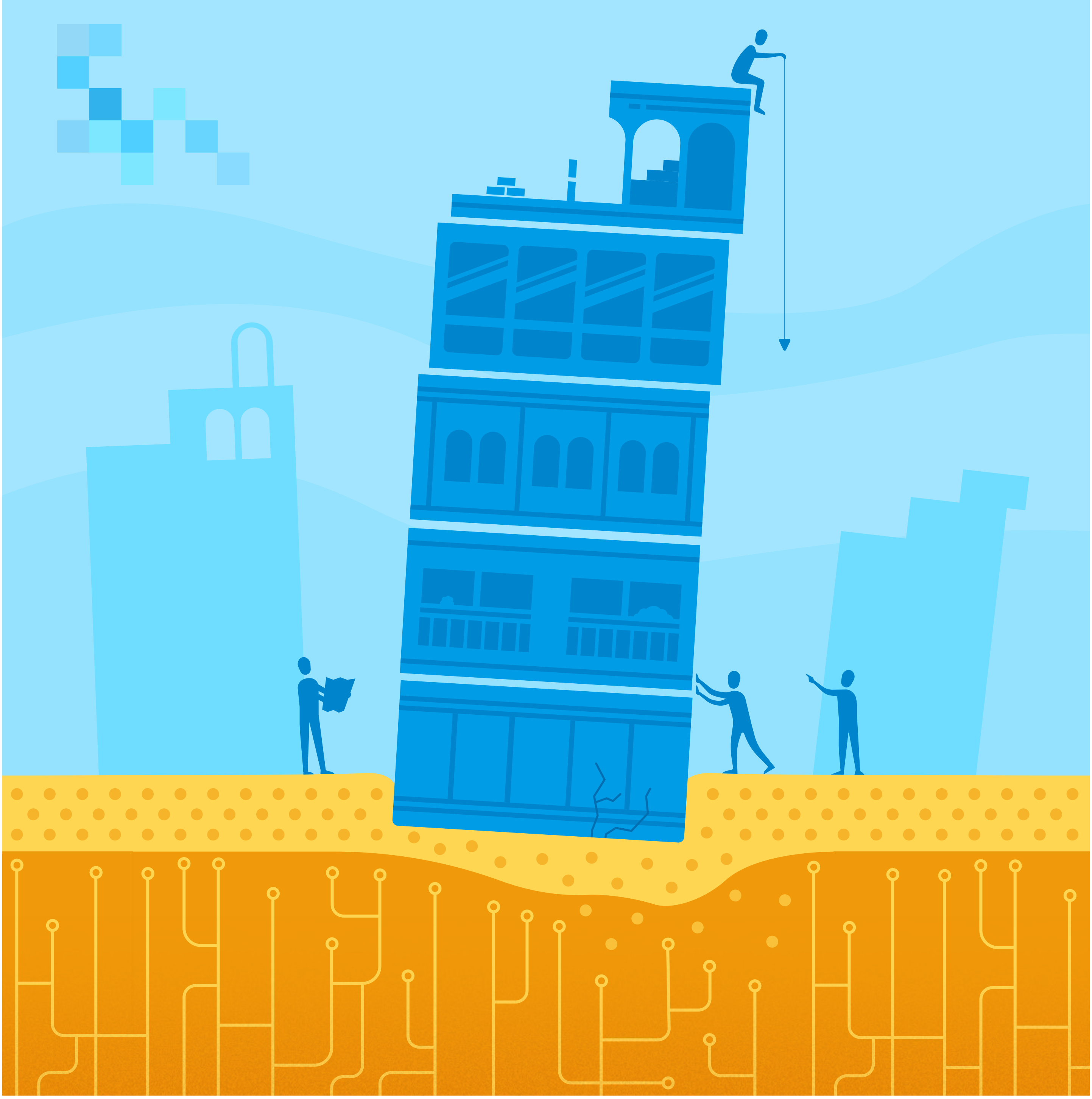
Fragile Foundations: Hidden Risks of Generative AI



BertelsmannStiftung

© Bertelsmann Stiftung, Gütersloh
September 2025

Publisher
Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
www.bertelsmann-stiftung.de

Author
Dr. Anne L. Washington, Duke University

Responsible for content
Teresa Staiger, Bertelmann Stiftung

Editing
Barbara Serfozo, Berlin

Recommended citation style
Washington (2025): Fragile Foundations: Hidden Risks of generative AI. Bertelsmann Stiftung. Gütersloh.

Layout and Typesetting
Paul Feldkamp

Image Credits
© Paul Feldkamp

DOI 10.11586/2025078

Fragile Foundations: Hidden Risks of Generative AI

Anne L. Washington

BertelsmannStiftung

Inhaltsverzeichnis

Preface

Foundation models – the underlying systems behind applications such as ChatGPT – are the backbone of generative artificial intelligence and increasingly shape the digital tools that permeate work and everyday life. These large-scale AI systems are trained on vast datasets and are versatile in their application. Language-based foundation models such as GPT *(Generative Pretrained Transformer)* and products based on them, such as ChatGPT or Gemini, are being deployed with growing frequency. Their appeal lies in promises of efficiency gains and freed-up capacity for other tasks, often amplified by the hype surrounding generative AI.

In recent years, mission-driven organizations have also begun adopting these models. These organizations pursue a clear mandate for the common good, whether in social services, environmental protection, or support for marginalized communities. Yet, the foundation models they rely on are trained on massive, unverified datasets scraped from the internet. These datasets reflect dominant power structures that are often Western, patriarchal, and discriminatory. Rather than reflecting diversity, the models reproduce and amplify existing biases. Their outputs may sound neutral or authoritative, but systematic distortions are embedded in the training data.

For mission-driven organizations, these risks are not abstract technicalities but immediate practical challenges. Precisely in contexts where vulnerable groups depend on accurate information and fair treatment, biased or erroneous AI outputs can cause real harm. A chatbot used in migration counseling might provide biased or skewed guidance on residency or labor rights. An automated tool in social work might deliver incorrect advice on social benefits, making access to support more difficult and undermining trust in the counseling process.

But these risks are no longer theoretical. Real-world examples already demonstrate the gravity of the consequences:

- A chatbot for people with eating disorders provided dieting tips, exacerbating the very problems the system was meant to address.
- A chatbot from Austria's employment agency gave stereotypical career advice, steering women toward cooking or nursing professions and men toward IT jobs.
- In California, parents reported that a chatbot reinforced their son's suicidal thoughts instead of directing him toward sources of help.

Although the underlying models are marketed as "general purpose models," experience shows that they are often ill-suited for general use. The result is erroneous, discriminatory, or unreliable outputs – most harmful precisely for those most dependent on trustworthy assistance.

This report highlights why it is essential to critically examine foundation models. It offers a starting point and an impetus: to use generative AI responsibly, we must first understand the foundations on which it rests and the questions this raises for mission-driven organizations.

The report examines the structural weaknesses of existing foundation models and how they impede innovation, fairness, and accountability. At the same time, it outlines concrete alternatives – technical, participatory, data-related, and collaborative approaches – that demonstrate another path is possible. The message is clear: if foundation models are to serve the common good, they must be developed, evaluated, and operated differently than they are today.

This report is directed primarily at practitioners and decision-makers in mission-driven organizations, but also at anyone committed to a responsible digital future. It invites us to critically examine and rethink the very foundations of generative AI, framing it as part of a digital infrastructure that is aligned with the social missions of its users.

After all, what happens if the public interest goals of these organizations conflict with the digital tools they use? What if they unknowingly adopt models built on data foundations that contradict their own values? This gap between aspiration and practice highlights an essential point. Just as a stable foundation ensures the safety of a building, the training data of foundation models must be robust and value-driven. Only then can the applications built on them – whether in counseling, translation, or health education – fulfill their purpose of supporting people rather than reinforcing existing inequalities.

We would like to express our gratitude to Anne L. Washington for her collaboration and authorship in preparing this publication.

We look forward to your feedback and any form of constructive criticism.

Teresa Staiger
Project Manager
Digitalisierung und Gemeinwohl
Bertelsmann Stiftung

Executive summary

Generative artificial intelligence, introduced only three years ago, has already become embedded into everyday life. Systems such as ChatGPT, Gemini, Dall-E, or CoPilot now shape search results, education, and workplace routines despite little scrutiny of their underlying technology. The foundation models that determine what information is available, amplified, or omitted from AI responses, remain largely unexamined and unregulated. The foundations of generative AI therefore rest on uncertain ground.

The report "Fragile Foundations: Hidden Risks of Generative AI" investigates the influence of this underlying technology through expert interviews and comparative system analysis. It shows that many shortcomings in artificial intelligence, can be traced to how they are built. Foundation models are trained on uncurated data, sustained by business practices that privilege profit over safety, and kept largely closed to external evaluation. As a result, foundation models reinforce unique historic patterns that risk failing the needs of contemporary society.

At the same time, the report highlights competing visions for future foundation models that could serve the public interest. Researchers, civil society, and industry leaders are experimenting with innovative approaches designed to critically engage inherent tensions in shared resources. Alternative strategies that recognize knowledge as a public good include collaborative data collection, participatory governance, and more sustainable computational techniques.

The core message is that foundation models need not inevitably reproduce existing imbalances of power and knowledge. With deliberate curation, continuous improvement, and a service orientation, foundation models could evolve into infrastructures that are more firmly aligned with the public interest.

At a glance

The case for scrutiny of foundation models	Alternative approaches	Key recommendations
Data problems: training on unknown data, embedded historical bias, insufficient representation	**Computational alternatives**: fine-tuning and post-processing Source alternatives: transparent, multilingual, geographically diverse datasets	**Avoid training data monocultures:** curate intentional and reliable datasets
Governance gaps: business models prioritize profit over safety, impossibility of external evaluation, lack of accountability	**Participation alternatives:** civic input, constitutional AI, alignment assemblies	**Build for continuous improvement:** enable routine evaluation and external scrutiny
Systemic risks: resource-intensive computing, cascading problems into downstream applications	**Collaboration alternatives:** open science initiatives, community-driven datasets, collective models	**Imitate a library:** structured, documented, and service-oriented collections

1. Introduction

Generative artificial intelligence was rapidly embedded into daily life, with little systemic evaluation of the technical infrastructure on which it is built. Since OpenAI debuted ChatGPT in November 2022, systems such as Microsoft CoPilot, Anthropic‘s Claude, and Google Gemini have become ubiquitous. These commercial systems shape search results, consumer interactions, business operations, research, and education, yet with little scrutiny of the underlying foundation models that too often fall short of serving the public interest.

This report takes a step back to ask pivotal questions about fundamental risks: What challenges do we face in pooling data in foundation models, and what would it take to critically engage the tensions inherent in designing shared resources? By unpacking how multiple foundation models are built, the risks they pose, and the incentives that drive their development, we aim to clarify what is at stake in shaping the future of artificial intelligence.

At the heart of this question is the recognition that the development of generative AI systems is far from neutral. Countless decisions directly influence the generated texts and images shown in response to language prompts. Many of generative AI‘s shortcomings are rooted in the foundation models that power them. The Stanford Institute for Human-Centered Artificial Intelligence‘s (HAI) defines a foundation model as a large-scale machine-learning model “trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks.” (Bommasani, 2022). Examples of foundation models include large language models like GPT-3, or text-image models like MidJourney. Further definitions of key terms such as generative AI and machine learning are provided in the appendix.

Rather than treating generative AI as a set of dazzling consumer products, this report looks beneath the surface to the foundation models that make them possible. Our research method unfolded in three stages. First, we surveyed academic and corporate research papers on foundation models. This literature review yielded a bibliography of core concepts, pioneering research, and critical perspectives. Second, we conducted interviews with experts with a broad range of perspectives from skeptics to practitioners and legal scholars. From these discussions, we identified recurring themes, gathered illustrative quotations, and compiled a list of systems for further examination. Third, we conducted a comparative analysis of foundation models, organized around the central challenges highlighted by experts. The analysis compared and contrasted the reasoning for developing new foundation models. The findings from these three tasks inform the challenges and recommendations presented later in this report.

This report examines how foundation models could be designed and governed so the benefits of large-scale data collection could be equally shared – across geographies, ideologies, populations, and languages. The foundation models we consider range from proprietary, commercially driven approaches to more collaborative, open-oriented alternatives. We found that foundation models today are built with uncurated data, driven by profit-first business models. Foundation models in the public interest could be designed to satisfy more than profit incentives and could genuinely serve the public and the interests of society as a whole.

