market power manifests not through exclusive ownership or proprietary access, but through the ability of dominant AI firms to disregard legal and policy instructionsset defaults, impose technical coupling, and limit meaningful opt-out. This layer is particularly important for understanding how dominant firms convert control over one market, such as search, into advantages in adjacent AI markets.

3.4 Category four: Synthetic data and recursive advantage

The fourth category consists of synthetic data generated by AI systems themselves, including model outputs used for further training, fine-tuning, evaluation, or data augmentation.[1] Synthetic data has attracted increasing attention as a potential response to constraints on access to human-generated or rights-restricted data, and is often framed as a substitute that could reduce dependence on scarce, costly, or legally contested real-world datasets.

In AI development, debates over synthetic data reflect a broader divide between approaches that emphasize data quantity and those that emphasize data quality, although there is general agreement that the purpose or objective of the system ultimately determines what kinds of data are most useful and valuable. Researchers and developers working within a data-quantity paradigm increasingly explore the use of AI-generated outputs to (re)train models, particularly as concerns grow that high-quality human-generated data may become more limited or degraded over time. (Cox, 2024; Ryan-Mosley, 2023; Villalobos, 2024) These concerns are frequently linked to the saturation of the open web with low-quality or machine-generated content, sometimes described as "AI slop," and to predictions that frontier models may soon exhaust easily accessible internet data suitable for training at scale (C. C. Radsch, 2024). Synthetic data is both a technical and strategic response to anticipated data scarcity. But within the layered data framework advanced in this paper, synthetic data is unlikely to function as an independent or equalizing input and more so as a derivative and reinforcing layer, whose value is tightly coupled to access to the other three categories of data.

While synthetic data is frequently described as less expensive and more accessible than other forms of training data, particularly when compared to the costs of collecting, licensing, cleaning, and labeling large volumes of real-world data, current methods to create synthetic datasets also include several shortcomings and risks that may limits its utility for more general purposes (Hao et al., 2024). High-quality synthetic data depends on high-functioning models capable of generating outputs that meaningfully approximate or extend real-world data distributions, complex patterns, and nuances. The performance of those base models, in turn, has thus far depended on access to large volumes of first-party proprietary data, third-party proprietary datasets, and external publicly accessible content. These dynamics render the competitive implications of synthetic more complex. Synthetic data can lower some barriers to

[1] Synthetic data, as used here, refers to data generated by AI systems and should be distinguished from simulation data generated in closed or rule-based environments such as physics simulations or synthetic mathematical data which may raise different competitive dynamics. The analysis in this paper focuses on synthetic data that is recursively linked to deployment, user interaction, and control over predictive generative models, and therefore implicated in layered data power.

experimentation, testing, and limited model development, but currently does not function as an independent substitute for real-world data and other data layers, in part because it is contingent on the quality, diversity, and recency of the underlying real-world data on which generative models were initially trained and continually grounded. As a result, the ability to generate (output) quality synthetic data can be considered a derivative input whose value is shaped by control over first-party proprietary data and access to category two third-party proprietary data, rather than as an independent substitute for other data layers.

4. Layered Data as Governance Power in Generative AI Ecosystems

Dominant technology firms at the forefront of general AI development increasingly exert market power in the AI ecosystem not only through the accumulation of data, but through their control over the terms on which data is accessed, licensed, withheld, or extracted across interconnected markets. This is a form of governance power that dominant corporations occupying strategic positions at points in the data supply chain (e.g. search, social platforms, app stores, browsers, chatbots, and AI assistants and agents). Some can combine gatekeeping power, licensing control, technical dependency, and contractual coercion to structure these supply chains in their favor. This enables them to shape the conditions under which data and content are generated, circulated, and governed within AI systems. And these ecosystem players can leverage their existing power in specific markets and/or advantage their own downstream or upstream products and development

Layered data supply chains can function as sites of exclusionary conduct and raise barriers to entry by shaping the legal, technical, and economic and technical conditions under which data can be used and how it is valued throughout the AI ecosystem and the economy more broadly. They also blur the line between legitimate product integration and anticompetitive tying, particularly where participation in AI systems is made a condition of access to essential digital infrastructure such as cloud, compute, search or provided through defaults and integrations within existing product lines with significant user bases.

