Safe, **P**roductive, **E**fficient, **A**ccurate, and **R**esponsible (**SPEAR**) AI Systems

Mark Montgomery
KYield, Inc.

Abstract

Since the first large language model (LLM) chatbot was released to the public, leading experts in AI, catastrophic risk, economics, and cybersecurity, among others, have warned about the unprecedented risks caused by interactive LLM bots trained on large-scale, unstructured data [1, 2]. Although theoretical methods have been proposed, and incremental improvements are occurring, to date none have proven as effective or comparable to other safety-critical industries such as biological contagions or nuclear power [3, 5]. We therefore have an urgent need to adopt AI systems based on the proven laws of physics and economics without sacrificing the many benefits of AI. This paper is focused on the inefficiencies, risks, and limitations of LLMs, the business and economic incentives influencing decisions, and the architecture required to provide safe, productive, efficient, accurate, and responsible (SPEAR) AI systems. One such system is described—our data and human-centric KOS (EAI OS).

KEYWORDS: Artificial Intelligence, Data Governance, Risk Management, Organizational Management, Systems Engineering, Safety, Sustainability, Disaster Management, Catastrophes, Existential Risk, Bioweapons, Large Language Models, Cybersecurity

1. Introduction

The recent popularity and media exposure about LLM chatbots has been so prevalent that LLMs have become synonymous with all of AI, which is false[1]. Many different AI methods and models exist, including several types of large-scale language models, small language models, and neurosymbolic AI [4]. LLMs were developed on the shoulders of giants in AI research over a period of seven decades. Most of the early research was funded by government in small academic or corporate labs. In recent years, the super majority of AI research investment has been sourced from a few large tech companies investing hundreds of billions of dollars attempting "dominance in AI", in part by attracting and retaining the majority of top talent, and in part by directing the majority of research towards models that align with their interests, such as LLMs [5, 6].

One example of strategic research is the Transformer developed by Google researchers, which has become the building block for LLMs [7]. The Transformer model includes encoders and decoders, which when trained and tuned provides text tokens that are assigned probabilities for generating text strings. Although the efficiency gains are small, at massive scale with data stores that contain trillions of words and hundreds of billions of parameters, it becomes possible to produce the LLM chatbots we see today [8].

[1] Example: In talk at Def Con 31, DoD CDAO Dr. Craig Martell revealed he had t-shirts made up with the message "LLMs ≠ AI".

Contact author: {markm} @kyield.com

As the dominant search engine with one of the largest data stores and cloud infrastructure investments, Google has a powerful incentive to invest heavily in natural language processing (NLP) research, related components, startups, and talent. Another example of strategic investment was by Nvidia in the optimization of Bidirectional Encoder Representations from Transformers (BERT) neural network architecture, improving speed in Google's Tensor Cores without losing accuracy [9].

Since LLM chatbots require vast amounts of computing resources, the revenue growth alone in cloud services can justify the approximately $20 billion investment in LLM startups to date by Microsoft, Google, and Amazon [10]. However, the large investments combined with integrating with their market leading products, and strategic agreements for technology infrastructure, such as the exclusive agreement in the OpenAI/Microsoft partnership, also provides the potential to "co-opt" competition, and extend and expand monopolies, duopolies, and the larger oligopoly over the rest of the economy [11, 12].

> "It means that industry domination of applied work also gives it power to shape the direction of basic research. Given how broadly AI tools could be applied across society, such a situation would hand a small number of technology firms an enormous amount of power over the direction of society". – Nur Ahmed, Munstasir Wahed, & Neil Thompson [13].

The scale required for the incremental improvements in efficiency is so vast that it causes large-scale economic and environmental risks and costs, including catastrophic risk discussed in this paper, and existential risk for vast sectors within the knowledge economy [14, 15], including industries relying on copyright, which alone contributed more than $1.8 trillion dollars to the U.S. economy in 2021 [16, 17].

Fortunately, safe, productive, efficient, accurate, and responsible AI is currently possible that can deliver critical functions by designing systems with integrity and applying similar principles as other safety-critical industries, which is the focus of this paper.

2. Safety and Security

The strategic interests of a few large technology companies do not necessarily align with the interests of many customers and society. LLMs for example, while aligned very well to the interests of a few large tech companies, are inherently unsafe. The so-called guardrails are false equivalencies, as guardrails imply a physical barrier, whereas LLM guardrails are text-based and easy to work around due to the nature of self-generating text models and the interactive dynamics of LLM chatbots.

> "We find that jailbreak prompts are introducing more creative attack strategies and disseminating more stealthily over time, which pose significant challenges to their proactive detection. Moreover, we find current LLMs and safeguards cannot effectively defend against jailbreak prompts in various scenarios. Particularly, we identify two highly effective jailbreak prompts which achieve 0.99 attack success rates on ChatGPT (GPT-3.5) and GPT-4, and they have persisted online for over 100 days"……"LLM vendors and adversaries have been engaged in a continuous cat-and-mouse game since the first jailbreak prompt emerged. As safeguards evolve, so do the jailbreak prompts to bypass them." – Shen, Chen, et al. [18]

3. Catastrophic Risk

Catastrophic risks such as assistance in developing biological weapons are of even greater concern than cybersecurity[2]. LLM chatbots are trained on many large data sets that include scientific journals in every discipline. Planning instructions and suggestions for biological terrorism and other catastrophic events that would require terrorist cells decades to research have been returned in seconds, demonstrating the ability to increase catastrophic risks at an exponential rate [19].

> "In less than 6 hours after starting on our in-house server, our model generated forty thousand molecules that scored within our desired threshold. In the process, the AI designed not only VX, but many other known chemical warfare agents that we identified through visual confirmation with structures in public chemistry databases. Many new molecules were also designed that looked equally plausible. These new molecules were predicted to be more toxic based on the predicted LD in comparison to publicly known chemical warfare agents. This was unexpected as the datasets we used for training the AI did not include these nerve agents." – Urbina, F., Lentzos, F., Invernizzi, C. et al. [20]

The trigger of the LLM arms race was the premature unleashing of LLM bots to the general public without benefit of rigorous safety engineering processes required of other industries offering safety-critical systems. Since one company took the risk to launch an LLM did not experience immediate government intervention due to safety risks, and then scaled rapidly, competitors felt the need to engage in similar behavior or risk being left behind. The arms race rapidly expanded to include nations where the LLM companies were headquartered, most recently in Europe where new LLM entrants reportedly successfully lobbied to block the proposed regulations for foundation models in EU AI Act, which includes LLMs.

One example of previous AI adoption is autonomous driving, which falls within the jurisdiction of the National Highway Traffic Safety Administration (NHTSA). The NHTSA adopted a six-level safety standard for autonomous vehicles, ranging from level 0 with no autonomous technology to level 5 for full automation. Most new cars today fall between level 1-3, offering augmentation that enhances safety. Automation technology employed by autonomous vehicles isn't restricted to one model like LLMs, but is rather a hybrid of many different types of technology across dozens of companies, ranging from new independent companies to incumbents and after-market products. Most importantly from a risk perspective, autonomous driving does not represent catastrophic risk, but rather limited individual events by comparison to LLM chatbots.