Throughout this report, we include quotations from an international group of experts who voiced concerns about the first generation of foundation models. The ideas raised in these interviews are supported by a comprehensive bibliography of research published before March 2025. In accordance with the Chatham House Rules, their comments are not attributed. We conducted expert interviews from October to December 2024. Expert comments appear in italics.

The next section addresses why foundation models demand scrutiny and explores the current failure to develop responsible foundation models. These challenges point to deeper structural problems that complicate the pursuit of foundation models in the public interest.

2. The case for foundation model scrutiny

Foundation models process training data to create representations even when the provenance of the source is unknown. In May 2024, a journalist traced the suggestion by Google's AI search to "cook pizza with glue" back to a single Reddit post more than a decade earlier (Koehler, 2024; McMahon & Kleinman, 2024). How did this possibly tongue-in-cheek comment surface as an appropriate recipe? Why did it become the foundation model's authoritative answer to a generic query? Such inexplicable connections reveal the fragility of foundation models.

With origins in academic research, foundation models now function as critical infrastructure for responses to generative AI prompts. Like most infrastructure, they run in the background with little intervention. Yet, unlike most infrastructure, they are governed by neither disclosure mandates nor regulatory oversight. Moreover, language evolves, as do social understandings of people, places, and ideas. Foundation models must therefore be designed to handle multiple interpretations of language. It is not entirely clear what information is available, amplified, or omitted from foundation models.

The evidence gathered for this report points to three central reasons for closer scrutiny of foundation models. First, foundation models are built on uncurated training datasets that fail to adequately reflect the full range of situations and populations where AI systems are active. Second, prevailing artificial intelligence business models privilege profit at scale, narrowing opportunities for feedback, and slowing subsequent innovation. Third, the models are not robust across all contexts. In addition, systematic tools for evaluating remain scarce, and AI users who want to mitigate risk have few alternatives. Taken together, these concerns highlight the urgent need for greater ethical and technical inquiry now that foundation models are deployed outside controlled academic experiments and into real-world, high-stakes settings.

2.1. Questionable data quality

Training data is typically treated as an afterthought by foundation model builders. As one expert put it: *"No one cares about the data."* Inadequate training data can produce not only factual inaccuracies but also harmful content that crosses legal or ethical boundaries. Equally concerning are misleading associations, distorted representations, systematic omissions, or political intervention. Many foundation models are built on illusions of objectivity and a shared reality where none exists. The examples that follow show how questionable data quality can translate directly into real-world risks, from cultural misrepresentation to copyright infringement.

2.1.1. Uneven representation in historical trends

The first foundation model representations tended to reproduce the past rather than reflect the present, encoding judgments or assumptions that are less accepted today (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016). For example, the word "professor" was more likely to appear near male-associated terms than female ones, despite clear demographic shifts in academia. As one expert observed: *"These are not moral engines. Foundation models are giant history models, and history is biased."*

The credibility of the system is based on the quality of the training data. When Meta released its first chatbot in 2022, it reproduced offensive language, absorbed from broad internet sources known for poor quality control (Belanger, 2023). The perspectives embedded in artificial intelligence may thus skew responses toward the informal and often antagonistic language that dominates online content today. As composites of multiple sources, training data inevitably reflect contradictions and, often, perspectives that are at odds with the goals of the generative AI system.

Generative AI responses often mirror outdated patterns because foundation models calculate occurrences. Although a model might contain recent information, the evidence base is thinner for contemporary realities than for historical precedence (Rivers, 2024; Yagoda, 2024). A researcher interested in unemployment policy, for instance, might seek information about the latest version of a law but receive responses about an older version since it appears more frequently in the training data. Because a newer law appears less often and may be considered an underrepresented idea, the foundation model might dismiss it in favor of its longer-standing predecessor. What is unrepresented, under-represented, or over-represented is not a point of discussion in most computer science research on foundation models.

Massive datasets tend to reflect majority cultures – typically socially dominant groups. For example, one study found that geotagged photos disproportionately represented outsider perspectives, because most pictures were predominantly from tourists or non-local photographers (Naggita, LaChance, & Xiang, 2023). Training data for foundation models reflect the perspective of the cultures where they are designed. The result is that much of the world – indeed, the majority of the global population – must contend with AI systems that fail to reflect their values and experiences. Foundation models do not emerge from neutral ground but from deeply uneven data landscapes. As one expert put it succinctly: *“Large training data represents who is in power.”*

One might expect a new technology to point toward the future and unlimited creativity. In practice, though, artificial intelligence often looks into a rear-view mirror, reflecting the status quo. Concerns about poor data quality thus extend beyond skewed representations to questionable claims to authority and the incorporation of copyrighted material without permission.

2.1.2. Authoritative but misleading

Most foundation models have no mechanisms for determining authoritative sources. When insufficient information exists, generative AI fills the gap by making arbitrary associations. These errors are frequently marketed to the public as “hallucinations,” a term that misleadingly suggests mystery or creativity. In reality, they are systematic errors – direct consequences of inadequate training data and flawed processing.

When trained on questionable data, models can produce bizarre or misleading responses to queries. One chatbot notoriously suggested that eating rocks could be a source of nutrition (McMahon & Kleinman, 2024). Unreliable responses like these are a predictable outcome of poor training data. As one expert remarked: *“How many rocks should I eat a day? No one has an answer to that query. But an LLM will respond with an answer, even if it is from a random user on Reddit.”*

Another major weakness of foundation models lies in the design choice to present a single response to any given prompt. By mimicking an authoritative response, models imply that one definitive answer exists, disregarding the fact that many questions invite multiple perspectives. Traditional search systems, by contrast, return pages of possibilities, acknowledging the contested nature of knowledge. Assuming that unbiased answers are always available fails to account for social tensions and conflicting perspectives. The people and businesses building foundation models often overlook how language is shaped by political and cultural forces. Language changes continually; new terms emerge while older ones evolve. But foundation models lack a way to contextualize these dynamics. As one expert noted: *“There are no unbiased LLM responses. Language is incredibly political.”*

Copyright law governs who can profit from a creative work, and many creators have argued that someone else has profited from their creativity. AI companies that have included authoritative sources without attribution or consent have faced lawsuits claiming copyright infringement. Getty Images sued Stability AI for training on its photographs (Brittain & Brittain, 2023); and the prominent author John Grisham is one of many who has sued OpenAI for misusing their books (Alter & Harris, 2023; Spangler, 2023). These cases often require plaintiffs to prove that foundation models contain their copyrighted works – a burden made heavier by companies' refusal to disclose training materials. As one expert put it: *"OpenAI will be regarded as a hero for making it all available – ChatGPT was 'let them all eat cake.' But they will be cast as villains for running roughshod over privacy and copyright."*

One reason data collection is handled so poorly lies not in technical limitations but in the commercial incentives shaping the development of foundation models. Artificial intelligence companies reward scale and speed over careful curation. As a result, training data are treated as a disposable input rather than as the consequential foundation on which upstream reliability and risk are built.

2.2. Business models stifle innovation and reinforce harms

The business model of the companies that create foundation models amplify existing technical challenges. Typically, these companies rely on free or inexpensive content to train new models, which are then kept static until the next, more advanced version can be sold at a premium. As one expert asked: *"Can I say that I cannot start a restaurant unless I rob a bank? These companies are worth hundreds of billions of dollars on the back of unpaid training data."*

2.2.1. Profit over innovation

Spending money on resources to train models is often noted as a low priority. The history of ImageNet illustrates this logic: its academic developers hired unskilled crowd workers rather than students to label images due to financial constraints (Gershgorn, 2021). Using tools such as Amazon's Mechanical Turk, academic researchers have been able to obtain labor at below minimum wage to train models (Ipeirotis et al., 2010; Irani and Silberman, 2013). Similarly, companies that depend on foundation models for their generative AI products prioritize stability and profitability. Fiscal concerns and reluctance to change show how the problems associated with training data are baked into both machine-learning models and the business models around them, as demonstrated by a landmark study on facial analytics (Buolamwini & Gebru, 2018). Once models are released, businesses have little incentive to revisit or correct questionable training data, even when it includes NSFW – Not Safe for Work – words or images (Birhane & Prabhu, 2021). Financial incentives encourage a reliance on free data sources, even in extreme cases where the material may be inappropriate for general audiences.

2.2.2. Invisible labor and exploitation

Another overlooked dimension of foundation models is the human labor that sustains them. Many verification tasks are carried out by low-paid workers, often in countries of the Global South. Holding labor costs low is part of the plan to maximize profits. These workers rarely benefit from the systems they help build and, in some cases, bear disproportionate harm (Gray and Suri, 2019). The human toll of this model has become increasingly clear. In Kenya, for example, workers who performed traumatic content moderation for Facebook filed a legal case citing severe mental health consequences (Kannampilly and Malalo, 2024). As one expert pointed out, *"Much of the affected workers are in the global south, the same communities that are already facing some of the greatest exposure to climate and other adverse consequences of data centers."* The comment underscores the fact that foundation models not only rely on extractive labor but also depend on resource-intensive infrastructures with uneven environmental costs.

2.2.3. Digital divide by design

First-generation foundation models continue to be sold at reduced prices, which results in a proliferation of mistakes to those who cannot afford improved versions. OpenAI, for instance, initially released ChatGPT 1.0 at no charge but placed subsequent versions behind a high monthly subscription fee. Anthropic has followed a similar path, offering advanced versions of its Claude model at additional cost. If the best systems remain available at a steep cost and inferior systems are easily accessible, only the wealthy will be able to fully benefit from these technologies. The result is a deepening of the digital divide.

The trajectory of OpenAI illustrates these tensions in stark terms. Once founded with altruistic ambitions, the company has shifted toward a for-profit orientation, a move that coincided with the departure of 11 of its 13 original founders. In 2023, CEO Sam Altman was briefly ousted, highlighting internal conflict over the organization's evolving mission (OpenAI, 2024). One interviewee offered a blunt assessment: *"OpenAI is a teenager. It is reckless. The business model is to survive. They can't compete with the cash reserves of Microsoft. So they just have to get attention until someone can buy it."*

2.2.4. Hyper-competitive trade secrecy obscures accountability

The intense competition – both among companies and between nation-states – over artificial intelligence means that transparency is becoming more rare. In an environment driven by winner-take-all incentives, firms are increasingly reluctant to disclose their methods or data sources, making external scrutiny difficult if not impossible. The prevailing ethos is to scale quickly with little regard for broader consequences such as amplifying disinformation campaigns or producing nonsensical misinformation. The lack of visibility into how the systems work limits avenues for policy interventions or mechanisms for consumer feedback.

The prevailing business models behind artificial intelligence prioritize rapid data collection and cost-cutting over innovation, feedback, and data curation. Compensation for data creators is absent, while rent-seeking behavior ensures that additional scrutiny is viewed as a threat to profitability. The result is a structural disincentive to improve quality or address harms. As these tools become essential to daily life, their commercial logic limits opportunities for accountability and innovation. This makes systematic evaluation and oversight not only desirable but essential.

Unfortunately, research on evaluating foundation models is scant because these immensely complicated systems are shrouded in trade secrecy. As a result, meaningful assessment often comes only after public failures, journalistic investigations or posts on social media have gone viral. The commercial competition around foundation models leads to businesses that strive for market domination without technical transparency or social accountability.

2.3 Real costs when AI goes wrong

Failure to anticipate risks from false or misleading information can have tangible financial consequences. In 2023, Alphabet lost an estimated $100 billion in market value after its first AI chatbot, Bard, gave an incorrect answer during a promotional event. The model falsely claimed that the James Webb Space Telescope, launched in 2021, captured the first images of an exoplanet – an achievement that actually occurred in 2004 (Coulter and Bensinger, 2023).

AI chatbots that give incorrect information can leave a company liable to lawsuits. In January 2024, Air Canada faced a lawsuit after its new chatbot provided outdated guidance on special fares, directing a passenger to apply under an old policy (Yagoda, 2024). Because chatbots prioritize the most frequently appearing information, they are more likely to draw on outdated policies. Air Canada argued that the passenger was responsible for not clicking through and reading the link the chatbot provided. The court sided with the passenger who was flying to attend a funeral and ordered Air Canada to retroactively apply the bereavement discount because of the chatbot's error. As one expert cautioned: *"Foundation models potentially amplify risks related to failures and inaccuracies that could impact and harm under-represented populations."*

Even firms that build systems that rely on foundation models may be unaware of underlying risks when systems changes made mid-stream, upending the expectations of those who rely on them. A case involving the National Eating Disorders Association (NEDA) offers a stark example here. The nonprofit rolled out a new chatbot to handle requests for help after ending its contract with human hotline staff. Working with a vendor, it built a decision tree of approved responses. But without the association's explicit consent, the vendor integrated generative AI. The problem surfaced when a person with an eating disorder posted examples of the chatbot's advice on social media (Wells, 2023). Instead of providing vetted responses to people already at starvation weights, the chatbot produced generic internet-style dieting tips aimed at the general population. People seeking help to eat regularly were told, for example, that skipping meals could be a way to lose weight. Consistent with research on automation bias (Skitka, Mosier, and Burdick, 2000), studies show that people are more likely to follow harmful computerized advice because they trust its reliability (Choi, Kim, Park, Kim, and Lee, 2024). With hotline staff disbanded and no adequate technical substitute, the association failed to deliver on its core mission of providing direct assistance to people with eating disorders (Thorbecke, 2023). Lacking standardized third-party evaluation tools, organizations cannot always tell when they are relying on unreliable foundation models, thereby putting individuals and communities at risk.

