YOU ARE WHAT YOU EAT: NURTURING DATA MARKETS TO SUSTAIN HEALTHY GENERATIVE AI INNOVATION

BY
STEFAN
HUNT

&
WEN
JIAN

&
AMAN
MAWAR

&
BARTLEY
TABLANTE

Keystone.ai. Keystone has worked with several AI platform firms including Amazon, Meta, and Microsoft; regulatory agencies in the U.S. and Europe; digital publishers; and other actors in the AI supply chain.

TechREG CHRONICLE
SEPTEMBER 2023

YOU ARE WHAT YOU EAT: NURTURING DATA MARKETS TO SUSTAIN HEALTHY GENERATIVE AI INNOVATION

By Stefan Hunt, Wen Jian, Aman Mawar & Bartley Tablante

Generative AI (GenAI) has enormous potential but raises thorny issues. To get to grips with these issues, it's important to focus on the underlying drivers of the problems and the solutions. This paper highlights the critical role of data as a "production input" across all three stages in the GenAI value chain: foundation models, fine-tuned models and grounded applications. We show how, in each stage, data is particularly important to developing and improving models, directly reducing misinformation and other concerns. But there are three emerging risks in the development of data markets. First, the data used to train models is becoming increasingly opaque – and it is increasingly hard to attribute value to data providers. Second, high-quality data is scarce and looks set to get scarcer – there is a great need to increase the supply of high-quality data. Third, unequal access to data may lead to network effects – making it harder for small firms to compete. These issues could profoundly impact and restrict GenAI competition, if not addressed. We encourage policymakers, many of which are considering regulating these markets, to look more closely at ways to actively nurture these crucial data markets and support the evolution of GenAI.

01
INTRODUCTION

Generative Artificial Intelligence ("GenAI") has enormous potential for good – from transforming jobs and helping reduce inequality, to creating widespread economic gains.[2] But to achieve the full promise of this technology requires innovation to mitigate challenging issues that undermine consumer trust, including consumer safety and misinformation. Advancing GenAI technology sustainably will require a concerted effort from industry and increasingly it is recognized from regulators, who must appreciate and work with the grain of the competitive dynamics.[3]

But the path towards regulation could be bumpy – given GenAI is fast-moving and highly complex. Too slow and regulation will be one step behind, inadequately addressing legitimate concerns. Too fast or too deep and it could damage a wellspring of innovation or even entrench and reinforce problems. Achieving the right balance requires insight into and thoughtful consideration of the technological processes and key inputs underpinning AI production. The Competition and Markets Authority ("CMA") expressed such thinking at the launch of their current inquiry into the development and use of AI foundation models.[4]

This paper examines data as an input for GenAI production and as a key aspect of innovation and competition between AI firms, focusing on large language models ("LLMs").[5] We consider the role of data in the GenAI value chain across three separate stages: foundation models, fine-tuned models, and grounded applications:

1. Foundation models are made by training a machine learning algorithm (called pre-training in this context) using huge datasets to produce a model that can be refined and used in many downstream applications.

2. Fine-tuned models are foundation models that are refined through additional training on a narrower set of use case specific data.

3. Grounded models have access to additional data sources, allowing the model access to information (e.g. real-time news) beyond the pre-training and fine-tuning data.

In each stage, data is a crucial production input, meaning that under-development of data markets could hinder competition. Through analyzing the technologies used in AI, we highlight issues such as data scarcity, data transparency, unequal access to data, and dynamic network effects, each of which can profoundly impact competitive dynamics. Our analysis can help to inform and target current policy efforts to promote healthy and competitive markets for the continued evolution and greater deployment of GenAI models. AI research shows data is critical for improving GenAI model performance.[6] The diversity, volume, and quality of data greatly affect LLMs' ability to understand and generate contextually relevant, high-quality output. An LLM will flourish or wither depending on the data it is trained on; and nothing can make up for that. Even with significant computational resources and top-tier talent, models cannot generate meaningful output without sufficient rich, varied, and relevant data. Specialized datasets are essential for the creation of fine-tuned models tailored to specific tasks or industries, e.g. AI models for medical diagnosis require medical records for training.

One issue with data is insufficient supply. This is already a limiting factor in model development, and data scarcity is projected to worsen as models massively expand in size.[7] Two factors exacerbate the issue. First and foremost, most GenAI models rely heavily on data scraped from the web, a prime example being the datasets constructed by Common Crawl. But data providers as it stands have limited incentives to make more data freely available online. The current growth rate of this data is too low to sustain LLM development. Second, it is hard to form data markets for AI because data providers often do not know what of their data

[2] Agrawal et al., *Do we want less automation?* Science (July 13, 2023), https://www.science.org/doi/abs/10.1126/science.adh9429. Goldman Sachs, *Generative AI could raise global GDP by 7%* (April 5, 2023), https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html.

[3] For example, the need for regulation has been argued by AI industry leaders such as Sam Altman, OpenAI's CEO. See Cecilia Kang, *OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing,* The New York Times, 16. May 2023 https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html.

[4] Competition and Markets Authority, *CMA launches initial review of artificial intelligence models,* Gov.UK (May 4, 2023), https://www.gov.uk/government/news/cma-launches-initial-review-of-artificial-intelligence-models.

[5] While we focus on LLMs, we also use the term Gen AI throughout the paper. Most of our points hold for other GenAI models as well.