Dominant AI firms have governance power, they control the pipelines and terms by which data flows into generative AI systems, in part because most big tech corporations have so many other lines of business and companies they are invested in (Blankertz et al., 2025; Cho, 2021). This governance power is exercised through integration, contracts, defaults, technical design, and legal insulation stemming from massive valuations, thus allowing incumbents to structure the future of AI competition in their favor even where formal market dominance is assessed narrowly or remedies are confined to a single product market. This was the case with the limited remedies imposed in the Google Search monopoly case which did not consider the impact of Google's illegal search dominance on its competitive advantage in the AI ecosystem.

Competition interventions focused on conduct within individual services, such as search or advertising, and de facto going to miss the way power works in the AI ecosystem. While such remedies may address specific harms, they often leave intact the underlying data flows that sustain dominance across markets. As a result, competitive advantage rooted in layered data power may persist even after market-specific interventions..

4.2.1 Vertical integration across data supply chains

Firms that control content platforms, user interfaces, cloud infrastructure, and AI models are able to recycle data across product lines, generating feedback loops that continuously improve their models while depriving rivals of comparable signals. Usage data from AI-enabled products, such as search queries, prompts, clicks, engagement metrics and the like feeds back into model refinement, ranking systems, and personalization engines (OECD, 2024a).

In this sense, data supply chains are recursive systems that structure future data generation. Control over the interface, or platform, becomes control over the next round of training and fine-tuning data. Licensing, extraction, and feedback thus operate together, creating self-reinforcing advantages that are poorly captured by market definitions or likely to be remedied after dominance is achieved, litigated, and determined to be illegal under antitrust law. The value and benefits that accrue to dominant firms in the layered data ecosystem compound rapidly, as evidenced by trillion-dollar market caps and billion-dollar funding rounds in the AI sector.

4.2.3 Forced participation and conditional access in platform-mediated data flows

Dominant tech platforms that enjoy monopolistic economic and governance power in various parts of the AI ecosystem are able to require participation in AI through their dominance over social media, search, cloud, and other digital platforms that millions and billions of people use, from mail, maps and instant messaging, to browsers, operating systems and devices. The case of Microsoft and Meta's use of data from users of their social media platforms discussed earlier is emblematic. Google's control over general search provides a particularly clear illustration of how data extraction can be conditioned on participation in downstream AI products. By the end of 2025, more than 60 percent of Google searches include AI overviews by default (Melton, 2025). As documented in multiple legal proceedings and competition reviews, publishers and website operators are effectively unable to opt out of Google's AI crawling and use of their content for generative AI features, such as AI Overviews, without jeopardizing their visibility in Google Search.

With upwards of 85 percent market share in general search in many jurisdictions, this lack of meaningful separation between Google's search crawler and its AI crawler creates a coercive choice: either allow content to be used for AI training and RAG or accept exclusion or demotion in the dominant search engine on which traffic and revenue depend (Montoya & Radsch, 2025). From a competition perspective and particularly considering a U.S. judge ruled in 2025 that Google held an illegal monopoly in search, raising serious concerns about tying and retaliation.

Conditioning access to search indexing on participation in AI crawling resembles a form of technological tying, enforced through technical architecture and programming choices that mirror other forms of platform coercion in digital advertising and news distribution (C. Radsch, 2025). Furthermore, in the absence of clear line rules restricting Google from retaliating against publishers who opt out of its AI products by demoting their search rankings, akin to how publishers that did not adopt the Accelerated Mobile Pages (AMP) protocol were disadvantaged in search results (Montoya & Radsch, 2025). Such practices also exemplify its governance power over data flows. Google does not merely collect data, it sets the rules under which others may refuse, negotiate, or exit from this data exchange.