When the inevitable accidents occur in autonomous vehicles, they are obvious, limited to a small number of people, and analyzed by expert third parties like police who file reports. If the autonomous system is found to be the cause, cars can be recalled and tested. In contrast, with the unique risk profile in LLMs, it may be years before catastrophic risk manifests, potentially impacting millions or even billions of people.

[2] I've warned about the LLM risks in biological weapons since November of 2022, in private communications, social media posts, and my enterprise AI newsletter. In 2019, I unveiled our synthetic genius machine (SGM) in a talk at the leading technical conference in New Mexico, 'Metamorphic transformation with enterprise-wide AI'. Shortly thereafter traffic patterns on our web sites convinced me to voluntarily restrict additional information on the SGM in public due in part to catastrophic risks similar in nature to LLMs. While our SGM employs strong security and compression and is much more accurate and environmentally friendly than LLMs, the intention of the system is to accelerate discovery. Our SGM is a dual-purpose system that would present similar risks as LLMs if made widely available to anyone in the public, hence the need to restrict access with strong security architecture.

Technology that contains potential catastrophic risk typically face rigorous requirements, including biological and nuclear risk. To transport the Ebola virus, for example, requires a special permit from the Department of Transportation, which has jurisdiction for transport, working "closely with CDC, OSHA, HHS, DOD, EPA, and state and local government to assure that our respective safety missions are adequately addressed in these scenarios". However, the influenza virus is subject only to CDC recommendations[3].

Nuclear power is regulated by the U.S. Nuclear Regulatory Commission (NRC). Created in 1974 as an independent agency, the NRC "regulates commercial nuclear power plants and other uses of nuclear materials, such as in nuclear medicine, through licensing, inspection and enforcement of its requirements". The International Atomic Energy Agency (IAEA) was created in 1957 as an autonomous organization within the United Nations to promote peaceful use of nuclear power and inhibit military use.

The catastrophic risks in LLMs are similarly well understood by objective scientists, but it's been more than a year now since LLM chatbots have been released to the public and very little has been done to address current risks beyond minor safeguards that have been proven easily breached [21].

4. Environmental Costs

LLMs trained on web-scale data cause many different types of risks and costs, including significant environmental costs, particularly in water and energy use [22, 23]. One new large datacenter can use as much energy as hundreds of thousands of new homes.

The world's largest data center market is Northern Virginia, with over 275 data centers, is experiencing a dramatic spike in energy use due to datacenters. CBRE, a leading commercial real-estate services firm, reported the amount of power available in the Northern Virginia market shrank to 38.4 megawatts earlier this year from 46.6 megawatts the previous year, despite an increase in inventory of 19.5% for a total of 2,131 megawatts [24].

Much of the worldwide increase in energy for datacenters is due to compute-intensive AI, particularly in datacenters owned by the leading cloud providers—AWS, Microsoft, and Google, all of which have experienced significant growth in recent years[4]. Although the volume of data continues to expand at a rapid pace, the exponentially expanding size of LLM models is driving the majority of demand for datacenters. LLM models have grown by 1000x in a few short years. The explosive growth in AI computing has caused serious shortages in GPUs [25], as well as increased demand for more efficient and less costly alternatives.

[3] Pathogens with various levels of risk ranging from near zero to near certain death is a good model for risk management in AI systems as some systems and applications contain near zero risk while others represent near certain death (e.g., autonomous weapons or the deadliest viruses). LLMs trained on web-scale data and made available to the general public with only minor restrictions represent among the greatest risks in AI systems, whereas rules-based AI systems trained on high quality curated data with strong security, provenance, and verification represent low risk.

[4] The pay-for-use elastic computing model known as the 'cloud' pioneered by Amazon (AWS) was a brilliant innovation that has proven to be one of the most successful business models in history. However, as the cloud model rapidly expanded, serious perverse incentives became more problematic, including investing heavily in research that would expand dependency on the cloud model, such as LLMs, and increased systemic risk due to an ever-increasing portion of critical functions in our economy hosted on the three top cloud providers. For example, while I supported the CIA's adoption of AWS a few years earlier as a wise decision, I warned the Department of Defense on the systemic risk in awarding a sole-source contract for $10 billion in the initial 'JEDI' contract. By the time JEDI was awarded, a large portion of the S&P 500 and critical government functions were hosted on AWS, representing unprecedented dependence on a single host. One well-designed hack or well-placed insider could threaten the entire system.

McKinsey is forecasting power consumption in datacenters to reach 35 gigawatts (GW) by 2030 in the U.S. alone—more than doubling from 2022, representing about 40 percent of the global market [26]. Compute-intensive datacenters training LLMs also require very large quantities of water to cool servers, so water demand is growing in a similar trajectory to energy[5]. The subject of water use is sensitive and often kept secret. However, researchers have reported that training GPT-3 in Microsoft's most efficient datacenters can evaporate 700,000 liters of clean freshwater, and that global AI demand for water is forecast to require 5-6 billion cubic meters of water in 2027, which is the equivalent of the annual use for Denmark [27].

5. Socioeconomic Costs

The same causal factors that result in safety failures in LLMs also cause widespread bias, hallucinations, inaccuracies, and enable deepfakes and widespread misinformation. Each of these weaknesses either unintentionally cause socioeconomic costs, or can be exploited to attack democratic institutions and erode the socioeconomic fabric of society [28, 29, 30, 31, 32].

For example, in a recently revealed query about the Covid-19 Pandemic, ChatGPT returned a response about a New York Times article that was completely fabricated [33], including title and links. Such extreme misinformation could have profound negative impacts either through the chatbots directly or repeated on social networks, which occurs at massive scale on a continuous basis.

This type of misinformation is unique to a class of self-generating algorithms like LLMs that were until November 2022 limited to controlled research labs. The same technical vulnerabilities can lead to misinformation about politics, race, religion, companies, government, non-profits, and individuals, opening up very broad potential socioeconomic damage that is already occurring [34]. A group of DeepMind researchers published a good review of social risks from LLMs just under a year before ChatGPT was launched (Weidinger, Laura, et al.,). All of the six risk categories included in their paper have since been realized in LLM chatbots.

6. Impact on the Knowledge Economy

An area of AI research conspicuous in its absence investigates the potential damage to the knowledge economy from LLMs. Presumably due to the specialty disciplines of the researchers, and lack of depth in economics, the paucity of research also likely reflects the influence of a few big-tech companies, so we must look to economics and business consulting to better understand these impacts [35, 36].

Fritz Machlup was responsible for initially categorizing and measuring the knowledge economy (KE) in his 1962 report, *The Production and Distribution of Knowledge in the United States*, when he estimated the knowledge economy represented 29 percent of U.S. GNP in 1958. The U.S. GNP is expected to be a bit over $23 trillion for 2023, so if the KE was still 29 percent of U.S. GNP, it would be $7-8 trillion. However, if employing similar criteria Machlup used in 1962 to today's economy, the KE would be a much larger portion

[5] Paul Churnock, an engineer at Microsoft, recently estimated Nvidia's popular H100 chips alone require as much electricity as Phoenix, Arizona, the fifth largest city in the U.S. The current retail price of a H100 on 1/2024 is reported to be $100,000.

of the U.S. economy, clearly representing a majority—perhaps even a supermajority. Any serious threat to the KE like that posed by unfettered LLM chatbots is a serious threat to the U.S. and its economy.