More importantly, someone or something could shift the probabilities within the model that would result in the model generating an entirely different set of "factual" responses, producing alternate versions of reality. The threat becomes especially acute if such systems are manipulated by malicious or self-interested actors seeking to shape public opinion. Building on the problems of evaluation in machine learning benchmarks, some scholars have rigorously evaluated large language models for completeness to an established standard (Raji et al., 2021), but these efforts remain rare.

2.4 When certainty is an illusion

Foundation models began as research projects – rapidly assembled prototypes built on inexpensive, often low-quality datasets. Yet market competition has since discouraged meaningful investments in improving data quality or enabling external oversight. The intertwined issues of data quality, business models, and social costs point toward deeper ethical challenges in the design of foundation models. Foundation models, despite their unstable foundations, are now embedded across digital infrastructures – from commercial search engines to nonprofit crisis hotlines – often with little or no oversight.

Many of the failures documented in this report stem from the inability of foundation models to demonstrate fallibility or uncertainty. They are rarely designed to expose their internal processes – such as revealing probabilities or explaining how discretionary choices are made. Instead, they are designed to deliver a single highly probable answer. One expert described overconfident, rarely nuanced generative AI as *"the perfect mansplainer."*

In the next chapter, we move from these observed harms to discuss the structural obstacles that prevent foundation models from being developed and deployed in ways that serve the public interest.

3. The structural barriers to responsible foundation models

The path toward responsible foundation models is riddled with obstacles. None are insurmountable, but they make the challenges far from easy to resolve. This section examines why building foundation models in the public interest is so difficult, focusing on curation, representation, cultural associations, computation, and risk.

3.1. The impact of poor data curation

Training datasets are rarely selected and assembled with intention. Curation – the deliberate selection of data – produces more reliable results, yet most training data are neither formally nor adequately curated. Current foundation models are focused more on quantity in the form of parameter size instead of curation. GPT-1, released in 2018, had 117 million parameters. Each subsequent version grew larger, such as GPT-4 released in 2023, which is estimated to have about 1.8 trillion parameters (Howarth, 2025). As one expert remarked: "It's the Cold War of parameter sizes." Artificial intelligence businesses compete on the number of parameters in their foundation models. The efficacy of parameter size, or quantity over quality, has not been determined in the research literature. Curation, however, has been shown to improve results and is rooted in research methods from linguistics.

3.1.1 Why curation matters

Many of these problems trace back to how training datasets were built. Data scientists have successfully made logical connections between disparate "big data" sets, sparking demand for ever-larger collections of digital material – often without regard for origin or quality (Washington, 2023). Often created solely for academic research and experimentation, large datasets were often made open and widely available, later becoming de facto benchmarks in both academia and industry.

Before foundation models, computational linguists curated research collections about word frequencies and language relationships. WordNET, as described in the original 1993 paper by George Miller, was a multi-year project designed to expand the availability of research-quality labeled linguistic data for measurement and analysis (Miller 1993). The WordNET research team manually tagged each word in intentionally selected and curated sources, making it possible to track how word use varied across genres and contexts.

In contrast, ImageNet – which has influenced many foundation models – did not apply the same level of conceptual rigor as WordNet. ImageNet's founder acknowledged that it was too expensive to hire students to manually tag three million files from Flickr (Gershgorn, 2021). ImageNet assembled its labeled image dataset by hiring low-wage labor to assign WordNet terms to photos at scale (Deng et al. 2009). Although started as a university effort, ImageNet soon attracted the attention of venture capitalists eager to fund startups (Gershgorn, 2021). Eventually, it became instrumental in academic competitions but continued to struggle with harmful representations and the model's lack of flexibility (Denton, 2021; Birhane & Prabhu, 2021). Compared with WordNet's structured and intentional design, the curation practices behind today's foundation models are often less transparent and less rigorous. Modern large language models have inherited more from ImageNet's scale-oriented logic than WordNet's focus on genre and curation.

3.1.2 Omissions create skewed and harmful responses

Uncurated collections can produce skewed or exclusionary outcomes. Without curation, training data overwhelmingly reflect majority perspectives. A well-known example is the Faces-In-the-Wild dataset, which advanced computer vision research by compiling expressive facial images (Learned-Miller, et al, 2016). But because it relied almost exclusively on photos of well-known men in sports and politics, it was widely criticized for its inability to automatically recognize expressions on other human faces (Birhane & Prabhu, 2021).

Foundation models that emphasize majority perspectives can cause unforeseen harm when a prompt forces an encounter with content that shows up only occasionally in the training data. For instance, YouTube users searching for religious songs in the Amharic language were instead shown violent content – material that would have been flagged or removed in English or other languages with Latin characters (Nigatu & Raji, 2024). AI-generated assumptions and inaccuracies can also damage reputations (Hsu, 2023). In one instance, a child protection algorithm flagged parents with disabilities as abusive for failing to conform to normative expectations (Belanger, 2023). Training data with little variety will thus continue to propagate the stereotypes majority populations hold against others.

3.1.3. Curation is the beginning

Curation is the first step toward ensuring that relevant perspectives, populations or languages are no longer systematically excluded. But the curatorial process itself can be shaped by a narrow group of actors whose influence has its own skewed preferences. Wikipedia, a free, multilingual, community-edited encyclopedia, is frequently incorporated into language models. It is widely regarded as well curated and well governed, yet it too suffers from monoculture. Longitudinal studies have found that early editors – disproportionately men from Western cultures – continue to hold the most editorial influence (Madanagopal & Caverlee, 2021; Paling, 2015; Reagle & Rhue, 2011). Wikipedia editors often limit discussions to mainstream perspectives, excluding the experiences or achievements of others (Greenstein & Zhu, 2012; Tripodi, 2021). Wikipedia thus illustrates the challenge: even data that appear well curated may not align with the broader public interest goals of representing a range of perspectives.

Curation is a starting point, not an endpoint. The examples above reveal a clear pattern – harm arises not only from what is included but also from what, or who, is missing. The next section explores how representation shapes outcomes.

3.2 Representation should be deliberate

Foundation models reflect the majority of whatever is in their training data. In many free internet sources, that majority consists of widely spoken languages and affluent consumers. Smaller languages, cultures, populations, and places may disappear entirely when foundation models focus on trends. Without a deliberate strategy to include a variety of training data, models will remain limited in their ability to serve the public interest.

3.2.1. Global sameness erases local nuance

The experts interviewed for this report lamented the sameness found in most foundation models. Current models draw heavily on American English-language content from the public internet, rather than incorporating perspectives from the global majority. This dominance risks erasing other languages spoken by millions and flattening regional variations, such as Australian or British English. As one expert warned: *"Everything may merge to a rather bland mid-Atlantic average of what's on the internet."* Another expert expressed the concern that unique regional cultures risk being outnumbered and then decontextualized by outside voices. For example, a Hindu sutra skillfully translated by a native speaker may be overshadowed by an interpretation from a popular yoga teacher in Los Angeles. Foundation models rarely capture the nuance between expertise and popularity.

3.2.2. Silence can distort and exclude

Data collections may be large but not representative of the language or the culture relevant to the task. Because language can differ significantly across linguistic and cultural contexts, an analysis of American English cannot always be transferred to other languages. To make this point explicit, linguist Emily Bender introduced the "Bender Rule," which requires computational linguists making universal claims to specify which language they studied (Bender, 2011).

More troubling still are "data voids," where the absence of reputable sources allows misleading or irrelevant associations to dominate. In one example, one small town that was suddenly thrust into the news, mistakenly became associated with fringe political views in multiple sources, including Wikipedia (Boyd & Golebiewski, 2019).

3.2.3. Completeness signals credibility

Insufficient representation is especially concerning for systems meant to serve people worldwide and now being applied across sectors such as healthcare, education and governance. When foundation models fail to represent groups equitably – as the examples above illustrate – they risk more than contributing to social exclusion. AI businesses also put their credibility at risk, prompting some to wonder what else the business is not aware of. Businesses can build credibility by demonstrating their awareness of how foundation models can be optimized to work well in their context.

The next section examines how under-representation and cultural imbalance become embedded in technical systems.

3.3. Automating cultural associations

Human culture, reflecting partiality toward some groups, ideas or practices over others, is inherently biased. Because foundation models are trained on human language and imagery, they inevitably mirror cultural associations. Recent research, for instance, shows that large language models often fail to capture the diversity of experience within socially subordinate groups. Language about minority populations enforces stereotypes by limiting descriptions to a small set of negative characteristic popular with the majority (Lee, Montgomery, & Lai, 2024). As foundation models become more widely deployed, the businesses building them will need to decide how to address tensions between groups protesting against or insisting on automated cultural associations.

3.3.1. Negative word associations

A central challenge is the presence of negative associations that stigmatize certain groups. Models learn probabilities of word associations, but those associations may encode stereotypes. For example, the words "dumb" and "blonde" may frequently appear together, reinforcing a derogatory stereotype that people with light-colored hair are less intelligent. Bias can also operate in the opposite direction, by unnaturally elevating certain groups or ideas through repeated positive associations. The effects scale quickly. Foundation models can reach users from Palo Alto to Potsdam to Phuket, spreading stereotypes at global speed (Noble, 2018; Williams, 2023). Left unaddressed, these associations circulate across languages and geographies – even among people with no direct exposure to the groups being misrepresented.

What counts as bias is culturally sensitive and varies across regions, communities, and historical contexts. Scholars studying bias often draw on legal frameworks to describe the treatment of protected or marginalized groups. Researchers of machine bias go further to describe how cultural associations are inscribed and can be mitigated in automated systems (Blodgett, Barocas, Daumé III, & Wallach, 2020; Gallegos et al., 2024). Trying to capture this fluidity can be challenging because attempts to address bias can themselves be contested, with responses interpreted as partisan by opposing sides.

3.3.2. Homogeneity risks

Datasets based on one genre of material are often too homogeneous, leaving models prone to failure when applied outside their original context. This is particularly evident in facial recognition systems trained on narrow ranges of facial traits (Khan & Fu, 2021) that fail to recognize a broader set of characteristics (Buolamwini & Gebru, 2018). The commercial success of academic datasets like ImageNet, encouraged private companies to conduct industrial-scale data scraping, also unfortunately repeating size goals over curatorial goals. For example, Clearview AI scraped more than three billion photos from social media sources under questionable conditions of consent (Hill, 2020). ClearviewAI's database has drawn controversy partly because it is sold to governments interested in matching personal online photos to surveillance footage.

Foundation models may function differently for different groups because of how negative associations may be amplified. These harms typically fall hardest on those with the least social and political power. As one expert observed: *"Fairness and bias have haunted AI from the start and will never go away. We are balancing competing needs, and some people will be disappointed."*

Ultimately, bias can mean different things to different people and is often defined at the level of a society or culture. This variability makes bias in machine systems difficult to identify and address. It is layered, often hidden, and resistant to simple solutions.

3.4 Resource-intensive computing

High processing costs help explain why businesses are reluctant to update or replace existing models. Making changes to a foundation model is always a question of sustainability tradeoffs. Continuous improvement of foundation models can be very expensive, driven largely by the demands of hardware and energy.

The hardware required for high-performance computing consumes vast amounts of energy, making the creation of a foundation model both costly and resource-intensive. For instance, the 2020 GPT-3 release with 175 billion parameters required approximately 10,000 NVIDIA chips. Its successor in 2023, GPT-4, demanded 25,000 of the same chips. Scaling model size and complexity requires commensurate growth in both hardware and software resources (Patterson, 2021). Only the wealthiest companies can shoulder these costs, which incentivizes them to build broad, general-purpose models rather than more specialized ones.