[6] For example, Hoffman et al. (2022) find that modern LLM models use significantly less data than is optimal for their performance. See Hoffmann, et al. *Training Compute-Optimal Large Language Models,* arXiv (March 29, 2022), https://arxiv.org/abs/2203.15556.

[7] Villalobos, Pablo, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn & Anson Ho. "Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning." arXiv preprint arXiv:2211.04325 (2022).

(if any) is being used to train AI models and face challenges assessing its value. This lack of transparency adds friction in the negotiations between major GenAI companies and data providers.

In addition to the overall limited supply of data, there are other issues. Smaller AI companies might lack resources or business connections to negotiate content deals with large data providers, limiting entry into the sector. Furthermore, superior access to data could give some players a significant advantage: the largest AI players have access to proprietary datasets, e.g. YouTube data by Google, and train their models on them.[8] And when GenAI products directly improve from the data created by user interaction, there can be direct network effects.

Given the potential for such challenges, markets for data for pre-training, tuning, and grounding might need nurturing from regulators, to preserve healthy GenAI competition and alleviate consumer protection issues. Agencies might need to start getting "under the hood" more actively, for instance by monitoring the data AI companies are accessing and using. Regulators should also consider making this information at least partially available to some parties through transparency requirements, as the EU is proposing with the AI Act. Other regulatory responses to nurture data markets could include monitoring for harmful exclusionary vertical agreements, requirements on data sharing or other rules.

We also consider other major factors driving GenAI development besides data, including developer talent and compute resources. Fundamentally data is likely to be the major source of both GenAI industry issues and solutions.

Section II outlines the three stages of the GenAI value chain in turn, highlighting key facts about how the value chain works, focusing on the underlying technology, and describing the role of data at each stage. Section III explores how markets for data for GenAI may evolve and suggests some emerging issues and policy considerations for fostering effective competition.

02

THE GENERATIVE AI VALUE CHAIN AND THE ROLE OF DATA

This section outlines the three stages of the GenAI value chain: foundation models, fine-tuned models, and grounded user-facing applications. We describe each stage focusing on the role of data as an input – discussing the different elements illustrated in Figure 1. We find that data scale in pre-training is the key for the performance of foundation models, that specialized data is at the heart of LLM fine-tuning, and that grounding brings unique, real-time, use-case-specific data into GenAI applications reducing performance issues including "hallucinations" – the tendency of AI models to produce confident answers not supported by facts.

A. Framework to Characterize Data and Its Value

To help elaborate the role of data, we set out a core framework for characterizing and valuing data in different contexts. Using Iansiti (2021), we consider three dimensions of data that can help us gauge value:[9]

Quality:

· <u>Usability</u> – how AI modelers can ingest and use data
· <u>Accuracy</u> – the degree to which the data reflects the underlying environment the AI model is designed to emulate
· <u>Relevance</u> – the degree of correspondence between the data and the user's use case
· <u>Time-dependency</u> – how long the data stays relevant for a specific use case

Scaling:

· How the size of a dataset improves model performance and cost.

Uniqueness:

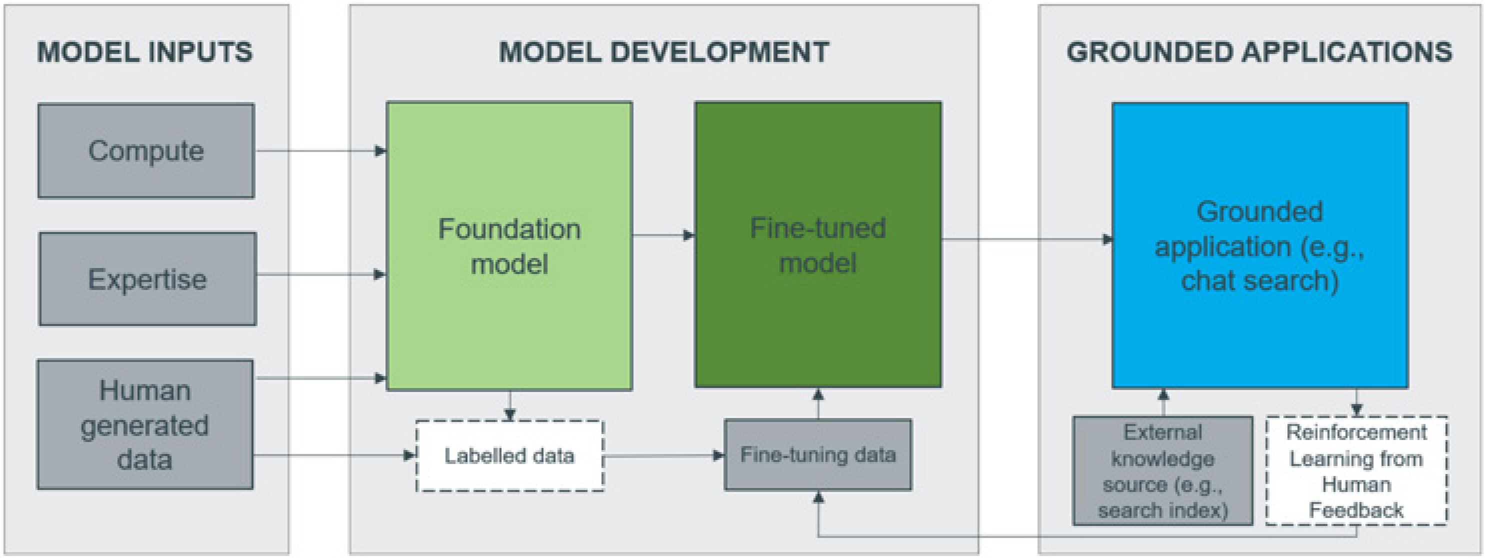
· <u>Exclusivity</u> – the degree to which other AI modelers can access data
· <u>Imitability</u> – whether dataset contents can be achieved through an alternative dataset

We draw on this framework throughout the paper.