Peter Drucker coined the term knowledge worker in his book, *The Landmarks of Tomorrow*, published in 1959, where he also introduced the concept of post-modernism, referring to the era as a transitionary period between the exponential increase in labor productivity increase and a more productive and rewarding economy consisting mainly of knowledge workers. Forty years later Drucker published an article on knowledge worker productivity with a subtitle "The Biggest Challenge", which provides a historic review of the study of labor leading to the post-modern era [37], with a specific focus on the contributions of Frederick Winslow Taylor [38]. Taylor had a profound impact on how people worked during the first half of the 20th century, during which time a fifty-fold increase in productivity was achieved in manufacturing automation.

It is this work on knowledge worker productivity near the end of Drucker's life when the dilemma facing employers and nations becomes obvious in the form of contradictions, conflicts, and misalignment of interests at the confluence of the knowledge economy and AI where we have been working for over three decades[6].

> "Knowledge workers are rapidly becoming the largest single group in the work force of every developed country. They may already compose two-fifths of the U.S. work force —and a still smaller but rapidly growing proportion of the work force of all other developed countries. It is on their productivity, above all, that the future prosperity—and indeed the future survival—of the developed economies will increasingly depend." – Peter Drucker (1999).

Drucker estimated that our collective understanding of knowledge worker productivity in the year 2000 was similar to the understanding of manual workers in the year 1900. He suggests knowledge workers must have autonomy, learn continuously, they should be treated as an asset rather than a cost, that quality of output is at least as important as quantity, and that employers must protect that asset. Further, he teaches that innovation must be the responsibility of knowledge workers and that they must manage themselves.

Although he was not referring to AI specifically, Drucker understood the challenges of adoption well before the technical ability to increase knowledge worker productivity was achieved: "Each of these requirements…is almost the exact opposite of what is needed to increase the productivity of the manual worker."

A quarter of a century later employers are faced with generative AI that is forecast to increase knowledge worker productivity representing $6-8 trillion annually, but also displace tens of millions of knowledge workers [39], possibly including the consultants making the forecasts today and their clients. It is perhaps unsurprising then that investment and adoption of generative AI has fallen well short of expectations [40].

7. Efforts to Improve LLMs

The decade preceding the commercialization of LLM chatbots experienced exponential growth in AI research, as evidenced by the number of attendees at conferences, papers published, graduate degrees,

[6] I operated a consulting firm that converted to a knowledge systems lab in the mid 1990s. We created several ventures including GWIN (Global Web Interactive Network), which was an experimental learning network for thought leaders. It was in 1997 during the operation of GWIN when I conceived the KYield theorem (yield management of knowledge), which is manifest as the KOS.

recruitment ads, and startups claiming to be AI companies. In the NeurIPS 2023 conference, for example, the number of papers accepted expanded from 411 in 2014 to 3,584 in 2023[7]. One encouraging sign is nearly twice as many papers were accepted in 2023 with the keyword "efficiency" than "generative"[8]. Focus areas for improving efficiencies are many, including data compression techniques, hardware innovations, fine-tuning parameters, data filtering and learning, improved model design, and new types of pre-training.

7.1 Pre-Training

One approach disclosed by a group of researchers (Lewis, Patrick, et al.) in a paper at the NeurIPS 2020 conference has since become widely deployed [41]. Retrieval-Augmented Generation (RAG) offers parametric and non-parametric memory components that are pre-trained and pre-loaded with extensive knowledge from external sources. RAG improved accuracy over state-of-the-art while reducing hallucinations, and is particularly beneficial when used in conjunction with an accurate, knowledge-intensive data source with frequent changes, such as Wikipedia tested by the researchers, news service, sensor monitoring, etc.

7.2 Fine-Tuning

Finetuning of parameters is a common technique to improve efficiency in LLMs for specific data sets. LLM chatbots have general knowledge but are terribly inefficient and inaccurate. Finetuning for a specific domain or corpus has proven to improve results, but is compute and energy expensive, hence the focus on parameter-efficient methods. Fine-tuning for efficiency can reduce the entire workload, including computing, storage, and costly time for engineering.

One example is Low-Rank Adaptation (LoRA), which is a finetuning method that reduces memory by using a small set of trainable parameters. A team at Microsoft Research found they could reduce the number trainable parameters by 10,000x and GPU memory requirements by 3x with no additional inference latency [42].

Another example is FLAN (Finetuned Language Net) proposed by a group of Google researchers. FLAN is employed to improve so-called zero-shot learning to support classification without annotated data. FLAN is an instruction-tuned version of a decoder-only language model. It can be more efficient due to lack of needing annotated data. FLAN was tested and found to outperform zero-shot GPT-3 on most datasets and GPT-3 few-shot performance on a few datasets [43]. The downside to fine-tuning is while it can improve efficiency, lower compute costs and reduce the need for manually annotating data, the descriptions may be less precise than annotations in some cases.

7.3 Simplifying Transformers

Complexity is one of many perverse incentives in LLMs. The larger the models and more functions introduced to improve those models; the more complexity created. Although complexity can be positive for top-tier talent in deep learning earning high six figure compensation packages, chip suppliers, and pay-for-

[7] An article by Jacob Marks provides a summary of analytics for NeurIPS 2023.

[8] Given the volume, no attempt is made to provide a comprehensive review. Rather, a few brief examples are provided to inform about the type of research recently published on improving LLMs.

what-you-use cloud services, it is not necessarily beneficial for customers or society, particularly when less effective than other methods and/or forced on customers and markets. One focus area by Bobby He and Thomas Hofmann at ETH is a method to reduce complexity in transformer blocks [44]. In experiments on both autoregressive decoder-only and BERT encoder-only models, the pair of researchers were able to accelerate throughput by 15% while using 15% fewer parameters. Although incremental in nature, the high energy and financial costs translate to significant savings.

7.4 Small Language Models

The scale needed for generalization in language models at this stage of technology evolution is what causes most of the risks and inaccuracies. Small language models (SLMs) combined with high quality data and precision data management offer a good alternative for many applications. SLMs provide the ability to perform many of the same functions as LLMs, including writing letters and reports, and can do so with higher levels of accuracy and security at 10% or less of the financial, environmental, social, and economic costs of LLMs[9]. Two significant SLMs released in 2023 include LLaMA by Meta AI and Phi by Microsoft Research.

LLaMA reduces memory usage and runtime with some methods, and then further improves training efficiencies with other methods, applying a variety of research techniques in classically incremental innovation for the purpose of efficiency [45]. The result was that LLaMA-13B outperformed GPT-3 despite being more than 10x smaller. LLaMA is available in several sizes (7B, 13B, 33B, and 65B parameters). Since Meta AI trained on publicly available datasets that may include copyrighted content, and some datasets are of poor quality, although much more efficient and less costly than larger LMs, LLaMA is still problematic for applied AI. However, Meta AI performed a valuable service by demonstrating what is possible when resources are focused on efficiency rather than only scale.