The energy burden is equally striking. GPT-3 consumed an estimated 1,287 megawatt hours of electricity – the equivalent of what about 120 average American homes use in a year (Strubell et al., 2019; U.S. Energy Information Administration, 2023). Environmental researchers have raised concerns about the long-term effects of data centers on surrounding communities (Slagowski and DesAutels, 2024). And the demands do not end once a model is trained. Post-training evaluations and refinements also rely on the same infrastructure, though at a smaller scale. Computing a foundation model therefore consumes significant financial and environmental resources, which in turn limits the extent to which feedback can realistically be incorporated.

3.5. Risks in recycling the trash

The technical flaws of foundation models migrate into the products and applications built on top of them. And foundation models certainly do not stand alone: they emerge from and become embedded into real-world applications. As discussed earlier, uncurated and unrepresentative data introduce persistent risks that often flow downstream. Addressing those risks is just as important as preventing problems at the source.

Computer scientists have long relied on free and open data which is why the same large-scale resources serve as training data for multiple foundation models. But the informal writing, inappropriate language, and toxic content found on scraped internet data now undermine the fairness and accuracy of foundation models trained on them. For instance, the Common Crawl dataset contains documented hate speech aimed at marginalized groups (Baack, 2024).

At the same time, the non-profit‘s freely available archives – scraped from the web since 2008 – remain an invaluable source for computer scientists. Common Crawl was a documented source for early versions of ChatGPT, LLaMA, and Gemini. The repetition of similar resources in foundation models, such as the Wikipedia or Common Crawl datasets, exacerbates risks. Problems in a single source can spread across multiple foundation models – and, in turn, many AI systems, creating the potential for systemic cascading risks.

Because projects move from research to commercialization so quickly (Gershgorn, 2021), many commercial systems are built on poorly sourced academic collections, often with little attention to quality or ethics. Researchers warn that the benefits of ever-larger models will plateau if the quality of underlying data is not maintained (Kaplan, 2020). Some refer to this problem of decreasing returns over time as “model collapse,” where performance stagnates or even declines as systems recycle the same low-quality material (Shumailov et al., 2024).

High-quality internet sources once played an important role in training, but many are no longer freely available after being mined by AI companies. A question-and-answer site provides an example of the changing role of open data. Stack Overflow maintained a permissive licensing policy until OpenAI began siphoning off its traffic. The two eventually negotiated a compensation deal granting OpenAI access to Stack Overflow’s valuable trove of user-generated questions and answers. In protest of the licensing deal, some Stack Overflow users began deliberately posting incorrect answers to prevent their knowledge from being harvested (Edwards, 2024). Users were willing to help out another person but not willing to feed correct answers to a profit-making company.

We see a similar dynamic underway in the gathering of visual data. Because many online photos are copyrighted, people looking for training data are more likely to rely on voyeuristic images from the beach or other non-consensual photos taken of people in public. In a striking example, Birhane & Prabhu (2021) tracked the identities of real people whose non-consensual photos were loaded into an image-based foundation model, thus raising serious privacy concerns.

Ultimately, foundation models have the potential of recycling data in a vicious circle of ideas. The biggest danger is when exploitive and harmful data seeps into the systems society relies on for documentation. Companies that integrate foundation models must reckon with the social, political, and legal contexts in which those systems operate and take responsibility for their broader impact.

3.6 Summary of obstacles

The obstacles to building responsible foundation models further support the case for scrutiny. The computing power required to generate foundation models is so resource-intensive that it discourages corrections and improvements. Foundation models easily cast some groups in a positive light and others in a negative one. Without knowing the scope of the training data, it is difficult to test the claim that general-purpose models are up to the task.

Scientific researchers are aware of these obstacles and are working on solutions to mitigate them. The next section examines how ongoing research and commercial systems are striving to build more balanced and reliable models that that could eventually serve as stable infrastructure.

4. Alternative solutions: Toward public-interest foundation models

The first generation of foundation models achieved unprecedented technical success but with significant room for improvement. In the few years since ChatGPT's release, a growing number of groups have introduced generative AI systems and alternative models. While commercial foundation models tend to evolve incrementally or remain static as firms seek to capture market share, numerous smaller efforts are pointing to different ways forward. Advances in computational science continue to push performance through cycles of refinement, while governance-focused initiatives – though fewer in number – underscore the importance of deliberately choosing how foundation models are used.

The experimental, research and startup efforts described in this section reflect competing visions for the future of AI. Some alternatives are developed by private actors targeting niche markets, such as Bloomberg's model for financial services. Others, such as an Irish-language LLM, emerge from collective initiatives driven by cultural or social needs. Together, these systems illustrate how models can be built with the public interest more clearly in mind.

This section examines alternative foundation models through four lenses: statistical and computational innovations, broader participation, new approaches to data, and data collaboratives.

4.1 Computational alternatives

Foundation models are just one layer in a broader stack: Applications built on top of them add further layers that together create the experience of generative AI. Advances in statistics and machine learning can reduce, obscure, or correct problematic word associations embedded in a model. The computational improvements described below refine the broad associations foundation models make, often in subtle but important ways.

Fine-tuning. Fine-tuning uses statistical methods to narrow what information is presented as relevant. The foundation model is trained on any set of unsupervised data with the expectation that it will later be refined through supervised data or automated adjustments tailored to a specific goal. For instance, Meta optimized its **Llama 2-Chat** model for short conversational exchanges (Touvron et al., 2023). Bloomberg took a different approach. It fine-tuned its model using its vast content archive to deliver domain-specific results for the financial services industry. Known as BloombergGPT, the system is proprietary and described only in a preprint paper. Still, Bloomberg reported that the model performed strongly on standard benchmarks and finance-specific tests (Wu et al., 2023). In general, fine-tuning produces superior results for targeted applications without significant losses in general abilities.

Reinforcement learning. Reinforcement learning guides a foundation model toward preferred responses, not by supplying the "correct" answer but by rewarding outputs that align with a given goal. Machine learning first made use of reinforcement learning to measure distance toward a preference. Artificial intelligence systems that are designed to produce more contextually appropriate responses rely on **reinforcement learning from human feedback, RLHF.** The foundation model imitates patterns built on human preferences that have been established by the model's creators. Instead of labeling responses as right or wrong, human reviewers assign relative preferences, shaping the model's behavior toward what appears more natural or socially acceptable. ChatGPT, for example, relies heavily on RLHF (Ziegler et al., 2020). A recent study comparing ChatGPT's responses with the World Values Survey found that they mirrored WEIRD values – Western, Educated, Industrialized, Rich, and Democratic – rather than the broader spectrum of global perspectives (Atari, 2023). The models may imitate human behavioral tendencies, but they do not reflect the social norms and concerns of a majority of the world's populations.

Efficiency gains. Early foundation models were computationally costly, spurring efforts to improve efficiency (Cottier et al., 2025). Computer scientists have since developed foundation models more efficiently by speeding up processing routines, or algorithms. Many systems that we describe in detail below have already achieved improvements in computational speed and made other gains. Models like **Mistral** (Jiang et al., 2023) and **Falcon** (Almazrouei et al., 2023) achieved notable performance gains while using similar training data. The **BLOOM** project explicitly sought to reduce environmental impact through design choices (BigScience Workshop et al., 2023).

Advances in computational speed are reshaping the market. In January 2025, a Chinese startup briefly upended the industry when its system, DeepSeek, surged to the top of download charts within a week. The launch eclipsed OpenAI's offerings and contributed to a sharp drop in Nvidia's valuation (Carew, Cooper, & Banerjee, 2025). Unlike foundation models that have relied on large supplies of expensive and powerful hardware, **DeepSeek** claimed to achieve competitive computing performance with fewer resources (DeepSeek-AI et al., 2025). Its entry challenged the dominance of established players such as OpenAI, Google and Microsoft.

Post-processing. Computational post-processing makes foundation models more useful for specific, narrowly defined tasks Approaches such as fine-tuning and reinforcement learning can deliver improvements at lower cost and with less environmental impact than building entirely new models from scratch.

4.2 Participation alternatives

Computer scientists rarely publish articles about improvements through non-technical means, but a few initiatives take advantage of social participation to improve foundation models.

One prominent example is **Anthropic's "constitutional AI."** Rather than relying solely on labeled data, Anthropic trains its models on broad principles drawn from documents such as the UN Declaration of Human Rights, DeepMind's Sparrow Principles, or industry safety guidelines. Its generative AI system, **Claude**, can produce responses grounded in a chain of reasoning that reflects these pre-defined principles (Anthropic, 2023). By embedding human documents at the core of training, Anthropic aims to open space for wider participation in how foundation models reason.

Governments have also experimented with participatory approaches. In 2023, Taiwan partnered with Anthropic and OpenAI to explore ways of fine-tuning foundation models through civic input (Ministry of Digital Affairs Taiwan, 2023, 2024). Taiwan's Minister of Digital Affairs, Audrey Tang, had routinely involved citizens in deliberations on government decisions like budgeting (Craveiro & Albano, 2015). Building on this tradition, the Collective Intelligence Project launched Alignment Assemblies, which are public forums designed to deliberate on ethical issues in AI policy (The Collective Intelligence Project, 2023). In May 2023, Taiwan began to hold **Alignment Assemblies** in an effort to strengthen democratic legitimacy in AI tools (Chen & Wang, 2024). Other initiatives also sought to empower impacted community members using participatory methods for supervised machine learning (Feffer, Skirpan, Lipton, & Heidari, 2023) and artificial intelligence (Birhane et al., 2022). These initiatives aim to give affected communities a direct voice in shaping how models operate, helping ensure they reflect diverse cultural contexts.

Constitutional AI represents a substantive shift in governance. However, without greater transparency, it is difficult to determine whether it is more than a marketing label for automated reinforcement learning. Unlike reinforcement learning from human feedback (Ziegler et al., 2020), Anthropic's method relies on **reinforcement learning from AI feedback, RLAIF** (Bai et al., 2022). Critics argue that constitutional AI lacks the transparency needed for true democratic legitimacy (Abiri, 2024). Anthropic has not released sufficient implementation details to enable third-party evaluation that could build public trust (Anthropic, 2025). Still, the company continues research on aligning foundation models with public input (Huang et al., 2024).

Participation-centered approaches show how foundation models can be shaped by reasoning and community principles rather than technical optimization alone. Unfortunately, the great uncertainty of potential use cases has stalled policy initiatives in governments around the world (G'Sell, 2024; Lawrence, Cui, & Ho, 2023). They make these systems more open to the scrutiny of people with little technical background who may have a high vested interest in their outcomes.

4.3 Source alternatives

The bar for improving data sources is easily overcome. Why rely on undesirable, and sometimes toxic, material when so much more is available? Scientists have begun moving away from the older data infrastructure created for machine learning. New foundation models are increasingly built on more reliable, credible and stable digital sources.

The most obvious path toward better data is simply avoiding questionable internet sources. A 2023 summit considered how to limit the exploitative use of public-domain data by commercial businesses and instead apply the principles of **Creative Commons** licensing (Hong, Walsh, Zehta, and Angell 2023). One alternative dataset developed for foundation models is The Pile (Gao et al., 2020), a curated collection of 825 GB of English text from business and academic sources. At the same time, research suggests that expanding the volume of data alone does not necessarily yield better results (Ji et al., 2023). Professionally written material by experts, however, is an obvious improvement over internet posts from authors with unknown motives or expertise.

Some foundation models are fully closed and proprietary designed to serve narrow, often high-paying, audiences. **HarveyAI**, built by and for lawyers, offers curated legal responses tailored to the profession (Harvey Team, 2024). Backed by an OpenAI startup fund and the UK law firm Macfarlanes, HarveyAI is run as a collaborative effort for clients seeking a carefully maintained model. As a private company without an active research arm, HarveyAI does not publish academic papers or preprints. Despite its claims, studies show its legal research system still produces inaccuracies at a significant rate (Magesh et al., 2024). Exclusivity does not necessarily translate into reliability, as most closed models are built on existing commercial foundations.

Other projects have emphasized source transparency. **Perplexity**, for example, cites its sources and even includes page references, adapting the logic of academic question-and-answer research for the consumer market (PerplexityAI, 2025). By drawing on multiple models – including DeepSeek, GPT-3, and Claude – Perplexity shows that transparency can be built into applications. Still, debates continue over what counts as "open," with some scholars warning against "open washing" in foundation models (Liesenfeld & Dingemanse, 2024).