[8] Jon Victor, *Why YouTube Could Give Google an Edge in AI,* The Information (June 14, 2023), https://www.theinformation.com/articles/why-youtube-could-give-google-an-edge-in-ai?rc=ui4kcg.

[9] Marco Iansiti, *The value of data and its impact on competition,* Harvard Business School NOM Unit Working Paper 22-002 (2021), https://www.hbs.edu/ris/Publication%20Files/22-002submitted_835f63fd-d137-494d-bf37-6ba5695c5bd3.pdf. We do not use a fourth dimension of data that the paper highlighted, scope.

Figure 1: Value chain of GenAI applications



B. Foundation Models: Pre-Training Relies on a Huge Volume of High-Quality Data

The first stage of the GenAI value chain is foundation models. Models such as the early BERT model by Google and GPT by OpenAI form the foundation (as their name suggests) of the GenAI production chain. Models, built by pre-training a machine learning algorithm on a broad dataset, seek to produce general-purpose, grammatically correct, and contextually coherent text output.

Figure 1 shows the three main inputs into foundation models: developer expertise, computing resources, and data. While the supply of talent who can optimize the development and training of GenAI models is limited, researchers and engineers can and do easily move between firms. Many established tech companies struggle to retain top talent: for example, all the eight authors of the 2017 seminal paper "Attention is all you need" that outlined the principal architecture of LLMs have since left Google.[10] In terms of compute, significant resources are required to train foundation models, e.g. the costs to train GPT-4 was reportedly above $100M.[11] However, improvements in cloud technologies and decline in prices of powerful hardware have improved the access to compute resources for companies.[12]

Based on the three main inputs, several system features determine the performance of foundation models and are key dimensions of competition.[13] We focus our discussion here on the relative contributions of the most important features, data quality, data size, and model size.

Data Quality Fuels AI Excellence

Data quality is a key consideration when selecting training data for foundation models. Amazon, Google, and UT Austin researchers note that:

[10] The original paper is: Vaswani et al., *Attention Is All You Need,* arXiv (August 2, 2023), https://arxiv.org/pdf/1706.03762.pdf. Article on where they are now: Madhumita Murgia, *Transformers: the Google scientists who pioneered an AI revolution,* Financial Times (July 23, 2023), https://www.ft.com/content/37bb01af-ee46-4483-982f-ef3921436a50.

[11] Will Knight, *OpenAI's CEO Says the Age of Giant AI Models Is Already Over,* Wired (April 17, 2023), https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.

[12] Kif Leswing, *Nvidia reveals new A.I. chip, says costs of running LLMs will 'drop significantly.'* CNBC (August 8, 2023), https://www.cnbc.com/2023/08/08/nvidia-reveals-new-ai-chip-says-cost-of-running-large-language-models-will-drop-significantly-.html.

[13] Including i) characteristics of the training data (primarily data quality and data size), ii) model size (number of parameters), and iii) the implementation and optimization of the pre-training process. Model architecture and training methods could also determine LLM performance but have been largely similar across the industry. In the latest generation of models, the overwhelming majority relies on a transformer architecture, used to understand contextual relationships among words or sentences. See Yule Wang, *An In-Depth Look at the Transformer Based Models,* Medium (March 17, 2023), https://medium.com/@yulemoon/an-in-depth-look-at-the-transformer-based-models-22e5f5d17b6b. And most developers use masked language modelling to train, where the model predicts missing words in a sentence using the contextual clues of surrounding words. But these could be important elements of dimensions in the future and some evidence points in that direction. For example, researchers have developed a multiscale decoder architecture that performs better than traditional transformer architectures on high-dimensional outputs, such as images or long sequences of text. This and similar model designs might lead to drastic increases in foundation model performance in the near future. See Yu et al., *MEGABYTE: Predicting Million-byte Sequences with Multiscale Transformers,* arXiv (May 19, 2023), https://arxiv.org/pdf/2305.07185.pdf.

> Real-world datasets are often "dirty," with various data quality problems and present the risk of "garbage in = garbage out" in terms of the downstream AI systems we train and test on such data.[14]

Quantitatively, AI research shows that the performance scores of text-based AI models trained on high-quality data can be significantly higher relative to models trained on data of average quality.[15] It could even increase model accuracy by as much as 71 percent for foundation models trained for code generation.[16] In addition, filtering text for quality improves the ability of a foundation model to learn from only a few examples, a key attribute of a generalist LLM.[17]

Pre-training data for LLMs comes from four sources:

- scraped public-facing data (e.g. text datasets such as Common Crawl),
- data in the public domain (e.g. old books),
- data already owned by GenAI firms (e.g. Bloomberg data, YouTube data or GitHub), or
- data licensed from third-party providers.