Phi-1 is a transformer-based model with only 1.3B parameters trained on a combination of high-quality data from the web and high-quality synthetic data [46], yet were able to surpass most open-source models despite employing a model that is 10x smaller and a dataset that is 100x smaller. Although the Phi research team used GPT-4 to generate code and synthetic data, high-quality data is pre-existing in many organizations, and can be produced by employing other methods, including data science tools and human expert curation. The primary contribution of Phi was to build from the previous work on "TinyStories" (Eldan & Li) [47], and demonstrate the benefit of combining high quality data and refined LMs for organizations with smaller budgets while dramatically reducing computing needs, power and water consumption for generative AI functions.

7.5 Hardware Improvements

The arms race in AI is perhaps nowhere as aggressive as hardware, complete with trade restrictions and large-scale industrial subsidies for semiconductors. Although hardware is slower to innovate than LLM

[9] Our Synthetic Genius Machine (SGM) is a new type of SLM with a hybrid neurosymbolic architecture designed to provide knowledge compression and security for efficient domain-specific acceleration of discovery. Since acceleration of discovery enables bad actors as well as good, we made the decision to restrict disclosures in our research. Although we lack the market power of big tech to incentivize sharing selective research and the resources to commercialize the SGM, it represents our second generation of R&D. Our intent is to introduce elements of the SGM to the KOS once it can be rigorously tested in a safe and secure manner.

models, once matured in the product line the efficiencies become more widespread. Google, Microsoft, Meta, AWS, Apple, and Tesla have custom chip programs underway, presumably to reduce costs, ensure supply, and reduce reliance on industry leader Nvidia while also providing a proprietary advantage tailored to their tech stacks. AI chips are customized to the needs of each company and their products. All of the leading semiconductor companies are investing heavily in AI and several startups have gained traction.

Nvidia is the clear market leader in AI chips with a suite of products and enjoy strong leadership in cloud services with its GPUs. Nvidia took an early lead by seizing the opportunity offered when its GPUs for gaming were discovered by Stanford researchers to be far more efficient for deep learning [48]. Their CUDA compute platform has become the preferred choice for a very large number of engineers and developers.

However, in response to the popularity, high cost of Nvidia's chips, and inventory shortages, competition is rapidly emerging, which bodes well for improving efficiency and impact. AMD launched its MI300X chip as a direct competitor to Nvidia's H100, and has early commitments from several large customers. Intel has responded with Gaudi 2, a competitor to H100, which outperforms H100 in some tasks[10], and is less expensive. Startup AI chip companies that have gained significant traction include SambaNova Systems, Cerebras, and Groq. Many others have interesting research and IP.

One of the most important trends in hardware is so-called edge computing, or on devices that are much more energy efficient and secure than LLMs hosted on cloud infrastructure to date. Qualcomm, for example, offers the Neural Processing SDK for developers to run NN models on Snapdragon mobile platforms. Apple's A12 Bionic is the first 7nm smartphone chip, containing 6.9bn transistors [49]. Apple has also been active in researching SMLs. Their UNO-EELBERT model is just 1.2 MB in size, but achieves General Language Understanding Evaluation (GLUE) benchmark within 4% of a model almost 15x its size [50].

8. Data Governance

"Organizations and their personnel defining, applying and monitoring the patterns of rules and authorities for directing the proper functioning of, and ensuring the accountability for, the entire life-cycle of data and algorithms within and across organizations." – Janssen, Marijn, et al.

The underlying commonality causing the majority of risks and much of the costs in AI today, beyond scale combined with the nature of deep learning, is primarily the lack of robust data governance, data provenance, and data security. This is in part due to the ultra-emphasis on superintelligence in AI research driven by financial incentives, but also the personal motivation of scientists and engineers. Google researchers published a paper in 2001 (Sambasivan, Nithya, et al.), confirming what was already widely understood, which is that "Everyone wants to do the model work, not the data work" [51]. Incentives, motivation, and elite cultures notwithstanding, LLMs have not fundamentally altered information theory or the laws of physics. The quality of the output still substantially depends on the quality of the input.

[10] Hugging Face tested Gaudi2 and Nvidia A100, finding Gaudi2 latencies were x2.84 faster than Nvidia A100 (2.63s versus 0.925s). Databricks also performed tests on Gaudi2, finding it had better training and inference performance-per-per-dollar than Nvidia chips, including the H100. Nvidia is expected to H200 in Q2 2024.

Consumer LLM chatbot firms offer open access to anyone in the public without security protocols, methods, or procedures developed over decades for data governance, provenance, security, or digital rights management. The majority of data sets LLMs train on lack the sufficient data structure necessary to mitigate significant risks and costs to customers and society, including catastrophic risks. Poorly designed data management systems make robust data governance difficult, which "can have profound legal, financial and social implications on the organizations involved, citizens and businesses, and society at large" [52].

9. Discussion on LLMs

The combination of self-generating models with vast financial and computing resources, which was then made available to the general public, triggered an unprecedented chain of events still in the early stages of expansion. While the benefits of LLMs were immediately obvious to individuals experimenting with the bots, the risks were understood by only a small group of experts who have studied these risks for decades.

Two out of the three Turing Award winners for deep learning (2018) have warned about the catastrophic risks (Bengio and Hinton). The third (LeCun) still works for a big tech (Chief AI Scientist at Meta). The majority of experts in a position to understand the technical risks work for conflicted companies, are bound by non-disclosure agreements, and have significant financial incentives that prevent disclosure. Multiple crises in LLM firms and AI units in big tech resulted in 2023 as tensions flared over safety and ethics. Most of the leading LLM companies and their big tech partners have admitted they can't self-regulate, and need to be regulated[11], yet as of today nothing has been done to mitigate current LLM chatbot risks, presumably due to the national arms race in AI. Trillions of dollars annually are potentially at stake.

On the economic front, the week I am writing this, a PwC survey of CEOs reports they expect a 5% reduction of workers in 2024 due to generative AI, the IMF issued a warning "AI will affect almost 40 percent of jobs around the world", and the Institute of International Finance warned about our "huge fiscal problem" in record levels of public debt. A deeply indebted world suffering from severe wealth gaps and concentration of market power needs increased productivity that leads to economic and environmental sustainability, which requires increased economic diversification, not further consolidation of wealth and market power.

There is clear consensus in the most qualified unconflicted experts in this specific type of technology and the specific types of risks they create, and that is LLMs are inherently unsafe and were prematurely released. LLMs present much greater risks when deployed in an interactive manner to the general public—including bad actors, which enables prompt injections, jailbreaking, and other methods to work around safeguards for nefarious purposes, such as expected attempts in bioterrorism and other types of weaponization.