BLOOM represents perhaps the most ambitious commitment to open sources. The **BigScience Large Open-science Open-access Multilingual Language Model** was trained on ROOTS, a 1.6-terabyte multilingual corpus spanning 59 languages (BigScience Workshop et al., 2023). **ROOTS, the Responsible Open-science Open-collaboration Text Sources**, was compiled with open-access publications, making peer-reviewed research available without paywalls across languages, scientific disciplines, and regions (Laurençon et al., 2022). The dataset also incorporated historical texts, inclusivity measures, curatorial oversight, and social impact evaluations. BLOOM remains one of the few models to make its training data explicitly known by releasing the open-source ROOTS that contains publicly available sources.

The BLOOM team also released a detailed training chronicle documenting datasets, architecture, computing infrastructure, training parameters, and tokenizers, which marks an unusual level of transparency, and invited public review. The chronicle, akin to a lab notebook, provided insights into trial-and-error decisions throughout development. It has since inspired similar efforts, such as **BloombergGPT's** decision to release its own training chronicle.

The next section considers another path to diversifying training data: intentional collaboration.

4.4 Collaboration alternatives

The alternatives discussed in this section demonstrate how people with shared values, language, or interests built foundation models to meet their needs. For example the **South East Asian Languages in One Network, SEA-LION**, based in Singapore, aims to advance language technology research for the region (AI Singapore, 2024). These efforts highlight the power of collective action to develop models that reflect community priorities, rather than the interpretations of a few individuals in commercial firms.

Organizing in groups is crucial. The **BLOOM** foundation model emerged from a meeting of the BigScience initiative and eventually listed more than 300 co-authors on its primary paper (BigScience Workshop et al., 2023; Laurençon et al., 2022). This broad collaborative process is typical of open science initiatives, which emphasize transparency and accountability to support shared goals. It also facilitated a bottom-up approach to data sourcing, as few people outside the communities represented, for instance, in the Masader dataset, Niger-Congo language groups, or Devanagari scripts have the necessary expertise. The BLOOM initiative not only engaged with institutions holding legal rights to the data, but also actively sought input from those described in the data – the data subjects themselves.

Importantly, many low-resourced language groups have created their own foundation models, testing the robustness of several popular models in the process. **UCCIX**, a large language model for the Irish language, relied on Irish benchmarking tools from **NLP, natural language processing** research to evaluate its performance across four foundation models before deciding on Llama (Tran, O'Sullivan, and Nguyen, 2024). Gaelic, though the official language of the Republic of Ireland, is classified as endangered by UNESCO. Native Gaelic speakers curated a dataset for the first Gaelic LLM to help preserve Ireland's culture through its language, drawing on multilingual open sets such as CulturaX and ParaCrawl while removing duplicates (Tran, O'Sullivan, & Nguyen, 2024).

Other population groups have also identified new digital sources for training foundation models. One analysis of NLP research defines low-resource languages at the intersection of sociopolitical conditions, human agency, and the availability of time, devices, and appropriate NLP resources (Nigatu, Tonja, Rosman, Solorio, and Choudhury, 2024). This typology shows that "low-resource" means different things in different contexts. For instance, Kalaallisut is considered low-resource because it is spoken by only about 50,000 people in Greenland, whereas isiZulu – spoken as a first language by millions in South Africa – remains low-resourced due to the scarcity of written and digital sources available. Some projects have created tools to facilitate collaboration. For example, the **Wikibench** project allows Wikipedia communities to curate data used to train artificial intelligence so that it better represents local ambiguities and community norms (Kuo et al., 2024).

Smaller language groups, often overlooked in large general-purpose models, now have alternative foundation models that represent their populations, languages, and cultures, making them visible in artificial intelligence. For instance, a research consortium across several Ethiopian universities built a multilingual large language model. **EthioLLM** covers five Ethiopian languages as well as English and has been benchmarked with a new **EthioBenchmark** (Tonja et al., 2024). Given that over 80 languages are spoken in Ethiopia, the consortium plans to expand the project to fully represent the region.

Collaborations and collective ownership of foundation models represent a form of AI sovereignty. Data and AI sovereignty reflect the need to control data and predictive associations without outside influence (Rodriguez-Lonebear, 2016). Without such collaborative initiatives, foundation models risk amplifying only the loudest voices, drowning out smaller populations and less widely spoken languages.

4.5 Summary of solutions

Limiting our review to systems in operation as of early 2025, the systems above improve computational efficiency, expand participation, diversify data sources, and foster new forms of collaboration. Together, these efforts show that computer scientists and community leaders are beginning to hold the technology to higher standards. Rather than focusing solely on profit or imposing one-size-fits-all designs, these initiatives point to the possibility of building models that serve the public good anywhere in the world.

5. Recommendations for building responsible foundation models

The foundations of the first foundation models are shaky. Rather than offering unquestioned support, the infrastructure behind generative AI reveals significant weaknesses and potential for systemic risks.

These models are designed for average situations, but they routinely fail in unexpected settings or with nuanced context. Probability estimates can be mistaken for inevitability. A large language model predicts the likelihood of the next word, but a high probability can be misread as certainty. This obscures the smaller, creative possibilities that fall outside the dominant pattern. Such errors become more consequential when deployed at scale across global populations.

In this report, we have examined competing visions for the technical and social infrastructure that drove generative artificial intelligence from late 2024 through early 2025. Drawing on findings in this report, we propose three recommendations for developing the next generation of foundation models:

1. Avoid training data monocultures. Improve training data by curating datasets to be more reliable, representative, and suitable for professional and formal contexts.

2. Design for continuous improvement. Build foundation models that anticipate feedback and ongoing evaluation instead of relying on static releases.

3. Imitate a library. Consider how libraries provide universal access to knowledge by intentionally collecting and disseminating library material that represents divergent ideas, populations, and geographies.

5.1 Avoid training data monocultures

The most pressing problem with foundation models is that their training datasets are rarely curated with intention. There is no single correct way to curate a training data collection because language is fluid and varies across time, place, and culture. But some deliberation is essential. Training sources should reflect the people, customs, religions, ideas, and languages relevant to future uses. Without knowing where the data comes from, it is impossible to direct benefits strategically or mitigate harms effectively.

A first step involves avoiding unreliable or undesirable sources. Much of the internet is dominated by majority voices from the United States. Such material is insufficient for building models in other languages, or for creating robust versions of English that account for different contexts and communities. When deprived of diverse and high-quality new data, AI systems risk recursion and stagnation (Shumailov et al., 2024). Without any efforts made to change the corpus of training data, models will simply reflect the status quo and can thus amplify existing negative associations and reinforce hostility toward stigmatized groups.

Internet text, beyond being informal and toxic in places, often reflects a monoculture that flattens the rich complex nature of human society (Farrell & Berjon, 2024). Monocultures overly constrain choices and limit alternatives (Scott, 1998). Even when businesses build foundation models that are deliberately aimed for use in general applications, unexpected associations will continue to appear in data sets that draw from similar sources.

Yet alternatives exist. Multilingual collections such as ROOTS (Laurençon et al., 2022), EthioLLM (Tonja et al., 2024), and Mozilla's Common Voice (Mozilla, 2024) provide more representative sources. Historic variants of English, preserved in open repositories like Project Gutenberg, offer further opportunities. Training data could – and should – be drawn from high-quality, traceable materials curated toward specific goals. The challenge lies in creating financial incentives to prioritize data quality. As one expert observed, even ethical actors face competitive pressure in a marketplace where *"the dominant business model is to train on anything that isn't nailed down."*

Failure to address data issues risks confusion at best and serious harm at worst. Pronoun use provides one example: the word "they" may function as a singular, gender-neutral pronoun for nonbinary individuals or as a plural pronoun for groups. Without care in training, models can misinterpret such distinctions. More alarming are cases like CharacterAI, which faced multiple lawsuits after its chatbot – trained on popular culture characters – generated violent responses, including alleged incitements to suicide (Samuel, 2025).

Uncurated datasets are, quite simply, a liability. Studies show that large language models are still trained on material that promotes violence (Xiang, 2023). While chatbots built on foundation models that can control for swear words mark an improvement over Microsoft's Tay 2016 Chatbot (Bright, 2016), guarding against hostile sentiments is much more difficult. For instance, LLMs produced negative associations that were deemed harmful to the mental health of teenagers (Wolfe, Dangol, Howe, & Hiniker, 2024). Better-quality sources are readily available, yet many models still lean on problematic material that surfaces as offensive or harmful outputs.

The bar for improvement is low, but the potential impact is high. Training foundation models on anger and other negative emotions poses particular risks, especially when these systems are used in mental health contexts or are accessible to young people. For responsible foundation models to advance, intentional curation must be taken seriously. Training data has downstream effects that only deliberate selection can anticipate. Building datasets that reflect the public interest means ensuring balanced, accurate, and positive representations for all.

5.2 Design for continuous improvement

Foundation models could be built to improve through ongoing user input and professional evaluation. Yet today, no such pathway exists. There is no standard for civic engagement, public feedback, or external assessment. Even basic questions – such as whether a model was trained and tested on the same dataset – often cannot be easily examined (Rhue & Washington, 2020).

After data curation, increased transparency would likely have the greatest impact on improving foundation models. One model for ethical practice is the Cultural Intellectual Property Rights Initiative, which rests on the "3Cs": consent, credit, and compensation (Bota-Moisin, 2017). Under this framework, derivative works should first seek consent from the original creator, ensuring the material is used appropriately. Consent is only beginning to gain traction in AI, following lawsuits by prominent novelists (Alter and Harris, 2023). Credit requires insight into training datasets and their sources. Compensation remains largely ignored, though a few publishers have secured nominal payments for authors whose work is used in training.

The 3Cs protocol was originally developed by Indigenous copyright scholars to protect cultural heritage from capitalist exploitation based on colonial logics of extraction. Ironically, a framework created in response to a centuries-old dispute over cultural sovereignty may be well-suited for users of cutting-edge technology. Artificial intelligence companies wield enormous influence through an ability to shape perceptions and limit opportunities. Some AI companies have been openly hostile to creators, such as encouraging users to generate fake material in the style of some of their greatest critics (Jacobs, 2025; McMillan-Cottom, 2025). In a competitive AI marketplace, voluntary adoption of the 3Cs is rather unlikely. Offering consent, credit, and compensation places firms at a disadvantage compared with competitors who bypass such obligations. As one expert stated: *"There is no ethical training data under capitalism."*

A zero-sum approach to profits leaves little incentive to pay for content. One expert expressed outrage at being offered only a few thousand dollars to include a recent book in a training collection – an arrangement that would undercut future sales and royalties. Attempts to enforce protections under the Digital Millennium Copyright Act have largely failed against commercial foundation models (Skadden, Arps, Slate, Meagher & Flom LLP, 2024). Ethics makes little financial sense in a marketplace where free, hastily grabbed data is defended in the courts.

Fair competition might be a realistic intervention given the current business reality. Ultimately, the public interest depends on adequate transparency (Washington & Cheung, 2024). At present, no tools exist to help people without technical backgrounds distinguish between foundation models of similar speed and power. Disclosures help to at least diagnose problems after the fact (Liu et al., 2023; Wolfe et al., 2024). Governments have not mandated any regulatory reporting requirements that could unleash third-party evaluators. In an ideal situation, access to comparative information would better position philanthropic organizations and public-interest organizations to determine whether a given foundation model meets their ethical standards.

Policymakers should also consider which layer of artificial intelligence infrastructure is the proper target. Recent legislative proposals have required models above a certain size to begin regulatory reporting (Johnson, 2024). Transparency efforts could address software code, foundation model weights, reinforcement learning, licensing, or publications (Liesenfeld & Dingemanse, 2024). Analyzing each underlying layer alongside the specific use case of harm is necessary to understand the overall risks (Kapoor et al., 2024). Without such analysis, downstream risks are pushed upstream, creating the possibility of systemic failure.

Decisions about the distribution of resources in society cannot be left solely to technologists and investors. The fact that the material is digital does not mean the decisions are not political (Washington, 2023). Citizens, lawmakers, and other experts must have a role in providing feedback to the creators of foundation models. The question is, will they be heard?

5.3 Imitate a library

The library offers an alternative vision for foundation models. Libraries and foundation models share common characteristics that start with gathering representations of human creativity. Both organize and analyze their collections to make them widely available and accessible. But where foundation models are built largely for commercial gain, libraries are built around a mission of service.

The mission of any library is to preserve and provide access to knowledge in the public interest. National libraries, such as the Biblioteca Nazionale Centrale di Firenze in Italy, curate collections of enduring cultural significance. Inclusion in such collections is often regarded as an honor. The National Recording Registry at the Library of Congress, for instance, preserves music and film deemed culturally, historically, or aesthetically important, drawing nominations from the public.