4.2.3 Indemnification and the insulation of dominant firms

Another underexamined dimension of data supply chain power is the role of legal and financial insulation as a competitive advantage. Dominant AI firms with market capitalization above a trillion dollars, like Microsoft and Google, have offered indemnification for copyright infringement claims to enterprise customers of their generative AI products to shield their clients from potential intellectual property claims arising from training data or outputs to mitigate risk and encourage adoption (Nowbar, 2023; *Protecting Customers with Generative AI Indemnification*, n.d.). By lowering implementation risk and accelerating integration of generative AI chatbots and content generation tools they further securing their customer base across cloud, productivity software, and thus generating more data to train its AI systems.

Billed by the tech corporations as a consumer protection measure, indemnification is a anticompetitive weapon available only to the biggest players. Smaller AI developers typically lack the resources necessary to absorb litigation risk at such a scale. As a result, indemnification further entrenches incumbents by making their AI products safer to adopt even as underlying data practices remain legally contested. The ability to internalize legal risk thus becomes part of the data advantage itself, reinforcing first-mover benefits gained through earlier extraction of unlicensed content.

5. Implications for Competition and Platform Governance: Takeaways from the Google Search Case

The U.S. Department of Justice's successful monopolization case against Google Search marked a rare and significant finding of illegal monopoly power in a core digital market. In his 2025 ruling, Judge Amit Mehta concluded that Google unlawfully maintained its dominance in general search through exclusionary agreements, most notably default placements, that foreclosed rival distribution and reinforced Google's position as the primary gateway to online information. The decision was notable for recognizing that Google's dominance was not merely the product of superior quality or innovation, but of conduct that distorted competitive conditions in the general search and search advertising market over time.

At the same time, the ruling illustrates the limits of current antitrust frameworks and competition enforcement when remedies are not aligned with the underlying sources of power or interconnectedness within contemporary technology markets, particularly generative AI. The court did not meaningfully address how Google's search monopoly and ensuing data advantages derived from it have not only enabled it to extend and secure dominance in adjacent markets, particularly generative AI development and deployment, but also to shape the underlying logic – or governance – of those markets.

From the perspective advanced in this paper, this omission is consequential and likely limits the effectiveness of the effort to remedy Google's illegal monopoly in search, which is increasingly inseparable from, and integrated with, its AI products and does nothing to mitigate Google's other monopolies. Google's dominance in search has long generated a continuous stream of high-value behavioral, linguistic, and contextual data at a scale unmatched by competitors and accounts for more than 50 percent of the company's revenue, which rose significantly following the antitrust ruling (Mickle, 2025). This data does not remain confined to the search market. It is repurposed across products and services, including advertising systems, knowledge graphs, language models, and AI-powered features increasingly embedded throughout Google's ecosystem. In this sense, the search monopoly functions as a foundational layer in a broader configuration of layered data power.

The remedies in the Google Search case focused on default placements and contractual restrictions but did little to disrupt these layered, cross-market data flows. Even if rivals gain improved access to underlying search index and related data, Google retains control over the accumulated data, governance structures, and feedback loops that continue to shape its AI capabilities. The result is that market power rooted in past exclusionary conduct can be translated into future dominance in other markets without additional unlawful acts that clearly fit within existing antitrust categories and narrowly defines markets (Montoya & Radsch, 2025).

This highlights a structural blind spot in market-specific remedies that are imposed after anticompetitive conduct, and the need for an ecosystemic approach to evaluating market dominance and digital competition in the AI era. By treating search as a self-contained market and focusing on conduct within that market, enforcement failed to address how monopoly-derived data advantages are leveraged to shape competition elsewhere. In generative AI markets, where model quality, deployment scale, and continuous learning are tightly linked to access to large, diverse, and continuously refreshed datasets, this omission is particularly problematic. Dominance secured in one market becomes an input into dominance in another.

From a governance perspective, the Google Search case underscored the difficulty of addressing layered data power through *ex post*, conduct-based remedies alone. The court's analysis implicitly assumes that restoring competition in search distribution will, over time, rebalance competitive conditions. Yet this assumption sits uneasily with data-centric AI markets characterized by cumulative, multilayered advantages of data-driven feedback loops and the governance power of dominant platforms.