Until recently, pre-training data for most LLMs almost exclusively came from scraped content available on the internet, including webpages, books, and dialogues from social media or online forums. Research estimates the total stock of this data to be 740 trillion words.[18] This volume is nearly 17 million times the Encyclopaedia Britannica, or a stack of these dense encyclopaedias 8,500 km in height! But crucially, high-quality language data is estimated to be just 1 percent of this total stock. It appears that discussion of cat memes dominates the words of Shakespeare.

AI researchers typically describe high-quality data as peer-reviewed and professionally written content and major sources include books, news articles, scientific papers, and Wikipedia. Such data scores highly on the dimensions of usability (as needs less pre-processing) and relevance (to users) in the Iansiti (2021) framework. Quantitatively, developers can measure quality based on how closely the data resembles a high-quality benchmark. For example, Meta researchers filtered Common Crawl using Wikipedia as the benchmark finding greatly improved performance.[19] Filtering training data to weed out low-quality content is a usual step in model training.[20]

What Size is Most Important? Size of High-quality Data?

The relative contributions of (high-quality) data size and model size on model value can be compared using "scaling laws." These laws can be expressed by a formula that describes the impact of model size (the number of model parameters) and training dataset size on a model's performance.[21] The formula allows us to directly assess the relative contribution of model size and data size. Put simply, it allows us to measure just how important data size is to these models. (Translating

[14] Aroyo et al., *Data Excellence for AI: Why Should You Care,* arXiv (February 25, 2022), https://arxiv.org/ftp/arxiv/papers/2111/2111.10391.pdf.

[15] Wenzek et al., *CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data,* European Language Resources Association (ELRA) (May 11-16, 2020), https://aclanthology.org/2020.lrec-1.494.pdf.
The authors texted the impact of data quality on AI performance for the set of fast-Text models designed for text classification (based on meaning, parts of speech, etc.) and representation. The 9 percent estimate is obtained from the statistics listed in Table 1 for English language.

[16] See Figure 2.1 in Gunasekar et al., *Textbooks Are All You Need,* arXiv (June 20, 2023), https://arxiv.org/pdf/2306.11644.pdf. The estimate of 71 percent is based on the comparison between 1.3B models trained on the Stack+ and Code textbooks.

[17] Du, Nan, et al., *GLaM: Efficient Scaling of Language Models with Mixture-of-Experts,* ArXiv, 2021, /abs/2112.06905. Accessed Sept. 1, 2023.

[18] Villalobos, Pablo, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn & Anson Ho. *Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning.* arXiv preprint arXiv:2211.04325 (2022).

[19] Wenzek et al., *CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data,* arXiv (November 15, 2019), arxiv.org/pdf/1911.00359.pdf.

[20] E.g. Brown et al., *Language Models are Few-Shot Learners,* arXiv (July 22, 2020), https://arxiv.org/pdf/2005.14165.pdf; Chowdhery et al., *PaLM: Scaling Language Modeling with Pathways,* arXiv (October 5, 2022), https://arxiv.org/pdf/2204.02311; Touvron et al., *LLaMA: Open and Efficient Foundation Language,* arXiv (February 27, 2023), https://arxiv.org/pdf/2302.13971.pdf.

[21] Given that foundation models are not intended for use in downstream applications without fine-tuning and/or grounding, indicators of safety, lack of bias, or other downstream performance measures are less relevant for the pretrained models, and the training loss represents a universal performance criterion. Researchers at Google DeepMind trained a set of language models with different compute budgets and training datasets to estimate the impact of data and model size on model loss. The researchers found a precise formula to describe the model loss based on data size and number of model parameters. Hoffmann et al., *Training Compute-Optimal Large Language Models,* arXiv (March 29, 2022), https://arxiv.org/pdf/2203.15556.pdf. Other studies rely on similar estimates to select their training data and model sizes. See Wu et al., *BloombergGPT: A Large Language Model for Finance,* arXiv (May 9, 2023), https://arxiv.org/pdf/2303.17564.pdf; Anil et al., *PaLM 2 Technical Report,* arXiv (May 17, 2023), https://arxiv.org/pdf/2305.10403.pdf.

to economics, the formula allows us to estimate the production function and compare the elasticities of production with respect to data size and model size).

Using scaling laws, we find that for GPT-3 the formula implies data size is roughly 3.4 times more important than the model size.[22] Similarly, for the BloombergGPT model, the ratio is approximately 2.1.[23] If model performance was determined exclusively by these inputs (and it largely is), data size matters far more. Assuming that value is proportional to model performance (and, for economists, applying shadow pricing methods), we find that data accounts for 68-77 percent of model value for BloombergGPT and GPT-3.

A corollary of the scaling laws research is that most foundation models prior to 2022 used pre-training datasets that were too small.

Summary

Our analysis of foundation models finds that data is the most important input. This data has come primarily from publicly available sources, scraped from the web. Web content needs to be carefully filtered to identify high-quality data, which is crucial for ensuring high-quality outputs that suffer less from consumer protection issues including false and misleading information. While the amount of data fed into the models is enormous, the models are also gargantuan, and research shows perhaps surprisingly that data is the scarce factor.

C. Fine-Tuned Models Require Highly Relevant and Accurate Data

This section describes fine-tuning, the second stage of the GenAI production chain. Pre-training a large-language model produces a model that understands text but is not tailored for any specific task. A fine-tuned model is a foundation model that has been trained on specific data to perform a specific task or function. In addition, the output of foundation models typically inherits issues present in the model's training data leading to discrimination or bias.[24] These drawbacks of foundation models are typically addressed during this stage.