Initiatives intent on gathering all human knowledge are not new. In the United States, libraries flourished after 1865, when formerly enslaved people were legally permitted to read and write. Charles Ammi Cutter's influential 1876 work standardized library science practices, while Andrew Carnegie began financing public libraries in 1888. Wealth from the gold rush and railroads fueled the construction of grand buildings that embodied the idea of knowledge as a public good.

The contemporary movement has amassed more knowledge but with less interest in dissemination. At best, current foundation models resemble a special collection of American English, stretched to stand in for global knowledge. The shifting mission of OpenAI illustrates the current trajectory away from the service-oriented work of librarians. Founded in 2015 with a mission resembling that of a library's – pledging to share research, licenses, and even patents – it soon transformed into a for-profit company that closely guarded its intellectual property (Jacobs, 2025). BLOOM (BigScience Workshop et al., 2023), by contrast, more closely reflects the spirit of the library by embedding human connection and community values into the model itself.

Imagining a foundation model as a library helps clarify what is missing in current conversations about public interest AI. Public libraries traditionally concentrate on local needs and serve anyone in the community. If training datasets were like digital public library collections, they could serve as structured repositories of knowledge. Perplexity AI comes close by blending generative AI probabilistic outputs with search engine results that cite definitive sources.

A foundation model modeled on a national library is perhaps a more appropriate comparison. The goal would be to preserve and expand access to a substantial body of knowledge over time. Training data would provide balanced and positive representations of everyone. Achieving that vision would take time and consensus, especially in today's contested information environment. But current practices – failing to credit or compensate creators – undermine both trust and the representation of ideas available.

The aim of these recommendations is to enable AI prompters, practitioners, and policymakers to make more informed choices that can improve the underlying technology of artificial intelligence.

5.4 Conclusion

This report has examined current research to imagine more broadly what is possible. We found that foundation models today are built with uncurated data, driven by profit-first business models, and deployed without reliable tools to test risks across contexts. The ubiquity of generative AI calls for urgency in understanding its underpinnings before even more systems are built on this fragile foundation.

Awareness of the issues outlined in this report is the starting point for necessary conversations. Anyone working with vulnerable populations must remain alert to how AI systems interact with their constituents. Non-commercial groups and businesses alike should regularly monitor how foundation models represent their views. Managers, in particular, should demand updates and explanations from technology vendors and be prepared to take action if agreements are violated.

Only a few years have passed since the release of the first generative AI system, and further computational and business innovations are on the horizon. Foundation models offer multiple opportunities for growth and renewal. The race to establish artificial intelligence infrastructure remains open. With better datasets, stronger evaluation, and improved access to specialized knowledge, foundation models can become safer for everyone.

Ideally, they could be built in the spirit of libraries, with a commitment to the public interest. Any digital representation of the world‘s knowledge requires a strong technical, ethical, and intellectual foundation.

References

Abiri, Gilad. 2024. “Public Constitutional AI.” arXiv. https://doi.org/10.48550/arXiv.2406.16696.

AI Singapore. 2024. “SEA-LION.AI – South East Asian Languages in One Network.” 2024. https://sea-lion.ai/.

Ali, Syed Mustafa, Stephanie Dick, Sarah Dillon, Matthew L. Jones, Jonnie Penn, and Richard Staley. 2023. “Histories of Artificial Intelligence: A Genealogy of Power.” BJHS Themes 8 (January):1–18. https://doi.org/10.1017/bjt.2023.15.

Almazrouei, Ebtesam, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, et al. 2023. “The Falcon Series of Open Language Models.” arXiv. https://doi.org/10.48550/arXiv.2311.16867.

Alter, Alexandra, and Elizabeth A. Harris. 2023. “Franzen, Grisham and Other Prominent Authors Sue OpenAI.” The New York Times, September 20, 2023, sec. Books. https://www.nytimes.com/2023/09/20/books/authors-openai-lawsuit-chatgpt-copyright.html.

Anthropic. 2023. “Claude’s Constitution.” Anthropic. May 9, 2023. https://www.anthropic.com/news/claudes-constitution.

Anthropic. 2025. “Anthropic’s Transparency Hub.” Anthropic. February 27, 2025. https://www.anthropic.com/transparency.

Atari, Mohammad, Mona J. Xue, Peter S. Park, Damián Blasi, and Joseph Henrich. 2023. “Which Humans?” OSF. https://doi.org/10.31234/osf.io/5b26t.

Baack, Stefan. 2024. “A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl.” In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2199–2208. FAccT ’24. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3630106.3659033.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2016. “Neural Machine Translation by Jointly Learning to Align and Translate.” arXiv. https://doi.org/10.48550/arXiv.1409.0473.

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, et al. 2022. “Constitutional AI: Harmlessness from AI Feedback.” arXiv. https://doi.org/10.48550/arXiv.2212.08073.

Belanger, Ashley. 2023. “AI Tool Used to Spot Child Abuse Allegedly Targets Parents with Disabilities.” Ars Technica. January 31, 2023. https://arstechnica.com/tech-policy/2023/01/doj-probes-ai-tool-thats-allegedly-biased-against-families-with-disabilities/.

Bender, Emily M. 2011. “On Achieving and Evaluating Language-Independence in NLP.” Linguistic Issues in Language Technology 6 (3). https://doi.org/10.33011/lilt.v6i.1239.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–23. FAccT ’21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Bergen, Mark. 2024. “UAE Releases New Falcon AI Model to Challenge Meta, OpenAI.” Bloomberg.Com, May 13, 2024. https://www.bloomberg.com/news/articles/2024-05-13/uae-releases-new-falcon-ai-model-11b-to-rival-meta-s-llama-openai-and-google.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, et al. 2023. “BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.” arXiv. https://doi.org/10.48550/arXiv.2211.05100.

Birhane, Abeba, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. “Power to the People? Opportunities and Challenges for Participatory AI.” In EAAMO ’22: Equity and Access in Algorithms, Mechanisms, and Optimization, 1–8. Arlington VA USA: ACM. https://doi.org/10.1145/3551624.3555290.

Birhane, Abeba, and Vinay Uday Prabhu. 2021. “Large Image Datasets: A Pyrrhic Win for Computer Vision?” In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2021), 1537–47, 1537–47. https://openaccess.thecvf.com/content/WACV2021/html/Birhane_Large_Image_Datasets_A_Pyrrhic_Win_for_Computer_Vision_WACV_2021_paper.html.

Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. “Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes.” http://arxiv.org/abs/2110.01963.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. “Language (Technology) Is Power: A Critical Survey of ‘Bias’ in NLP.” In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5454–76. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485.

Bolukbasi, Tolga et al. 2016. “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings.” In Advances in Neural Information Processing Systems 29, eds. D. D. Lee et al. Curran Associates, Inc., 4349–57.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 2022. “On the Opportunities and Risks of Foundation Models.” arXiv. https://doi.org/10.48550/arXiv.2108.07258.

Bota-Moisin, Monica. 2017. “The 3Cs - Get Weaving!” Cultural Intellectual Property Rights Initiative. https://www.culturalintellectualproperty.com/the-3cs.

Boyd, Danah, and Michael Golebiewski. 2019. “Data Voids.” Data & Society. October 29, 2019. https://datasociety.net/library/data-voids/.

Bright, Peter. 2016. “Tay, the Neo-Nazi Millennial Chatbot, Gets Autopsied.” Ars Technica (blog). March 26, 2016. https://arstechnica.com/information-technology/2016/03/tay-the-neo-nazi-millennial-chatbot-gets-autopsied/.

Brittain, Blake. 2023. “Getty Images Lawsuit Says Stability AI Misused Photos to Train AI.” Reuters, February 6, 2023, sec. Legal. https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. “Language Models Are Few-Shot Learners.” In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), 33:1877–1901. NeurIPS. Curran Associates, Inc. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Burrington, Ingrid. 2015. “The Environmental Toll of a Netflix Binge.” The Atlantic (blog). December 16, 2015. https://www.theatlantic.com/technology/archive/2015/12/there-are-no-clean-clouds/420744/.

Carew, Sinéad, Amanda Cooper, and Ankur Banerjee. 2025. “DeepSeek Sparks AI Stock Selloff; Nvidia Posts Record Market-Cap Loss.” Reuters, January 28, 2025, sec. Technology. https://www.reuters.com/technology/chinas-deepseek-sets-off-ai-market-rout-2025-01-27/.

Chen, Fan-Yu, and Chia-Ying Wang. 2024. “Using AI to Strengthen Democracy: Audrey Tang on Taiwan’s Global Role.” Taiwan Business Topics (blog). December 23, 2024. https://topics.amcham.com.tw/2024/12/using-ai-to-strengthen-democracy-audrey-tang-on-taiwans-global-role/.

Choi, Ryuhaerang, Taehan Kim, Subin Park, Jennifer G. Kim, and Sung-Ju Lee. 2024. “Private Yet Social: How LLM Chatbots Support and Challenge Eating Disorder Recovery.” arXiv. https://doi.org/10.48550/arXiv.2412.11656.

Cottier, Ben, Robi Rahman, Loredana Fattorini, Nestor Maslej, Tamay Besiroglu, David Owen, and Epoch AI. 2025. “The Rising Costs of Training Frontier AI Models.” arXiv. https://doi.org/10.48550/arXiv.2405.21015.

Coulter, Martin, and Greg Bensinger. 2023. “Alphabet Shares Dive after Google AI Chatbot Bard Flubs Answer in Ad.” Reuters, February 9, 2023, sec. Technology. https://www.reuters.com/technology/google-ai-chatbot-bard-offers-inaccurate-information-company-ad-2023-02-08/.

Craveiro, Gisele da Silva, and Cláudio Sonáglio Albano. 2015. “Budgetary Data (in an Open Format) Benefits, Advantages, Obstacles and Inhibitory Factors in the View of the Intermediaries of This System: A Study in Latin American Countries.” In Open and Big Data Management and Innovation, edited by Marijn Janssen, Matti Mäntymäki, Jan Hidders, Bram Klievink, Winfried Lamersdorf, Bastiaan van Loenen, and Anneke Zuiderwijk, 223–35. Lecture Notes in Computer Science 9373. Springer International Publishing. https://doi.org/10.1007/978-3-319-25013-7_18.

Crawford, Kate, and Trevor Paglen. 2021. “Correction to: Excavating AI: The Politics of Images in Machine Learning Training Sets.” AI & SOCIETY 36 (4): 1399–1399. https://doi.org/10.1007/s00146-021-01301-1.

Dave, Paresh. 2023. “Stack Overflow Will Charge AI Giants for Training Data.” Wired, April 20, 2023. https://www.wired.com/story/stack-overflow-will-charge-ai-giants-for-training-data/.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, et al. 2025. “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.” arXiv. https://doi.org/10.48550/arXiv.2501.12948.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. “ImageNet: A Large-Scale Hierarchical Image Database.” In 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–55. https://doi.org/10.1109/CVPR.2009.5206848.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), edited by Jill Burstein, Christy Doran, and Thamar Solorio, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Edwards, Benj. 2024. “Stack Overflow Users Sabotage Their Posts after OpenAI Deal.” Ars Technica (blog). May 9, 2024. https://arstechnica.com/information-technology/2024/05/stack-overflow-users-sabotage-their-posts-after-openai-deal/.

Fairly Trained. 2024. “Fairly Trained Launches Certification for Generative AI Models That Respect Creators’ Rights.” Fairly Trained. January 17, 2024. https://www.fairlytrained.org/blog/fairly-trained-launches-certification-for-generative-ai-models-that-respect-creators-rights.

Farrell, Maria, and Robin Berjon. 2024. “We Need To Rewild The Internet,” Noema, April 2024. https://www.noemamag.com/we-need-to-rewild-the-internet.

Feffer, Michael, Michael Skirpan, Zachary Lipton, and Hoda Heidari. 2023. “From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research.” In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 38–48. AIES ’23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3600211.3604661.

Gadiraju, Vinitha, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. “‘I Wouldn’t Say Offensive but...’: Disability-Centered Perspectives on Large Language Models.” In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 205–16. FAccT ’23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3593013.3593989.

Gallegos, Isabel O., Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. “Bias and Fairness in Large Language Models: A Survey.” arXiv. https://doi.org/10.48550/arXiv.2309.00770.

Gao, Leo, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, et al. 2020. “The Pile: An 800GB Dataset of Diverse Text for Language Modeling.” arXiv. https://doi.org/10.48550/arXiv.2101.00027.

Gershgorn, Dave. 2021. “The Data That Transformed AI Research, and Possibly the World.” Quartz, July 26, 2021. https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world.

Greenstein, Shane, and Feng Zhu. 2012. “Is Wikipedia Biased?” American Economic Review 102 (3): 343–48. https://doi.org/10.1257/aer.102.3.343.