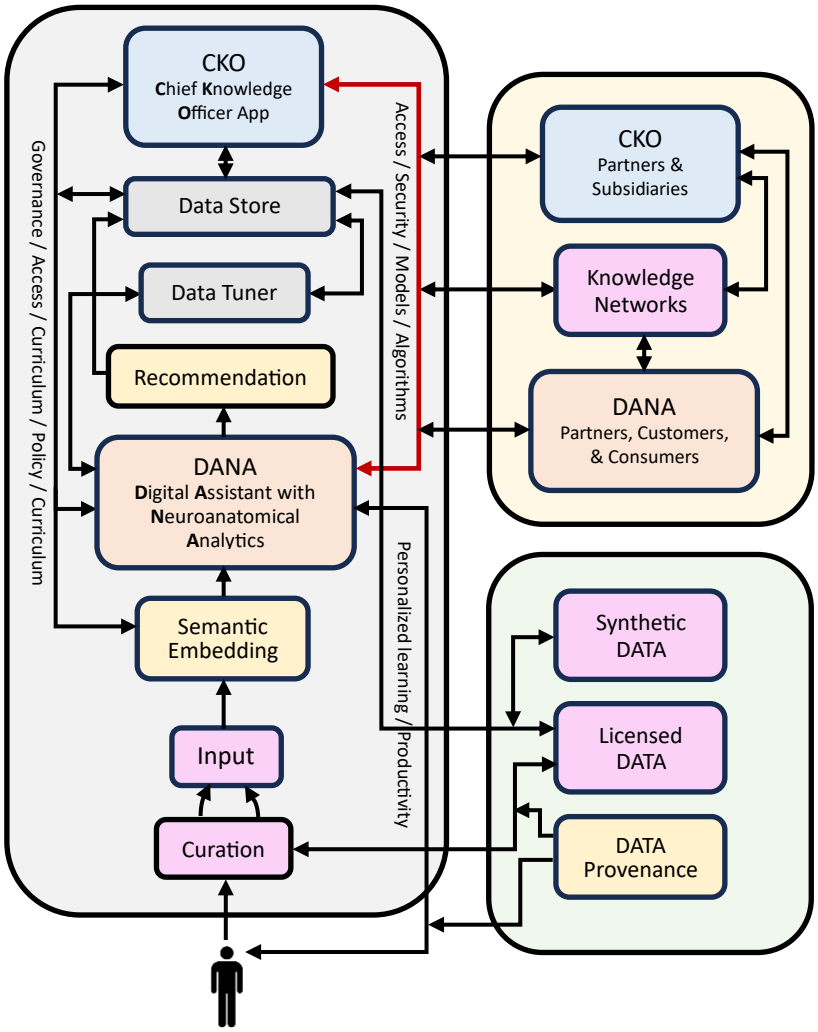
However, when managed within safety-critical systems architecture and trained on high quality data, language models of various sizes can be effective tools for both generalized and specialized productivity growth at much higher levels of accuracy, and avoid the many types of risks caused by consumer LLM chatbots trained on web-scale data, including catastrophic risks, which also potentially represent trillions of dollars annually.

[11] An old engineer friend who was on the founding team of a market leading tech company in a private email said: "the inventors don't know how to make LLMs safe and are asking the least qualified people to make decisions" (about safety and security).

10. A SPEAR AI System: KOS (EAI OS)

The KOS is an organizational operating system with a data-centric architecture and human-centric design. It is an end-to-end AI system with governance, safety, and security designed-in from inception of the underlying theorem 27 years ago (1997). The priority of the precision data management core is high-quality information production and precision presentation, tracked from the source wherever possible, with data provenance and security in every step. Refined over an extended period of R&D, the system incentivizes accuracy and provides multiple types of productivity improvement, representing a good example of a SPEAR AI system[12].

KOS Architecture

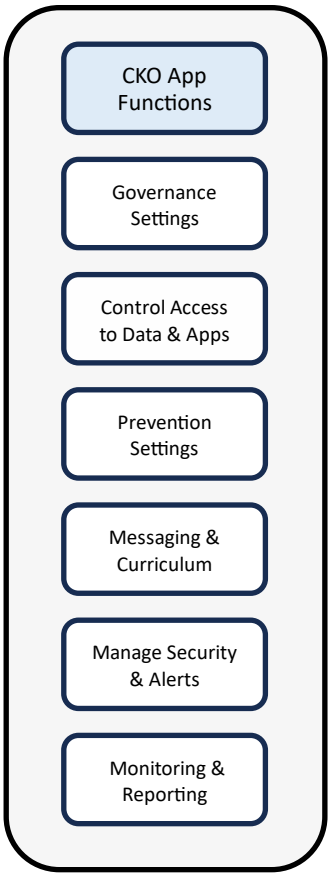


[12] Portions of the KOS as described here is patented (M. Montgomery. "Modular system for optimizing knowledge yield in the digital workplace". USPTO #8,005,778, 23 August 2011).

10.1 System Components

The KOS is a distributed modular enterprise AI OS. The system components consist primarily of two modules; the enterprise administrative application (CKO), and a digital assistant for every person (DANA). Both modules are administered through a simple natural language interface. By prioritizing strong governance, precision data management, and high-quality data, computing resources and associated financial and environment costs can be reduced by about 90%, with much higher accuracy levels than is currently possible with LLM chatbots trained on web-scale data. I will briefly describe the two primary modules.

10.1.1 CKO Module



The CKO Module (or app) is a flexible administration tool for management of the organization, business unit, or network with the ability to administer the entire system, including controlling access to digital assistants, integration with other organizations, manage licensed content, curriculum, and multiple types of security settings, among other functions. The CKO app is also the compliance mechanism for the KOS, which can be adapted to each jurisdiction. It's a universal system that can be tailored to the needs of each organization.

The intent is to achieve a CALO (Continuously Adaptive Learning Organization) for the purpose of creating a competitive advantage, maintain sovereignty, and adapt to change. Ideally, the CKO app is optimized to provide as much autonomy as practicable for the specific organization, generate precision accuracy in high quality data, and achieve high levels of productivity while maintaining a unified mission for the organization and each business unit.

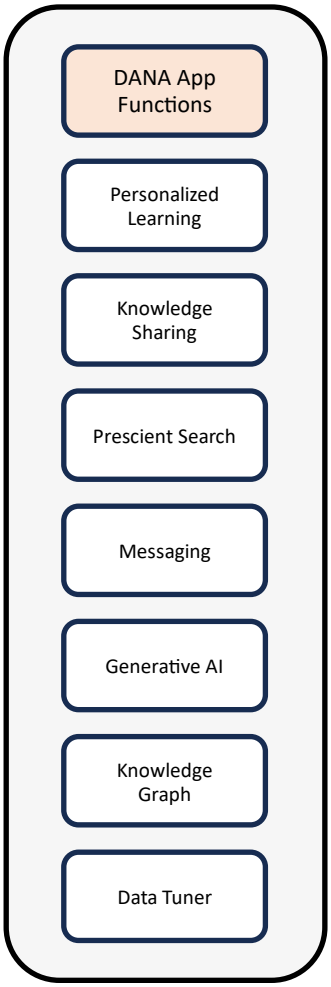
The CKO app is automated with the exception of semi-automated controls for administration changes. Reports and alerts are provided to the administrators of the CKO app for security, risk management, and for optimization of the KOS, which can be further tailored to the organization by administrators.

10.1.2 DANA (Digital Assistant with Neuroanatomical Analytics)

DANA is the personal digital assistant made available to every employee of an organization with a KOS enterprise license. DANA assists individuals in personalized learning and several different types of productivity, and also works dynamically with the CKO app to automate precision data management across the digital workplace. The interaction between the CKO app and DANA is important for safety and security, protecting intellectual property, ensuring high-quality data, and enabling productivity.