The case therefore illustrates a broader challenge for digital competition policy. Even when enforcement succeeds in establishing liability, remedies that do not account for data governance and cross-market reuse risk leaving the most consequential sources of power intact. For generative AI, this means that illegal monopoly power in foundational markets can continue to shape innovation, entry, and competition in downstream and adjacent markets and the broader AI ecosystem, effectively locking in advantage under the guise of lawful technological progress.

6. Conclusion

Emerging regulatory frameworks outside the United States increasingly recognize ecosystemic and structural power, cross-market effects, and ex ante obligations. Layered data supply chains raise concerns about concentration of power that are difficult to address through the dominant competition paradigm in the U.S., which is where the majority of the dominant AI corporations are based. Whereas the European Union and the U.K. consider strategic market status and permit proactive, ex ante remedies, they remain insufficient for adequately and appropriately addressing the ecosystemic power of big tech corporations in AI. Using an ecosystem analysis drawn from both organizational and information theory, these approaches better address the dynamics of AI markets are better aligned with the realities of layered data power.

Generative AI intensifies longstanding challenges in digital competition by amplifying the role of data governance and cross-market integration while underscoring the need to adopt an ecosystemic approach to assessing market power in the AI economy. Market power in these contexts cannot be understood solely through static measures of data possession or product-level dominance since it emerges from layered and multidimensional configurations of data, governance, and feedback that shape both present competition and future market dynamics.

By articulating a framework of layered data power, this paper contributes to ongoing debates on digital competition, AI, and internet governance, offering analytical frameworks for regulators seeking to understand and address data-rooted dominance without stifling innovation.

Abbott, A., & Marar. (2025). *Is Data Really a Barrier to Entry? Rethinking Competition Regulation in Generative AI*. Mercatus Center. https://www.mercatus.org/research/working-papers/data-really-barrier-entry-rethinking-competition-regulation-generative-ai

Blankertz, A., Rock, B., & Shaxon, N. (2025). *Google's Hidden Empire* (No. arXiv:2511.02931v1). arXiv. https://arxiv.org/abs/2511.02931

Cho, C. Y. (2021). *Mergers and Acquisitions in Digital Markets* (CRS Report No. R46739; p. 26). Congressional Research Service. https://crsreports.congress.gov

Comment to the French Competition Authority on Competition and Generative AI. (2024, March 22). Center for Journalism & Liberty at Open Markets Institute. https://www.journalismliberty.org/publications/comment-france-autorite-concurrence-competition-generative-ai

Cox, J. (2024, January 18). Google News Is Boosting Garbage AI-Generated Articles. *404 Media*. https://www.404media.co/google-news-is-boosting-garbage-ai-generated-articles/

Crémer, J., Montjoye, Y.-A. de, & Schweitzer, H. (2019). *Competition policy for the digital era.* (Working Paper No. 6; p. 127). European Commission. Directorate General for Competition. https://data.europa.eu/doi/10.2763/407537

Duffy, K. (2025, September 4). A Complete List of Publishers and Their AI Licensing Deals. *Futureweek*. https://futureweek.com/a-complete-list-of-publishers-striking-ai-content-licensing-deals/

Elon Musk's X will use public data to train AI models. (2023, September 1). Engadget. https://www.engadget.com/elon-musks-x-will-use-public-data-to-train-ai-models-184924197.html

Fielding, S. (2024, April 5). *YouTube CEO warns OpenAI that training models on its videos is against the rules*. Engadget. https://www.engadget.com/youtube-ceo-warns-openai-that-training-models-on-its-videos-is-against-the-rules-121547513.html

Goodwin, D. (2024, February 14). *Reddit shown excessively in Google product review search results, study finds*. Search Engine Land. https://searchengineland.com/reddit-dominates-google-search-discussions-forums-437501

Guaglione, S. (2026, January 1). A timeline of the major deals between publishers and AI tech companies in 2025. *Digiday*. https://digiday.com/media/a-timeline-of-the-major-deals-between-publishers-and-ai-tech-companies-in-2025/