Fine-Tuning Often Relies on Human Feedback and Reduces Misinformation.

Fine-tuned models are trained on narrower, task-specific datasets for use-cases like dialogue with users, legal advisories, customer service, or medical consultations. Fine-tuned models are created by firms that own foundation models or by other AI companies; and enterprises now are increasingly fine-tuning models themselves, often using third-party services, e.g. Scale was just announced as OpenAI's fine-tuning partner. An example of a fine-tuned model is OpenAI's ChatGPT – which was tweaked from the GPT-3.5 foundation model to perform specialized chatbot functionalities using a narrower set of chat-specific training data. A second example is the ResNet foundation model, fined-tuned to let users detect specific diseases from X-ray and MRI scans.

> *Our analysis of foundation models finds that data is the most important input*

Fine-tuning typically uses labelled data, a type of structured dataset in which data elements are characterized with a data tag, for example pictures of animals (data) may be labeled with "dog," "cat," etc., or students' academic essays labeled with "A," "B," etc. It uses a process called supervised learning, which adjusts the model's parameters to make the model's output fit the data labels better, improving the model's performance for the specific use case.[25] Fine-tuning can enhance safety and reduce the biases picked up from the foundation model training data.

One of the most important methods used in fine tuning is reinforcement learning with human feedback ("RLHF"). During RLHF, the model learns to perform tasks by optimizing its actions based on feedback, which is often in the form of rewards or penalties determined by human workers. RLHF can use the data generated by humans

[22] Authors' calculations based on Hoffmann et al., *An empirical analysis of compute-optimal large language model training,* Google DeepMind (April 12, 2022), https://www.deepmind.com/publications/an-empirical-analysis-of-compute-optimal-large-language-model-training; Brown et al., *Language Models are Few-Shot Learners,* arXiv (July 22, 2020), https://arxiv.org/abs/2005.14165.

[23] Bloomberg designed this model using scaling laws for financial application and this model was constrained by limited high-quality financial data. See Wu et al., *BloombergGPT: A Large Language Model for Finance,* arXiv (May 9, 2023), https://arxiv.org/abs/2303.17564.

[24] As noted by a joint statement from several federal agencies. FTC, *FTC Chair Khan and Officials from DOJ, CFPB and EEOC Release Joint Statement on AI* (April 25, 2023), https://www.ftc.gov/news-events/news/press-releases/2023/04/ftc-chair-khan-officials-doj-cfpb-eeoc-release-joint-statement-ai.

[25] In contrast, foundation model training (pretraining) is unsupervised as data here is unlabeled with no additional context provided to the model.

directly or use a reward algorithm trained on human feedback. For instance, "OpenAI WebGPT" data is based on AI model answers that were rated by humans. Similarly, OpenAI Summarization's fine-tuning data is comprised of examples of human worker feedback regarding the summaries generated by the model.[26] Google's LaMDA model's fine-tuning data also contains rankings of model's responses on a set of performance metrics.[27] Data from the interactions of human users with AI models and user feedback are also important sources of fine-tuning data for dialogue-based models like ChatGPT.[28] So the more people that use it, the better the model will become, i.e. a network effect.

Fine-tuning is therefore critical in reducing misinformation in AI output and leading to significant improvements in model accuracy, safety, and other performance metrics. It can improve model performance drastically. For example, InstructGPT generated 37 percent higher user satisfaction relative to GPT-3 and evaluators noted that the fine-tuned model generated "truthful and informative answers about twice as often"[29]

Data Quality and Uniqueness Matter for Fine-Tuned Models

The main characteristics of data that matter during the fine-tuning stage include the data quality, specifically, the relevance of data for a specific application or domain, and its usability. High quality and relevance are typically achieved by building custom datasets, e.g. selecting high-quality materials within a narrow domain or collecting data from human workers/users. In addition, synthetic or machine generated data from a more powerful model, such as GPT-3, has been used to fine-tune several models.[30] However, this only works if the model generating the synthetic data performs better than the target model.

Fine-tuned models do not benefit from the volume of data as much as pre-trained models do, e.g. some reports claim that fine-tuning LLaMA, which is a model with a relatively small number of parameters, on a small dataset with only 1000 examples can result in a performance similar to large models such as Open AI's GPT-4 or Google's Bard.[31]

Uniqueness – both exclusivity and lack of imitability – of content can add tremendous value to a fine-tuning dataset. Unique domain-specific data can allow AI researchers to develop models that are better suited to a particular application, ultimately gaining a competitive advantage in the market for AI applications.

Summary

We find that fine-tuning of models requires specialized datasets that come from human feedback, other more powerful foundation models, web-scraped data, or private content. Human feedback is the primary source and as specific applications are increasingly used, feedback increases which can improve the performance of the applications, positively impacting other users. Fine-tuning can reduce misinformation in AI output and drive significant improvements in model accuracy, safety, and other performance metrics. Where user interaction creates a feedback loop, improving AI performance, there can be direct network effects.

D. Grounded Models Rely on Timely and Relevant Data

The final stage of the GenAI value chain is the user-facing applications that integrate foundation or fine-tuned models with end-user interfaces. Well-known GenAI applications include ChatGPT, Bard, and search chatbots like New Bing or Google's Search Generative Experience ("Google SGE").