G’Sell, Florence. 2024. “Regulating Under Uncertainty: Governance Options for Generative AI.” Stanford. https://cyber.fsi.stanford.edu/content/regulating-under-uncertainty-governance-options-generative-ai.

Harvey Team. 2024. “Introducing BigLaw Bench.” Harvey. August 9, 2024. https://www.harvey.ai/blog/introducing-biglaw-bench.

Hill, Kashmir. 2020. “The Secretive Company That Might End Privacy as We Know It.” The New York Times, January 18, 2020. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

Hong, Shannon, Kat Walsh, Timid R. Zehta, and Nate Angell. 2023. “AI and the Commons: Outcomes from the 2023 CC Global Summit Alignment Assembly.” Creative Commons Global Summit. Mexico City, Mexico: Creative Commons.

Howarth, Josh. 2025. “Number of Parameters in GPT-4 (Latest Data).” Exploding Topics (blog). February 24, 2025. https://explodingtopics.com/blog/gpt-parameters.

Hsu, Tiffany. 2023. “What Can You Do When A.I. Lies About You?” The New York Times, August 3, 2023, sec. Business. https://www.nytimes.com/2023/08/03/business/media/ai-defamation-lies-accuracy.html.

Huang, Saffron, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. 2024. “Collective Constitutional AI: Aligning a Language Model with Public Input.” In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 1395–1417. FAccT ’24. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3630106.3658979.

Ipeirotis, Panagiotis G., Foster Provost, and Jing Wang. 2010. “Quality Management on Amazon Mechanical Turk.” In Proceedings of the ACM SIGKDD Workshop on Human Computation, 64–67. Washington DC: ACM. https://doi.org/10.1145/1837885.1837906.

Irani, Lilly C., and M. Six Silberman. 2013. “Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk.” In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 611–20. Paris France: ACM. https://doi.org/10.1145/2470654.2470742.

Jacobs, Harrison. 2025. “OpenAI CEO Sam Altman Continues to Completely Miss the Point With AI Art.” ARTnews.Com (blog). April 9, 2025. https://www.artnews.com/art-news/opinion/openai-ceo-sam-altman-ai-art-generators-1234738240/.

Ji, Yunjie, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Lei Zhang, Baochang Ma, and Xiangang Li. 2023. “Exploring the Impact of Instruction Data Scaling on Large Language Models: An Empirical Study on Real-World Use Cases.” arXiv. https://doi.org/10.48550/arXiv.2303.14742.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, et al. 2023. “Mistral 7B.” arXiv. https://doi.org/10.48550/arXiv.2310.06825.

Johnson, Khari. 2024. “Why Silicon Valley Is Trying so Hard to Kill This AI Bill in California.” CalMatters, August 12, 2024, sec. Technology. http://calmatters.org/economy/technology/2024/08/ai-regulation-showdown/.

Kannampilly, Ammu, and Humphrey Malalo. 2024. “Kenya Court Finds Meta Can Be Sued over Moderator Layoffs.” Reuters, September 20, 2024, sec. Africa. https://www.reuters.com/world/africa/kenya-court-rules-meta-can-be-sued-over-layoffs-by-contractor-2024-09-20/.

Kapoor, Sayash, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, et al. 2024. "On the Societal Impact of Open Foundation Models." arXiv. http://arxiv.org/abs/2403.07918.

Khan, Zaid, and Yun Fu. 2021. "One Label, One Billion Faces: Usage and Consistency of Racial Categories in Computer Vision." In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 587–97. FAccT '21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445920.

Koebler, Jason. 2024. "Google Is Paying Reddit $60 Million for Fucksmith to Tell Its Users to Eat Glue." 404media. May 23, 2024. https://www.404media.co/google-is-paying-reddit-60-million-for-fucksmith-to-tell-its-users-to-eat-glue/.

Kuo, Tzu-Sheng, Aaron Lee Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. 2024. "Wikibench: Community-Driven Data Curation for AI Evaluation on Wikipedia." In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 1–24. CHI '24. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3613904.3642278.

Laurençon, Hugo, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, et al. 2022. "The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset." Advances in Neural Information Processing Systems 35 (December):31809–26.

Lawrence, Christie, Isaac Cui, and Daniel Ho. 2023. "The Bureaucratic Challenge to AI Governance: An Empirical Assessment of Implementation at U.S. Federal Agencies." In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 606–52. AIES '23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3600211.3604701.

Learned-Miller, Erik, Gary B. Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua. 2016. "Labeled Faces in the Wild: A Survey." In Advances in Face Detection and Facial Image Analysis, edited by Michal Kawulok, M. Emre Celebi, and Bogdan Smolka, 189–248. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-25958-1_8.

Lee, Messi H.J., Jacob M. Montgomery, and Calvin K. Lai. 2024. "Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans." In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 1321–40. FAccT '24. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3630106.3658975.

Leslie, David, and Antonella Maia Perini. 2024. "Future Shock: Generative AI and the International AI Policy and Governance Crisis." Harvard Data Science Review, no. Special Issue 5 (May). https://doi.org/10.1162/99608f92.88b4cc98.

Liesenfeld, Andreas, and Mark Dingemanse. 2024. "Rethinking Open Source Generative AI: Open Washing and the EU AI Act." In The 2024 ACM Conference on Fairness, Accountability, and Transparency, 1774–87. Rio de Janeiro Brazil: ACM. https://doi.org/10.1145/3630106.3659005.

Liu, Zhengzhong, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, et al. 2023. "LLM360: Towards Fully Transparent Open-Source LLMs." arXiv. https://doi.org/10.48550/arXiv.2312.06550.

Lyons, Michael J. 2021. "'Excavating AI' Re-Excavated: Debunking a Fallacious Account of the JAFFE Dataset." arXiv. https://doi.org/10.48550/arXiv.2107.13998.

Madanagopal, Karthic, and James Caverlee. 2022. "Improving Linguistic Bias Detection in Wikipedia Using Cross-Domain Adaptive Pre-Training." In Companion Proceedings of the Web Conference 2022, 1301–9. WWW '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3487553.3524926.

Magesh, Varun, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools." arXiv. https://doi.org/10.48550/arXiv.2405.20362.

McMahon, Liv, and Zoe Kleinman. 2024. “Glue Pizza and Eat Rocks: Google AI Search Errors Go Viral.” BBC. May 24, 2024. https://www.bbc.com/news/articles/cd11gzejgz4o.

McMillan-Cottom, Tressie. 2025. “The Tech Fantasy That Powers A.I. Is Running on Fumes.” The New York Times, March 29, 2025, sec. Opinion. https://www.nytimes.com/2025/03/29/opinion/ai-tech-innovation.html.

Merler, Michele, Nalini Ratha, Rogerio S. Feris, and John R. Smith. 2019. “Diversity in Faces.” arXiv. https://doi.org/10.48550/arXiv.1901.10436.

Ministry of Digital Affairs Taiwan. 2023. “The Ministry of Digital Affairs Has Partnered with the International Organization, the ‘Collective Intelligence Project’ (CIP), in Fostering Consensus on the Needs and Risks Associated with Artificial Intelligence.” Press Release. Ministry of Digital Affairs. May 31, 2023. https://moda.gov.tw/en/press/press-releases/5243.

Ministry of Digital Affairs Taiwan. 2024. “Utilizing AI to Enhance Information Integrity: Citizens’ Deliberative Assembly.” Ministry of Digital Affairs Taiwan. 2024. https://moda.gov.tw/en/major-policies/alignment-assemblies/2024-deliberative-assembly/1521.

Mozilla. 2024. “Common Voice 20 Is Now Available.” Mozilla Foundation Blog (blog). December 11, 2024. https://foundation.mozilla.org/en/blog/common-voice-20-is-now-available/.

Naggita, Keziah, Julienne LaChance, and Alice Xiang. 2023. “Flickr Africa: Examining Geo-Diversity in Large-Scale, Human-Centric Visual Data.” In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 520–30. AIES ’23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3600211.3604659.

Narayanan, Arvind, and Sayash Kapoor. 2024. AI Snake Oil: What Artificial Intelligence Can Do, What It Can’t, and How to Tell the Difference. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691249131/ai-snake-oil.

Narayanan, Arvind, and Vitaly Shmatikov. 2007. “How To Break Anonymity of the Netflix Prize Dataset.” arXiv, cs/0610105. https://doi.org/10.48550/arXiv.cs/0610105.

Nigatu, Hellina Hailu, John Canny, and Sarah E. Chasins. 2024. “Low-Resourced Languages and Online Knowledge Repositories: A Need-Finding Study.” In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 1–21. CHI ’24. New York, NY, USA: Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3613904.3642605.

Nigatu, Hellina Hailu, and Inioluwa Deborah Raji. 2024. ““I Searched for a Religious Song in Amharic and Got Sexual Content Instead”: Investigating Online Harm in Low-Resourced Languages on YouTube.” In The 2024 ACM Conference on Fairness, Accountability, and Transparency, 141–60. Rio de Janeiro Brazil: ACM. https://doi.org/10.1145/3630106.3658546.

Nigatu, Hellina Hailu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. “The Zeno’s Paradox of `Low-Resource’ Languages.” In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, 17753–74. Miami, Florida, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.983.

Noble, Safiya. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press. https://nyupress.org/9781479837243/algorithms-of-oppression/.

Noveck, Jocelyn, and Matt O’Brien. 2023. “Visual Artists Fight Back against Artificial Intelligence Companies for Repurposing Their Work.” PBS News Hour (blog). August 31, 2023. https://www.pbs.org/newshour/arts/visual-artists-fight-back-against-artificial-intelligence-companies-for-repurposing-their-work.

Paling, Emma. 2015. “Wikipedia’s Hostility to Women.” The Atlantic. October 21, 2015. https://www.theatlantic.com/technology/archive/2015/10/how-wikipedia-is-hostile-to-women/411619/.

Patterson, David, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. “Carbon Emissions and Large Neural Network Training.” arXiv. https://doi.org/10.48550/arXiv.2104.10350.

Paullada, Amandalynne, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. “Data and Its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research.” Patterns 2 (11). https://doi.org/10.1016/j.patter.2021.100336.

Penedo, Guilherme, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. “The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only.” arXiv. https://doi.org/10.48550/arXiv.2306.01116.

perplexityAI. n.d. “Getting Started with Perplexity.” Accessed April 9, 2025. https://www.perplexity.ai/hub/getting-started.

Power, Alethea, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. “Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets.” arXiv. https://doi.org/10.48550/arXiv.2201.02177.

Rhue, Lauren, and Anne L. Washington. 2020. “AI’s Wide Open: Premature Artificial Intelligence and Public Policy.” Boston University Journal of Science and Technology Law 26 (2): 353–78. https://ssrn.com/abstract=3720944.

Rivers, Christopher. 2024. “Moffatt v. Air Canada.” Civil Resolution Tribunal. February 14, 2024. https://decisions.civilresolutionbc.ca/crt/crtd/en/item/525448/index.do.

Rodriguez-Lonebear, Desi. 2016. “Building a Data Revolution in Indian Country.” In Indigenous Data Sovereignty, edited by Tahu Kukutai and John Taylor. ANU Press. https://press.anu.edu.au/publications/series/caepr/indigenous-data-sovereignty.

Russell, Stuart, and Peter Norvig. 2021. Artificial Intelligence: A Modern Approach. 4th ed. Pearson. https://www.pearson.com/en-us/subject-catalog/p/artificial-intelligence-a-modern-approach/P200000003500/9780137505135?tab=table-of-contents.

Samuel, Sigal. 2025. “AI Is Impersonating Human Therapists. Can It Be Stopped?” Vox, February 10, 2025. https://www.vox.com/future-perfect/398905/ai-therapy-chatbots-california-bill.

Scott, James C. 1998. Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed. New Haven, Conn.; London: Yale University Press.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2024. “The Curse of Recursion: Training on Generated Data Makes Models Forget.” arXiv. https://doi.org/10.48550/arXiv.2305.17493.

Skadden, Arps, Slate, Meagher & Flom LLP,. 2024. “Digital Millennium Copyright Act Claims in AI-Training Cases – Recent Developments.” Skadden Insights (blog). December 3, 2024. https://www.skadden.com/insights/publications/2024/12/recent-decisions-on-whether-ai-training-violates-the-digital-millennium-copyright-act.

Skitka, Linda J., Kathleen Mosier, and Mark D. Burdick. 2000. “Accountability and Automation Bias.” International Journal of Human-Computer Studies 52 (4): 701–17. https://doi.org/10.1006/ijhc.1999.0349.