Hagiu, A., & Wright, J. (2025). Artificial intelligence and competition policy. *International Journal of Industrial Organization*, *103*, 103134. https://doi.org/10.1016/j.ijindorg.2025.103134

Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., Zhou, Z., & Tang, H. (2024). *Synthetic Data in AI: Challenges, Applications, and Ethical Implications* (No. arXiv:2401.01629). arXiv. https://doi.org/10.48550/arXiv.2401.01629

Heath, A. (2024, July 31). Reddit CEO says Microsoft needs to pay to search the site. *The Verge*. https://www.theverge.com/2024/7/31/24210565/reddit-microsoft-anthropic-perplexity-pay-ai-search

Hovenkamp, H. (2023). Structural Antitrust Relief Against Digital Platforms. *Journal of Law & Innovation*, *7*(1), 57–105. https://doi.org/10.2139/ssrn.4616175

Iansiti, M. (2021). The Value of Data and Its Impact on Competition. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3890387

Jacobides. (2021). *Ecosystems and competition law in theory and practice* [Submission to the Platform Strategy 2021 Symposium]. https://questromworld.bu.edu/platformstrategy/wp-content/uploads/sites/49/2021/07/PlatStrat2021_paper_14.pdf

Jenny, F. (2021). Competition law enforcement and regulation for digital ecosystems: Understanding the issues, facing the challenges and moving forward. *Concurrences Review*, *N° 3-2021*, 38–62. https://www.concurrences.com/en/review/numeros/no-3-2021/articles/frederic-jenny

Just, N. (2018). Governing Online Platforms: Competition Policy in Times of Platformization. *Telecommunications Policy*, *42*(5), 386–394. https://www-sciencedirect-com.gate3.library.lse.ac.uk/science/article/pii/S0308596117303415

Khan, L. M. (2017). Amazon's Antitrust Paradox. *The Yale Law Journal*, 96.

Kira, B., Sinha, V., & Srinivasan, S. (2021). Regulating digital ecosystems: Bridging the gap between competition policy and data protection. *Industrial and Corporate Change*, *30*(5), 1337–1360. https://doi.org/10.1093/icc/dtab053

Lazzaro, S. (2024, September 12). Meta admits users will only get an opt-out from AI training if governments force it to offer one. *Fortune*. https://web.archive.org/web/20240912170851/https://fortune.com/2024/09/12/meta-admits-in-australia-hearing-that-it-trains-ai-on-user-data-unless-governments-force-it-not-to/

Lebow, S. (2024, September 5). *Reddit is at the top of Google Search results and growing ad revenues. Should marketers be taking note?* https://www.emarketer.com/content/reddit-top-google-search-results-growing-ad-revenues-should-marketers-take-notes

Lynn, B. C., von Thun, M., & Montoya, K. (2023). *AI in the Public Interest: Confronting the Monopoly Threat* (p. 59). Open Markets Institute. https://www.journalismliberty.org/publications/report-ai-in-the-public-interest-confronting-the-monopoly-threat

Maiberg ·, E. (2024, July 24). *Google Is the Only Search Engine That Works on Reddit Now Thanks to AI Deal*. 404 Media. https://www.404media.co/google-is-the-only-search-engine-that-works-on-reddit-now-thanks-to-ai-deal/

Melton, W. (2025, November 16). *New Data: Google AI Overviews Now Appear in 60% of Searches*. https://xponent21.com/insights/google-ai-overviews-surpass-60-percent/

Mickle, T. (2025, October 29). Alphabet Revenue Jumps 16% With Strong Cloud Sales. *The New York Times*. https://www.nytimes.com/2025/10/29/technology/alphabet-google-earnings.html

Montoya, K., & Radsch, C. C. (2025). Beyond court remedies in the Google Search case: A competition reform for the search ecosystem. *Concurrences: Competition Law Review*, *1*, 52–57.