Grounding Enables Models to Access Live Datasets

The production of downstream apps often requires additional techniques. For example, grounding is a technology that provides fine-tuned or foundation AI models with access to external use-case-specific knowledge that is not originally part of the training data. Because fine-tuned models cannot access real-time data, grounding is extremely useful for AI applications in search, news aggregator apps, and

[26] Abbeel et al., *Koala: A Dialogue Model for Academic Research,* MKAI (April 3, 2023), https://mkai.org/koala-a-dialogue-model-for-academic-research/.

[27] Thoppilan et al., *LaMDA: Language Models for Dialog Applications,* arXiv (February 10, 2022), https://arxiv.org/pdf/2201.08239.pdf.

[28] Michael Schade, *How your data is used to improve model performance,* OpenAI (2023), https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance.

[29] See Ouyang et al., *Training language models to follow instructions with human feedback,* arXiv (March 4, 2022), https://arxiv.org/pdf/2203.02155.pdf. Similarly, fine-tuned LaMDA-137B displayed 28 percent higher performance across several metrics relative to the pre-trained version. See Table 28 in Thoppilan et al., *LaMDA: Language Models for Dialog Applications,* arXiv (February 10, 2022), https://arxiv.org/pdf/2201.08239.pdf.

[30] Taori et al., *Alpaca: A Strong, Replicable Instruction-Following Model,* Stanford University (2023), https://crfm.stanford.edu/2023/03/13/alpaca.html.

[31] Touvron et al.,*LLaMA: Open and Efficient Foundation Language Models,* arXiv (February 27, 2023), https://arxiv.org/abs/2302.13971. Zhou et al., *LIMA: Less Is More for Alignment,* arXiv (May 18, 2023), https://arxiv.org/abs/2305.11206.

writing assistants. For example, Microsoft's search engine Bing, which licenses the GPT-4 foundation model, launched a search chatbot feature that answers users' questions by querying Microsoft's search index engine data after the user submits their prompt.

Grounding allows AI models to verify their responses against external benchmarks and reduces the propensity to hallucinate. Grounding therefore helps protect consumers by considerably mitigating (many but not all) the problems of misinformation from AI model output.

Grounding has been shown to make model responses more informative and improve their quality. For example, a grounded LLM scored 32.3 percent higher on usefulness and 13.9 percent higher on humanness relative to non-grounded ChatGPT.[32]

Data Quality and Uniqueness also Matter in Grounding

During grounding, the main characteristics of data are quality – specifically, accuracy, time-dependency, relevance, and uniqueness. While pre-training and fine-tuning can alter the style and structure of LLMs' responses, sources of external knowledge are primarily used by LLMs to retrieve facts. Thus, the role of data accuracy is paramount. Some grounded applications, like search chatbots or AI news aggregators, also benefit from access to the information on latest events that is not included in models' training or fine-tuning data. Time-dependency of grounding data is crucial for such applications because news or other real-time information attracts a large amount of user traffic. Similar to fine-tuning, uniqueness of grounding data can serve as a source of a competitive advantage for companies that have access to it.

Summary

Grounding allows fine-tuned and foundation models to access external sources of information. By incorporating additional, often time-sensitive information, grounding reduces hallucinations and makes outputs more informative and useful for consumers. But to the extent that information is not available, then the quality of applications decreases, such as the quality of answers to generative search queries.

03

FUTURE MARKET DYNAMICS, EMERGING ISSUES, AND POLICY CONSIDERATIONS

The final section of this paper builds on the analysis of the value chain and considers how GenAI data markets are evolving, and emerging issues that may need to be addressed. We conclude with some developing implications for policymakers such as competition agencies, and suggestions for what they should focus on going forward.

A. The Evolution of Data Markets for GenAI

As has become increasingly evident following the splash made by the public launch of ChatGPT, GenAI markets are dynamic and evolving rapidly. Vast sums of venture capital are gushing into the sector, an estimated $15.2 billion worldwide in the first half of 2023.[33] For example, French firm Mistral AI raised $113m in European's largest ever seed-funding round. Developments can happen at an astounding pace. For instance, the openness of Meta's LLaMA model led to rapid innovation as various research groups sped to build atop the model.[34]

Nonetheless, despite the dynamism, fundamental issues in GenAI data markets are likely to persist. Three trends stand out as relevant to data markets.

First, which datasets foundation models are pre-trained on is becoming increasingly opaque. As described in the foundation model section above, we know what data early models, such as GPT-2/GPT-3 and LaMDA, were trained on because publicly available research papers disclosed the sources in reasonable detail. More recent models, such as PaLM 2 (2023) and GPT-4 (2023), have not released training data details. This is an issue because non-transparency of training datasets exacerbates information asymmetry between content creators and AI firms.

[32] Peng et al., *Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback,* arXiv (March 8, 2023), https://arxiv.org/abs/2302.12813. Grounded LaMDA produces responses that are 23 percent more informative and 8 percent more "grounded" than the responses of non-grounded fine-tuned model. See Table 28 in Thoppilan et al., *LaMDA: Language Models for Dialog Applications,* arXiv (February 10, 2022), https://arxiv.org/pdf/2201.08239.pdf. To derive the statistics, compare the rows corresponding to "FT quality-safety (137B)" (non-grounded model fine-tuned for quality and safety) and LaMDA (137B) (fine-tuned and grounded model).

[33] Anna Cooban, *AI investment is booming. How much is hype?* CNN (July 23, 2023), https://edition.cnn.com/2023/07/23/business/ai-vc-investment-dot-com-bubble/index.html.