Slagowski, Naomi, and Christopher DesAutels. 2024. “Environmental and Community Impacts of Large Data Centers.” Gradient Corp. 2024. https://gradientcorp.com/trend_articles/impacts-of-large-data-centers/.

Smeaton, Alan F. 2024. “Understanding Foundation Models: Are We Back in 1924?” arXiv. https://doi.org/10.48550/arXiv.2409.07618.

Spangler, Todd. 2023. “Sarah Silverman Sues Meta, OpenAI for Copyright Infringement of Her Memoir ‘The Bedwetter.’” Variety (blog). July 10, 2023. https://variety.com/2023/digital/news/sarah-silverman-sues-meta-openai-copyright-infringement-1235665185/.

Sparck-Jones, Karen. 1994. “Natural Language Processing: A Historical Review.” In Current Issues in Computational Linguistics: In Honour of Don Walker, edited by Antonio Zampolli, Nicoletta Calzolari, and Martha Palmer, 3–16. The Association for Computational Linguistics. Dordrecht: Springer Netherlands. http://link.springer.com/10.1007/978-0-585-35958-8_1.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. “Energy and Policy Considerations for Deep Learning in NLP.” In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, edited by Anna Korhonen, David Traum, and Lluís Màrquez, 3645–50. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1355.

Technology Innovation Institute. n.d. “Falcon Models.” FalconLLM. Accessed December 1, 2024. https://falconllm.tii.ae/falcon-models.html.

The Collective Intelligence Project. 2023. “Alignment Assemblies.” The Collective Intelligence Project. 2023. https://www.cip.org/alignmentassemblies.

Thorbecke, Catherine. 2022. “It Didn’t Take Long for Meta’s New Chatbot to Say Something Offensive.” CNN. August 11, 2022. https://www.cnn.com/2022/08/11/tech/meta-chatbot-blenderbot/index.html.

Thorbecke, Catherine. 2023. “National Eating Disorders Association Takes Its AI Chatbot Offline after Complaints of ‘Harmful’ Advice.” CNN Business. June 1, 2023. https://www.cnn.com/2023/06/01/tech/eating-disorder-chatbot/index.html.

Tonja, Atnafu Lambebo, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Ah Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, et al. 2024. “EthioLLM: Multilingual Large Language Models for Ethiopian Languages with Task Evaluation.” In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, edited by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, 6341–52. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.561.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023. “LLaMA: Open and Efficient Foundation Language Models.” arXiv. https://doi.org/10.48550/arXiv.2302.13971.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. “Llama 2: Open Foundation and Fine-Tuned Chat Models.” arXiv. https://doi.org/10.48550/arXiv.2307.09288.

Tran, Khanh-Tung, Barry O’Sullivan, and Hoang D. Nguyen. 2024. “UCCIX: Irish-eXcellence Large Language Model.” arXiv. http://arxiv.org/abs/2405.13010.

Tripodi, Francesca. 2023. “Ms. Categorized: Gender, Notability, and Inequality on Wikipedia.” New Media & Society 25 (7): 1687–1707. https://doi.org/10.1177/14614448211023772.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). Vol. 30. NeurIPS. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Voorhees, Ellen, and Donna Harman. 2005. TREC: Experiment and Evaluation in Information Retrieval. The MIT Press. https://mitpress.mit.edu/9780262220736/trec/.

Voorhees, Ellen M., and Hoa Trang Dang. 2005. “Overview of the TREC 2005 Question Answering Track.” In Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005, edited by Ellen M. Voorhees and Lori P. Buckland. Vol. 500–266. NIST Special Publication. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec14/papers/QA.OVERVIEW.pdf.

Washington, Anne L. 2023. Ethical Data Science: Prediction in the Public Interest. Oxford Technology Law and Policy. Oxford University Press. https://doi.org/10.1093/oso/9780197693025.001.0001.

Washington, Anne L., and Joanne Cheung. 2024. "Towards Defining the Public Interest in Technology: Lessons from History." Journal of Integrated Global STEM 1 (2): 67–74. https://doi.org/10.1515/jigs-2024-0008.

Wells, Kate. 2023. "An Eating Disorders Chatbot Offered Dieting Advice, Raising Fears about AI in Health." NPR, June 9, 2023, sec. Health Reporting in the States. https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea.

Widder, David, and Mar Hicks. 2024. "Watching the Generative AI Hype Bubble Deflate." arXiv. https://doi.org/10.48550/arXiv.2408.08778.

Williams, Damien Patrick. 2023. "Bias Optimizers." American Scientist, June 21, 2023. https://www.americanscientist.org/article/bias-optimizers.

Williams, Damien Patrick. 2024. "Disabling AI: Biases and Values Embedded in Artificial Intelligence." In Handbook on the Ethics of Artificial Intelligence, 246–61. Sociology, Social Policy and Education 2024. https://www.elgaronline.com/edcollchap/book/9781803926728/book-part-9781803926728-22.xml.

Wolfe, Robert, Aayushi Dangol, Bill Howe, and Alexis Hiniker. 2024. "Representation Bias of Adolescents in AI: A Bilingual, Bicultural Study." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7 (1): 1621–34. https://doi.org/10.1609/aies.v7i1.31752.

Wolfe, Robert, Isaac Slaughter, Bin Han, Bingbing Wen, Yiwei Yang, Lucas Rosenblatt, Bernease Herman, et al. 2024. "Laboratory-Scale AI: Open-Weight Models Are Competitive with ChatGPT Even in Low-Resource Settings." In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 1199–1210. FAccT '24. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3630106.3658966.

Wu, Shijie, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. "BloombergGPT: A Large Language Model for Finance." arXiv. https://doi.org/10.48550/arXiv.2303.17564.

Xiang, Chloe. 2023. "'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says." VICE, March 30, 2023. https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/.

Yagoda, Maria. 2024. "Airline Held Liable for Its Chatbot Giving Passenger Bad Advice - What This Means for Travellers." BBC. February 23, 2024. https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know.

Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. "PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization." In Proceedings of the 37th International Conference on Machine Learning, 119:11328–39. ICML'20. JMLR.org.

Ziegler, Daniel M., Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. "Fine-Tuning Language Models from Human Preferences." arXiv. https://doi.org/10.48550/arXiv.1909.08593.

About the author

Anne L. Washington, PhD, is the Rothermere Associate Professor of Technology Policy at Duke University, where she directs the Digital Interests Lab in the Sanford School of Public Policy. Her research examines the governance of emerging digital technologies, with a particular focus on how power is distributed between individuals and the organizations that control digital record-keeping systems. International and domestic funders have recognized her research with multiple awards, including fellowships with the Peter Pribilla Foundation of Munich (Germany), and the Data & Society Research Institute of New York City, in addition to a National Science Foundation CAREER award. She currently serves as a faculty associate with the Berkman Klein Center for Internet & Society at Harvard University. She holds a doctorate in Information Systems and Technology Management from The George Washington University, an MLIS in Library and Information Science from Rutgers University, and a bachelor's degree in Computer Science from Brown University. Her most recent book, Ethical Data Science: Prediction in the Public Interest, was published in December of 2023 by Oxford University Press.

Acknowledgments

I would like to thank Tanya Neheta, a 2024 master's in Data Science New York University graduate, for her outstanding research assistance. I am also deeply grateful for the many conversations at the Center for Advanced Study in the Behavioral Sciences at Stanford University with my 2024-2025 colleagues, whose insights – shared over beautifully designed CASBS mugs – were indispensable to this project.

Appendix

Terms and concepts

We emphasize the research history of machine learning infrastructures in this section because they have directly shaped the development of foundation models. The foundation models at the center of this report rely on the same technological and organizational ecosystem.

Artificial intelligence (AI) is a broad concept that refers to machines performing tasks that typically require human intelligence. The meaning of AI has shifted subtly over time (Ali et al., 2023; Russell & Norvig, 2021), and this evolution is reflected in how marketing materials often use the label "AI" to suggest better outcomes for any automated task (Widder and Hicks, 2024). Today, AI is frequently used as a form of marketing hype that implies greater value for computerized tasks (Narayanan & Kapoor, 2024). In this report, we confine the scope of artificial intelligence to relatively narrow and specific outcomes documented in published scientific research. Contemporary scientific research on AI encompasses a wide variety of subfields, including machine learning, natural language processing, and computer vision, which are described below.

Generative AI refers to systems that can create entirely new text, images, or music given a single prompt. For example, a prompt such as "write a poem about spaceships in the style of Sterling K. Brown" will produce multiple stanzas in the author's style, though not identical to published work. Prompts are directive, but the models can interpret them in many ways, so the output may differ each time. It is worth noting the related concept of AGI, or Artificial General Intelligence. AGI is based on the belief that there is a single body of knowledge that can be fully known, and that a machine could be trained to perform any intellectual task a human can do. This is more a philosophical argument than a funded research program and is therefore not considered in this report.

Most generative AI systems are built on foundation models. The Stanford Institute for Human-Centered AI introduced this term to describe large-scale models trained on vast amounts of data that can be adapted to a wide range of tasks (Bommasani et al., 2022). As an aggregate category, foundation models encompass both computer vision for images and large language models for text. Examples include Anthropic's Claude for language generation and Stable Diffusion's DALL·E for image generation. A specific mechanism for building the current generation of foundation models is the **generative pre-trained transformer**, or GPT (Devlin, Chang, Lee, & Toutanova, 2019; Vaswani et al., 2017). Built on transformer architecture, GPT-3 was the foundation model behind the first release of ChatGPT (Brown et al., 2020). Parameters – statistical associations within the model – can be adjusted through fine-tuning, parameter manipulation, bias reduction, fairness mitigation, and other post-processing computational methods.

Large language models (LLMs) are foundation models that capture patterns in human language. Each parameter in an LLM represents the probability of associations between words or phrases. For example, by reviewing millions of emails, an LLM would learn that in American English the word "thank" is most often followed by "you." One parameter could represent this specific statistical relationship. LLMs contain billions of such parameters to estimate the likelihood of one word following another. The scale of these models is immense. As an early example, Google's BERT (Bidirectional Encoder Representations from Transformers) generated 340 million parameters (Devlin et al., 2019). Research on LLMs builds directly on earlier work in natural language processing.

Natural language processing (NLP) analyzes associations between words, phrases, or entire sets of documents. NLP has a long tradition in information retrieval, library science, statistics, and linguistics dating back to the early twentieth century (Spark-Jones, 1994). By examining both statistical and linguistic elements of texts, NLP algorithms can translate, summarize, interpret, suggest, or generate human language (Voorhees & Harman, 2005). Summarization research is particularly relevant to generative AI, as it was the first attempt to condense long texts into summaries that made sense to humans (Zhang et al., 2020). NLP researchers have also explored translation services, enabling words in one language to be rendered in another (Bahdanau et al., 2016). Much of this research underpins user-facing technologies such as chatbots and voice assistants. For example, NLP work on question answering (Voorhees & Dang, 2005) serves as the foundation for customer service chatbot applications. While these efforts focused on patterns in language, researchers soon extended similar methods to other domains – most notably, to uncovering patterns in images.

Computer vision enables computers to analyze and process visual information from still images, photos, films, and long-form videos. While early research focused on comparing images or locating specific objects within a photo, contemporary research extends to computationally interpreting the built environment – for example, driving an autonomous vehicle (Gu et al., 2024). Controversial for their implications for privacy, facial recognition applications remain popular despite their limited ability to reliably identify human faces (Hill, 2020). More practically, computer vision has become essential in radiography and other forms of medical imaging, where comparing images to detect similarities and differences is critical. More recently, multimodal models demonstrate how computer vision has evolved: these models combine multiple modalities, enabling the analysis of different kinds of data, such as text and images. Originally, computer vision was referred to as visual machine learning.

Machine learning (ML) combines computational science and statistics to draw conclusions or make predictions from patterns in data. For instance, a machine-learning algorithm can infer characteristics shared by every pitcher who has won the World Series by analyzing past trends. Two forms of machine learning are especially relevant to generative AI. Unsupervised learning identifies patterns independently and largely mirrors how foundation models operate. Supervised learning, by contrast, compares new data against specific criteria and shapes the production of some training datasets that foundation models rely on. Machine learning rose to prominence through competitions focused on improving a single dataset, such as the 2006 Netflix prize (Narayanan & Shmatikov, 2007). It is therefore the core mechanism for training foundation models that enables them to recognize patterns.

Address I Contact

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
Telefon +49 5241 81-0
bertelsmann-stiftung.de

Teresa Staiger
Project Manager
Digitalization and the Common Good
Phone +49 30 275788-160
teresa.staiger@bertelsmann-stiftung.de

BertelsmannStiftung