DANA is intended to be the primary tool for learning and productivity in the digital workplace, so great care was made to design DANA to be easy to use, ultra-personalized, and productive. High-quality data is produced in the natural daily work process within the governance and security parameters set in the CKO app and further refined by individuals in their DANA profile. AI functions are trained on the data individuals produce, augmented with high-quality licensed data, vastly reducing unnecessary risk, waste, and costs.

Personal assistant functions in DANA



1. Continuous, Ultra-personalized Learning

DANA has one primary mission, and that is to assist the individual human it is assigned to within parameters set by the CKO app. Continuous learning tailored to and as determined by each individual is a high priority for DANA.

2. Selective Knowledge Sharing

Learning is achieved by sharing first-person knowledge, curated posts, and through links and licensed content. Every individual and file in the KOS network is rated by a proprietary system that weighs heavily on track record, experts, and peers. Incentives for sharing valuable knowledge can be integrated.

3. Prescient Search

End-to-end data structure in DANA enables the ability to deliver relevant knowledge. If an individual has an upcoming project or needs to monitor a topic, it can be added to the prescient search settings in the DANA profile. The results are displayed prominently on the home page of DANA as it becomes available.

4. Work-related Networking

The CKO app controls access to networking for the organization, which can then be further restricted by each individual. Some documents may be accessible by all employees, contractors and vendors, whereas others can be restricted to only those with approved access. Automating this process improves productivity and provides security, including for training the generative AI function.

5. Limited Communications

Limited communications through messaging provides a necessary and valuable addition to the DANA app. Due to tight security and focus on quality, DANA messsaging is free from the external spam and distractions of social networking and email.

6. Generative Productivity (GenAI)

We announced generative AI integration in the summer of 2023 for the KOS. Organizations with large amounts of pre-existing high-quality data and/or license large, high-quality datasets can train SMLs and be up and running quickly, whereas others may want to wait until they have enough data to train SMLs.

7. Knowledge Graphs (KG)

KGs are an excellent natural addition to DANA due to the high-quality pre-structured data. Data structure, security, networks and relational data is naturally provided in the KOS and DANA work process.

8. Data Tuner

The data tuner allows individuals to increase or decrease the quality and quantity of data consumption for most of the data functions, offering a unique productivity tool that can be set for specific time periods, such as deadlines or periods of immersive learning. Mandatory content is not included in the DANA data tuner.

10.1.3 Testing the KOS

The KOS has been tested in four phases in an ultra-lean manner:

1. **GWIN**: I conceived the underlying theorem 'yield management of knowledge' while operating GWIN (Global Web Interactive Network) from 1997-2000. GWIN was an experimental learning network for thought leaders with a membership consisting primarily of professors, senior executives, analysts, boards of Fortune 500 companies, editors from leading publications, and investors. The GWIN platform allowed us to experiment with early concepts for a prospective enterprise system, including cross-disciplinary learning, prevention, and a then primitive digital assistant. The primary obstacles at the time for achieving the theorem later manifest in the KOS was poor scalability, high compute costs, and ineffective machine learning for most of our tasks.
2. **Components:** As the architecture began to come together in the 2000s, I tested components of the KOS and DANA in my small lab. Although still far from commercially viable, sufficient improvements were observed to complete the initial design and submit a patent application. By 2012 we began contacting prospective customers, limited to those few with supercomputers.
3. **Commercial Viability:** By 2016 we could see a path to sustainability in a commercially viable manner if we could find a large customer willing to risk a custom installation. We focused in particular on accelerating discovery with pharma companies, and had many discussions, but couldn't find a prototype sponsor, so we pulled back and continued internal R&D. In 2021 and 2022 we hired a senior engineer to perform the back-end functions, which were confirmed without any significant problems.
4. **MVP:** After many discussions on custom installs with leading companies, it became apparent we needed a working prototype. Otherwise, we risked losing control of the technology. In 2022, we decided to hire a small team of specialists to build a minimally viable product, which continued into early 2023. The initial focus was on DANA. Although the admittedly low-budget MVP prevented achieving the quality needed for our customer pipeline, the MVP demonstrated basic technical viability at a very attractive price-point.

10.1.4 Use Cases

In hundreds of discussions over the past 15 years, engaging with a significant portion of the leading companies based in the U.S. and EU, as well as trade groups, startups, and various potential strategic partners, we invested many human years studying use cases, including but not limited to the following:

1. R&D companies, organizations, and business units for accelerating discovery.
2. Personalized healthcare where DANA would serve as the patient's personal digital assistant.
3. Very large banks for internal use and a limited version for customers. We explored midmarket enterprise customers to lower risk profiles, wealth management clients, and others.
4. Enterprise-wide corporate installs in most industries, including extending the KOS to partners in supply chains and partners. Some included extending DANA to end consumers.
5. University systems, non-profits and government entities at the city, state, and national level, including very large agencies hoping to increase productivity and reduce bureaucracy.

11. Conclusion

Nearly seven decades after John McCarthy coined the term artificial intelligence, we find ourselves in the midst of a global AI arms race, which was triggered by unleashing an LLM chatbot to the general public for generating text, which rapidly converted to multimodal. The LLM bots are trained on the content owned by millions of people and organizations without their permission, representing trillions of dollars of investment and great personal effort to develop over centuries.

A review of AI research for this paper confirmed all known existing safeguards for LLMs are easily breached. This is no surprise to many of the pioneers of this very technology, hence the consistent warnings from Hinton, Bengio and many others, including scientists from the LLM firms themselves and their big-tech enablers.

As briefed in this paper, LLM chatbots in use today are nowhere close to meeting the rigorous safety standards required of other industries. These risks include significant cybersecurity risks, social and psychological risks, economic risks, and catastrophic risks, among others, yet over a year after achieving the most rapid adoption of any product in history, the U.S. Government still hasn't taken the action required to mitigate these risks to levels comparable to other safety-critical technologies in other industries.

I can only speculate based on behavior of other arms races that a perception of national competition is one reason for the unprecedented lapse in safety governance. However, given the state of technology markets in the U.S. today, and the amount of money at stake, I postulate that unhealthy influence from the big-tech sponsors of LLM chatbot firms are also contributing to the hands-off policy for these very serious risks.

However, regardless of what courts and legislatures do or don't do in response to what I believe was a recklessly premature release of high-risk technology, individuals and organizations are nevertheless faced with their own risks to consider, as well as opportunities, and must make decisions. As the unprecedented live experiment on society has now confirmed, the supermajority of the problems and risks caused by LLM bots are due to the lack of effective data governance, which is due in part to the very large scale necessary to provide mimicry of generalized intelligence to consumers on any topic.

Fortunately, strong data governance, safety, and security can be provided to individuals and organizations without sacrificing the majority of productivity benefits offered by LLM chatbots. Moreover, precision data management can provide much higher levels of accuracy at a small fraction of the financial and environmental costs created by LLM chatbots. As many companies have now demonstrated, even generative AI functions can be executed within systems with strong data governance by employing small language models and other techniques, such as licensing additional data for larger models free from copyright and reputational liability.