Narayan, D. (2022). Platform capitalism and cloud infrastructure: Theorizing a hyper-scalable computing regime. *Environment and Planning A: Economy and Space*, *54*(5), 911–929. https://doi.org/10.1177/0308518X221094028

Nowbar, B. S., Hossein. (2023, September 7). *Microsoft announces new Copilot Copyright Commitment for customers*. Microsoft On the Issues. https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/

OECD. (2024a). *Artificial intelligence, data and competition*. OECD. https://doi.org/10.1787/e7e88884-en

OECD. (2024b, February 21). *A new expert group at the OECD for policy synergies in AI, data, and privacy*. https://oecd.ai/en/wonk/expert-group-data-privacy

Pandey, M. (2023, March 24). *OpenAI, a Data Scavenging Company for Microsoft*. Analytics India Magazine. https://analyticsindiamag.com/openai-a-data-scavenging-company-for-microsoft/

Patel, N. (2023, August 22). Google and YouTube are trying to have it both ways with AI and copyright. *The Verge*. https://www.theverge.com/2023/8/22/23841822/google-youtube-ai-copyright-umg-scraping-universal

Protecting customers with generative AI indemnification. (n.d.). Google Cloud Blog. Retrieved July 1, 2024, from https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification

Radsch, C. (2024a). *What is the Value of Journalism to AI? A Framework for Establishing Journalism's Value in Artificial Intelligence Systems* (Expert Brief, p. 9). Center for Journalism & Liberty at Open Markets Institute. https://openmarketsinstitute.squarespace.com/publications/value-of-journalism-to-ai

Radsch, C. (2024b, December). Adaptive partnerships in the AI data marketplace. *Nieman Lab*. https://www.niemanlab.org/2024/12/adaptive-partnerships-in-the-ai-data-marketplace/

Radsch, C. (2025, July 9). Cloudflare Wades into the Battle Over AI Consent and Compensation. *Tech Policy Press*. https://www.techpolicy.press/cloudflare-wades-into-the-battle-over-ai-consent-and-compensation/

Radsch, C. C. (2024, October 29). AI Needs Us More Than We Need It. *Washington Monthly*, *Nov/Dec 2024*. http://washingtonmonthly.com/2024/10/29/ai-needs-us-more-than-we-need-it/

Radsch, C. C., von Thun, M., & Nie, M. (2025). *AI and Market Concentration* [Expert Brief]. Open Markets Institute. https://www.journalismliberty.org/publications/expert-brief-ai-and-market-concentration

Ryan-Mosley, T. (2023, June 27). Next-gen content farms are using AI-generated text to spin up junk websites. *MIT Technology Review*. https://www.technologyreview.com/2023/06/27/1075545/next-gen-content-farms-ai-generated-text-ads/

Tan, E. (2024, June 26). When the Terms of Service Change to Make Way for A.I. Training. *The New York Times*. https://www.nytimes.com/2024/06/26/technology/terms-service-ai-training.html

Thomas, A., & Kretschmer, M. (2025, December 1). The AI licensing economy. *CREATe*. https://www.create.ac.uk/blog/2025/02/24/the-ai-licensing-economy/

Times, D. L. L. is B. O. U. technology columnist H. was previously a correspondent for the F., & News, B. B. C. (2025, July 28). Google Is Reaping the Rewards of Its Unfair AI Advantage. *Bloomberg.Com*. https://www.bloomberg.com/opinion/articles/2025-07-28/google-is-reaping-the-rewards-of-its-unfair-ai-advantage

van der Vlist, F., Helmond, A., & Ferrari, F. (2024). Big AI: Cloud infrastructure dependence and the industrialisation of artificial intelligence. *Big Data & Society*, *11*(1), 1–16. https://doi.org/10.1177/20539517241232630

Villalobos, P. (2024, June 6). *Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data*. Epoch AI. https://epochai.org/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data

Wiggers, K. (2022, March 22). Companies are commercializing multimodal AI models to analyze videos and more. *Venture Beat*. https://www.twelvelabs.io/blog/companies-are-commercializing-multimodal-ai-models-to-analyze-videos-and-more

Wrona, A. (2024, June 13). *Why Opting Out of Meta's Use of Facebook, Instagram Posts for AI Training Isn't Easy*. Snopes. https://www.snopes.com//news/2024/06/13/meta-ai-training-user-data/