[34] Wayne Xin Zhao et al., *A Survey of Large Language Models,* arXiv (June 29, 2023), https://arxiv.org/abs/2303.18223.

Without information disclosure, it is challenging (and may be currently impossible in some cases) to identify all the data that has been used in training. Reverse engineering whether certain data has been used to train a model, especially for language data, can be extremely difficult. Without more disclosure, firms, regulators, and the public risk being left in the dark.

Second, data scarcity is likely to worsen, potentially rapidly. The growth of the stock of high-quality data is growing by an estimated 7 percent per year currently. In contrast, the size of datasets used for pre-training has been growing at a rate of 50 percent per year, and are now an appreciable share of the total available stock. For example, LaMDA's training dataset is approximately equal to 17 percent of today's estimated stock of high-quality language data.[35] A recent paper estimates that the stock of high-quality data will be exhausted by 2027 at the latest. Thus, the supply of high-quality data may soon fall short of the AI industry's demand.

Evidence suggests the supply of data is already a restrictive factor. Researchers behind BloombergGPT[36] state that they were "limited in the amount of domain-specific training data," relative to the optimal size. DeepMind researchers[37] similarly emphasise that "high quality datasets will play a key role in any further scaling of language models" and find most LLMs trained before March 2022 used too small training datasets.

Synthetic data, i.e. data generated by AI software, has been proposed as a solution for the issue of growing data scarcity (and is being widely experimented with).[38] However, researchers have shown that models trained on synthetic data undergo "model collapse — a degenerative process whereby, over time, models forget the true underlying data distribution."[39] And while efforts are underway to use existing datasets more efficiently, currently the growth of human-generated quality data is of critical importance to further advancements of GenAI models.[40]

Third, consistent with the expanding AI markets and growing need for data, markets for such data have started to spring up, with reports of proprietary licensed data being used to train major LLMs and negotiations between technology companies and publishers.[41]

There are risks that could arise as these data markets emerge. For example, it could lead to unequal access to high-quality datasets: with smaller players subsisting disproportionately on web-scraped data while the largest firms are making deals with high-quality data providers, as they are on the cutting edge of model size and volume of data and so can get the benefits of higher accuracy.

> ***There are risks that could arise as these data markets emerge***

[35] Thoppilan et al., *supra* note 27, at 13.

[36] Wu et al., *supra* note 21, at 11.

[37] Hoffman et al., *supra* note 5, at 3.

[38] Madhumita Murgia, *Why computer-made data is being used to train AI models,* Financial Times (July 19, 2023), https://www.ft.com/content/053ee253-820e-453a-a1d5-0f24985258de.

[39] Ilia Shumailov et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget,* arXiv (May 31, 2023), https://arxiv.org/pdf/2305.17493v2.pdf. Model collapse takes place because synthetic data typically reproduces text (words and tokens) that is more common in the training data. Synthetic data does not retain less frequent combinations of words and tokens. Thus, repeated re-training of models on synthetic data reduces the variance of output, to the extent that the model output becomes irrelevant to the user and does not resemble the original training data. Notably, synthetic data can still be used in model training and fine-tuning, but only if it is combined with human feedback or other human-generated inputs. Madhumita Murgia, *Why computer-made data is being used to train AI models,* Financial Times (July 19, 2023), https://www.ft.com/content/053ee253-820e-453a-a1d5-0f24985258de.

[40] The Economist, *The bigger-is-better approach to AI is running out of road* (June 21, 2023), https://www.economist.com/science-and-technology/2023/06/21/the-bigger-is-better-approach-to-ai-is-running-out-of-road.

[41] E.g. OpenAI, *GPT-4 Technical Report,* arXiv (March 27, 2023), https://arxiv.org/pdf/2303.08774.pdf. For example, in July 2023, OpenAI made a deal with The Associated press to license part of its news archive data and in exchange will provide "technology and product expertise" (Matt O'Brien, *ChatGPT-maker OpenAI signs deal with AP to license news stories,* AP (July 13, 2023), https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a.). Microsoft, Google, and Adobe have similarly been reported to have engaged with news executives to discuss copyright issues and potential negotiations relating to the use of news data in GenAI models. Google has reportedly been conducting such meetings with UK news outlets including The Guardian and NewsUK. (Cristina Criddle et al., *AI and media companies negotiate landmark deals over news content,* Financial Times (June 17, 2023), https://www.ft.com/content/79eb89ce-cea2-4f27-9d87-e8e312c8601d.)

B. Emerging Issues and Policy Considerations

Data transparency for content owners and consumer protection

As it stands, it is difficult for providers of critical high-quality content to ascertain the value of their data. The overall value of a trained model is hard to break down and ascribe at the individual content creator level, especially from the outside, so there is a knowledge gap between content creators and foundation model firms.

Limited disclosure about training datasets exacerbates this problem. If firms do not know if their content is being used or not, data value is even harder to estimate. Ultimately this lack of visibility may lead to the following:

> a. Content creators could increase or create barriers to sharing their data. They might implement more strict copyright protections or restrictive licensing agreements, or charge fees to access content. For example, Reddit began charging for its API to stop tech companies from scraping their data for free.[42] News producers like The New York Times and NBC News are exploring how to stop AI use without compensation.[43] Such reactions may reduce access to web content for consumers.
> b. Content owners may become disincentivised to release new high-quality data. Insufficient information to price data could lead to valuable content being mispriced and inefficient use of content. Data providers may invest less time and resource into releasing new high-quality content, making available non-public content, or digitizing data stores, such as old books. This could lead to undersupply of content, impacting the size and representativeness of AI training data.