Well-designed SPEAR AI systems are technically viable today. I offer a briefing on the KOS as an example of one such system. By adopting a refined SPEAR AI system like the KOS, organizations can benefit from nearly three decades of R&D, save significant time and money, improve accuracy with precision data management, increase productivity, and reduce risks while avoiding unnecessary and wasteful social, economic, and environmental impacts.

References

[1] Bengio, Hinton, Yao, Song, et al. “Managing AI Risks in an Era of Rapid Progress.” *ArXiv*. (2023). https://arxiv.org/pdf/2310.17688.pdf

[2] Weidinger, Laura, et al. "Ethical and social risks of harm from language models." *arXiv preprint arXiv:2112.04359* (2021). https://arxiv.org/pdf/2112.04359.pdf

[3] Eujeong Choi, Jeong-Gon Ha, Deagi Hahm, Min Kyu Kim. “A review of multihazard risk assessment: Progress, potential, and challenges in the application to nuclear power plants”, *International Journal of Disaster Risk Reduction*, Volume 53. (2021). https://www.sciencedirect.com/science/article/pii/S2212420920314357

[4] Mark Montgomery. “The Power of Neurosymbolic AI.” (2023). https://www.linkedin.com/pulse/power-neurosymbolic-ai-mark-montgomery

[5] “Big tech and the pursuit of AI dominance”, *The Economist*. (2023) https://www.economist.com/business/2023/03/26/big-tech-and-the-pursuit-of-ai-dominance

[6] Courtney Radsch. Written testimony submitted to the Canadian Parliament’s Standing Committee on Canadian Heritage (CHPC) on big tech abuse and manipulation. (2023). https://www.openmarketsinstitute.org/publications/cjl-director-courtney-radsch-testifies-before-the-canadian-parliaments-standing-committee

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Jakob, Llion Jones, Aidan Gomez, Łukasz Kaiser; Illia Polosukhin. "Attention is All you Need" (PDF). *Advances in Neural Information Processing Systems*. (2017). https://arxiv.org/abs/1706.03762

[8] Kaplan, Jared et al. “Scaling Laws for Neural Language Models.” *ArXiv*. (2020). https://arxiv.org/pdf/2001.08361.pdf

[9] Christopher Forster, Thor Johnsen, Swetha Man- dava, Sharath Turuvekere Sreenivas, Deyu Fu, Julie Bernauer, Allison Gray, Sharan Chetlur, and Raul Puri.” BERT Meets GPUs”. Technical report, NVIDIA AI. (2019). https://medium.com/future-vision/bert-meets-gpus-403d3fbed848

[10] Jin Berber, & Tom Dotan. “Tech Giants Spend Billions on AI Startups—and Get Just as Much Back.” *WSJ*. (2023). https://www.wsj.com/tech/ai/ai-deals-microsoft-google-amazon-7f624054

[11] Gerrit Vynck. “How Big Tech is co-opting the rising stars of artificial intelligence”. *Washington Post*. (2023). https://www.washingtonpost.com/technology/2023/09/30/anthropic-amazon-artificial-intelligence/

[12] Amba Kak, Sarah Meyers West, & Meredith Whittaker. “Make no mistake—AI is owned by Big Tech.” *MIT Technology Review.* (2023). https://www.technologyreview.com/2023/12/05/1084393/make-no-mistake-ai-is-owned-by-big-tech

[13] Nur Ahmed, Munstasir Wahed, & Neil Thompson. “The growing influence of industry in AI research”, *Science*, VOL 379 ISSUE 6635. (2023). https://ide.mit.edu/wp-content/uploads/2023/03/0303PolicyForum_Ai_FF-2.pdf

[14] Machlup, Fritz. *The production and distribution of knowledge in the United States*. Vol. 278. Princeton university press. (1962).

[15] Drucker, Peter. *The age of discontinuity: Guidelines to our changing society*. Routledge. (2017).

[16] Robert Stoner, Jéssica Dutra. “Copyright Industries in the U.S. Economy”. (2022). https://www.iipa.org/files/uploads/2022/12/IIPA-Report-2022_Interactive_12-12-2022-1.pdf

[17] Mark Montgomery. “What’s your GAI plan if copyrighted material is disallowed by SCOTUS?” (2023). *Enterprise AI*. https://www.linkedin.com/pulse/whats-your-gai-plan-copyrighted-material-disallowed-mark-montgomery-trx8c

[18] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, & Yang Zhang. “Do Anything Now”: Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. *ArXiv* August, 2023. https://arxiv.org/pdf/2308.03825.pdf

[19] Mouton, Christopher A., Caleb Lucas, and Ella Guest. “The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach.” RAND Corporation. (2023). https://www.rand.org/pubs/research_reports/RRA2977-1.html

[20] Urbina, F., Lentzos, F., Invernizzi, C. et al. “Dual use of artificial-intelligence-powered drug discovery.” *Nat Mach Intell* **4**, 189–191. (2022). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9544280/

[21] Hubinger, Evan, et al. "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training." *arXiv preprint arXiv:2401.05566* (2024). https://arxiv.org/pdf/2401.05566.pdf

[22] Carole-JeanWu,RamyaRaghavendra, UditGupta, BilgeAcun, NewshaArdalani, KiwanMaeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. “Sustainable AI: Environmental implications, challenges and opportunities.” In *Proceedings of Machine Learning and Systems*, volume 4, pages 795–813. (2022). https://proceedings.mlsys.org/paper_files/paper/2022/file/462211f67c7d858f663355eff93b745e-Paper.pdf

[23] Strubell et al. "Energy and Policy Considerations for Deep Learning in NLP." *ArXiv*. (2019) https://arxiv.org/pdf/1906.02243.pdf

[24] Angus Loten. "Rising Data Center Costs Linked to AI Demands". *WSJ*. (2023). https://www.wsj.com/articles/rising-data-center-costs-linked-to-ai-demands-fc6adc0e

[25] Don Clark. "Nvidia Revenue Doubles on Demand for A.I. Chips, and Could Go Higher". *The New York Times*. (2023). https://www.nytimes.com/2023/08/23/technology/nvidia-earnings-chips.html

[26] McKinsey & Company. "Investing in the rising data center economy." (2023). https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/investing-in-the-rising-data-center-economy#

[27] Pengfei Li, Jianyi Yang, Mohammad A. Islam, Shaolei Ren. "Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models". *ArXiv*. (2023). https://arxiv.org/pdf/2304.03271.pdf

[28] Schramowski, Patrick, et al. "Large pre-trained language models contain human-like biases of what is right and wrong to do." *Nature Machine Intelligence* 4.3 (2022): 258-268. https://arxiv.org/pdf/2103.11790.pdf

[29] Rawte, Vipula, Amit Sheth, and Amitava Das. "A survey of hallucination in large foundation models." *arXiv preprint arXiv:2309.05922* (2023). https://arxiv.org/pdf/2309.05922.pdf

[29] Nguyen, Thanh Thi, et al. "Deep learning for deepfakes creation and detection: A survey." *Computer Vision and Image Understanding* 223 (2022): 103525. https://arxiv.org/pdf/1909.11573.pdf

[31] Melissa Heikkilä. "Three ways AI chatbots are a security disaster." *MIT Tech Review*. (2023). https://www.technologyreview.com/2023/04/03/1070893/three-ways-ai-chatbots-are-a-security-disaster/

[32] Bender, Emily M., et al. "On the dangers of stochastic parrots: Can language models be too big? 🦜." *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. (2021). https://dl.acm.org/doi/pdf/10.1145/3442188.3445922

[33] New York Times Company V. Microsoft, OpenAI, Inc., et al. (2023) https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf

[34] Blodgett, Su Lin, et al. "Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021. https://aclanthology.org/2021.acl-long.81.pdf

[35] John Burn-Murdoch. "Here's what we know about generative AI's impact on white collar work". *The Financial Times*. (2023) https://www.ft.com/content/b2928076-5c52-43e9-8872-08fda2aa2fcf

[36] Dell'Acqua, Fabrizio, et al. "Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality." *Harvard Business School Technology & Operations Mgt. Unit Working Paper* 24-013 (2023). https://www.hbs.edu/ris/Publication%20Files/24-013_d9b45b68-9e74-42d6-a1c6-c72fb70c7282.pdf

[37] Drucker, Peter F. "Knowledge-worker productivity: The biggest challenge." *California management review* 41.2 (1999): 79-94. https://www.iriscrm.com/app/uploads/2021/05/knowledge_workers_the_biggest_challenge.pdf

[38] Taylor, Frederick Winslow. *The principles of scientific management*. Harper & brothers. (1919)

[39] Chui, Michael, et al. "The economic potential of generative AI." (2023). https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

[40] "What happened to the artificial-intelligence investment boom?" The Economist. (2024). https://www.economist.com/finance-and-economics/2024/01/07/what-happened-to-the-artificial-intelligence-investment-boom

[41] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in Neural Information Processing Systems* 33 (2020): 9459-9474. https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[42] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021). https://arxiv.org/pdf/2106.09685.pdf

[43] Wei, Jason, et al. "Finetuned language models are zero-shot learners." *arXiv preprint arXiv:2109.01652* (2021). https://arxiv.org/pdf/2109.01652.pdf

[44] Bobby He and Thomas Hofmann. "Simplifying Transformer Blocks." *ArXiv*. (2023). https://arxiv.org/pdf/2311.01906.pdf

[45] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023). https://arxiv.org/pdf/2302.13971.pdf

[46] Gunasekar, Suriya, et al. "Textbooks Are All You Need." *arXiv preprint arXiv:2306.11644* (2023). https://arxiv.org/pdf/2306.11644.pdf

[47] Eldan, Ronen, and Yuanzhi Li. "TinyStories: How Small Can Language Models Be and Still Speak Coherent English?." *arXiv preprint arXiv:2305.07759* (2023). https://arxiv.org/pdf/2305.07759.pdf

[48] Raina, Rajat, Anand Madhavan, and Andrew Y. Ng. "Large-scale deep unsupervised learning using graphics processors." *Proceedings of the 26th annual international conference on machine learning*. (2009). http://robotics.stanford.edu/~ang/papers/icml09-LargeScaleUnsupervisedDeepLearningGPU.pdf

[49] Tanya Singh. "Top AI Chip-Making Companies For Smartphones". *Mobile App Daily.* (2024). https://www.mobileappdaily.com/knowledge-hub/ai-chip-making-companies-for-smartphones

[50] Cohn, Gabrielle, et al. "EELBERT: Tiny Models through Dynamic Embeddings." *arXiv preprint arXiv:2310.20144*(2023). https://arxiv.org/pdf/2310.20144.pdf

[51] Sambasivan, Nithya, et al. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI." *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. (2021). https://research.google/pubs/everyone-wants-to-do-the-model-work-not-the-data-work-data-cascades-in-high-stakes-ai/

[52] Janssen, Marijn, et al. "Data governance: Organizing data for trustworthy Artificial Intelligence." *Government Information Quarterly* 37.3 (2020): 101493. https://repositorium.sdum.uminho.pt/bitstream/1822/69192/1/JBEBJ20.pdf

Additional References

Hoffmann, Jordan, et al. "Training compute-optimal large language models." *arXiv preprint arXiv:2203.15556* (2022). https://arxiv.org/pdf/2203.15556.pdf

Rae, Jack W., et al. "Scaling language models: Methods, analysis & insights from training gopher." *arXiv preprint arXiv:2112.11446* (2021). https://arxiv.org/pdf/2112.11446.pdf

Wan, Weier, et al. "A compute-in-memory chip based on resistive random-access memory." *Nature* 608.7923 (2022): 504-512. https://www.nature.com/articles/s41586-022-04992-8

Ellingrud, Kweilin, et al. "Generative AI and the future of work in America." (2023). https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond. *Generative AI at work*. No. w31161. National Bureau of Economic Research, 2023. https://www.nber.org/system/files/working_papers/w31161/w31161.pdf

Hui, Xiang, Oren Reshef, and Luofeng Zhou. "The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market." *Available at SSRN 4527336* (2023). https://www.econstor.eu/bitstream/10419/279352/1/cesifo1_wp10601.pdf

Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in Neural Information Processing Systems* 33 (2020): 9459-9474. https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Longpre, et al. "The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI ." November, 2023. https://arxiv.org/pdf/2310.16787.pdf

Jianping Gou, Baosheng Yu, Stephen Maybank, & Dacheng Tao . "Knowledge Distillation: A Survey". (2021). https://arxiv.org/pdf/2006.05525.pdf

Omiye, J.A., Lester, J.C., Spichak, S. *et al.* "Large language models propagate race-based medicine." *npj Digit. Med.* **6**, 195 (2023). https://doi.org/10.1038/s41746-023-00939-z

Yan, X., Qian, J.H., Ma, J. *et al.* "Reconfigurable mixed-kernel heterojunction transistors for personalized support vector machine classification." *Nat Electron* **6**, 862–869 (2023). https://doi.org/10.1038/s41928-023-01042-7 https://www.nature.com/articles/s41928-023-01042-7

Verdecchia, Roberto, June Sallou, and Luís Cruz. "A systematic review of Green AI." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2023): e1507. https://wires.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/widm.1507

Alizadeh, Keivan, et al. "LLM in a flash: Efficient Large Language Model Inference with Limited Memory." *arXiv preprint arXiv:2312.11514* (2023). https://arxiv.org/pdf/2312.11514.pdf

Chen, Jiaoyan, et al. "Contextual semantic embeddings for ontology subsumption prediction." *World Wide Web* (2023): 1-23. https://arxiv.org/pdf/2202.09791.pdf

Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." *arXiv preprint arXiv:2305.14314* (2023). https://arxiv.org/pdf/2305.14314.pdf

Wu, Shijie, et al. "BloombergGPT: A Large Language Model for Finance. 2023." *ArXiv preprint: https://arxiv. org/pdf/2303.17564. pdf*. https://arxiv.org/pdf/2303.17564.pdf\

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, Xiaomo Liu. “DOCLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding. 2024. https://arxiv.org/pdf/2401.00908.pdf