Increasing transparency, such as the requirement in the EU AI Act, might meaningfully reduce information asymmetries.[44] Transparency requirements can increase incentives for data creators to create more content, result in more parties trading data, and relieve the impending data bottleneck. But as ever, the details matter. Simultaneously, any requirement needs to consider its burden and impact on GenAI innovation. Balancing these factors appropriately will result in more competition on data quality and encourage investment.

Lack of information about data used for training also affects downstream markets for AI applications. Each production stage of a GenAI application affects model output and influences the risks of misinformation and bias. Thus, transparency would help developers of downstream products to identify model biases and solutions. Downstream application developers can better understand if a model requires further fine-tuning and can better identify appropriate grounding material.

The media has widely reported concerns around how GenAI can contribute to misinformation, and public sentiment towards AI has included uncertainty, lack of trust and unease (as well as amazement as to what is now possible). Moreover, GenAI tools can be used to produce and spread intentionally misleading information. AI industry leaders have acknowledged these issues, including Sam Altman, CEO of OpenAI, who stated that "[t]he general ability of these models to manipulate and persuade, to provide one-on-one interactive disinformation is a significant area of concern."[45]

> ***Lack of information about data used for training also affects downstream markets for AI applications***

Downstream, data transparency may facilitate the development of consumer safety standards and increase attention on potential consumer protection issues. Transparency requirements could also make at least some information available to regulators, content creators, other interested third parties, and researchers. It may help these parties bring to light potential consumer safety problems.

But disclosure requirements obviously have costs as well, to AI firms directly and through potential pro- or anti-competitive effects. And costs could of course be borne by consumers. Regulators will therefore need to carefully calibrate any requirements – striking the right balance between the benefits and costs involved.

[42] Rohan Goswami, *Reddit will charge hefty fees to the many third-party apps that access its data,* CNBC (June 1, 2023), https://www.cnbc.com/2023/06/01/reddit-eyeing-ipo-charge-millions-in-fees-for-third-party-api-access.html.

[43] Alex Sherman and Lillian Rizzo, *A.I. poses new threats to newsrooms, and they're taking action,* CNBC (June 6, 2023), https://www.cnbc.com/2023/06/06/news-organizations-ai-disinformation.html.

[44] European Parliament, *EU AI Act: first regulation on artificial intelligence* (June 14, 2023), https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[45] Dan Milmo & Alex Hern, *Elections in UK and US at risk from AI-driven disinformation, say experts,* The Guardian (May 20, 2023), https://www.theguardian.com/technology/2023/may/20/elections-in-uk-and-us-at-risk-from-ai-driven-disinformation-say-experts.

Ensuring Markets Are Competitive

As discussed above, unequal access from ownership or frictions impeding smaller players from negotiating might result in significant advantages accruing to certain players. Such advantages can grow over time.

Network effects can originate from the feedback loop between the user base, data derived from user interactions, and improvements in product performance. While this directly benefits users, markets may tip because of this dynamic. Major products, e.g. Google Search, might retain or develop new unassailable advantages with its Search Generative Experience compared to other products. This form of tipping is something the competition community is all too aware of already – it's akin to the barriers that arise from asymmetric access to query-and-click data in internet search.

To address this, we recommend that regulators monitor the data used for training and fine-tuning and the contracts between data providers and AI firms. This information could help them detect growing barriers to entry occurring through reinforcement learning for human feedback, harmful exclusivity issues (noting that exclusivity can often be fine), or other means.

We believe monitoring the following would help regulators ensure they are on top of key issues:

- Which datasets are used for training, fine-tuning, and grounding, and the demand for and supply of data, given the predictions that the demand of data will outstrip supply;
- Ongoing research on how data quality and size relate to accuracy, safety, reliability, truthfulness, and other key AI performance metrics;
- Whether there are any impediments to new high-quality data being created or valuable stores of offline data being digitized; and
- The nature of the deals struck between third-party data providers and AI firms – what data is being traded and on what terms.

More generally, GenAI technology is advancing and spreading rapidly. Agencies will need to keep up with technological developments and the potential implications. Properly understanding these markets requires data science and engineering expertise, which underlie our analysis, in addition to economics and law. Many agencies have invested in these capabilities, including the UK Competition and Markets Authority, the U.S. Federal Trade Commission and Department of Justice, the Australian Competition and Consumer Commission, the Canadian Bureau of Competition, the French Autorité de la Concurrence, and others.[46] To support evolution and greater deployment of GenAI models, we encourage policymakers to maintain an active and watching brief on GenAI markets, maintaining active dialogue with AI firms, content owners and others. Agencies should consider whether they may need to take action, to nurture markets for data to improve consumer outcomes, promote healthy and competitive AI markets, and preserve the remarkable pace of innovation. ■

> “*We recommend that regulators monitor the data used for training and fine-tuning and the contracts between data providers and AI firms*

[46] Stefan Hunt, *The technology-led transformation of competition and consumer agencies: The Competition and Markets Authority's Experience,* CMA (June 14, 2022), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1085931/The_technology_led_transformation_of_competition_and_consumer_agencies.pdf; Stephanie Nguyen, *A Century of Technological Evolution at the Federal Trade Commission,* FTC (February 17, 2023), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/02/century-technological-evolution-federal-trade-commission.